UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D11/D1.3 – Report on second set of FSTP projects

Nature	Report	**Work Package**	WP1
Due Date	30/09/2025	**Submission Date**	30/09/2025
Main authors	Wilker Aziz (UVA)		
Co-authors			
Reviewers	Maryam Hashemi		
Keywords	survey, languages, resources		
Version Control			
v1.0	**Status**	Final	30/09/2025

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

List of Figures

Abstract

Within UTTER we have allocated EUR 895 206.00 to run an FSTP programme, which is a key component of UTTER's impact. We have split our programme in two separate calls for project proposals, one running in 2023/2024, the other running in 2024/2025. In this document, we describe the entire programme. This document hence includes parts of D1.2, but this is marked clearly for any reader already familiar with that document; we opted to have the entire programme in one document as this allows for a better comparison of the two calls.

Besides fully documenting the structure of our programme and its results, in this document, we also document deviations of action in Section 5 and reflect on ongoing exploitation 6.

Our two FSTP calls attracted 54 and 50 submissions, respectively, of which 15 were selected for funding (8 in the first call, 7 in the second). All 15 projects started and finished (execution and dissemination) successfully on time. For completeness, Appendix B and Appendix C contain the final reports from the 15 project teams.

1 UTTER's FSTP Programme

The content of this section is copied from D1.2,[1] *with only minor differences in structure of presentation. A reader familiar with that document can safely skip this section.*

1.1 Infrastructure

To ensure full compliance with our data management plan (D1.1),[2] we managed our FSTP call (both submission and review process) via a UVA-hosted instance of HotCRP[3] available from: https://utter-fstp.science.uva.nl.

At time of writing, September 2025, the site is no longer active, but we have stored in our UVA-hosted GitLab instance complete snapshots of the website's database, one for each of the two FSTP calls in UTTER. The complete data (submissions, feedback, and any interventions by the pilot board, etc.) are carefully stored in UVA-hosted servers, in compliance with our DM plan.

We hosted the call documentation on GitHub (https://github.com/utter-project) and linked it from our website (https://he-utter.eu) and from the European commission Funding and tender opportunities website.

For communication, we created a UVA-hosted mailing list: utter-fstp@list.uva.nl.

1.2 Dissemination

We posted our FSTP calls on our own website (https://he-utter.eu/#fstp) and on the European commission Funding and tender opportunities website. We further advertised the FSTP call in a number of ways:

- During UTTER's user days;
- By joining FSTP-disseminating events organised by third parties;
- on UTTER's social media channels (e.g., Twitter and Linkedin);
- and within the networks of UTTER's PIs.

1.3 Call documentation

The complete call is described in a core Call Documentation document along with 4 annexes:

- **A1 Guide for applicants.** A document with detailed instructions and screenshots to guide applicants through the submission process;
- **A2 Third party agreement.** The agreement that awardees would have to sign in order to execute their project under our funding scheme;

[1] D1.2: https://projectnetboard.absiskey.com/viewdocument/0b9209-69e20e-a46a17-0758bf-000044
[2] D1.1: https://projectnetboard.absiskey.com/viewdocument/5078a3-3d1fd5-a89ee2-cbe72c-000033
[3] HotCRP (https://hotcrp.com) is an open-source software for managing conference review processes.

- **A3 Project proposal template.** A docx proposal template, with the required fields and guidelines of how to fill them in;
- **A4 Evaluation criteria.** The description of the evaluation process and criteria.

In addition, we also shared UTTER's GA and CA. For ease of reference, we also prepared a key facts sheet, mostly used for dissemination. For transparency, all these documents are hosted at https://github.com/utter-project/fstp and they were clearly linked from our own website as well as from the corresponding call on the EC funding and tender opportunities website.

1.4 Key Parameters

Objectives. Develop and/or pilot applications using XR models (i.e., pre-trained neural network models adaptable to a large variety of forms of expression, interaction, languages, domains, styles and intent) in new sectors, with a focus on enabling new types of human-human and human-machine interaction. Examples of welcome project objectives include:

- Improving or demonstrating efficiency of XR model inference;
- Improving or demonstrating efficiency of XR model training;
- Designing interfaces for usability;
- Extending XR models to new languages, domains or modalities;
- Applying XR models to new tasks;
- Building resources for XR models;
- Evaluation of XR models.

Proposals.

- Maximum budget per project: 60,000 euro
- Project duration: 6–9 months
- Applicant: SME or research organisation from a Horizon Europe eligible country

Project execution.

- Development
- Dissemination

1.5 Proposal Template

The templated collected the following information, in structured format.

- Project identification
- Applicant identification
- Project description
- Project team
- Budget
- Ethics self-assessment
- Detailed Budget
- Consent to process personal data
- Declaration of Honour

1.6 Evaluation

1.6.1 Programme committee

We organised a programme committee (PC) to carefully assess the proposals. Structure of PC:

- Coordinator: Wilker Aziz (UTTER/UVA).
- Manager: Maryam Hashemi (UTTER/UVA). Within our process the manager assists with formal checks as well as monitoring the mailing list, arranging payment for external reviewers, and various other tasks.
- Chairs: UTTER research personnel. Within our process a PC chair will a) assess proposals for adequacy, b) perform full qualitative reviews, c) invite external reviewers to contribute full qualitative reviews, and d) contribute to final decisions.
- External reviewers: experts with a PhD (or senior PhD candidates) who are not part of UTTER nor in active collaboration with UTTER personnel.

We refer to Coordinator and Chairs collectively as the Pilot Board.

1.6.2 Conflict of Interest (CoI)

There are two types of CoI that are relevant for our process:

- The applicant *is* in active collaboration with UTTER partners. We only welcome proposals that are free of this kind of CoI. Applicants self-declare this form of CoI, as part of the submission form, and the Coordinator verifies that no such CoI went unnoticed as part of a formal requirements check.

- The applicant *has been* in some form of collaboration with UTTER partners.
  We treat this as the regular kind of CoI in conference reviewing processes, namely, this form of CoI rules out the UTTER partner in conflict as a Chair for the proposal.

1.6.3 Criteria

We have three sets of evaluation criteria, all of which are clearly described in the call documentation Annex 4. They are summarised next.

Formal requirements. Every proposal must comply with a number formal requirements. These mostly ensure:

- the proposal is written in English, submitted on time using the proposal template;
- the applicant signed a DoH;
- the applicant's legal status fits the call;
- the applicant is legally established in a Horizon Europe eligible country;
- there's no CoI (i.e., no active collaboration with UTTER partners);
- the applicant filled in an ethical self-assessment.

Formal requirements are treated as Yes/No criteria, checked by the Manager and the Coordinator. All formal requirements must be met.

Eligibility criteria. Proposals that meet formal requirements are then assessed for their eligibility to the call, which we prescribed in terms of the following:

- *Relevance.* Does the proposal address the objectives of the call?
- *Uniqueness.* Does the proposal break new ground?
- *Completeness.* Does the proposal include both project phases (i.e., Development and Dissemination)?

These are treated as Yes/No questions, an eligible proposal meets all 3 criteria. Our goal is to filter out clear cases where the applicant misunderstands the goals of the call, or the proposal is obviously redundant (i.e., with clear known examples to support this assessment), or the proposal does not address (or address arguably superficially) the two project phases. Eligibility criteria are assessed by Chairs aiming at having high recall (i.e., we prefer to mark a borderline proposal as eligible than to mark a potentially eligible proposal as ineligible).

Qualitative criteria. Finally, eligible proposals undergo a complete qualitative assessment. The critearia are listed below.

- *Objective fit (2).* Are the project goals and planned achievements in line with the overall objectives of UTTER? Is it likely that the project will deliver added value to UTTER?
- *Technical approach (2).* Are the planned activities feasible and facilitate the achievement of project outputs? Does the proposal push the boundaries of existing XR technology?
- *Business, Integration and Dissemination (BID) plan (3).* Is the business plan reasonable and ambitious? How well is the integration of project outputs planned? Are the dissemination and promotion activities planned adequately?
- *Budget adequacy (1).* Does the budget correspond to all planned activities and outputs?
- *Team (1).* Is the applicant's team capable of executing the project and delivering its outputs (in required time, quality and with estimated budget)?
- *Ethics (1).* Is the ethical self-assessment thoughtful and thorough? Does it provide convincing justification that the applicant will ensure the work will be done ethically?

In bracket, for each criteria, we indicate their weight towards a final numerical score used for ranking. To map the qualitative criteria to a numerical score, we use the scale listed in Table 1.

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

Table 1: Rating scale

1.6.4 Procedure

The evaluation of FSTP proposals is divided in phases, which we describe next. Figure 1 gives an overview of the process.

Formal requirements. In this phase, we assess proposals on basic formal requirements. This check is conducted by both the Manager and the Coordinator.

Projects that do not comply with one or more formal requirements are marked for desk-rejection, with the following exceptions:

- Invalid or missing PIC
- Missing abstract

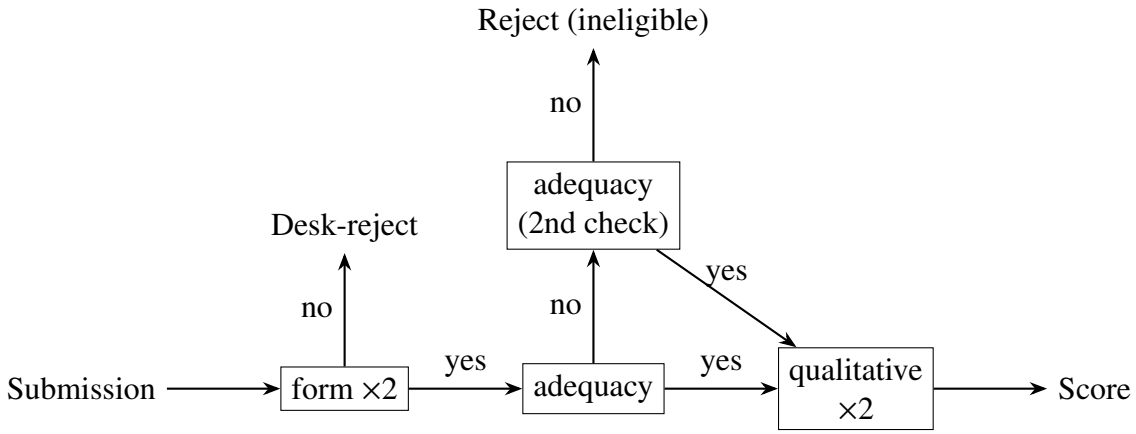


Figure 1: Overview of assessment phases (rectangles).

- Error in budget (e.g., wrong categories, typos, unexpected totals)

We assume errors of these sort signal a lesser degree of experience with FSTP calls and, given that these errors are easy to amend in case the proposal be (conditionally) selected for funding, we opted for not desk-rejecting such cases.

Eligibility checks. Proposals that pass the formal check are then assessed for their eligibility to the call. The Coordinator assigns proposals to Chairs observing two criteria: a) topic alignment (where possible), b) lack of CoI. At this point the Coordinator's assignment is based on educated guesses, hence, this first assignment is adjusted based on the Chair's self-declared CoIs. After these adjustments, each Chair manages a few proposals free of CoI (in our two FSTP calls, each chair cared for 4-5 proposals), assessing them along the three eligibility criteria for the call. Proposals flagged as ineligible (*i.e.*, failing at one or more of the three criteria) are then independently assessed by a second Chair (again, free of CoI). The proposal is marked for rejection in case both Chairs agree on the ineligibility outcome (not necessarily on the specific assessment of each criteria, but on the overall ineligibility outcome). In case of disagreement, we opted to regard the proposal as *eligible* and hence have it moved to the final phase of assessment.

Qualitative assessment. Eligible proposals receive a full qualitative assessment by two independent evaluators (both free of CoI), a Chair and an External Reviewer (external reviewers self-declare CoI). Each evaluator evaluates all individual criteria using qualitative rubrics, which are then mapped to points. They also provide free-text comments to corroborate their views. The points from all evaluators are averaged by criterion. Points by criterion are then multiplied by the criterion's weight and summed up in order to get the proposal's *overall score*. The Pilot Board can change the total number of points assigned to a proposal in the range of at most 30 points (up or down) of all the points the proposal received from the evaluators. The total overall score of an individual proposal is 130 points: maximum 100 points from evaluators + maximum 30 points from Pilot Board.

Decisions. We rank the proposals for quality and select those that receive highest scores for funding subject to i) not selecting more than the budget available for the call, and ii) not selecting

proposals that appear to be significantly worse than others being selected (e.g., should there be a clear top cluster followed by one or more second-tier clusters). With (i) in mind, we hoped to find 6–8 strong proposals forming a clear top cluster in each of our two FSTP calls.

1.6.5 Review Forms

We designed review forms on HotCRP with all necessary information for reviewers, as well as rubrics to guide their assessment. There are 3 forms:

- Formal Requirements – Appendix A Figure 11
- Adequacy to Call (that's the form for eligibility criteria) – Appendix Figure 12
- Qualitative Assessment – Appendix A Figures 13 (Key Parameters), 14 (Objective fit), 15 (Technical approach), 16 (BID), 17 (Team and Budget), 18 (Ethics and Comment for PC).

1.7 Execution

Each project was assigned a 'Sponsor', a contact person within UTTER, who oversees the project execution. At a minimum, the sponsor met with the project team 3 times: at kickoff, halfway through the project duration, and at the end. In preparation for the midterm and final meetings, the project team shared a progress report covering the following:[4]

- Project execution:
  - Deviations from plan
  - Development
  - Dissemination
  - Ethics
- Summary of results and plans
  - Results
  - Business plan
  - Future plans

The sponsor assesses the performance of the project team with respect to the proposed plan.

They received the funding in two instalments, one at the beginning and another at the end, the latter being conditioned on a positive assessment from the Sponsor.

[4] Template available at https://raw.githubusercontent.com/utter-project/fstp/main/2023/UTTER_FSTP1_Report_Template.pdf.

	First Call	Second Call
Budget available (EUR)	895, 206.00	424, 240.28
Open	July 31, 2023	May 1, 2024
Closed	October 15, 2023	July 31, 2024
Posted on EC funding and tender opportunities	July 31, 2023	May 1, 2024
Decisions communicated to authors	December 20, 2023	December 9, 2024
Decisions posted on UTTER's website	December 21, 2023	December 16, 2024
Submissions received	54	50
Projects selected for funding	8	7
Awarded funding (EUR)	470, 965.72	409, 643.75
Projects successfully concluded	8	7
Execution started	January 2024	February 2025
Execution ended	September 2024	August 2025
Remaining budget (EUR)	424, 240.28	14, 596.53

Table 2: Overview of the execution of UTTER's FSTP programme.

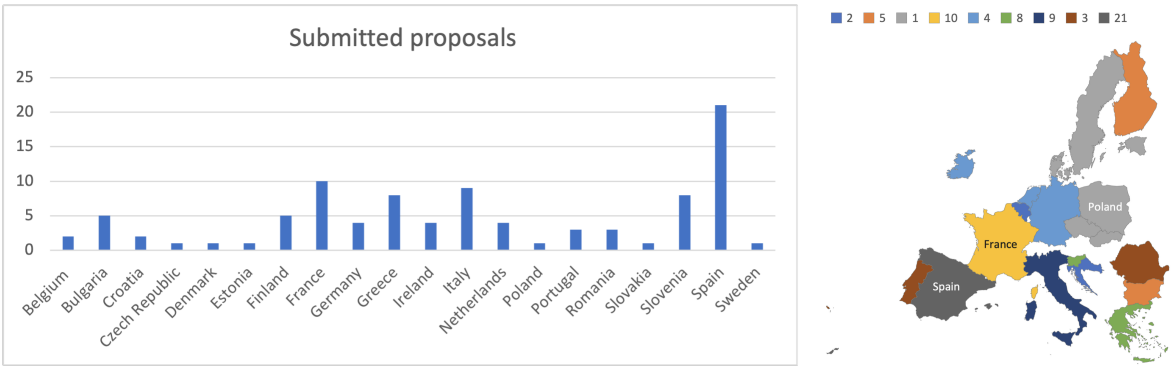


Figure 2: Number of proposals per country, submitted to our FSTP calls.

2 Overview of the Entire Programme

In UTTER, we split the budget of our FSTP programme in two calls. This section briefly summarises some facts about the two calls, so they can be appreciated side by side. The next two sections will detail each of the calls independently.

Table 2 is an overview of some key facts about the execution of the two calls. Figure 2 shows where we received proposals from. And, last, Figure 3 compares the quality of the proposals we received in the two calls.

All 15 projects (8 in FSTP1 and 7 in FSTP2), see Table 3, concluded successfully, as we report in later sections.

Call	Project	Recipient	Website
1	MaLA - Massive Language Adaptation of LLMs	University of Helsinki, Finland	https://huggingface.co/MaLA-LM/emma-500-llama2-7b
	PenGUIn - Prototype of an ExteNded reality Graphical User INterface	Re:LAB Srl, Italy	https://www.re-lab.it/projects/penguin
	HR-XR-XTEND - Croatian XR Extensions	University of Zagreb, Croatia	https://hr-xr-xtend.ffzg.unizg.hr/
	SignReality - Extended Reality for Sign Language Translation	DFKI GmbH, Germany	https://www.dfki.de/en/web/research/projects-and-publications/project/signreality
	DeMINT - Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts	University of Alicante, Spain	https://github.com/transducens/demint
	SURE-GB - Identifying under-representational, stereotypical and algorithmic gender bias in machine translation	ICCS-NTUA, Greece	https://ailswp.ails.ece.ntua.gr/suregb/
	InCroMin - Interactive Crosslingual Minutes	Charles University, Czech Republic	https://github.com/ELITR/incromin-test-calls
	pyannote.mobile	Université Toulouse III – Paul Sabatier, France	https://pyannote.ai/
2	DETOEX - DEtection of TOxic and hateful speech with EXplanations	Datoptron, Greece	https://github.com/ails-lab/detoex?tab=readme-ov-file
	Cognifit Harmony: Home-based Mixed Reality Therapy for Dementia	University of Modena and Reggio Emilia, Italy	https://github.com/ARSControl/app_HoloLens
	TEASE - TExt And Schematic for Education	LISN, France	https://gitlab.lisn.upsaclay.fr/nlp/corpora/tease
	INFINITY - Inclusive Networking Framework for Immersive AR Technology Integration	DASKALOS APPS, France	https://infinity-xr.netlify.app/
	VISIXR - Vision AI for XR	ZAUBAR UG, Germany	https://about.zaubar.com/en/blog/utter-x-zaubar-assistants-for-xr
	SwarmChat: Enabling Intuitive Human-Swarm Robot Conversation	Inventors Hub, The Netherlands	https://swarmchat.github.io/
	EOLAS: E-Learning Of Language Augmented Services	Walton Institute for Information and Communication Systems, Ireland	https://waltoninstitute.ie/projects/eolas

Table 3: Third-party projects funded under UTTER's FSTP programme.

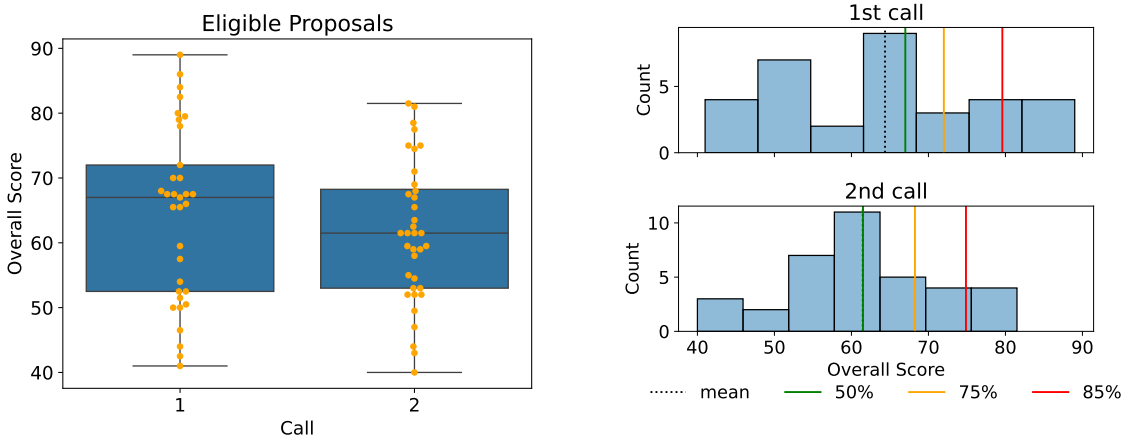


Figure 3: Distribution of Overall Score for Eligible proposals in our 2 FSTP calls. FSTP1 selected for funding the 8 proposals in the top cluster, FSTP2 selected the 7 proposals in the top cluster.

3 First Call

This entire section is copied from D1.2, if the reader is familiar with that document, this section can be safely skipped.

3.1 General

Infrastructure. We received submissions through https://utter-fstp.science.uva.nl from July 31, 2023 to October 15, 2023.

Dissemination. In accordance with the GA (MS6), we posted the first FSTP call on our own website on July 26, 2023 and on the European commission Funding and tender opportunities website on July 27, 2023.[5] We further advertised the FSTP call in a number of ways:

- On July5 , 2023 UTTER hosted its 1st User Day, a recording of which is accessible from https://www.youtube.com/watch?v=3Bm_3C9HrP0, there we advertised our first FSTP call;
- On September 15, 2023 UTTER joined a remote dissemination event targetting contemporary FSTP calls organised by Sploro,
  see https://sploro.eu/cascade-funding-opportunities-october-2023/;
- we advertised the call on social media:
  e.g., https://x.com/UTTERProject/status/1691031765530386432
  and https://x.com/bazril/status/1688468973522755584

Documentation. The complete documentation was linked from our website (see Figure 4) and from the EC website. For transparency, all documents are hosted on GitHub, including call documentation (and appendices), the grant and consortium agreements, as well as templates for reports: https://github.com/utter-project/fstp (see Section *First call (2024)* therein).

[5] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/competitive-calls-cs/3722

Figure 4: Screenshot of the call documentation package, as disseminated through our website on July 6, 2023.

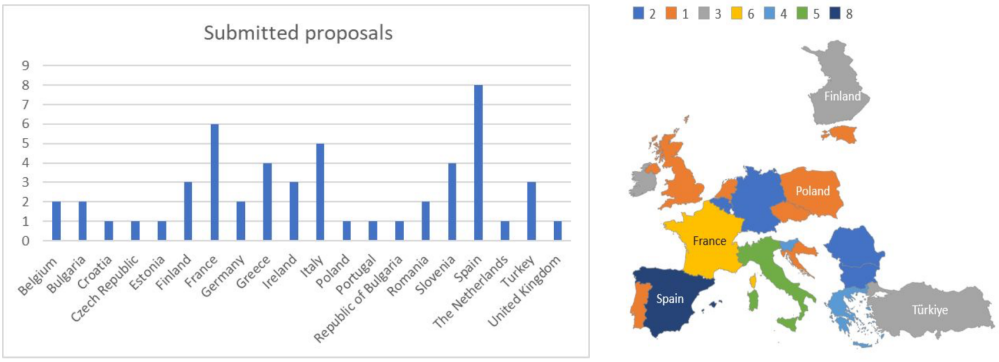


Figure 5: Number of proposals per country, submitted to our fist FSTP call.

Key parameters. No deviations from the key parameters as reported in Section 1.4. The total budget for FSTP in UTTER is 895, 206.00 EUR and we aimed at allocating about 50% of that to projects selected in our first FSTP call, hence leaving enough budget for a similarly-sized second call.

Selection. We selected 8 projects out of 54 submissions; the awardees were notified on December 20, 2023. See details in Section 3.2.

Execution. All projects started in January 2024 and were successfully completed by September 30, 2024, see details in Section 3.3

3.2 Selection

Here we summarise the outcome of our review process:

- Submissions: 54 (see Figure 5 for an overview of where they were submitted from).
- Desk-reject due to formal requirements: 5
- Ineligible after first check: 17
- Ineligible after second check: 16
- Qualitatively assessed: 33

Desk-rejected projects. 3 projects were desk-rejected because they failed to address the vast majority of required fields of the proposal template. 1 project was from an ineligible country, for this call. 1 proposal was a (likely accidental) duplicate—we only desk-rejected the copy (the original remained under consideration).

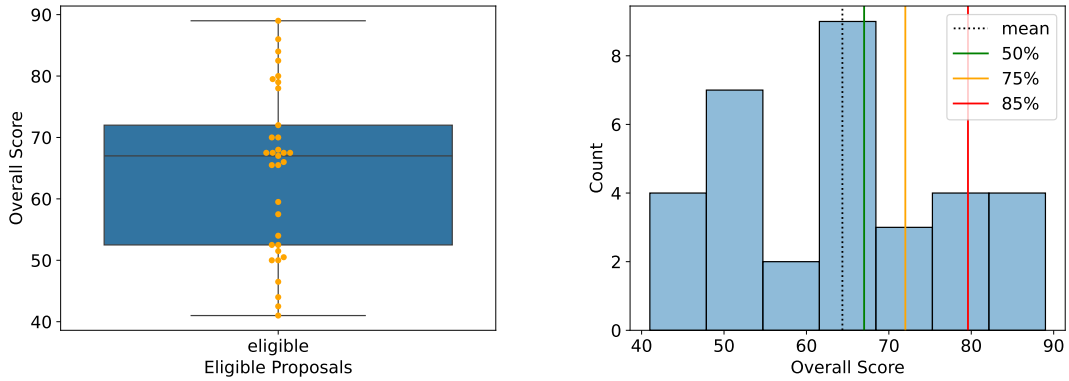


Figure 6: Distribution of Overall Score for Eligible proposals. The top cluster of 8 proposals were selected for funding.

Ineligible projects. 15 projects were judged to fail along the *Relevance* dimension, 9 projects were judged to fail along the *Uniqueness* dimensions, and 1 project was judged to fail along the *Completeness* dimension. One project was flagged as potentially ineligible in the first check and then considered potentially eligible in a second check—this was the only case were 2 Chairs disagreed, this project was then treated as eligible. These are the most frequent criticism that Chairs identified to support their assessment: *wrong XR / confused XR for VR* (10), *unclear impact of XR* (8), *unclear goals* (7), *unrealistic dependencies* (4).

Decisions. In total, 33 proposals received two complete qualitative reviews. In no case did the Pilot Board identify a need for making adjustments to scores. Hence, we ranked projects on their overall quality scores.

Selected for funding. We selected the top 8 proposals for funding, as that amounts to about half of the total budget allocated for FSTP calls in UTTER and those 8 proposals formed a cluster reasonably separated from the rest (see Figure 6).

Table 4 summarises the total funding requested.

	Number of proposals	Funding requested (EUR)
Proposals received	54	2 905 965
Eligible proposals	33	1 854 529
Selected proposals	8	470 966

Table 4: Number of proposals and funding requested.

Notification of decision. Applicants were notified of our decisions by email (sent from our HotCRP instance) on December 20, 2023 and published on UTTER's website on December 21, 2023. Besides the decisions, applicants received the complete feedback gathered throughout the evaluation procedure.

Complaints. We received one complaint via email concerning a desk-rejected submission. The applicant was based in the UK, which at the time was not considered an eligible country. Even though we had already consulted with the PO about this condition, prior to the beginning of the evaluation procedure (on August 21, 2023), we did consult our PO (on January 8, 2024) to address this complaint. The PO indeed confirmed the decision was sound, given the call documentation. This was communicated to the applicant, and the complaint was settled.

3.3 Projects and Results

Next, we introduce the projects, a summary of their key results and a qualitative remark from the Sponsor based on the project's final report and performance. The complete reports (attached to Appendix B) contain much more detailed information on project execution, results and plans.

3.3.1 MaLA - Massive Language Adaptation of LLMs

- Recipient: University of Helsinki (Research Organisation)
- Country: Finland
- Project duration: 9 months
- Funding Awarded (EUR): 56 273.00
- Website: https://huggingface.co/MaLA-LM/emma-500-llama2-7b

Project Description. This project explores language model adaptation across multiple languages and domains to improve human-machine interaction, especially for underrepresented languages. It aims to expand language models' capabilities by collecting and fusing data in over 500 languages in various domains, addressing challenges of language diversity. It delves into continual learning methods and adaptation techniques based on existing successful model architectures and open models, increasing the accessibility and applicability of large language models, particularly for low-resource languages.

Summary of Results. The UTTER FSTP has made significant strides in advancing multilingual language models with the creation of the MaLA corpus[6] and the development of the EMMA-500 model[7]. The MaLA corpus is a diverse dataset encompassing 939 languages, 546 of which were used to train EMMA-500, a cutting-edge multilingual model. EMMA-500 has demonstrated improved performance on various language tasks such as machine translation, commonsense reasoning, and text classification across multiple languages, including low-resource languages. A pre-print is available Ji et al. (2024).

[6] https://huggingface.co/collections/MaLA-LM/mala-corpus-66e05127641a51de34d39529
[7] https://huggingface.co/MaLA-LM/emma-500-llama2-7b

Recommendation by Sponsor. The goals of this project were to collect a massively multilingual corpus and use this to train an LLM supporting a large number of languages. This has been achieved, the MaLA corpus was released and the EMMA model created by fine-tuning Llama 2 7B on this corpus. The evaluation results show strong performance, especially in MT. The data and model have both been made publicly available and there is a preprint describing them on Arxiv.

3.3.2 PenGUIn - Prototype of an ExteNded reality Graphical User INterface

- Recipient: RE:LAB Srl (SME)
- Country: Italy
- Project duration: 9 months
- Funding Awarded (EUR): 57 395.00
- Website: https://www.re-lab.it/projects/penguin

Project Description. The aim of PenGUIn is to support user experience through an innovative, inclusive, adaptive and usable Graphical User Interface for XR platforms, and study the most appropriate information design framework to support agent tasks and the relative cognitive load in the presented UTTER's use cases - and beyond (e.g., virtual learning, virtual healthcare), to support the achievement of the task objectives. The proposed solution will converge innovativeness, usability and content design in a dynamic of functionality, effectiveness, and ergonomics, according to RE:LAB's methodology "Interaction Engineering". PenGUIn aims to design and test a user-centred design approach to prototype a library of graphical elements that are intuitive, cross-cutting, and compatible with usability criteria.

Summary of Results. The project, spanning 9 months, aimed to enhance user experience through an intuitive, inclusive, and adaptive Graphical User Interface (GUI) for online platforms. This was done by studying the most appropriate information design framework and applying suitable interaction strategies to support user's tasks in the context of two case studies: a customer assistant platform and an online meeting platform. PenGUIn's concept was driven by innovation and usability to achieve functionality, effectiveness, and ergonomic experience, building on RE:LAB's user-centric methodology, "Interaction Engineering". The purpose of PenGUIn's design effort was to guide the user through the multiple platforms' functionalities, from the multilingual translation to the AI-assistant. PenGUIn UI supported transparent and task-oriented dialogue and interaction between users of these virtual platforms. The project focused on customization flexibility, going through several design iterations, and validating the prototypes through expert analysis, focus group, and testing. The work carried out in the project has represented an additional opportunity to experiment RE:LAB original proposition and new research purposes, to consolidate the team expertise in creating and testing novel user experiences. The final prototypes are available as interactive demos: Customer Assistant Interface Prototype and Meeting Assistant Interface Prototype.

Recommendation by Sponsor. The project proposal was aiming at "design and test a user-centred design approach to prototype a library of graphical elements that are intuitive, cross-cutting, and compatible with usability criteria" considering the UTTER use cases. The work

resulted in two user interface prototypes that have been tested in focus groups to evaluate their usability. These user interfaces were made available as Figma templates that could be used as a base for developing graphical user interfaces using any desired front-end framework. The project delivered on what has been proposed. To the best of our knowledge the project has been disseminated on social media channels and in the company's webpage. The project team documented business plans and possible future works including possible opportunities collaboration with one of the institutions that belong to UTTER (NAVER). Based on this, the recommendation is to approve the final payment to the project Awardee.

3.3.3 HR-XR-XTEND - Croatian XR Extensions

- Recipient: University of Zagreb, Faculty of Humanities and Social Sciences (Research Organisation)
- Country: Croatia
- Project duration: 9 months
- Funding Awarded (EUR): 60 000.00
- Website: https://hr-xr-xtend.ffzg.unizg.hr

Project Description. The project is to develop a large language model (LLM) for the Croatian language and it will be trained on a massive dataset of Croatian text. The project is aligned with the objectives of the call, as it aims to build resources for XR models, extend XR models to new language, and evaluate XR models. The project goals are to collect at least 6 billion tokens of Croatian text and prepare that data for LLM training, create a LLM for the Croatian language using monolingual data only, and evaluate the LLM for downstream tasks. The experimental phase will focus on developing and evaluating the model architecture and training process. The integration phase will involve integrating the LLM into the UTTER platform. The dissemination phase will involve disseminating the project results to the research community and the public.

Summary of Results. The project aimed to create a large-scale monolingual Croatian language model (HR-GPT Beta). A significant training dataset was collected and cleaned from existing mono- and multilingual resources that include texts in Croatian. The preprocessing featured also advanced deduplication techniques, resulting in a final training dataset of 7.72 billion tokens. Three training scenarios were used: training from scratch, continued pretraining on a monolingual model, and continued pretraining on a multilingual model. The evaluation was performed using several benchmark datasets, and fine-tuning with the Alpaca dataset improved model performance. Larger models, like "gemma-7b", outperformed smaller ones, and fine-tuning enhanced results further. Key results include multiple model versions (160M, 350M, 410M, and 1.4B parameters) and a cleaned training dataset. Future work involves additional data collection, additional model training, further NLP task evaluations, and more training experiments. The HR-GPT Beta and training material (partially) will be publicly accessible under permissive licenses from the HR-CLARIN repository (https://clarin.hr). More information can be found on the project website.[8]

[8] https://hr-xr-xtend.ffzg.unizg.hr

Recommendation by Sponsor. The project planned the collection of Croatian datasets and training a large language model on Croatian and they delivered on this objective. This project was successfully disseminated. UTTER could use these datasets for training the EuroLLM language model. This project has completed successfully.

3.3.4 SignReality - Extended Reality for Sign Language Translation

- Recipient: DFKI GmbH (Research Organisation)
- Country: Germany
- Project duration: 9 months
- Funding Awarded (EUR): 59 995.22
- Website: https://www.dfki.de/en/web/research/projects-and-publications/project/signreality

Project Description. SignReality will create a 3D sign language interpreter displayed in Augmented Reality glasses. It will serve as an extension of the UTTER online/hybrid interfaces, aiming at usability and accessibility for deaf and hard-of-hearing people. The app will be based on an XR model consisting of a pre-trained sequence-to- sequence neural network, connected to a framework for geometrical transformations for synthesizing an animated avatar. This will follow a client-server architecture, connected with the SDK of the AR device and via an API to other apps. Participatory design and evaluation in co-operation with the user community is planned. Results will be disseminated to the user and scientific communities, to UTTER and parallel research projects and will be used to initiate further research.

Summary of Results. The project achieved significant milestones in bridging sign language technology with Extended Reality. Key results include the development of an engine for avatar animation, accompanied by device-specific implementation on two AR devices (Hololens 2 and XReal Light). Translation from spoken language to a textual sign language representation (German → DGS) was enabled through an encoder-decoder translation model, whereas further improvement of relevant models will benefit from the work on corpus acquisition and alignment. The implementation was tested for intelligibility, user experience and acceptance in a user study with native sign language users at the Centre for Culture and Visual Communication of the Deaf Berlin Brandenburg (ZFK) providing valuable feedback. The project has been integrated into several academic theses and university workshops, and research findings will be submitted in relevant academic venues.

Recommendation by Sponsor. The project clearly achieved all of its goals, with very minor deviations from the original plan, both along the scientific and dissemination dimensions. The project team has experience with the ethical considerations behind the experimental setup and did a remarkable job both at complying with all relevant guidelines and regulations but also at clearly documenting the scope of their findings and technology.

3.3.5 DeMINT - Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts

- Recipient: University of Alicante (Research Organisation)
- Country: Spain
- Project duration: 9 months
- Funding Awarded (EUR): 57 567.50
- Website: https://github.com/transducens/demint

Project Description. This project focuses on developing an AI chatbot to serve as a tutoring assistant for non-native English speakers, enhancing their language skills through post-meeting analysis of meeting transcriptions. This effort aligns with UTTER's objectives, particularly its interest in harnessing language models for video conferencing applications. Following recent advances in LLM-based chatbots and agents, our system will exploit pre-trained large language models, refined for the tutoring task through a mix of in-context learning, external non-parametric memory retrieval, efficient parameter fine-tuning, and tool exploitation. Human evaluation will be conducted through individual debriefings after simulated, scenario-based video conferences with small test groups.

Summary of Results. DeMINT has developed a prototype of a conversational system designed to enhance non-native English speakers' language skills through post-meeting analysis of the transcriptions of video conferences in which they have participated. The code of the system is already available as open-source software on https://github.com/transducens/demint, and a paper (Pérez-Ortiz et al., 2024) has been published at the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning. Future plans include developing a more engaging and speech-based interaction with the chatbot and knowledge from theories of second language acquisition.

Recommendation by Sponsor. The DeMINT team has done an excellent job at delivering all results promised in the project proposal. A publication at a relevant workshop and an open-source codebase for the tool developed have been disseminated. The project has not deviated in any major way from what was proposed. The project sponsor, therefore, makes a positive payment recommendation for the DeMINT project.

3.3.6 SURE-GB - Identifying under-representational, stereotypical and algorithmic gender bias in machine translation

- Recipient: Institute of Computer and Communication Systems, National Technical University of Athens, ICCS-NTUA (Research Organisation)
- Country: Greece
- Project duration: 9 months

- Funding Awarded (EUR): 60 000.00
- Website: https://ailswp.ails.ece.ntua.gr/suregb/

Project Description. SURE-GB aims to build an automated service that identifies occupation-related under representational, stereotypical, and algorithmic gender bias in machine translation, in English and French, as well as low resource languages like Greek. The proposed method involves creating a curated knowledge graph that a) encodes standardised knowledge and data for occupations (based on data and hierarchies from EU- LFS1, the ESS2, and the International Classification of Occupations-ISCO3), b) incorporates statistics for occupation-related gendered language usage derived from linguistic corpora. Our goal is to develop a ready-to-use machine learning toolkit, that utilises the above knowledge to detect and categorise gender biases for: a) providing actionable recommendations for improvement, b) establishing guidelines for unbiased language translation, c) raising awareness of gender biases in machine translation systems.

Summary of Results. The SURE-GB project has made significant strides in understanding and addressing occupational gender biases in machine translation (MT) systems. Our research has resulted in the creation of a curated Knowledge Graph that encapsulates essential statistics on gender representation in various occupations across Greece, France, and the UK, as well as linguistic biases in corresponding textual corpora. We have developed an automated system to detect and classify occupational gender bias in MT systems. By revealing the disparities between real-world statistics and their representation in MT systems, we have identified critical flaws in current technologies and paved the way for more equitable and accurate translations. Through this project, we enable a more nuanced study of machine learning biases by disentangling the real world, the data, and the systems, while still recognizing their interconnectedness. Explore our findings through our website.

Recommendation by Sponsor. Given the clear achievements, adherence to the proposal, and the potential for future impact and expansion, the SURE-GB team has shown more than satisfactory performance and already produced several outputs. Hence, it is recommended that the SURE-GB project receives the final funding part as originally planned.

3.3.7 InCroMin - Interactive Crosslingual Minutes

- Recipient: Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics (Research Organisation)
- Country: Czech Republic
- Project duration: 9 months
- Funding Awarded (EUR): 59 826.25
- Website: https://github.com/ELITR/incromin-test-calls

Project Description. The goal of the project is to (1) expand existing interactive meeting summarization tools (such as our MinuteMan, or those developed by UTTER) to facilitate cross-lingual access to meeting content (live transcripts and live minutes) and (2) make these tools benefit from human interpretation, if available in the meeting. As a necessary prerequisite, the project will prepare a test set and rigorously evaluate the underlying models of speech transcription, translation and summarization in this setting.

Summary of Results. InCroMin examined and carefully evaluated the applicability of recent state-of-the-art speech-to-text translation tools in real cross-lingual calls, i.e. calls between parties that do not have a common language. The project adapted MinuteMan (https://github.com/fkmjec/minuteman) for this purpose and collected a corpus of such calls. The deidentified part of the corpus is available here: https://github.com/ELITR/incromin-test-calls. Additional results of InCroMin include an evaluation of latency metrics for speech translation, translation of ELITR-Bench (https://github.com/utter-project/ELITR-Bench) into Czech to allow evaluation of cross-lingual access to past meeting content or translation of a part of MultiWOZ dialogues into Czech and German to assess translation quality of dialog-critical features such as participants' gender preservation. All the outputs are detailed in InCroMin Final Report.

Recommendation by Sponsor. The project exceeded expectations. In summary, they i) adapted MinuteMan to support cross-lingual calls, ii) collected a new and potentially valuable corpus of simulated cross-lingual meetings, and iii) conducted practical tests to assess the usability of the extended MinuteMan and identified areas for improvement. For well-supported languages, MinuteMan appears close to being fully operational. Additionally, the potential founding of a spin-off for MinuteMan is under consideration, with FSTP funding playing a crucial role in bringing the system closer to production-ready. Finally, even though it wasn't initially planned, InCroMin developed a Czech version of the ELITR-Bench meeting, which will soon be added to the UTTER/ELITR-Bench repository. This could also spark future collaboration between Naver and Charles University on cross-lingual QA on long documents (meeting transcripts).

3.3.8 pyannote.mobile

- Recipient: Université Toulouse III – Paul Sabatier (Research Organisation)
- Country: France
- Project duration: 9 months
- Funding Awarded (EUR): 59 908.75
- Website: https://pyannote.ai

Project Description. pyannote.mobile aims at extending pyannote speaker diarization open-source toolkit in two complementary directions. The first one is to add streaming speaker diarization support, as it currently only supports offline/batch processing. The second one is to investigate the feasibility of "on device" streaming speaker diarization (as opposed to cloud-based processing): we will develop a streaming speaker diarization proof-of-concept running on mobile

(iOS or Android). For both directions, we will aim for the best compromise between accuracy and (algorithmic and computational) latency.

Summary of Results. pyannote.mobile project led to the extension of the pyannote.audio open-source speaker diarization toolkit to perform speaker diarization in real-time while controlling the trade-off between latency and accuracy. It also led to the creation of an iOS/macOS streaming speaker diarization SDK which will be handed over to interested parties through the local university tech transfer office.

Recommendation by Sponsor. As the sponsor of this project, we confirm that the project successfully delivered its planned results. The dissemination efforts were effective, including an iOS application soon to be available, and a scientific paper published at Interspeech 2024. The project lead also provided comprehensive documentation and effective communication throughout the duration of the project, which were appropriate and well-aligned with the project's objectives. Overall, we recommend this project positively, as it has met its key objectives and demonstrated potential for future impact.

4 Second Call

4.1 General

Infrastructure. We received submissions through https://utter-fstp.science.uva.nl from May 1, 2024 to July 31, 2024.

Dissemination. In accordance with the GA (MS15), we posted the first FSTP call on our own website and on the European commission Funding and tender opportunities website on May 1, 2024.[9] We further advertised the FSTP call in a number of ways:

- On July5 , 2024 UTTER hosted its 2nd User Day, a recording of which is accessible from https://www.youtube.com/watch?v=-ZhoLLhQYcc, there we advertised our second FSTP call;
  https://www.eventbrite.com/e/utter-user-day2024-tickets-936485271657?aff=oddtdtcreator
- we advertised the call on social media:
  e.g., https://www.linkedin.com/feed/update/urn:li:activity:7221877511444451329/
  https://x.com/UTTERProject/status/1795825409860493402
  https://x.com/UTTERProject/status/1816121358876655858

Documentation. The complete documentation was linked from our website (see Figure 7) and from the EC website. For transparency, all documents are hosted on GitHub, including call documentation (and appendices), GA, CA and templates for reporting: https://github.com/utter-project/fstp (see Section *Second call (2024)* therein).

[9] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/myarea/projects/cascade-funding-details/details/6464/43108390/101070631

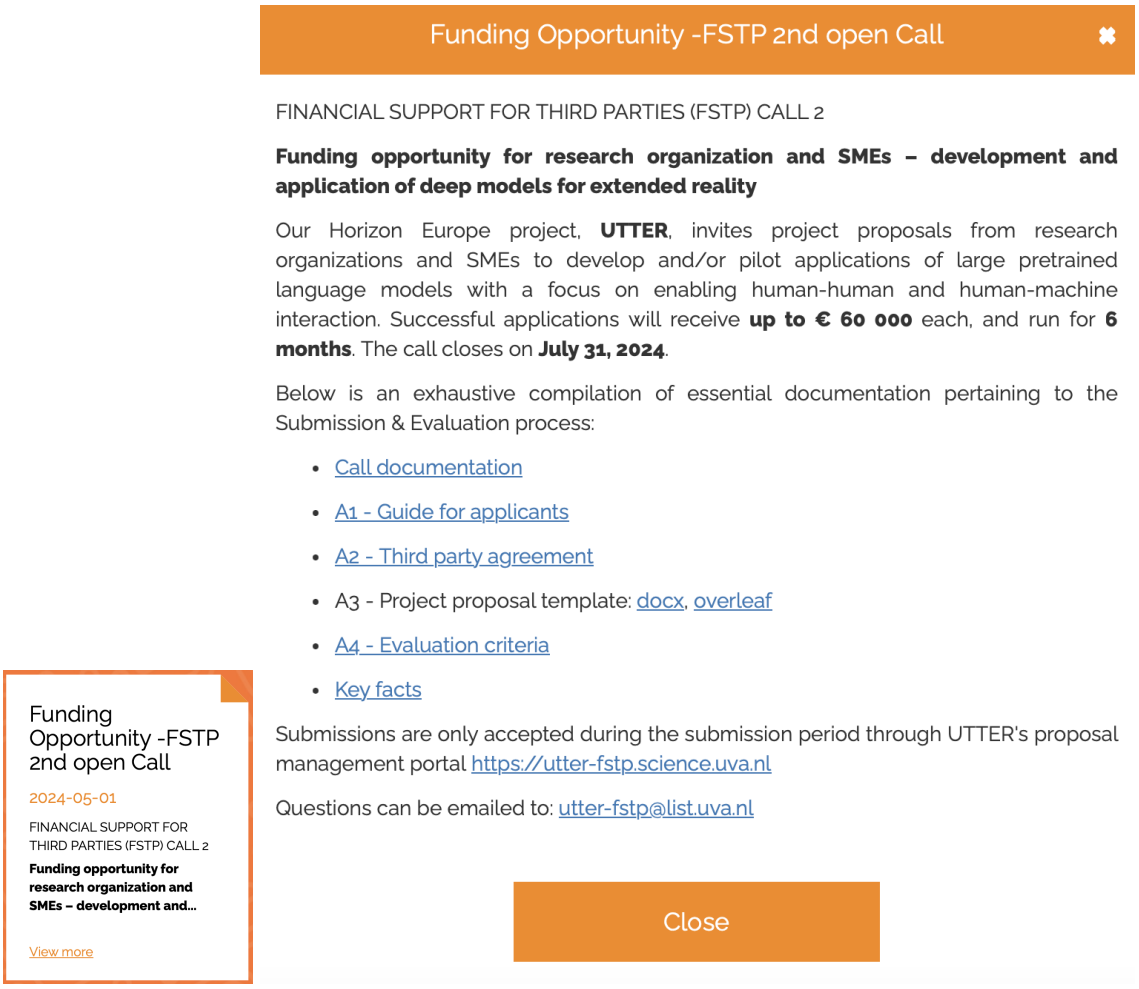


Figure 7: Screenshot of the call documentation package, as disseminated through our website on May 1, 2024.

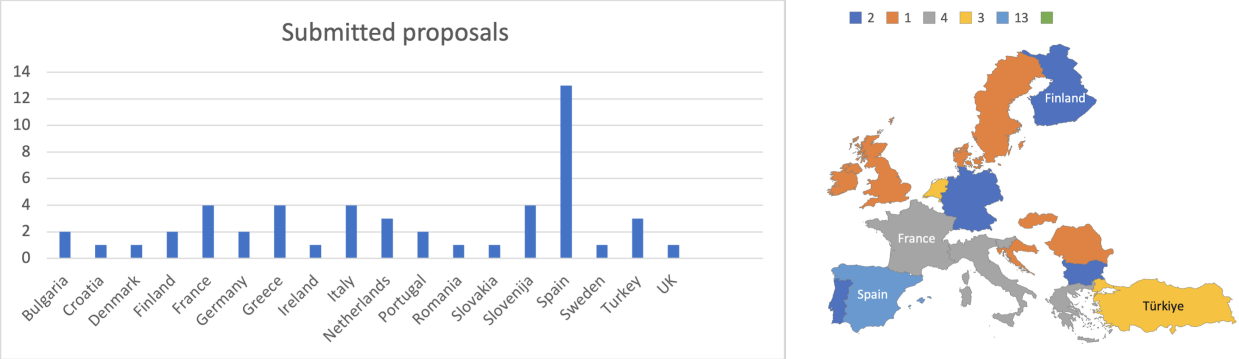


Figure 8: Number of proposals per country, submitted to our second FSTP call.

Key parameters. The only deviation from the key parameters as reported in Section 1.4 is that in our second call we limited project duration to 6 months (rather than 6–9). After the first call, the remaining budget available for FSTP was 424, 240.28 EUR and we aimed at allocating all of it to projects selected in our second FSTP call.

Selection. We selected 7 projects out of 50 submissions; the awardees were notified on December 9, 2024. See details in Section 4.2.

Execution. All projects started in February 2025 and were successfully completed by August 30, 2025, see details in Section 4.3.

4.2 Selection

Here we summarise the outcome of our review process:

- Submissions: 50 (see Figure 8 for an overview of where they were submitted from).
- Desk-reject due to formal requirements: 1
- Ineligible after first check: 15
- Ineligible after second check: 13
- Qualitatively assessed: 36

Desk-rejected projects. 1 project was desk-rejected because the application failed to identify a valid *Legal Form* for the applicant (in *Section 2 Applicant Identification* of the proposal template)[10] as one of the two eligible options (namely, SME or Research Organisation) for this FSTP programme. See Figure 9 for a screenshot of the relevant field of the proposal template.

[10]Template: https://raw.githubusercontent.com/utter-project/fstp/main/2024/UTTER_FSTP2_A3_Proposal_Template.docx

APPLICANT IDENTIFICATION (required)	
Organisation	Text (max. 100 characters) (Name of the organisation)
National VAT Number	Text (max. 100 characters)
Year of Foundation	Text (max. 100 characters)
Number of Employees	Text (max. 100 characters)
Legal Form	Selection Entry (SME or Research Organisation)
Turnover of last fiscal year	Text (max. 100 characters) (For SMEs only)

Figure 9: A screenshot of the relevant part of the proposal template. The highlight is not part of the template, but used here to draw attention to the field relevant to the desk-rejected proposal.

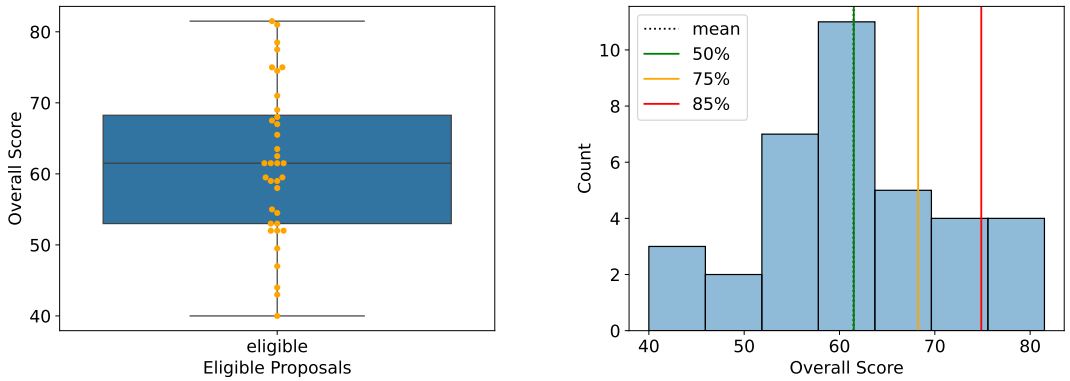


Figure 10: Distribution of Overall Score for Eligible proposals. The top cluster of 7 proposals were selected for funding.

Ineligible projects. 11 projects were judged to fail along the *Relevance* dimension, 3 projects were judged to fail along the *Uniqueness* dimensions, and 6 projects was judged to fail along the *Completeness* dimension. Two projects were flagged as potentially ineligible in the first check and then considered potentially eligible in a second check—these projects were then treated as eligible.

Decisions. In total, 36 proposals received two complete qualitative reviews. In *one* case the Pilot Board motivated an adjustment of +8 points, in order to calibrate the harshness of the external reviewer along *objective fit* and *technical approach*. After this one adjustment, we proceeded to rank projects on their overall quality scores.

Selected for funding. We selected the top 7 proposals for funding, and that was the most we could fit within the total budget left for FSTP2. Besides, and still importantly, those 7 proposals formed a cluster reasonably separated from the rest (see Figure 10).

Table 5 summarises the total funding requested.

Notification of decision. Applicants were notified of our decisions by email (sent from our HotCRP instance) on December 9, 2024 and published on UTTER's website on December 16,

	Number of proposals	Funding requested (EUR)
Proposals received	50	2 977 959 185
Eligible proposals	36	2 137 148 185
Selected proposals	7	409 644

Table 5: Number of proposals and funding requested.

2024. Besides the decisions, applicants received the complete feedback gathered throughout the evaluation procedure.

Complaints. On December 11, 2024, we received a complaint by email. Their proposal had been desk-rejected due to an invalid Legal Form. In their email, the applicant argued that they indeed qualified as an SME, despite having filled the form differently—the applicant self-declared *Einzelunternehmer*—for the relevant field (shown in Figure 9). On December 18, 2024, and again on December 19, 2024, we reaffirmed and communicated our rationale for desk rejecting the proposal. We double-checked the rationale and decision internally and also with our PO.

4.3 Projects and Results

Next, we introduce the projects, a summary of their key results and a qualitative remark from the Sponsor based on the project's final report and performance. The complete reports (attached to Appendix C) contain much more detailed information on project execution, results and plans.

4.3.1 DETOEX - DEtection of TOxic and hateful speech with EXplanations

- Recipient: Datoptron (SME)
- Country: Greece
- Project duration: 6 months
- Funding Awarded (EUR): 59750
- Website: https://github.com/ails-lab/detoex?tab=readme-ov-file

Project Description. DETOEX presents a novel approach to hate speech detection that combines large language models (LLMs) with a curated vocabulary of derogatory language and traditional NLP techniques. The method focuses on identifying language that is offensive or derogatory toward groups or their members based on identity-related characteristics or beliefs. It also provides contextualized explanations for why certain expressions are considered offensive. The system integrates outputs from two complementary pipelines. The first is a term-based pipeline designed to detect terms that are inherently offensive towards certain groups of people. To support this, a vocabulary of toxic terms and accompanying usage descriptions is developed and used to guide and ground an LLM that disambiguates the use of the term within a specific context. The second pipeline focuses on identifying expressions explicitly directed against groups or individuals defined

by particular traits or beliefs. DETOEX has been implemented to analyze text in Greek, French, and English. Its outputs - both the detection decisions and the accompanying explanations - are evaluated by human participants. The feedback is further analyzed to draw insights into the tool's accuracy, potential biases, and the inherently subjective nature of the task.

Summary of Results. The project has led to the following main outputs:

- Three vocabulary with politically-charged toxic terms and accompanying descriptions in Greek, English and French.
- The tool for detecting politically-charged toxic language in Greek, English and French, made available as source code, Docker container and API.
- The evaluation results, providing information about the perception of the tool's results by humans and comparative insights about the use of different LLMs across the three considered languages.
- A scientific paper describing the overall methodology and achieved results.

The project's results have been disseminated in a number of ways. Scientific dissemination: a scientific paper has been written and will be soon submitted to a peer-reviewed venue. Dissemination to the IT and research communities: a) the source code of the detection system has been documented and made available on GitHub (it is also offered as a Docker container with the tool's functionalities exposed via an API); b) the multilingual semantic vocabularies are published on Zenodo, to facilitate its further reuse. Communication to broader audiences: the HomoDigitalis mailing list has been used to inform its network's members about the project results, a virtual event (in Greek) to inform members of the HomoDigitalis network about the project's objectives and engage participants into the evaluation activities has been organised (the recording is available here). The online evaluation process has also acted as a means of outreach, by mobilising more than 15 participants in the process of detecting and understanding different types of hate speech. A LinkedIn post informing about the project's paper has also been made.

Recommendation by Sponsor. The DETOEX project set out to build a multilingual resource and system for detecting politically-charged toxic language in three languages, Greek, French, and English. All contractual milestones were reached within the six-month window, accounting for the switch in the human evaluation campaign format, *i.e.* the decision to switch from the Crowd-Heritage platform to the use of formatted spreadsheets. Future iterations should broaden language coverage, benchmark against generic toxicity detectors, and release the annotation materials to maximise community benefit.

4.3.2 Cognifit Harmony: Home-based Mixed Reality Therapy for Dementia

- Recipient: University of Modena and Reggio Emilia (Research Organisation)
- Country: Italy
- Project duration: 6 months

- Funding Awarded (EUR): 60000
- Website: https://github.com/ARSControl/app_HoloLens

Project Description. Cognifit Harmony is a Mixed Reality (MR) solution designed to promote active ageing by combining physical and cognitive training for elderly individuals, including those with early cognitive decline. Developed using Microsoft HoloLens 2 and tools like Unity, MRTK, and World Locking Tools, the system includes three interactive MR games for physical engagement, along with an LLM-powered storytelling module for cognitive stimulation. The project included a home-based user study involving elderly volunteers to assess usability, comfort, and feasibility.

Summary of Results. Results were highly promising: participants found the system intuitive, non-intrusive, and engaging, even when used in domestic environments. Usability and workload assessments using NASA-TLX and SUS scales confirmed the system's accessibility and user-friendliness. Dissemination efforts include a peer-reviewed publication accepted at the IEEE RO-MAN 2025 conference: Gabbi, M., Villani, V., Sabattini, L. "Towards User- Friendly MR Solutions for Cognitive and Motor Stimulation in Active Ageing". The MR application source code is publicly available at: https://github.com/ARSControl/app HoloLens Although no immediate commercialization is planned, future work includes extended testing with users affected by cognitive impairment and the integration of adaptive features for broader accessibility. The project highlights the potential of MR and AI-powered systems for home-based digital therapies and age-friendly innovation.

Recommendation by Sponsor. The project achieved all of its goals both in terms of results and dissemination. I see it as successfully completed.

4.3.3 TEASE - TExt And Schematic for Education

- Recipient: Laboratoire Interdiscplinaire des Sciences du Numérique (LISN) (Research Organisation)
- Country: France
- Project duration: 6 months
- Funding Awarded (EUR): 60000
- Website: https://gitlab.lisn.upsaclay.fr/nlp/corpora/tease

Project Description. The project TEASE aims to produce a new corpus of text-image for the development of visual-textual question answering for the education domain. education. Especially the project leverages automatic annotation mechanisms and LLMs as judge methods. The work is particularly focused on evaluation and how models can align to human judgment.

Deviations from Original Plan. The original objectives have not been fully met, mostly in light of unforeseen blockers, which consumed more time than the team had envisioned. In particular, the team is still working to advance two of their objectives: (a) the creation of a high-quality question- answering dataset and (b) methods for automatic evaluation. Nonetheless, the team showed transparency and agility to plan accordingly, and the project still developed to satisfactory results.

Summary of Results. Currently, we do not have exploitable results; we cannot yet share the question-answering corpus. At the time, we selected the different resources we planned to use for the corpus; we produced the first annotation subset by leveraging generative visual language models, and we started experimenting with automatic evaluation approaches. Nonetheless, we will present the project and the work already completed to the conference TALN5.[11]

Recommendation by Sponsor. This project broadly delivered on its intended outcomes. The dissemination results (a data- set, an automatic evaluation procedure and research paper) are available in a preliminary unpublished format in the form of this report and a privately shared download link to the present version of the dataset. The current project results will also be presented on July 4th at the CORIA-TALN conference. The project team also intends to make the dataset available to the broader public and a submit a research paper in the near future, continuing development with a trainee hired outside of UTTER funding. The results of this project could be useful for the training and evaluation of multimodal systems in UTTER. The project sponsor recommends this project for payment.

4.3.4 INFINITY - Inclusive Networking Framework for Immersive AR Technology Integration

- DASKALOS APPS (SME)
- Country: France
- Project duration: 6 months
- Funding Awarded (EUR): 59250
- Website: https://infinity-xr.netlify.app

Project Description. The INFINITY project successfully develops a multimodal AI-driven translation system integrated within an immersive XR environment, making multilingual communication more accessible for deaf, hard-of-hearing, and linguistically diverse individuals.

Summary of Results. The INFINITY project has successfully progressed towards the development of an AI-driven multilingual translation system integrated into an Extended Reality (XR) environments, making language learning more accessible for deaf, hard-of-hearing, and linguistically diverse individuals. The system combines real-time speech-to-text transcription,

[11] The team shared proof of progress with the Sponsor. In order not to compromise their potential for future exploitation, this was shared on a private basis, but, for now, they cannot be directly available from this report.

text-to-text translation using Facebook's No Language Left Behind (NLLB) model, and International Sign Language (ISL) interpretation. Users are able to interact with AI-powered avatars in VR learning scenarios, where both spoken and sign language translations are provided, enhancing engagement and inclusivity. To disseminate the project's results, INFINITY was showcased at TEDx Patras, highlighting the role of AI in breaking communication barriers. A peer-reviewed scientific paper has also been accepted and was presented as POSTER presentation at Salento XR 2025 https://www.xrsalento.it/ ("AI-Driven Sign Language and Multilingual Translation for Inclusive XR Learning"; June 17–20, 2025), providing an in-depth look at the system's development and impact. The solution was also being pitched to technology leaders at startup summits in Paris (https://vivatechnology.com/) and London (https://londontechweek.com/) in June 2025, promoting partnerships for commercial adoption. Additionally, ongoing dissemination efforts include a dedicated project website, social media engagement on LinkedIn (https://www.linkedin.com/company/infinity-xr-app/about), and regular updates on AI-driven accessibility innovations. From a business perspective, INFINITY is exploring pay-per-use and monthly licensing models to integrate the system into language schools, and professional training centres. The goal is to provide a scalable and sustainable solution while offering continuous support for educators and students. Moving forward, the focus will be on expanding adoption, integrating in XR environment the AI models and optimising the learning experience to further enhance accessibility in multilingual education.

Recommendation by Sponsor. The Infinity project has successfully delivered its main expected outcomes by developing the first version of an AI-based multilingual communication tool within an immersive XR environment. The team effectively disseminated results through multiple channels, including publications, presentations at leading startup summits, a dedicated project website, and key partnerships. They have also outlined clear future plans, including improvements to avatar generation and active pursuit of new collaborations. I don't have access to their application, but I have read their paper, and watched their very short video that they have published on their youtube channel. I think that, given the difficulty of the task and the number of modules involved in building the app, they have successfully met the criteria for me to approve their project. I therefore recommend proceeding with the payment of the remaining portion of the FSTP funding for this project.

4.3.5 VISIXR - Vision AI for XR

- ZAUBAR UG (HAFTUNGSBESCHRAENKT) (SME)
- Country: Germany
- Project duration: 6 months
- Funding Awarded (EUR): 56088
- Website: https://about.zaubar.com/en/blog/utter-x-zaubar-assistants-for-xr

Project Description. This project introduces an innovative XR chatbot in the domains of customer support and product demonstrations through advanced AI and image analysis technologies. The system combines sophisticated image segmentation, vision-based AI analysis, and a dynamic

knowledge base to create an interactive, visually-driven customer experience. Users can explore product images through clickable segments, interacting with a customizable spatial agent in a web-based interface. By addressing critical gaps in current e-commerce and customer support solutions, the project introduces immersive, context-aware interactions that set new standards for engaging, informative, and personalized customer support in the digital age. Initially web-based, the system is designed for future expansion into full XR experiences with 3D models and specialized hardware, demonstrating the practical and powerful application of XR principles in everyday commercial interactions.

Summary of Results. The project has yielded significant results, culminating in an integrated, end-to-end system that demonstrates the core vision of the product. The most critical result is the successful development of a functional prototype that combines a sophisticated back-end AI pipeline with an interactive front-end user experience. This includes a fully operational workflow for image segmentation where images are uploaded, analyzed by an AI to identify key features, segmented using Grounded-SAM, and made available for interaction. A key technical achievement is the real-time multimodal interaction pipeline, which effectively processes user speech, understands user intent, queries a dynamic knowledge base with a high-performance LLM, and generates a spoken response, creating a fluid conversational experience. The successful integration with a Unity-based WebGL front-end makes the experience accessible on a wide range of devices and demonstrates the feasibility of delivering XR-like interactions through a standard web browser. The iterative testing and optimization cycles led to important improvements, such as switching to more efficient models and refining the LLM pipeline to significantly reduce response latency. These results validate the project's technical approach. The project dissemination so far included:innovation days https://v-i-r.de/events/vir-online-innovationstage/, public outreach (Open House at the Federal Ministry for Economic Affairs; https://www.bundesregierung.de/breg-de/schwerpunkte/tag-der-offenen-tuer/tag-der-offenen-tuer-2373830) and on social media (LinkedIn) and https://about.zaubar.com/en/blog/utter-x-zaubar-assistants-for-xr).

Recommendation by Sponsor. The VISIXR team completed all the work as promised and delivered the expected results. The project developed an AI-powered tool that lets users interact with product images in real time, providing instant information and support directly through a web interface. The software meets the requirements and was made available for testing on 4 August 2025 via the following demo link: https://zaubar.dev/spatialagent?image=498. The project achieved its goals, with the developed solution working as planned and matching the initial objectives.

4.3.6 SwarmChat: Enabling Intuitive Human-Swarm Robot Conversation

- Inventors Hub (SME)
- Country: The Netherlands
- Project duration: 6 months
- Funding Awarded (EUR): 60000
- Website: https://swarmchat.github.io

Project Description. SwarmChat is an open-source, multilingual pipeline that converts everyday speech or text into validated XML behaviour trees (BTs) and streams them to swarms of autonomous robots. Built under the EU-funded UTTER programme, it couples state-of-the-art transcription (SeamlessM4T / EuroLLM), safety filtering (Llama-Guard), and a LoRA-fine-tuned Falcon-10B behaviour-tree generator..

Summary of Results. A variery of models and benchmark results, covering all the component parts of the system. The prototype achieved full operation; the full system transforms spoken or typed commands in nine EU languages into validated XML behavior trees (BTs) and executes them within a 50-agent simulator. The project has been widely disseminated on social media (LinkedIn, X), company website, medium, YouTube, and code and models hosters (GitHub, Huggingface). A paper for scholarly dissemination is in preparation.

Recommendation by Sponsor. At midterm, the project was very well on-track, EuroLLM9B model is used in the pipeline and a demo is planned at the end of the project. Bu the end-of-project: all key milestones have been achieved. The full system now transforms spoken or typed commands in nine EU languages into validated XML behavior trees (BTs) and executes them within a 50-agent simulator. I watched a video of the demo prototype and also tested it myself. It worked well, although it was a bit slow due to being hosted on Hugging Face's basic (free) infrastructure. Finally, the FSTP team provided correct documentation of the ethical implications and risks associated with their project. Overall, the report and presentation were both clear, and all aspects of the project—model, demo, and blog posts—were made readily accessible in the dissemination materials. I recommend proceeding with the payment of the remaining portion of the FSTP funding for this project.

4.3.7 EOLAS: E-Learning Of Language Augmented Services

- Walton Institute for Information and Communication Systems Science (Research Organisation)
- Country: Ireland
- Project duration: 6 months
- Funding Awarded (EUR): 59688
- Website: https://waltoninstitute.ie/projects/eolas and https://github.com/Ian-Mills/EOLAS

Project Description. E-Learning Of Language Augmented Services (EOLAS) is aimed at teaching of the Irish language through an XR enhanced platform. Through a fusion of LLM interaction and object recognition technologies to identify and translate everyday objects into centre of the conversation, EOLAS delivers a novel means of language learning open to all learning levels. EOLAS is bolstered by the addition of an open source grammatical database, enabling the LLM to become your own personal language tutor. With EOLAS, which in Gaeilge (Irish) means knowledge, we leverage an Irish Large Language Model (LLM) to facilitate learning through interaction. EOLAS creates a dynamic and interactive language learning environment, making Irish language education more engaging and effective for learners of all levels.

Summary of Results. EOLAS has developed an AR application which allows the end user to observe the world around them, highlight a particular object and receive the translated recognition object. EOLAS's object recognition model is based on a vision transformer model called FastViTMA36F16 and is designed to take input from the user selection and interpret the contents of the image. The Eolas LLM can then use this input to translate and provide additional sentences/information for the recognised object. The output from the translation is visible on screen to the end user and a chat bar is available to send additional queries about the recognised output. For dissemination, EOLAS produced marketing materials, event promotion, engaged in internal trials and planned a Gaeltacht region event in Sneem (https://www.sneemfestivals.ie). A post on the Walton Institute website and social media posts is currently under review and will be published shortly on the main website and social media.

Recommendation by Sponsor. Given the release of the language learning app leveraging an XR-based LLM and the planned activities for dissemination in both academic and non-academic contexts, the project has delivered the expected results in all aspects. Therefore, it is recommended that EOLAS receive the final funding part as originally planned.

5 DoA

MS15. Executing our first FSTP call taught us that certain administrative tasks (e.g., obtaining external reviews, having agreements signed, etc.) require more time than initially anticipated. To have a smooth timeline for the execution of our second FSTP call, on April 8, 2024, we requested PO to allow us to launch the second FSTP call earlier, on May 1, 2024, instead of July 31, 2024. On April 11, 2024 the PO approved the following change:

- MS 15 Second call for FSTP project proposals published: advanced to May 1, 2024.

MS13 and D6/D1.2. Due to the following three main reasons meeting the original deadlines for MS13 and possibly D6/D1.2 would no longer be feasible:

1. A prior delay of approximately one month for the prerequisite milestone (MS7: FSTP project proposal decisions for the first call), caused by longer-than-anticipated external review processes.
2. The signing of agreements took longer for some partners, leading to a later start of the projects than originally planned.
3. All FSTP1 awardees opted for a nine-month project duration (which was permitted in the first call).

On June 24, 2024, we communicated this to the PO. On June 25, 2024, the PO approved the following changes:

- MS13 First set of FSTP projects finish: Extended to October 31st, 2024.
- D6/D1.2 Report on first set of FSTP projects: Extended to October 31st, 2024.

6 Exploitation

Our FSTP programme extends UTTER's impact. Here we highlight a few examples where the results of UTTER-funded FSTP projects are being exploited beyond their lifetime.

FSTP1.

- SURE-GB currently use UTTER's EuroLLM, the project team joined a subset of UTTER's partners (UVAa dn IT) in a new Horizon Europe collaboration (ASCLEPIUS), a submission to call HORIZON-HLTH-2025-01-CARE-01.
- HR-XR-XTEND provided data for future versions of UTTER's EuroLLM as well as for the related project OpenEuroLLM (https://openeurollm.eu).
- MaLA led to a shared preprint Ji et al. (EMMA-500; 2024), which in turn brought TU Darmstadt, LMU Munich into a partnership with a subset of UTTER's partners (UEDIN, UVA, IT, NAV) towards a new Horizon Europe collaboration (AARC), a submission to call HORIZON-CL4-2025-04-DIGITAL-EMERGING-07.

FSTP2. Though it is early to measure impact of FSTP2, it's worth highlighting that Cognifit Harmony's team joined a subset of UTTER's partners (UEDIN, UVA, IT, NAV) in a new Horizon Europe collaboration (AARC), a submission to call HORIZON-CL4-2025-04-DIGITAL-EMERGING-07.

7 Conclusion

Our FSTP programme was composed of two large calls (receiving over 100 submissions) and funded 15 excellent projects. All selected projects were executed on time and successfully, they have been appropriately disseminated and their outputs are in almost every instance directly available to the public. In a few instances, their outputs contributed back to UTTER (for example, in the form of training data) or to future collaborations amongst the UTTER partners (for example, in the form of project proposals submitted to Horizon Europe in 2025, around the time of UTTER's completion or soon after). Most projects are likely to outlive UTTER, with project teams executing their own business plans.

We documented everything about our FSTP programme thoroughly and with transparency, and hope this deliverable and accompanying documentation can be a valuable resource for future Horizon projects running FSTP programmes.

References

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. Emma-500: Enhancing massively multilingual adaptation of large language models, 2024. URL https://arxiv.org/abs/2409.17892.

Juan Antonio Pérez-Ortiz, Miquel Esplà-Gomis, Víctor M. Sánchez-Cartagena, Felipe Sánchez-Martínez, Roman Chernysh, Gabriel Mora-Rodríguez, and Lev Berezhnoy. A conversational intelligent tutoring system for improving English proficiency of non-native speakers via debriefing of online meeting transcriptions. In Thomas Gaillat, Cyriel Mallart, Fabienne Moreau, Jen-Yu Li, Griselda Drouet, David Alfter, Elena Volodina, and Arne Jönsson, editors, *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 187–198, Rennes, France, October 2024. LiU Electronic Press. URL https://aclanthology.org/2024.nlp4call-1.14.

A Screenshots of Review Forms

Free of CoI *

I have no conflict of interest with this submission.

Formal requirements - checklist *

Formal requirements check is the first step in the evaluation process. If one of the formal requirements is not fulfilled, the proposal is rejected.

1. Language: Proposal is in English in all required parts.
2. Submission: Proposal delivered on time, through the designated system, using the requested template.
3. Declaration of Honour: Declaration of Honour is signed.
4. Legal Status: Applicant is an SME or research organisation (incl., but not limited to, higher education organisations, independent research organisations and NGOs).
5. Country: Applicant is legally established in a Horizon Europe eligible country.
6. Number of Proposals: Maximum of one proposal per applicant.
7. Conflict of Interest: No conflict of interest.
8. Complete: All required sections of the proposal are filled in.

Comments on formal requirement

Use this field to make remarks about violated formal requirements and missing information.

Amendments

You can list amendments that must be done (or at least considered) in case the proposal goes on to the **contract signing** phase.

Formal requirements - recommendation *

We desk-reject submissions that fail to comply with the criteria in the formal requirement checklist. If a proposal complies with those criteria, but requests more budget than the maximum allowed in this call (i.e., €60,000 per pilot project), it may be moved to the next phase of the assessment, but, should it get to the contract signing phase, it will require amendments. Note that, the proposal may still be deemed unviable in the next assessment phase (even before any amendment is requested).

1. Desk-reject due to failure to comply with formal requirements
2. Move to next phase, but mark for amendment.
3. Move to next phase

Figure 11: Screenshot of review form for Formal Requirements.

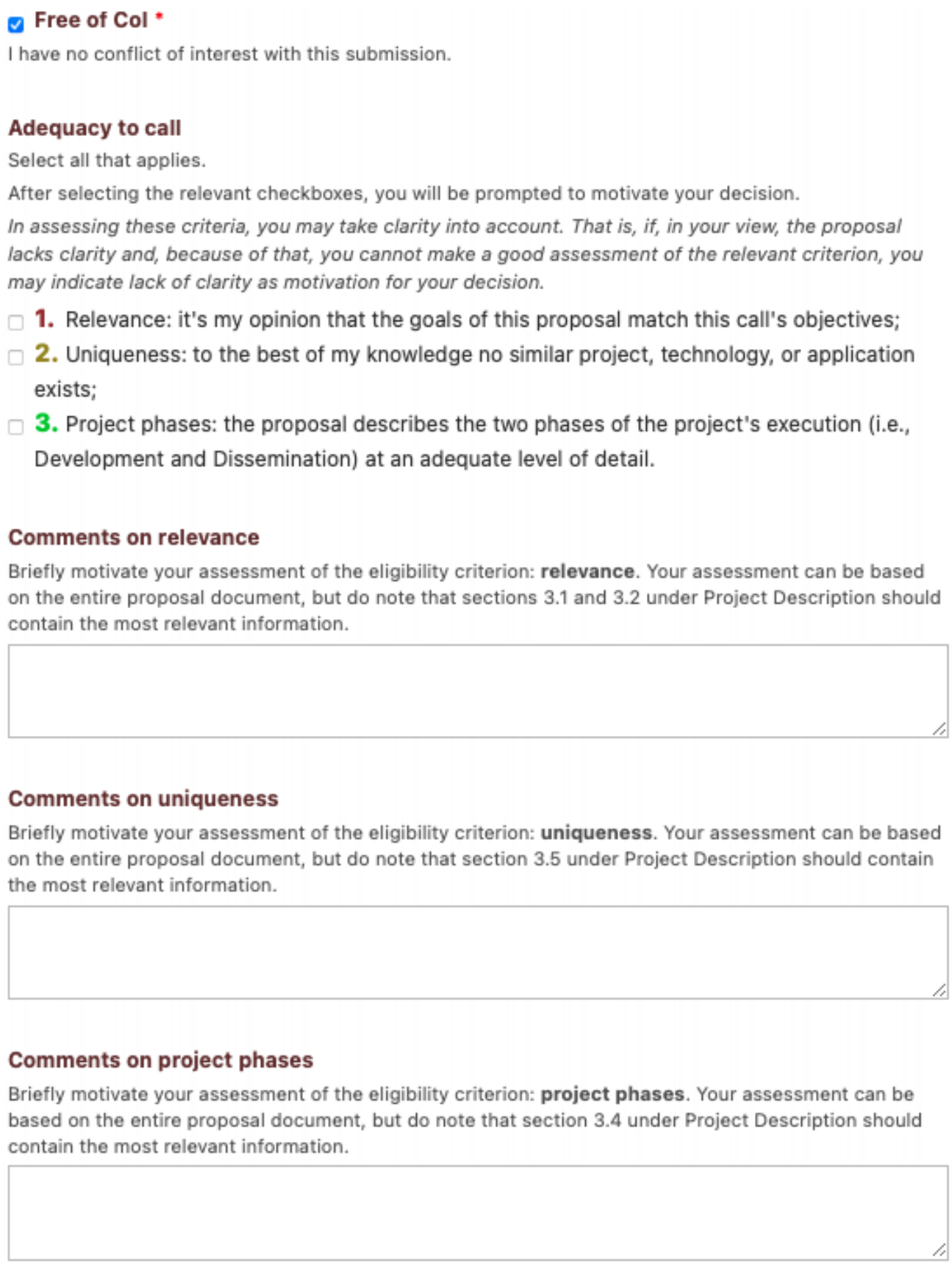


Figure 12: Screenshot of review form for Adequacy to Call (Eligibility).

☑ **Free of CoI** *

I have no conflict of interest with this submission.

☐ **Key parameters** *

I am familiar with the key parameters of this call:

Objectives

Develop and/or pilot applications using XR models (i.e., pre-trained neural network models adaptable to a large variety of forms of expression, interaction, languages, domains, styles and intent) in new sectors, with a focus on enabling new types of human-human and human-machine interaction. Examples of welcome project objectives include:

- Improving or demonstrating efficiency of XR model inference;
- Improving or demonstrating efficiency of XR model training;
- Designing interfaces for usability;
- Extending XR models to new languages, domains or modalities;
- Applying XR models to new tasks;
- Building resources for XR models;
- Evaluation of XR models.

Proposals

- Maximum budget per project: 60,000 euro
- Project duration: 6 months
- Applicant: SME or research organisation for a Horizon Europe eligible country

Project execution

- Development
- Dissemination

Evaluation criteria

- Objective fit
- Technical approach
- Business, Integration and Dissemination (BID) plan
- Budget adequacy
- Team
- Ethics
- Evaluation of XR models.

Criteria fulfilment

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

Figure 13: Screenshot of review form for Qualitative Assessment (1/6) - Key Parameters

Objective fit - clarity *

Are the project goals clear?

- **1.** The goals aren't explicitly outlined, I had to infer what the goals might be by studying their proposed plan.
- **2.** The proposal outlines its goals, but I do not understand them well (e.g., the proposal lacks detail, it is difficult to appreciate the key arguments without specialised knowledge I don't have).
- **3.** The proposal clearly outlines its goals, they are explained well and I understand them.

Objective fit - adequacy *

Are the project goals and planned achievements in line with the overall objectives of UTTER?

- (●) **1.** The proposal treats its goals as somewhat intuitively obviously aligned with UTTER's goals for the call, and I don't agree with this assumption.
- **2.** The proposal justifies how its goals are aligned with UTTER's objectives for this call, but I do not agree with some of the arguments. For example, I disagree with certain premises or predictions.
- **3.** The proposal justifies how its goals are aligned with UTTER's objectives for this call, I find the justification reasonable and I agree with the arguments in the proposal.

Objective fit - impact *

Is it likely that the project will deliver added value to UTTER? There are various ways to add value. Here are some examples: i) resources (e.g., a dataset, a UI, a set of requirements), ii) methods (e.g., a technique for training, or inference), iii) results (e.g., observations about tools, users, or datasets), or iv) a position (e.g., a critical investigation of key premises, a careful outline of ethical considerations, a discussion about broader impact or implications of technology) that UTTER (or the larger body of work around UTTER) can build upon; or through v) an original demonstration of the impact that XR technology (developed by or relevant to UTTER) can have outside academia. You may also recognise some other mechanism which you believe has a similarly important value for UTTER.

- **1.** I do not think the project will add value to UTTER. Potential contributions are marginally relevant.
- **2.** There is potential for some added value along one or more dimensions such as (i-v) above. The contributions are, however, uninspiring and may go unnoticed, or they are unlikely to affect UTTER and the larger body of work around UTTER within UTTER's lifespan.
- **3.** The project will clearly add value to UTTER along one or more dimensions such as (i-v) above, UTTER and the larger body of work around UTTER will likely benefit from it within UTTER's lifespan.

Objective fit: overall score *

Assign an overall score along the *objective fit* dimension. Base your scores on your choices for the sub-criteria *objective fit: clarity*, *objective fit: adequacy* and *objective fit: impact*. As a guideline, you can multiply the marks you assigned for the sub-criteria and find a suggested overall score range below. You can deviate from this suggestion, but major deviations might be a sign that you need to reconsider your assessment of the sub-criteria.

Product	Score range
1	1-3
2-3	2-4
4	3-5
6-9	4-6
12	5-7
18	6-8
27	8-10

(Choose one) ▾

Comments on objective fit

You can use this box to make comments on objective fit. For example, to justify an extreme score, to explain a deviation from the rubrics, or to document a specific interpretation of a rubric, etc. Your comments will be visible to authors.

Figure 14: Screenshot of review form for Qualitative Assessment (2/6) - Objective fit

Technical approach - feasibility *

Are the planned activities feasible and facilitate the achievement of project outputs? As a reminder, this call invited proposals to develop project ideas over a period of 6-9 months with a maximum budget of 60 thousand Euro.

- **1.** The activities are documented insufficiently or the proposal lacks clarity. Or, the planned activities aren't realistic given the allocated resources. Or, there's an incongruence between the planned activities and the intended project output.
- **2.** The planned activities are reasonably aligned with the project goals, they may however be somewhat incompatible with the allocated resources.
- **3.** The activities are described clearly, they are aligned with the project outputs and compatible with the allocated resources.

Technical approach - originality *

Does the proposal push the boundaries of existing XR technology?

- **1.** The approach was discussed unclearly or at an insufficient level of detail. Or, I find the approach uninteresting, trivial or redundant.
- **2.** The approach is presented at a reasonable level of detail and I recognise some potentially original elements.
- **3.** The approach is presented at a reasonable level of detail, it contributes creatively to XR technology and/or application.

Technical approach: overall score *

Assign an overall score along the *technical approach* dimension. Base your scores on your choices for the sub-criteria *Technical approach: feasibility* and *Technical approach: originality*. As a guideline, you can multiply the marks you assigned for the sub-criteria and find a suggested overall score range below. You can deviate from this suggestion, but major deviations might be a sign that you need to reconsider your assessment of the sub-criteria.

Product	Score range
1	1-3
2	2-4
3	3-5
4	4-6
6	6-8
9	8-10

(Choose one)

Comments on technical approach

You can use this box to make comments on technical approach. For example, to justify an extreme score, to explain a deviation from the rubrics, or to document a specific interpretation of a rubric, etc. Your comments will be visible to authors.

Figure 15: Screenshot of review form for Qualitative Assessment (3/6) - Approach

BID plan: business *

Is the business plan reasonable and ambitious?

- 1. No (the business plan is unclear, missing, or unrealistic).
- 2. Somewhat (it might lack some ambition, or be a little too ambitious, or lack detail).
- 3. Yes.

BID plan: integration *

How well is the integration of project outputs planned?

- 1. Poorly. There are too many missing pieces, or it builds on non-existing resources and technology without a clear mitigation strategy, etc.
- 2. Decently, but I would have appreciated more detail, or I doubt the feasibility of some aspects and the plan did not discuss any contingencies.
- 3. Good plan presented at a good level of detail including contingencies where needed.

BID plan: dissemination *

Are the dissemination and promotion activities planned adequately?

- 1. No. The strategies aren't effective or too vague.
- 2. Reasonably well, it will probably reach the relevant target audiences.
- 3. Remarkably well, it's clear who the target audiences are and how they will be approached.

BID plan: overall score *

Assign an overall score along the *BID plan* dimension. Base your scores on your choices for the sub-criteria *BID plan: business*, *BID plan: integration* and *BID plan: dissemination*. As a guideline, you can multiply the marks you assigned for the sub-criteria and find a suggested overall score range below. You can deviate from this suggestion, but major deviations might be a sign that you need to reconsider your assessment of the sub-criteria.

Product	Score range	Rubric
1		1-3
2-3		2-4
4		3-5
6-9		4-6
12		5-7
18		6-8
27		8-10

(Choose one)

Comments on business, implementation and dissemination plan

You can use this box to make comments on BID plan. For example, to justify an extreme score, to explain a deviation from the rubrics, or to document a specific interpretation of a rubric, etc. Your comments will be visible to authors.

Figure 16: Screenshot of review form for Qualitative Assessment (4/6) - BID

Team *

Is the applicant's team capable of executing the project and delivering its outputs (in required time, quality and with estimated budget)?

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

(Choose one)

Comments on team

You can use this box to make comments on team. Your comments will be visible to authors.

Budget *

Does the budget correspond to all panned activities and outputs?

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

(Choose one)

Comments on budget

You can use this box to make recommendations regarding the budget. Your comments will be visible to authors.

Figure 17: Screenshot of review form for Qualitative Assessment (5/6) - Team and Budget

Ethics *

Is the ethical self-assessment thoughtful and thorough? Does it provide convincing justification that the applicant will ensure the work will be done ethically?

Score	Rubric
0	Not at all: The proposal fails to address the criterion or cannot be assessed due to missing or incomplete information
3	Limited: The criterion is inadequately addressed or there are significant weaknesses.
7	Good: The proposal addresses the criterion well, but some shortcomings are present.
10	Excellent: The proposal successfully addresses all relevant aspects of the criterion, any shortcomings are minor.

(Choose one)

Comments on ethics

You can use this box to make remarks and/or recommendations regarding the ethics self-assessment. For example, if you don't think the self-assessment is thorough, you can highlight and defend this here. If you believe the assessment is thorough but the mitigation strategies aren't adequate, do highlight this here. Use this space for any other advice and/or recommendation. Your comments will be visible to authors.

Comments for PC (hidden from authors)

Figure 18: Screenshot of review form for Qualitative Assessment (6/6) - Ethics and Comment for PC

B Reports from Project Teams from First Call

B.1 MaLA

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – MaLA

Massive Language Adaptation

Nature	Final Report	**Work Package**	WP1
Project start date	15/01/2024	**Project end date**	07/10/2024
Interim meeting	28/05/2024	**Report submission Date**	07/10/2024
Main authors	Barry Haddow (UEDIN)		
Co-authors	Shaoxiong Ji (University of Helsinki)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

There are no significant changes or deviations from orginal plan.

1.2 Development

Compilation of the MaLA corpus

The MaLA, **Ma**ssive **L**anguage **A**daptation, corpus has been successfully compiled, containing data from 939 languages. Of these, 546 languages with over 100,000 tokens have been selected for training the EMMA-500 model. The corpus offers diverse data types, including code, books, scientific papers, and instruction data, with more than 100 billion whitespace-delimited tokens. Four versions of the corpus have been made available, meeting different processing needs: noisy, cleaned, deduplicated, and split versions.

Extension of the MaLA corpus

The MaLA corpus[1] has been extended by integrating multiple curated datasets. This augmentation resulted in a rich, diverse data mix that supports the continual pre-training of large language models, ensuring a comprehensive dataset for enhanced language adaptation across a broad range of linguistic contexts.

Continual pre-training of the EMMA-500 model

The continual pre-training of the EMMA-500 model[2] has been completed using the Llama 2 7B model (Touvron et al., 2023). The training involved 546 languages and a massive multilingual corpus, leading to the development of a model that has been rigorously evaluated across various tasks.

Evaluation and benchmarking

The EMMA-500 model has been evaluated against other multilingual and decoder-only LLMs on a wide range of tasks, including commonsense reasoning, machine translation, open-ended generation, text classification, and natural language inference. We also composed a novel multilingual benchmark, called PolyWrite[3] in this work for evaluating open-ended generation in 240 languages.

1.3 Dissemination

We release a preprint on arXiv (Ji et al., 2024), a model and different versions of the datasets on Huggingface. We disseminated these in Helsinki-NLP twitter[4], and received 17 retweet, 65 likes,

[1] https://huggingface.co/collections/MaLA-LM/mala-corpus-66e05127641a51de34d39529
[2] https://huggingface.co/MaLA-LM/emma-500-llama2-7b
[3] https://huggingface.co/datasets/MaLA-LM/PolyWrite
[4] https://x.com/HelsinkiNLP/status/1840669891101172149

and 3.1k views as of Oct 3, 2024.

1.4 Ethics

Multilingual language models trained on large, diverse datasets risk inheriting and amplifying societal biases present in the data. There is a danger that these biases might manifest in harmful ways, particularly in sensitive applications such as content generation, machine translation, and customer support. Despite our focus on low-resource languages, there remains a risk that certain languages or dialects may still be underrepresented, leading to poorer performance or exclusion from language technology advancements.

To mitigate the risk of underrepresentation, the MaLA corpus is continuously expanded to include more data from low-resource languages. By employing continual pre-training and bilingual datasets, we enhance the model's performance on these languages. We also emphasise that our released model is a "foundation model", intended for research purposes, in that it has not undergone the extensive safety testing that would be required for a deployed model. In fact, safety testing of LLMs for diverse languages is an active area of research.

2 Summary of Results and Plans

2.1 Results

In a comparison with decoder-only LLMs, including Llama 2-based continual pre-trained models and LLMs that are designed to be multilingual, the EMMA-500 model has outperformed several baselines, including Llama 2-based models, in most tasks, and outperformed some strong baselines in some tasks, showing strong progress toward improved multilingual performance. Our model achieves strong results:

- Out of models with parameter sizes from 4.5B to 13B, our model with 7B parameters has the lowest negative log-likelihood according to an intrinsic evaluation.
- Our model remarkably improves the performance of commonsense reasoning, machine translation, and open-ended generation over Llama 2-based models and multilingual baselines, and outperforms the latest advanced models in many cases.
- Our model improves the performance of text classification and natural language inference, outperforming all Llama 2-based models and LLMs designed to be multilingual.
- While math and machine reading comprehension (MRC) tasks are challenging for the Llama 2 7B model and other multilingual LLMs, our model remarkably enhances the Llama 2 base model. Our model yields improved performance on MRC over the base model but still produces quasi-random results similar to other multilingual baselines.
- We demonstrate that massively multilingual continued pre-training does not necessarily lead to regressions in other areas, such as code generation, if the data mix is carefully curated. Our model surpasses the Llama 2 7B base model's code generation abilities.

2.2 Business plan

We will need to apply for funding to support our future plans in Section 2.3.

2.3 Future plans

The next phase of this project will focus on expanding the training data and improving model evaluation processes. The planned steps are as follows:

1. **Preparation of Data Mix #2**: We are in the process of preparing Data Mix #2, which will include the content from Mix #1 used in training the released model along with additional bilingual texts and extra code/reasoning data. This dataset will contain approximately 300 billion tokens in total and will be used for the continued training of the Llama 3 model (Dubey et al., 2024).
2. **Preparation of Data Mix #3**: We also plan to work on preparing Data Mix #3, which will primarily feature more bilingual texts. The total token count for this dataset is to be determined, and the data will be used for training the Llama 3.1 model.
3. **Implementation of Additional Evaluation Tasks**: We will implement evaluation codes for two additional machine translation (MT) benchmarks, namely NTREX and Flores+, as well as an additional task of token classification. These evaluations will further assess the performance of our multilingual models across different linguistic and computational tasks.

These next steps aim to significantly advance the multilingual capabilities of our models, further improving the representation of low-resource languages and enabling better cross-lingual transfer in tasks such as machine translation and token classification.

2.4 Blurb for public dissemination on UTTER's website

The UTTER FSTP has made significant strides in advancing multilingual language models with the creation of the MaLA corpus[5] and the development of the EMMA-500 model[6]. The MaLA corpus is a diverse dataset encompassing 939 languages, 546 of which were used to train EMMA-500, a cutting-edge multilingual model. EMMA-500 has demonstrated improved performance on various language tasks such as machine translation, commonsense reasoning, and text classification across multiple languages, including low-resource languages.

3 Recommendation by Project Sponsor

The goals of this project were to collect a massively multilingual corpus and use this to train an LLM supporting a large number of languages. This has been achieved, the MaLA corpus was released and the EMMA model created by fine-tuning Llama 2 7B on this corpus. The evaluation results show strong performance, especially in MT. The data and model have both been made publicly available and there is a preprint describing them on Arxiv.

[5] https://huggingface.co/collections/MaLA-LM/mala-corpus-66e05127641a51de34d39529
[6] https://huggingface.co/MaLA-LM/emma-500-llama2-7b

References

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. EMMA-500: Enhancing massively multilingual adaptation of large language models. *arXiv preprint 2409.17892*, 2024. URL https://arxiv.org/abs/2409.17892.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*, 2023.

B.2 PenGUIn

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – PenGUIn

PenGUIn

Nature	Final Report	**Work Package**	WP1
Project start date	09/02/2024	**Project end date**	30/09/2024
Interim meeting	19/06/2024	**Report submission Date**	15/10/2024
Main authors	José Souza (UNB), Pedro Martins (UNB)		
Co-authors	Stefania Aguzzi (RE:Lab)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	09/10/2024
v1.0	**Status**	Final	09/10/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

This is the Final report of the PenGUIn, presenting the progress made since the previous report delivered in June 2024 and later presented to the UTTER team at the interim meeting. The aim of the project is to support UTTER by designing an innovative and user-friendly Graphical User Interface for the two use cases (UCs), the online customer service assistant and the meeting assistant. The work carried out by PenGUIn will, on one side, improve the usability and the user experience, while enhancing, on the other, the functionalities of the UTTER platforms. In this later phase, from June to September, RE:LAB has focused on implementing the feedback received from UTTER about the latest changes to the prototypes, leading to their finalization. In terms of dissemination, the final prototypes were also published on RE:LAB social media channels.

1.1 Deviations from original plan

No deviations occurred.

1.2 Development

Development of the Use Case 1 - Customer Service Assistant

Our work here was dedicated to the implementation of a colour code scale to visualize the range of values of the COMET index. The colour code has replaced the values from the previous versions of the prototypes to be clearer and intuitive. In fact, this was one of the feedback items emerging from the focus groups with experts held in June. Figure 1 shows the last version of the interface for this use case.

Figure 2 shows how the colour code works, with the interface providing a colour line under each chat box corresponding to the specific scale value.

Development of the Use Case 2 - Meeting Assistant

The updates included the implementation of the light mode of the interface and the reduction of the number of meeting lines listed in the dashboard. As per the previous use case, they reflected the comments shared during the focus group. The figures below show the latest modifications (Figure 3 and Figure 4).

1.3 Dissemination

RE:LAB has implemented a series of promotional activities aiming to disseminate and give visibility to the project. Examples of the promotion activities include:

- Project website in the company website
- Article dedicated to the project published on the company website
- Post on RE:LAB LinkedIn page
- Post on RE:LAB Instagram page

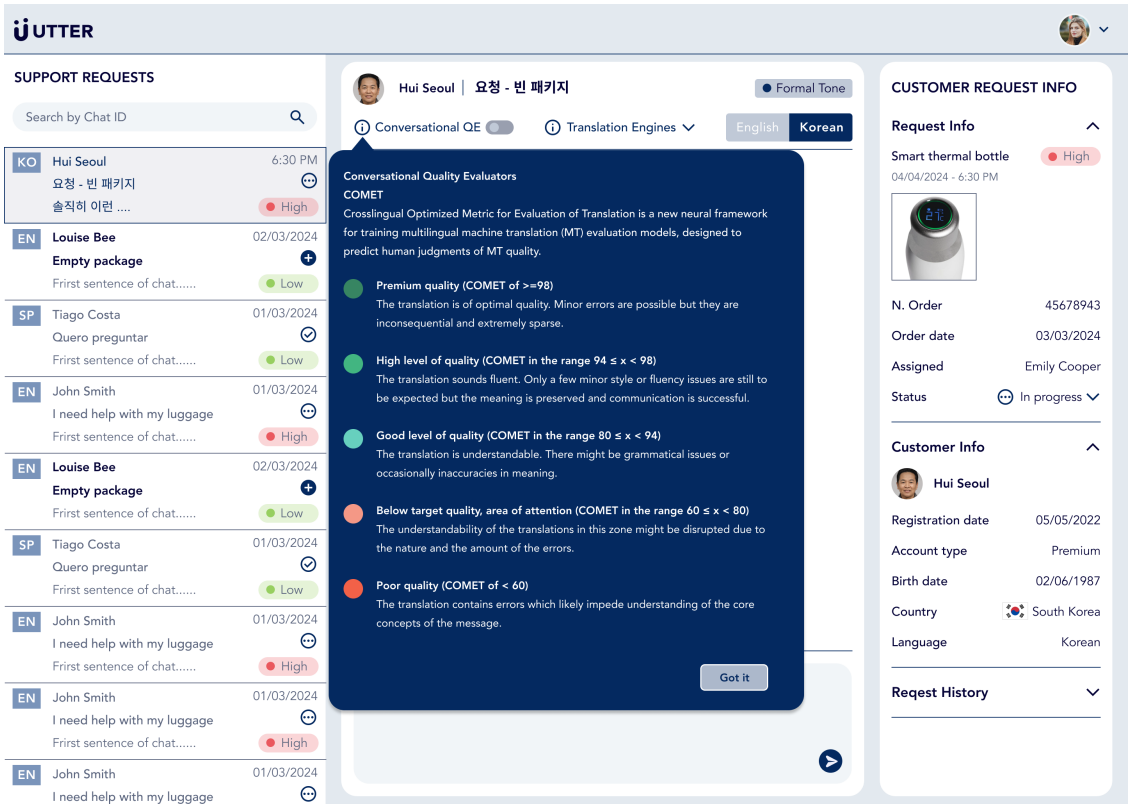


Figure 1: Customer Service Assistant – Last version.

1.4 Ethics

Although the ethical implications of the activities carried out in this project were limited due to their nature, RE:LAB applied standard compliance and prevention strategies to ensure the protection of personal data as well as the monitoring of potential risks.

In terms of project data management, the project has not produced relevant or sensitive data itself. All project materials have been stored in a company's secure virtual environment that is open only to the project team.

As far as it concerns the prototyping of the two interfaces, no particular requirements of privacy/ data protection were preliminarily identified nor raised, so the design activities have followed the standards applicable to these kinds of platforms.

With regards to the organisation of the focus group with UX/UI experts, RE:LAB applied the standard data protection strategies for the event. This was done by informing the participants about how their personal information would have been used and asking for consent to share information and event materials. Consent was also asked in order to recorder the online meeting.

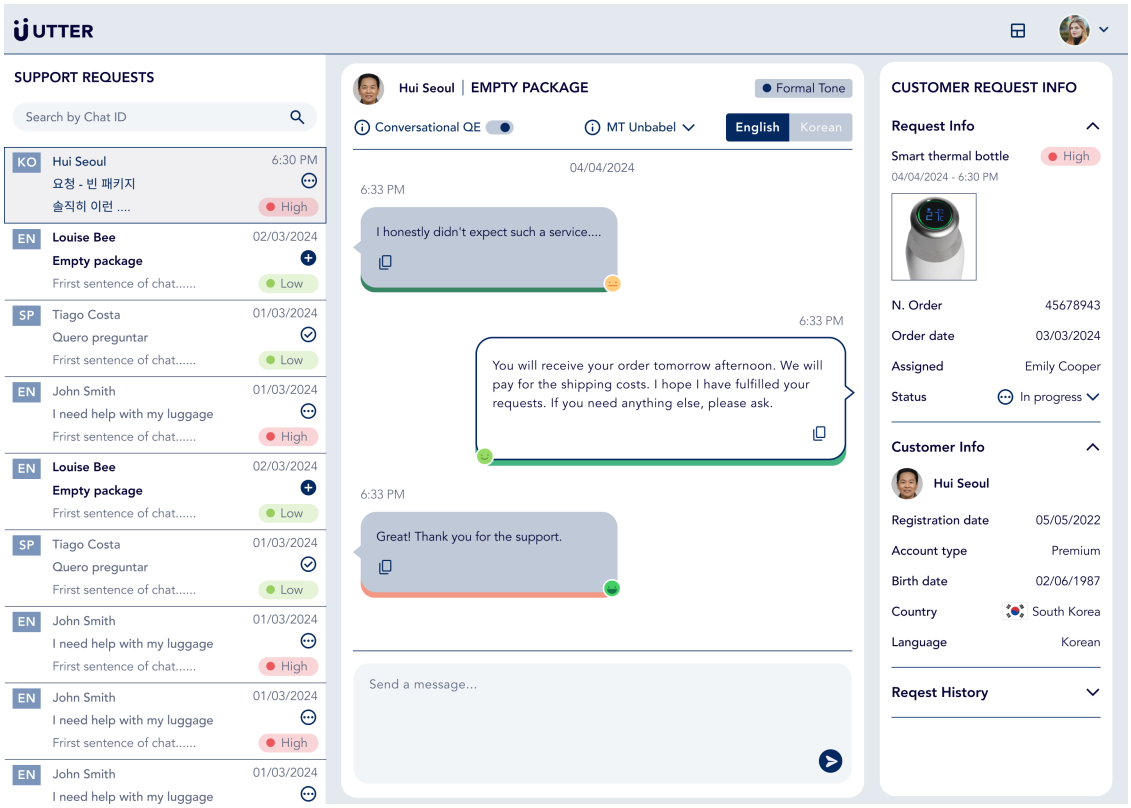


Figure 2: Customer Service Assistant – Colour code implemented.

2 Summary of Results and Plans

2.1 Results

As presented in the previous paragraphs, the final results of the PenGUIn project are the finalized prototypes of two user interfaces designed to support the UTTER use cases in terms of usability and user experience. The outputs are delivered as Figma files and demos, available at these two links:

- Customer Assistant Interface Prototype
- Meeting Assistant Interface Prototype

2.2 Business plan

In terms of business exploitation, RE:LAB will build on the experience and knowledge gathered in the context of the PenGUIn project to target opportunities in business and in research. RE:LAB will therefore seek collaborations on this topic with existing clients and partners and also by establishing new contacts. As an example of potential opportunities for collaboration resulting from the project, RE:LAB has had an introductory meeting with a research team in Naver Labs Europe to discuss potential research synergies. Leveraging its strong focus on research, RE:LAB – as part of a standard practice - will also pursue opportunities for the technical and scientific exploitation of the project results, for example by submitting scientific articles on the methodology and approach applied to this research theme.

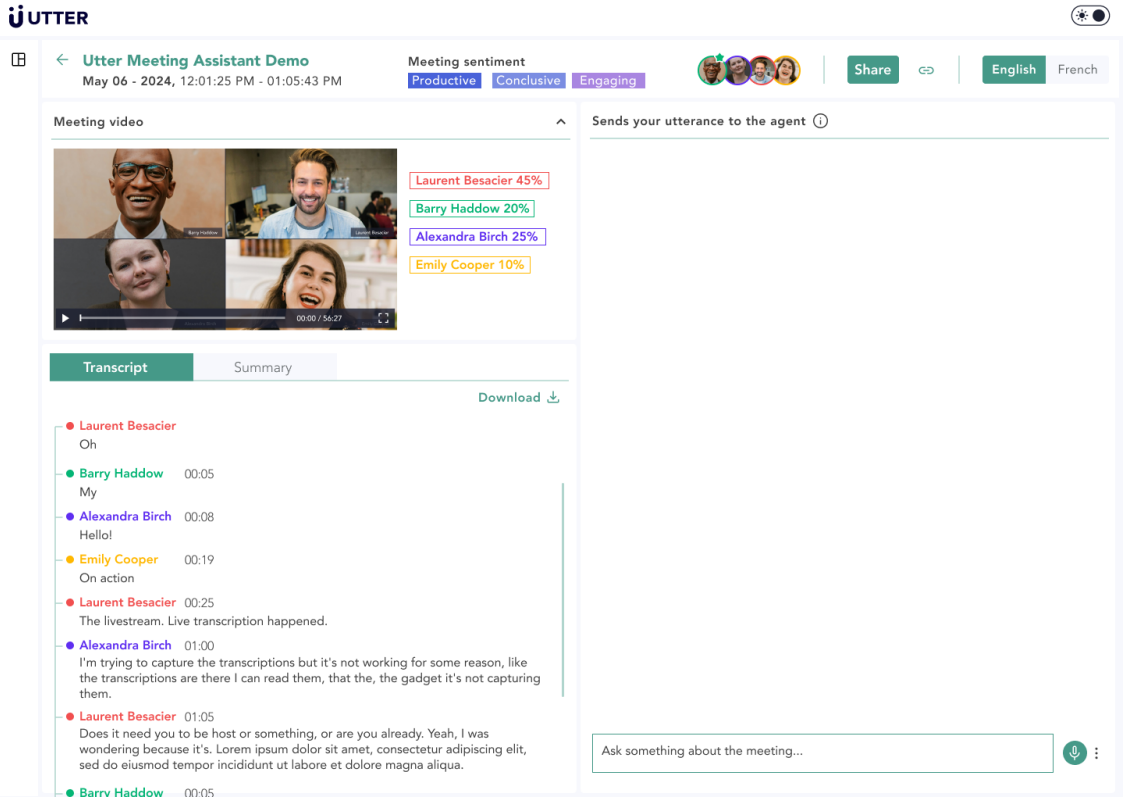


Figure 3: Use Case 2 Meeting Assistant – Light mode option

2.3 Future plans

As described in the PenGUIn project proposal, from a business perspective the project has represented an additional opportunity to strengthen our positioning in the design of innovative User Interfaces across different domains and applications, and to consolidate the expertise and RE:LAB research portfolio of technological solutions for user-centric HMIs. PenGUIn is therefore a company case study for UI and UX projects for online platforms, showcasing our goal to improve user experience in the context of automated and intelligent applications (AI assistants), and multilingual services. In terms of sustainability and exploitation, other funding avenues, from public and private sources, will be explored building on the results produced by the project, to further expand our research and development competences. In the context of Horizon Europe and other European funding programmes, RE:LAB, through its dedicated R&D team, will seek for specific calls and partnerships to leverage the PenGUIn experience and perform new research activities based on it.

2.4 Blurb for public dissemination on UTTER's website

RE:LAB was selected as one of the successful submissions to the UTTER's First Open Call with its project PenGUIn. The project, spanning 9 months, aimed to enhance user experience through an intuitive, inclusive, and adaptive Graphical User Interface (GUI) for online platforms. This was done by studying the most appropriate information design framework and applying suitable interaction strategies to support user's tasks in the context of two case studies: a customer assistant platform and an online meeting platform. PenGUIn's concept was driven by innovation and usability to achieve functionality, effectiveness, and ergonomic experience, building on RE:LAB's

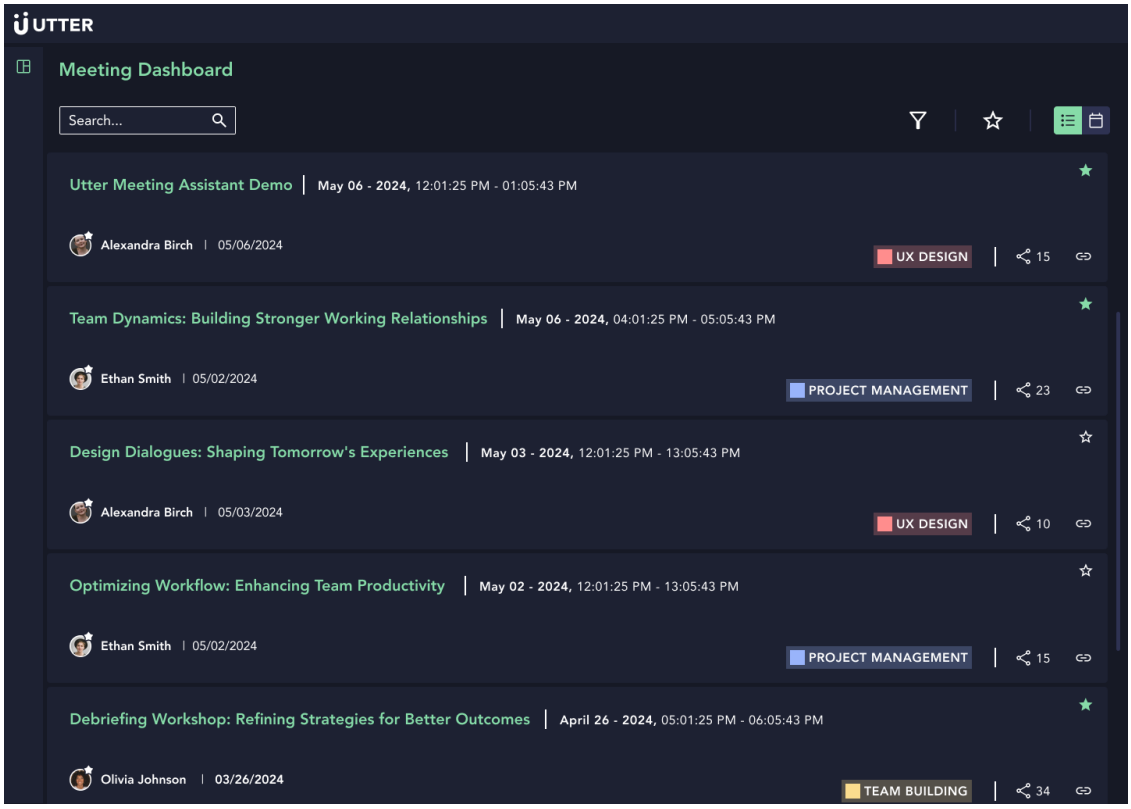


Figure 4: Meeting lines view.

user-centric methodology, "Interaction Engineering". The purpose of PenGUIn's design effort was to guide the user through the multiple platforms' functionalities, from the multilingual translation to the AI-assistant.

PenGUIn UI supported transparent and task-oriented dialogue and interaction between users of these virtual platforms. The project focused on customization flexibility, going through several design iterations, and validating the prototypes through expert analysis, focus group, and testing. The work carried out in the project has represented an additional opportunity to experiment RE:LAB original proposition and new research purposes, to consolidate the team expertise in creating and testing novel user experiences. The final prototypes are available as interactive demos at these links:

- Customer Assistant Interface Prototype
- Meeting Assistant Interface Prototype

3 Recommendation by Project Sponsor

The project proposal was aiming at "design and test a user-centred design approach to prototype a library of graphical elements that are intuitive, cross-cutting, and compatible with usability criteria" considering the UTTER use cases. The work resulted in two user interface prototypes that have been tested in focus groups to evaluate their usability. These user interfaces were made available as Figma templates that could be used as a base for developing graphical user interfaces using

any desired front-end framework. The project delivered on what has been proposed. To the best of our knowledge the project has been disseminated on social media channels and in the company's webpage. The project team documented business plans and possible future works including possible opportunities collaboration with one of the institutions that belong to UTTER (NAVER). Based on this, the recommendation is to approve the final payment to the project Awardee.

B.3 HR-XR-XTEND

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – HR-XR-XTEND

Croatian XR Extensions

Nature	Final Report	**Work Package**	WP1
Project start date	15/01/2024	**Project end date**	15/10/2024
Interim meeting	18/06/2024	**Report submission Date**	11/10/2024
Main authors	Gaurish Thakkar, Marko Tadić		
Co-authors	Matea Filko, Daša Farkaš, Vanja Štefanec		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	04/10/2024
v1.0	**Status**	Final	11/10/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

From call documentation (section 7.2). *The final evaluation of a project will be performed by the Project Sponsor after the dissemination activities took place. The project team is required to report their results, business plans, secured venture capital for further development and future plans. The Pilot Board will assess the finished projects and evaluate the immediate results. It will also formulate recommendations for sustainability and future operation of the project. The Project Sponsor will then prepare a short report (to be made public) and recommend to the Pilot Board to approve (or not) the final payment to the project Awardee.*

How to complete this report. *The Sponsor asks the Project Team to fill in Sections 1 and 2 prior to the meeting (this entire report is likely no longer than 2–4 pages). After the meeting, the Sponsor writes a recommendation to the Pilot Board (Section 3).*

1 Project Execution

1.1 Deviations from original plan

There were no significant deviations in project execution, other than slight changes to the schedule. Data collection took more time than initially anticipated, but the effects were mitigated by employing iterative data deduplication and training methods. For the evaluation objective, we initially planned to translate the Alpaca dataset[1] into Croatian, but we dropped the idea as we found already translated Alpaca dataset into Croatian in an online repository.

Namely, in the initial steps we selected a subset from Alpaca, translated it automatically using Google Translate, and used native speakers of Croatian to proofread and adjust the translations. The texts required significant adjustments, which was impossible to complete due to the size of the dataset and available person hours. Thus we decided to use the MMLU dataset, which uses text from a wide range of different domains, and was available as already translated to Croatian using translation by GPT3. We have also compared the sample of the MMLU dataset translated with GPT3, Google Translator and the translator available at the National Language Technology Platform Hrvojka (https://hrvojka.gov.hr/). The overall accuracy of Google Translator and Hrvojka is similar. However, we opted for the GPT3 translation. Although it may be a bit worse at the individual sentence level, the overall texts, especially longer ones, appear more coherent. Lastly, it should be noted that in some cases the problem of the translation can be the result of the sample itself: the original English sentences are not actually of the highest quality, and they have a lot of noise (correct answers, incorrect answers, problems with questions etc.).

1.2 Development

Objective 1. Collecting the training corpus

For the purpose of training the monolingual Croatian LLM, a large-scale data set was composed from the available monolingual corpora of Croatian language, parallel corpora containing Croatian as one of the languages, as well as several multilingual corpora composed of, other than Croatian, closely related South-Slavic languages, i.e., Serbian, Bosnian, and Montenegrin. Texts in other languages were filtered out of the multilingual corpora, and only texts in Croatian were used. Filtering was performed using the available metadata assigned to corpus samples. In some cases,

[1] https://huggingface.co/datasets/yahma/alpaca-cleaned

samples were already labelled for language, and in others, other metadata attributes were used, such as the domain URL. In Table 1, we list down all the data sources.

Name	Approx Size
CLASSLA Hr Web corpus 1.0	2.5 billion
CC100-Hr Dataset	2.27 billion
Corpus of Croatian News Feeds	2.25 billion
Parallel data for En-Hr on OPUS Resources*,	1.48 billion
Hr-news from XLM-R-BERTić dataset	1.4 billion
Croatian news/legal corpus	175 million
Corpus of Croatian Academic Theses	312 million
ParaCrawl*	69.96 million
Riznica from XLM-R-BERTić dataset	69.51 million
MARCELL Croatian legislative subcorpus	56 million
CURLICAT Croatian corpus	49 million
MARCELL Croatian-English Parallel Corpus of Legislative Texts*	14.3 million
Romance-Croatian Parallel Corpus* (literary works)	2.5 million
Total	8.9 billion

Table 1: Non-exhaustive list of largest data sources used for training the HR-GPT (Beta version) with approximate size in tokens. *Croatian texts only

We used the datatrove (Penedo et al., 2024) library to perform the near deduplication with MinHashLSH and a threshold of 0.72, following the advice that LLMs trained on deduplicated data are better and memorise less of their data (Lee et al., 2022). After deduplication, the deduplicated dataset is approximately 7.72B tokens in size, compared to the original dataset, which contains 8.9B tokens. The dataset was divided into a training and evaluation and test subset in 94:5:1 ratio.

Objective 2. Training the language model

We divided the language model training into three cases.

1. **Training from scratch:** In this case, we trained the model using the existing training configurations from the "GPT-NeoX" library (Andonian et al., 2023). We chose the following parameters based on the number of tokens available for training: 160M, 350M, 410M, and 1.4B parameters. The models are based on the GPT-2 tokenizer. We trained an additional model with 160M parameters, which relied on a newly trained tokenizer (hr-tok) from the training set. We conducted this study to examine the impact of a tokenizer specifically trained on Croatian texts.

2. **Continued pretraining on the monolingual model:** We investigated the effect of continued pretraining on the publicly available GPT-2[2] model. The model is originally monolingual, and we trained using the same Croatian text as in the previous scenario.

[2] https://huggingface.co/openai-community/gpt2

3. **Continued pre-training on multilingual model:** To perform continued pretraining, we used the quantised version of Gemma 7b, i.e., "unsloth/gemma-7b-bnb-4bit". We investigated the results of using the existing multilingual large language model as a backbone for further training.

Objective 3. Evaluation

There are three parts to the evaluation:

- Benchmark datasets for zero-shot evaluation
- Supervised instruction tuning with the Alpaca dataset
- Sentiment and choice of plausible alternatives datasets for supervised fine-tuning.

We conducted the evaluation using the evaluation-harness library (Gao et al., 2024). We used the following benchmarks: TruthfulQA (Clark et al., 2018), Multilingual ARC (Clark et al., 2018), Belebele (Bandarkar et al., 2024), Multilingual HellaSwag (Zellers et al., 2019), and the MMLU (Hendrycks et al., 2021a,b) dataset. In addition, we used two tasks from the Benchich[3] benchmarking dataset, namely sentiment analysis (SA) and choice of plausible alternatives (COPA). The sentiment task associated with sentiment identification in parliamentary proceedings and COPA evaluates cause and effect of premise and hypothesis in Croatian. We conducted evaluations in a zero-shot, three-shot, and ten-shot setting. Two linguists manually checked the MMLU dataset (en) and its Croatian translations, which the University of Oregon had translated using GPT-3.5-turbo[4]. We compared a total of 150 samples (75 per linguist) by checking them with their corresponding Google translations. Additionally, we performed supervised fine tuning on the Alpaca dataset (Croatian version) and evaluated the trained models using the benchmarking datasets.

No supervised training (zero-shot evaluation)										
		Pretraining					Vanilla		CPT	
benchmark	metric	160M	350M	410M	1.4B	160M+hrtok	gpt2	gemma-7b	gpt2-en-cpt-hr	gemma-7b-cpt
arc_hr	acc	18.91	20.96	20.36	20.44	20.87	19.85	**32.34**	18.82	21.81
	acc_norm	23.44	25.49	25.06	24.89	23.95	23.87	**36.53**	23.44	24.55
belebele_hrv_Latn	acc	22.78	23	23.11	22.67	22.78	23.44	**52.67**	21.33	23
	acc_norm	22.78	23	23.11	22.67	22.78	23.44	**52.67**	21.33	23
hellaswag_hr	acc	28.43	29.87	30.08	31.36	28.63	26.27	**38.5**	26.44	24.38
	acc_norm	30.07	32.74	33.38	35.52	30.63	29.42	**50.11**	28.14	24.24
m_mmlu_hr	acc	22.65	25.21	22.8	22.54	22.63	22.59	**41.5**	22.67	25.02
truthfulqa_hr_mc1	acc	25.88	24.58	25.75	26.27	25.49	22.24	**28.61**	26.01	18.34
truthfulqa_hr_mc2	acc	43.82	42.21	42.34	42.52	43.03	40.8	46.6	**46.79**	-

Table 2: Benchmarking evaluation (zero-shot) results for a variety of models without the use of any supervised training. The table displays scores for various models that did not utilise any supervised training (instruction fine tuning). ACC: accuracy and acc_norm: normalised accuracy.

The following are the key observations:

[3] https://github.com/clarinsi/benchich
[4] https://huggingface.co/datasets/alexandrainst/m_mmlu

Trained on benchich training data										
dataset	metric	160M	350M	410M	1.4B	160M+hrtok	gpt2	gemma-7b	gpt2-en-cpt-hr	gemma-7b-cpt
SA-Parlasent(hr-only)	acc	68.86	72.98	72.53	71.03	71.18	36.68	72.46	53.74	**74.48**
COPA	acc	50	49.4	49.6	47.8	48.8	48.4	**79.8**	50.2	79.6

Table 3: The model scores (accuracy) for supervised tasks related to sentiment analysis and choice of plausible alternatives (COPA).

Supervised training (instruction fine tuning)									
Alpaca									
benchmark	160M	350M	410M	1.4B	160M+hrtok	gpt2-en-cpt-hr	gemma-7b-cpt	gpt2	gemma-7b
arc_hr	21.21	19.76	22.75	23.1	20.19	19.08	35.76	19.67	**35.93**
	26.26	25.32	25.75	27.12	23.18	24.64	37.81	24.12	**39.95**
belebele_hrv_Latn	22.67	22.89	23.44	22.67	22.67	23.78	**58**	23.89	45.33
	22.67	22.89	23.44	22.67	22.67	23.78	**58**	23.89	45.33
hellaswag_hr	28.56	30.14	30.74	31.64	28.96	26.84	40.35	26.27	**41.1**
	30.19	32.76	33.75	35.46	30.39	27.62	53.56	27.7	**53.67**
m_mmlu_hr	22.69	23.05	22.82	22.76	22.79	22.63	**43.12**	22.62	33.12
truthfulqa_hr_mc1	24.19	22.63	23.67	26.92	25.23	25.1	**31.73**	24.45	30.04
truthfulqa_hr_mc2	42.58	40.8	39.2	42.87	42.93	41.1	**50.04**	40.08	47.68

Table 4: Benchmarking evaluation results for a variety of models trained with the Alpaca instruction tuning dataset.

1. **Performance Variation Across Models:** Larger models like "gemma-7b" tend to perform better than smaller ones in most tasks, both before and after Alpaca fine-tuning. For example, gemma-7b-cpt achieves high accuracy on belebele_hr, m_mmlu_hr and truthfulqa_hr tasks, indicating better performance with more complex pretraining strategies.

2. **Impact of Fine-Tuning:** Alpaca fine-tuning (+ALP) often improves performance, as seen with gemma-7b-cpt+ALP, which achieves the highest accuracy on several tasks.

3. **Comparison Between Vanilla and CPT:** Models with CPT generally show enhanced performance over their Vanilla counterparts, suggesting that CPT might be more effective for these tasks.

1.3 Dissemination

Dissemination activities were organised according to the project plan. First the project logo and website design was produced.

Figure 1: HR-XR-XTEND project logo

The project webpage was opened under the FFZG domain: https://hr-xr-xtend.ffzg.unizg.hr. The relevant project news were published on the webpage, particularly the news about the project presentation at different conferences.

The project and its results were presented at the following conferences:

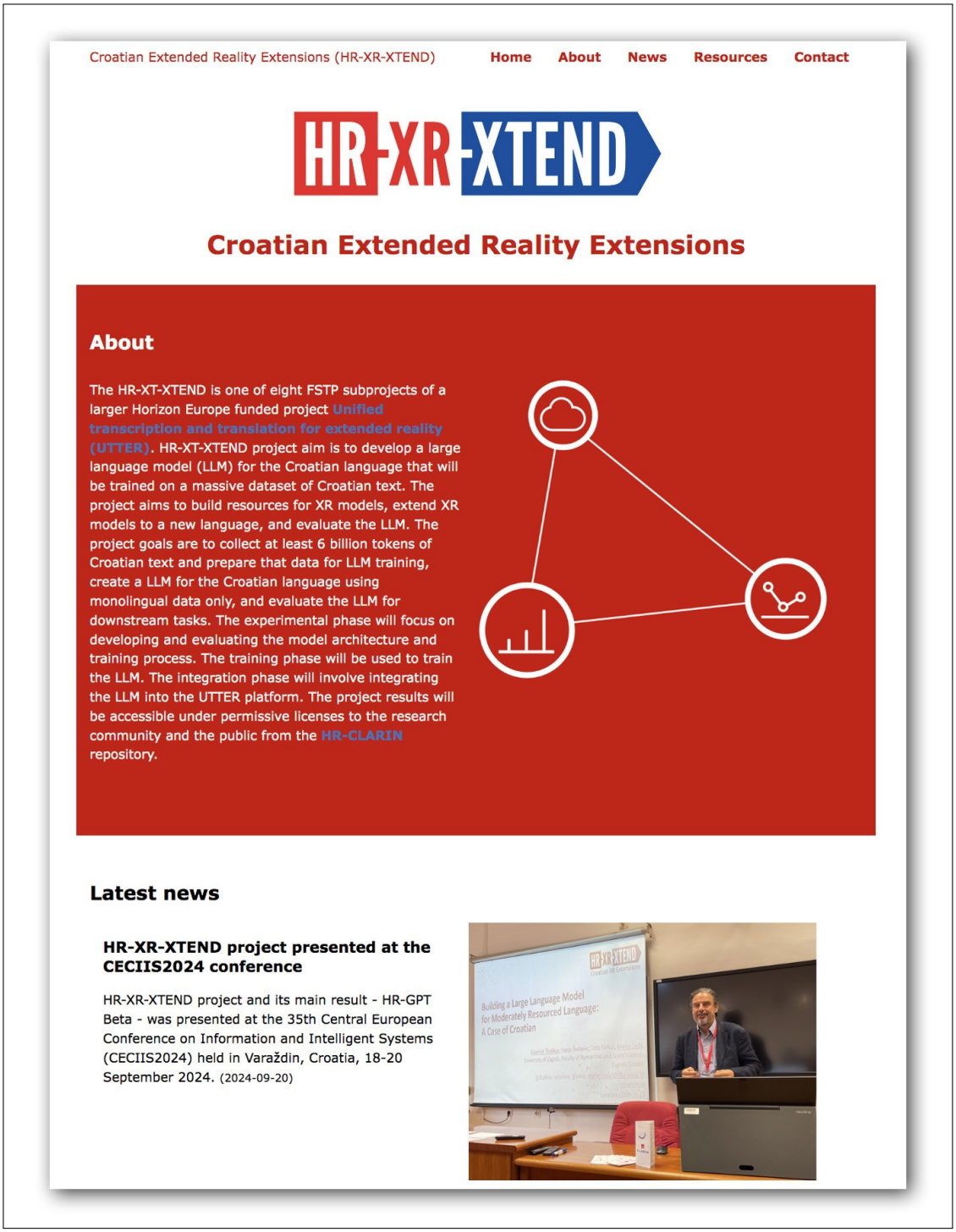


Figure 2: Snapshot of the project webpage

- **Dani e-infrastrukture 2024 (DEI) / Days of e-infrastructure 2024** (https://dei.srce.hr), Zagreb, Croatia, 16th, 18th and 19th April 2024, organiser: University of Zagreb Computing Centre. The project was presented by a poster [https://dei.srce.hr/2024/izabrane-poster-prezentacije];

- **Joint conference on Language Resources and Evaluation and Computational Linguistics (LREC-COLING2024)**, [https://lrec-coling-2024.org/], Turin, Italy, 20th to 25th May

2024, organiser: European Language Resources Association. The project was presented with a paper and poster presentation of an experiment in Sentiment Analysis [https://aclanthology.org/2024.lrec-main.946/];

- **New Trends in Translation Technology (NeTTT2024)**, [https://nettt-conference.com/], Varna, Bulgaria, 3th to 65h July 2024, organisers Lancaster University, UK and Association for Computational Linguistics, Bulgaria. The project was presented with a paper and oral presentation [https://acl-bg.org/proceedings/2024/NeTTT%202024/pdf/2024.nettt-1.17.pdf].
- **Central European Conference on Inteligent Information Systems (CECIIS2024)**, [https://ceciis.foi.hr/], Varaždin, Croatia, 18th to 20th September 2024, organiser: University of Zagreb, Faculty of Organisation and Informatics in Varaždin. The project was presented with a paper and oral presentation.
- **Festival of Languages 2024**, [https://croatia.representation.ec.europa.eu/events/festival-jezika-2024-2024-09-27_hr], Zagreb, Croatia, 27th September 2024, organiser: Representation of the European Commission in Croatia. The project was presented with an oral presentation.
- **Metaphorical Collocations (MetaKOl2024)**, [https://metakol.uniri.hr/en/konferencija-2024/], Dubrovnik, Croatia, 3rd and 4th October 2024, organiser: University of Rijeka. The HR-GPT Beta was presented with an oral presentation.
- **21st EURALEX International Congress Lexicography and Semantics (EURALEX2024)**, [https://euralex.jezik.hr/], **Workshop Large Language Models and Lexicography (LLM-Lex)**, Cavtat, Croatia, 8th to 12th October 2024, organiser: Institute for the Croatian Language. The HR-GPT Beta was presented by the oral presentation [https://www.cjvt.si/en/research/community/llm-lex-2024/].

1.4 Ethics

Since the training data set was composed from publicly available data, we expect that ethical issues have been sorted out by the data providers. All results of the project available under permissive licences in the HR-CLARIN repository, will always have a link to the original source of data and possible users will be able to check the status of ethical issues directly at the data source.

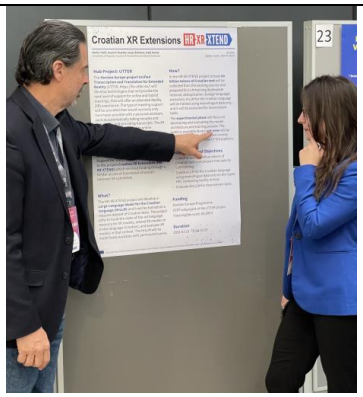
2 Summary of Results and Plans

2.1 Results

The project results and training data will be available at the Croatian CLARIN repository (https://www.clarin.hr/) under permissive licenses by the end of the project.

The key results of the project are:

- **HR-GPT Beta** trained in four sizes: 160M, 350M, 410M i 1.4B parameters;
- **cleaned training data set for HR-GPT Beta**, available only partially because for some of the training data we couldn't reach an agreement with data providers about distribution to the third parties;



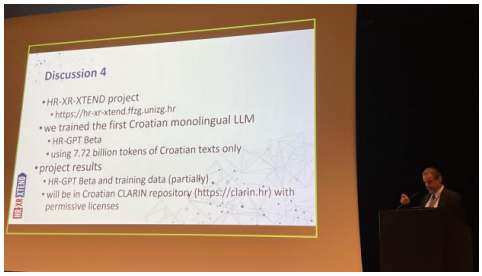
DEI 2024

LREC-COLING 2024

NeTTT 2024

CECIIS 2024

MetaCol 2024



EURALEX 2024

Figure 3: Photos of the project presentations at conferences

- **four sets of training data** for training the high-precision language identifier between Bosnian, Croatian, Montenegrin and Serbian languages (100 million tokens for each language). This language identifier will be trained by the Charles University, a coordinator of the High Performance Language Technology (HPLT) project.

2.2 Business plan

The fundamental prerequisite for the usage of HR-GPT Beta in different business environments is its availability in an digital repository with persistent identifiers. This prerequisite is fulfilled by depositing project results in the HR-CLARIN repository. The permissive licenses will allow open access usage of the results by researchers and developers for different purposes.

Since the research team has good connections with a number of similar research teams in Europe, we are aware of interest for our results that has been expressed already by different projects, e.g. Charles University with HPLT project, Tilde within the Large AI Grand Challenge, etc.

Our previous collaboration with a number of translation and localisation companies in Croatia opens also the possibility of deployment of the HR-GPT Beta in the post-processing of the MT output. The company Ciklopea Ltd. already expressed their interest in inclusion of this LLM in their work process.

2.3 Future plans

The future work can be divided into two directions.

In the first direction, we will collect and clean additional data for Croatian and perform additional filtering in terms of quality, but with much slower pace since the funding from HR-XR-XTEND expires. The non-exhaustive list of already collected and processed data combined with available additional data is presented in the Table 5.

In the second part, we would like to evaluate the models for various other NLP tasks like autocompletion and error correction.

Also, since one of the models we trained didn't converge on train loss for the 160M parameter model's default training configuration, more research is necessary for models that use the "hr-tok" tokenizer. We believe additional training is required.

2.4 Blurb for public dissemination on UTTER's website

The "Croatian XR Extensions" project aimed to create a large-scale monolingual Croatian language model (HR-GPT Beta). A significant training dataset was collected and cleaned from existing mono- and multilingual resources that include texts in Croatian. The preprocessing featured also advanced deduplication techniques, resulting in a final training dataset of 7.72 billion tokens. Three training scenarios were used: training from scratch, continued pretraining on a monolingual model, and continued pretraining on a multilingual model. The evaluation was performed using several benchmark datasets, and fine-tuning with the Alpaca dataset improved model performance. Larger models, like "gemma-7b", outperformed smaller ones, and fine-tuning enhanced results further. Key results include multiple model versions (160M, 350M, 410M, and 1.4B parameters) and a cleaned training dataset. Future work involves additional data collection, additional

Name	Approx Size
CC100-Croatian Dataset 1.0	3.3 billion
CLASSLA Hr Web corpus 1.0	2.5 billion
Corpus of Croatian News Feeds	2.25 billion
HPLT Croatian/Bosnian/Serbian Corpus, Croatian texts only	4 billion
Parallel data for En-Hr on OPUS Resources, Croatian texts only	1.48 billion
Corpus of Croatian Academic Theses	312 million
Joel Niklaus, Multi Legal Pile, Croatian	258 million
Leipzig Corpora	182.40 million
ParlaMint 4.0, Croatian texts only	88.16 million
ParaCrawl, Croatian texts only	79.06 million
hrWikipedia	66.48 million
MARCELL Croatian legislative subcorpus	56 million
Total	14.57 billion

Table 5: Non-exhaustive list of already collected data sources with approximate size in tokens for sources with 50+ million tokens.

model training, further NLP task evaluations, and more training experiments. The HR-GPT Beta and training material (partially) will be publicly accessible under permissive licenses from the HR-CLARIN repository (https://clarin.hr). More information can be found on the project website https://hr-xr-xtend.ffzg.unizg.hr.

3 Recommendation by Project Sponsor

This space is intended for the assessment by the project Sponsor. Did the project deliver its planned results? Was it successfully disseminated? Did the project team document business plans and future plans, and do these seem appropriate? Are there opportunities for exploitation of results within UTTER?

The project planned the collection of Croatian datasets and training a large language model on Croatian and they delivered on this objective. This project was successfully disseminated. UTTER could use these datasets for training the EuroLLM langaugae model. This project has completed successfully.

References

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, Phil Wang, and Samuel Weinbach. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch, 9 2023. URL https://www.github.com/eleutherai/gpt-neox.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.acl-long.44.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*, 2018.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 07 2024. URL https://zenodo.org/records/12608602.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021a.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021b.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577.

Guilherme Penedo, Hynek Kydlíček, Alessandro Cappelli, Mario Sasko, and Thomas Wolf. Datatrove: large scale data processing, 2024. URL https://github.com/huggingface/datatrove.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472.

B.4 SignReality

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – SignReality

Extended Reality for Sign Language translation

Nature	Final Report	**Work Package**	WP1
Project start date	dd/mm/2024	**Project end date**	dd/mm/2024
Interim meeting	dd/mm/2024	**Report submission Date**	dd/mm/2024
Main authors	Sponsor (PARTNER)		
Co-authors	Awardees (ORG)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

The following deviations have occurred at the development:

- The API for the communication between the augmented reality devices, the animation engine and the translation engine (WP2) is not fully implemented due to platform-specific incompatibilities. Back-off solution: the users have to click on Hololens 2 to play pre-recorded animations, generated at the animation engine at an earlier stage.
- The animation loading time on Hololens 2 (WP1) is unsuitable for real-time communication.
- XReal implementation (WP1) remains prototypical, and it is not possible to invoke animations, due to lack of documentation and the steep learning curve.
- Open source licensing and public distribution is not available for all parts of the pipeline and corpora due to licensing reasons beyond the control of the project and could not be solved within such a short duration. We are working to resolve them in near future.
- Scientific publications will be submitted in the near future due to the short project duration

Key results are ready (translation module, animation framework, avatar adjustment, evaluation study, demo). We see the above issues rather as software engineering problems than a scientific problem.

1.2 Development

1.2.1 WP1: Avatar animation and representation

Avatar animation engine During the development of the SignReality prototype, we have developed two new features of our sign language synthesis system[1]:

- We have developed a remote HTTP API for remote submission of MMS data and retrieval of avatar animation data. Essentially, the main rendering engine has been wrapped in a Flask server that receives MMS instances and runs the animation engine. The result, a JSON file containing a full animation of the SL sentence, is then returned.
- We improved the quality of the motion synthesis by revising the coordinate systems of the inflection of the hand motion. Originally, the motion of the hands was inflected relatively to the avatar center of the body. After several experiments and observation, we realized that it was going to be more intuitive and stable to perform inflections relatively to its torso.

[1] A demo can be found at:

Avatar representation Our work involves the display of the avatar on two XR devices:

- **Hololens 2**: We have improved prior implementation by adding user interaction features, through which the user can move and place the avatar in the augmented space. Such positioning mechanism has then been used to run the user studies.
- **XReal Light**: We ported a pre-existing Unity-mobile version to XReal Light, due to its low weight and cost. The implementation (incl. Android app) has remained into a prototypical stage. The avatar has been ported, but the animations have not been connected to a control panel, and therefore it is not possible to invoke them.

1.2.2 WP2: Translation model from text to sign language representation

Implementation of translation model An encoder-decoder model has been trained on the AVASAG corpus in the domain of train announcements (Nunnari et al., 2021; Bernhard et al., 2022), by continual learning of an NLLB model[2] (Team et al., 2022). In order to support MMS annotation as a supplementary annotation over the produced glosses, we have developed a custom PyTorch-based code, which also allows for executing an XML-RPC server to provide translations on demand. Although during the development, the model exhibited high performance (>30 BLEU score) on a limited test set of similar domain with cross-validation, small-scale human evaluation indicated that the model hallucinates occasionally and may have overfitted due to the small training corpus.
Results: basic model, XML-RPC server

Corpus acquisition and curation Due to the lack of existing resources, we continued fundamental research on making more parallel corpora available and usable. We have collected big amounts of subtitled sign language TV shows (DGS, various South American sign languages) from online sources and are acquiring permissions to use them for research, and we have also been recording our own corpus of German fairy tales, as a continuation of recent effort (DGS-Fabeln-1 Nunnari et al., 2024). We worked on two methods for aligning video (signed content) with subtitles:

1. Sign language segmentation based on temporal features (optical flow; Kishore et al., 2016; Zhang et al., 2019). The first ever sentence-level segmentation models were built for BSL and ASL.
   Result: The work is being finalized and will be submitted to an academic venue.
2. Video-text alignment based on i3D features using a Transformer (Bull, 2023; Bull et al., 2021), with the aim to fine-tune existing models for BSL to DGS and other sign languages.
   Result: data preprocessing, human annotation, first statistics on the average timeshift.

1.2.3 WP3: Participatory design and evaluation

Extending previous research (Nguyen et al., 2021; Nolte et al., 2022, 2023), we performed a usability study with the participation of 9 fluent speakers of the German Sign Language (8 of them deaf)

[2] Pre-trained NLLB model: `nllb-200-distilled-600M`

at the Centre for Culture and Visual Communication of the Deaf Berlin Brandenburg (ZFK). The study lasted about 9 hours and focused on *intelligibility, user experience* and *acceptance* focusing on HoloLens 2, and particularly the user-based interaction mechanisms to adjust and position the avatar in space. The participants tried out the avatar application on the HoloLens 2 for themselves in two sessions with variable scenaria. Every session was followed by a questionnaire following the *Raw Task Load Index* (TLX; Hart and Staveland, 1988) and the *Short Version of the User Experience Questionnaire* (UEQ-S; Schrepp et al., 2017). Finally, we collected the participants' opinions in the form of a short interview. According to the study, the currently implemented positioning features did not provide better user experience, but the users insisted on the need of the adjustment features and indicated valuable feedback for further improvements, such as difficulty with the gestures, unpractical size of the avatar after placement, missing interaction feedback from the control panel.
Result: Analysis, publication of results in a CHI-related conference.

1.3 Dissemination

The following dissemination tasks have been completed:

- **Community engagement:** Communication of the idea and user study at ZFK. Internal newsletter post at DFKI. Presentation of project concept and participation at the network of the Berlin XR Lab[3].
- **Web pages:** Project descriptions at the DFKI website[4], department websites[5] and personal websites[6].
- **Social media**: News about the project and user evaluation posted on the Affective Computing group LinkedIn page[7].
- **Academic activities:** The project is aligned (partially or entirely) with one BSc thesis, 4 MSc theses and one student coding workshop at the Technical University of Berlin and the Saarland University. Upcoming publications for text segmentation, sign language animation, evaluation study.

1.4 Ethics

Our present experiments on DGS are part of a broader research aiming to provide equal access to language technology for sign language users. The users of a sign language form a linguistic and cultural minority and the fact that the project is led by researchers that are hearing people entails the risk of developments that are not in accordance with the will of the former and lead to complaints for cultural misuse and appropriation. For this reason, in our broader research we have included interpreters and members of the deaf and hard of hearing communities as part of the

[3] Berlin XR Lab: https://www.berlin-xrlab.de/
[4] DFKI website: https://www.dfki.de/en/web/research/projects-and-publications/project/signreality
[5] Design Research Lab: https://www.dfki.de/en/web/research/projects-and-publications/project/signreality, Berlin Open Lab: https://berlin-open-lab.org/portfolio/signreality-extended-reality-for-sign-language-translation/
[6] Fabrizio Nunnari: https://www.dfki.de/~fanu01/
[7] LinkedIn: https://www.linkedin.com/company/affective-computing-group2021/

research team, consultants and participants in user studies and workshops, and we have been in constant co-operation with related unions and communication centers.

DGS is analysed and depicted is in a preliminary stage and results should by no means be presented as a functional product, which the respective communities might find offensive and diminishing. In particular, glosses are known to be inferior to the full linguistic capacity of the sign languages and are only seen as a methodological tool to aid further research. The signing avatar is lacking several elements of the sign language (smooth hand movements, facial expressions, mouthings) and should be seen as work in progress.

The users have been informed and consented as per GDPR about the storage of personal information and the video-recordings of the user study (which will only be processed internally to the project and won't be published). The results are anonymized. Users have participated in the study as part of their working time in ZFK and additionally received a compensation coupon. ZFK was compensated with a lump sum. A DGS interpreter was available for the entire duration of the study.

2 Summary of Results and Plans

2.1 Results

The results of the project have been uploaded to a cloud folder.[8] They include:

- Translation server and trained model (code and model)[9]
- Full translation pipeline with avatar animation engine (code and demo)
- Hololens 2 implementation with adjustment features (code and demo)
- XLReal Light port of the avatar, Android app (code)
- Sign language segmentation method (report)
- User study results and feedback (report)

2.2 Future plans

The results of the research will be reformatted as academic papers and will be submitted to relevant venues (e.g. *ACL, CHI, IVA, SLTAT). Successful components will be extended and integrated into relevant research projects (e.g. BIGEKO, Federal German Ministry of Education and Research, 2023-2026). Further research funding will be sought from European, federal and industrial sources.

2.3 Blurb for public dissemination on UTTER's website

The project "SignReality" achieved significant milestones in bridging sign language technology with Extended Reality. Key results include the development of an engine for avatar animation, accompanied by device-specific implementation on two AR devices (Hololens 2 and XReal Light). Translation from spoken language to a textual sign language representation (German→DGS) was enabled through an encoder-decoder translation model, whereas further improvement of relevant

[8] Project results: https://cloud-affective.dfki.de/s/o5SCGE8JZn4wsfN
[9] Translation server code: https://github.com/DFKI-SignLanguage/text-to-gloss-machine-translation

models will benefit from the work on corpus acquisition and alignment. The implementation was tested for *intelligibility, user experience* and *acceptance* in a user study with native sign language users at the Centre for Culture and Visual Communication of the Deaf Berlin Brandenburg (ZFK) providing valuable feedback. The project has been integrated into several academic theses and university workshops, and research findings will be submitted in relevant academic venues.

3 Recommendation by Project Sponsor

The project clearly achieved all of its goals, with very minor deviations from the original plan, both along the scientific and dissemination dimensions. The project team has experience with the ethical considerations behind the experimental setup and did a remarkable job both at complying with all relevant guidelines and regulations but also at clearly documenting the scope of their findings and technology. I recommend payment of the second installment.

Glossary

ASL American Sign Language. 4

BSL British Sign Language. 4

DGS German Sign Language (Deutsche Gebärdensprache). 4–6

MMS Multimodal Signstream, (consists of the annotated sentences augmented with the sign inflection parameters). 3, 4

References

Lucas Bernhard, Fabrizio Nunnari, Amelie Unger, Judith Bauerdiek, Christian Dold, Marcel Hauck, Alexander Stricker, Tobias Baur, Alexander Heimerl, Elisabeth André, Melissa Reinecker, Cristina España-Bonet, Yasser Hamidullah, Stephan Busemann, Patrick Gebhard, Corinna Jäger, Sonja Wecker, Yvonne Kossel, Henrik Müller, Kristoffer Waldow, Arnulph Fuhrmann, Martin Misiak, and Dieter Wallach. Towards Automated Sign Language Production: A Pipeline for Creating Inclusive Virtual Humans. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments*, PETRA '22, pages 260–268, New York, NY, USA, July 2022. Association for Computing Machinery. ISBN 978-1-4503-9631-8. doi: 10.1145/3529190.3529202. URL https://doi.org/10.1145/3529190.3529202.

Hannah Bull. *Learning sign language from subtitles*. PhD thesis, Université Paris-Saclay, Paris, France, 2023. URL https://theses.hal.science/tel-04055873v1/file/112750_BULL_2023_archivage.pdf.

Hannah Bull, Triantafyllos Afouras, Gül Varol, Samuel Albanie, Liliane Momeni, and Andrew Zisserman. Aligning Subtitles in Sign Language Videos. *Proceedings of the IEEE International Conference on Computer Vision*, pages 11532–11541, May 2021. ISSN 15505499. doi: 10.1109/ICCV48922.2021.01135. URL https://arxiv.org/abs/2105.02877v1. arXiv: 2105.02877 Publisher: Institute of Electrical and Electronics Engineers Inc. ISBN: 9781665428125.

Sandra G. Hart and Lowell E. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, volume 52, pages 139–183. Elsevier, 1988. URL https://www.sciencedirect.com/science/article/pii/S0166411508623869.

P.V.V. Kishore, M.V.D. Prasad, D. Anil Kumar, and A.S.C.S. Sastry. Optical Flow Hand Tracking and Active Contour Hand Shape Features for Continuous Sign Language Recognition with Artificial Neural Networks. In *2016 IEEE 6th International Conference on Advanced Computing (IACC)*, pages 346–351, February 2016. doi: 10.1109/IACC.2016.71. URL https://ieeexplore.ieee.org/abstract/document/7544860.

Lan Thao Nguyen, Florian Schicktanz, Aeneas Stankowski, and Eleftherios Avramidis. Evaluating the translation of speech to virtually-performed sign language on AR glasses. In *2021 13th International Conference on Quality of Multimedia Experience (QoMEX)*, pages 141–144, June 2021. doi: 10.1109/QoMEX51781.2021.9465430. URL https://ieeexplore.ieee.org/abstract/document/9465430. ISSN: 2472-7814.

Amelie Nolte, Karolin Lueneburg, Dieter P. Wallach, and Nicole Jochems. Creating Personas for Signing User Populations: An Ability-Based Approach to User Modelling in HCI. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '22, pages 1–6, New York, NY, USA, October 2022. Association for Computing Machinery. ISBN 978-1-4503-9258-7. doi: 10.1145/3517428.3550364. URL https://doi.org/10.1145/3517428.3550364.

Amelie Nolte, Barbara Gleißl, Jule Heckmann, Dieter Wallach, and Nicole Jochems. "I Want To Be Able To Change The Speed And Size Of The Avatar": Assessing User Requirements For Animated Sign Language Translation Interfaces. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, pages 1–7, New York,

NY, USA, April 2023. Association for Computing Machinery. ISBN 978-1-4503-9422-2. doi: 10.1145/3544549.3585675. URL https://doi.org/10.1145/3544549.3585675.

Fabrizio Nunnari, Judith Bauerdiek, Lucas Bernhard, Cristina España-Bonet, Corinna Jäger, Amelie Unger, Kristoffer Waldow, Sonja Wecker, Elisabeth André, Stephan Busemann, Christian Dold, Arnulph Fuhrmann, Patrick Gebhard, Yasser Hamidullah, Marcel Hauck, Yvonne Kossel, Martin Misiak, Dieter Wallach, and Alexander Stricker. AVASAG: A German Sign Language Translation System for Public Services (short paper). In Dimitar Shterionov, editor, *Proceedings of the 1st International Workshop on Automatic Translation for Signed and Spoken Languages (AT4SSL)*, pages 43–48, Virtual, August 2021. Association for Machine Translation in the Americas. URL https://aclanthology.org/2021.mtsummit-at4ssl.5.

Fabrizio Nunnari, Eleftherios Avramidis, Cristina España-Bonet, Marco González, Anna Hennes, and Patrick Gebhard. DGS-fabeln-1: A multi-angle parallel corpus of fairy tales between German Sign Language and German text. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4847–4857, Torino, Italy, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.434.

Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S). 2017. ISSN 1989-1660. doi: 10.9781/ijimai.2017.09.001. URL https://idus.us.es/handle/11441/107084. Accepted: 2021-04-14T11:12:39Z Publisher: UNIR.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. No Language Left Behind: Scaling Human-Centered Machine Translation, August 2022. URL http://arxiv.org/abs/2207.04672. arXiv:2207.04672 [cs].

Shujun Zhang, Weijia Meng, Hui Li, and Xuehong Cui. Multimodal Spatiotemporal Networks for Sign Language Recognition. *IEEE Access*, 7:180270–180280, 2019. ISSN 2169-3536. doi: 10.1109/ACCESS.2019.2959206. URL https://ieeexplore.ieee.org/abstract/document/8932517. Conference Name: IEEE Access.

B.5 DeMINT

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – DeMINT

Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts

Nature	Final Report	**Work Package**	WP1
Project start date	15/01/2024	**Project end date**	15/10/2024
Interim meeting	31/05/2024	**Report submission Date**	27/09/2024
Main authors	Juan Antonio Pérez-Ortiz (University of Alicante)		
Co-authors			
Reviewers			
Version Control			
v0.1	**Status**	Draft	06/09/2024
v1.0	**Status**	Final	27/09/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

This is the final report of project DeMINT ("Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts"). The project is funded after a Financial Support for Third Parties (FSTP) call[1] by the EU project UTTER ("Unified Transcription and Translation for Extended Reality").[2] DeMINT started on January 15, 2024 and is expected to end on October 15, 2024.

1.1 Deviations from original plan

There are no substantial deviations from the original plan. As a minor deviation, we can mention a low-level technical issue such as the decision to not use LanguageTool for grammar checking as the open source version was not as accurate enough for our purposes. This was not a big issue, as there exist tools based on neural models to replace it.

1.2 Development

DeMINT has developed a prototype of a conversational system designed to enhance non-native English speakers' language skills through post-meeting analysis of the transcriptions of video conferences in which they have participated. Following recent advances in chatbots and agents based on large language models (LLMs), the tutoring system leverages pre-trained LLMs within an ecosystem that integrates different techniques, including in-context learning, external non-parametric memory retrieval, efficient parameter fine-tuning, grammatical error correction models, and error-preserving speech synthesis.

In addition to the research team of faculty members, 3 graduate computer science students have joined the project as technicians.

A pilot human evaluation was designed and carried out under the supervision (and after the approval) of the University's ethics committee.[3] For the pilot study, 7 students from different degrees in our University and with a level of English proficiency of B2 or C1 were selected. The students were paid for their participation according to the budgeted amount. Each student filled a previous questionnaire to collect sociodemographic information and data regarding their familiarity with information and communication technologies and chatbots.

Each student was asked to engage in 5 videoconferences with other students. In total, 10 videoconferences with two participants and 5 videoconferences with three participants were held. To provide a topic for videoconferences, we proposed a role play extracted from a popular manual for English as a second language for each of them.[4] Finally, students were asked to interact with the chatbot to have a debriefing corresponding to each of the videoconferences in which they partitipated; participants finally provided feedback on their user experience by means of a questionnaire which can be found in appendix B.

[1] https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/competitive-calls-cs/3722
[2] Grant agreement number 101070631.
[3] https://web.ua.es/en/vr-investigacio/comite-etica
[4] Pitts, L. (2015). ESL Role Plays: 50 engaging role plays for ESL and EFL classes. ECQ Publishing.

1.3 Dissemination

Three main dissemination activities have been carried out so far.

- First, the main components of the system were presented at UTTER 2nd User Day, an online event held on July 5, 2024. The video of DeMINT's presentation is available on YouTube.[5]
- A paper describing the chatbot with the title "A Conversational Intelligent Tutoring System for Improving English Proficiency of Non-Native Speakers via Debriefing of Online Meeting Transcriptions" has been accepted to the 13th workshop on Natural Language Processing for Computer-Assisted Language Learning (NLP4CALL). The paper provides a detailed description of the system's architecture. It will be presented as an oral presentation during the conference to be held in Rennes (France) on 25–26 October 2024.
- The project's code is hosted on a GitHub repository.[6] The repository contains the code for the chatbot and the preprocessing pipeline, together with the scripts to fine-tune the models. The `README` file contains additional information and instructions on how to run the system. Code availability will contribute to the dissemination of the project.
- A dataset consisting of the audio recorded in the videoconferences held for the pilot evaluation of the chatbot, together with their transcription, will be published under an open license. A fine-tuned error-preserving speech-to-text Whisper-based model and the corresponding training dataset will be published in the HuggingFace hub.

We plan to disseminate the project results within our university, and also to write a second paper focused on the outcomes of the human evaluation.

1.4 Ethics

Since the human evaluation involves collecting and distributing data from participants, special care has been taken to adhere to relevant ethical guidelines and applicable data protection laws. Specifically, the research ethics committee of University of Alicante has overseen the experimental process. Each participant was informed about how their interaction with the model would be used and disseminated, and they signed a consent form. Additionally, participants' personal information has been pseudonymized in the released data.

A data management plan was elaborated. It addresses issues such as data collection, data generation, data sharing, property rights and privacy, and long-term preservation and re-use, in compliance with national and EU legislation. A copy of the data management plan can be found in appendix A.

[5] https://www.youtube.com/watch?v=TzEK9JlxVH4
[6] https://github.com/transducens/demint

2 Summary of Results and Plans

2.1 Results

The development of the entire system pipeline has been completed, and the project's code has been released as open-source software. The system is capable of analyzing video conference transcriptions and providing feedback to students.

Feedback from participants in the human evaluation was gathered from two perspectives: first, the overall user experience with the chatbot, and second, the chatbot's effectiveness as an intelligent English tutor. Participants rated their responses to the evaluation questionnaire on a Likert scale from 1 to 5, with 5 representing the highest score for all aspects evaluated.

Regarding the first aspect, general user experience, participants were generally satisfied with the tool's performance and response time. In response to the question, *"Did you enjoy interacting with the chatbot?"*, all participants gave positive feedback, with a score of 4 or 5. However, fluency emerged as the system's main area requiring improvement, with a average score being 3.

In terms of the chatbot's performance as an intelligent English tutor, the overall evaluation was positive, though some areas still require enhancement. The main concern of the participants in this evaluation was the accuracy of the chatbot in identifying speech errors, which received with an average of being 3. Other aspects, such as the chatbot's ability to understand their queries, or the usefulness of examples and resources provided by the chatbot, were rated with an average score of 3.3 The clarity of the chatbot's error explanations received a slightly higher average score of 3.4. Notably, most participants agreed that the chatbot helped improve certain aspects of their English, with five out of seven giving a score of 4 for this question. Additionally, when asked whether they would be interested in using a similar chatbot in future video conferences, all participants but one gave scores of 4 or 5, demonstrating a general interest in such tools.

2.2 Business plan

Currently, there is no business plan for the project as it is in a prototype stage. In case additional funding is secured for a long-term project, a second phase of development could lead the system to a more mature stage, and a business plan could be considered.

2.3 Future plans

The system is planned to be improved via students' master theses and other local projects. Potential funding opportunities will also be explored.

Potential work for an improved version of the system includes the following:

- Supporting voice cloning to fine-tune Whisper with each student's voice before using the tool. The student will speak a few sentences and the fine-tuning data coming from C4-200M and COREFL will be used to train a customized speech-to-text model. Models such as XTTS-v2[7] could be used for this purpose.
- Considering new models such as Tower,[8] a model that also performs, among other tasks,

[7] https://huggingface.co/coqui/XTTS-v2
[8] https://unbabel.com/announcing-tower-an-open-multilingual-llm-for-translation-related-tasks

grammar error correction (GEC).

- Making the interaction with the chatbot more engaging and speech-based.
- Improving the error detection capabilities of the system, and the heuristics used to prioritize the errors to be discussed.
- Incorporating human teachers to either evaluate the error detection capabilities of the system or the interaction between chatbot and students from the point of view of the teacher.
- Integrating knowledge from theories of second language acquisition to improve the system's effectiveness.

2.4 Blurb for public dissemination on UTTER's website

DeMINT ("Automated Language Debriefing for English Learners via AI Chatbot Analysis of Meeting Transcripts") has developed a prototype of a conversational system designed to enhance non-native English speakers' language skills through post-meeting analysis of the transcriptions of video conferences in which they have participated. The code of the system is already available as open-source software on https://github.com/transducens/demint. Future plans include developing a more engaging and speech-based interaction with the chatbot and knowledge from theories of second language acquisition.

3 Recommendation by Project Sponsor

The DeMINT team has done an excellent job at delivering all results promised in the project proposal. A publication at a relevant workshop and an open-source codebase for the tool developed have been disseminated. The project has not deviated in any major way from what was proposed. The project sponsor, therefore, makes a positive payment recommendation for the DeMINT project.

A Data Management Plan

A.1 Introduction

DeMINT (Automated Language **De**briefing for English Learners via AI Chatbot Analysis of **M**eet**in**g **T**ranscripts) is a project funded in 2024 by the EU project UTTER (Unified Transcription and Translation for Extended Reality) via the Financial Support to Third Parties (FSTP) feature, also known as cascade funding. This Data Management Plan (DMP) provides an analysis of the main elements of the data management policy that have been used in DeMINT with regard to all the datasets collected for or generated by the project.

This document addresses issues such as data collection, data generation, data sharing, property rights and privacy, and long-term preservation and re-use in accordance with national and EU legislation. We describe the types of data that are collected and generated. We also look at data storage and retention, as well as data protection and the implications of the regulations to be enforced on project data, particularly with regard to the protection of personal data.

A.2 Data collected

DeMINT is developing an AI chatbot to act as a tutoring assistant for non-native English speakers. Data will be collected from both open repositories on the web with a license allowing their use for research purposes, and from user interactions through an online meeting application and through the tutoring assistant itself.

The data downloaded from open repositories will be used to train the models used by some of the individual components used by the DeMINT pipeline. They will also be used to automatically evaluate these components during their development.

Data collected from user interactions through an online meeting application will be used to conduct a human evaluation of the DeMINT tool. For this purpose, users will be recorded (both audio and video) during online meetings in which they will perform role-playing activities to mitigate the need for anonymizing the data later. After the communication is recorded, each user will follow up with a session with the chatbot to analyze the mistakes made during their meeting. Additionally, these written conversations between the user and the chatbot will also be recorded.

A.3 Data generated

We distinguish four main categories of data that will be generated during the project:

- Data generated from existing datasets to be used for training and evaluation of the individual components of the DeMINT tool.
- Audio and transcription of the online role-playing meetings.
- Debriefing text generated from the interaction with the DeMINT assistant.
- Software, models, algorithms, etc.
- Academic research publications.

A.4 Data storage, preservation and re-use

The project data will be stored in private or public repositories — depending on the privacy level of the data, see next section. The **private repositories** will reside on a machine at the Department of Software and Computing Systems of the Universitat d'Alacant and will only be accessible from the department's local network by people involved in the DeMINT project. The **public repositories** will be on GitHub, and data that is relevant will be linked from CLARIAH-ES[9] (the Spanish node of CLARIN[10]).

All data in public repositories produced during the project will be made available under free/open-source licenses.

[9] https://www.clariah.es/
[10] https://www.clarin.eu/

A.5 Privacy: levels of access and sharing

There are different categories of data collected or generated during the project, with different levels and conditions for access and sharing:

Audio and video: The raw data collected during user interactions for evaluation purposes (both audio and video) will be private. This data will be used to generate the audio and text data to be distributed under a free/open-source license. Once the data to be distributed has been generated, original data will be removed from our servers.

Identification data: This data will be associated with the audio and video collected, and will be private. It will consist of name, family name, e-mail address, gender, age, mother tongue, academic background, and level of English.

Software and models: Software, models and algorithms will be public and released under free/open-source licenses.

Debriefing text: The text of the learning interactions with the DeMINT tool will likely be made public and released under a free/open-source license. However, if it is ultimately deemed not useful for the community, it will remain private.

Data for training/testing: Data used to train or test the DeMINT tool may directly come from public repositories or be generated from other public datasets. In the latter case, we will make it public and release it under a license as open as possible, compatible with the license of the original dataset.

Academic research publications: Academic publications will be made available as "green" open-access via institutional repositories.

A.6 Personal data protection measures

This section sets out how we will identify where personal data is involved and how that personal data will be protected. Security and privacy issues will be taken into account when designing the architecture and information flow. The processing of personal data is in accordance with the data protection regulations of the Universitat d'Alacant and Spain.

- Participants will be informed of how the data will be collected, processed, and stored. To participate, they will need to be of legal age and sign an informed consent form. Identification data provided by participant (see above) will be properly safeguarded on a file in a private repository.
- To protect personal data in the data collected during user interactions (audio and video), individuals participating in such interactions (which will be recorded for evaluation purposes) will be instructed not to reveal any personal data. They will participate in role-playing activities that are unrelated to their real identities which will make the restriction considerably easier to attain.
- In case a pilot demo is put online for dissemination purposes, the interactions with users as well as the text generated by the DeMINT tool will only be temporarily stored for the purpose of functioning on the tool during the debriefing session. After that, all data will be permanently deleted.

B Feedback form filled by participants in human evaluation

General instructions: In this form, we ask you to evaluate your experience based on your interaction with the conversation bot from the DeMINT project. All questions are answered by assigning a score from 1 to 5 to assess your experience across different aspects of the bot's functionality. If any of the responses you provide are published, they will be done anonymously and aggregated with the responses of other participants in this evaluation.

In your opinion, was the conversation with the bot smooth?

Give a score from 1 to 5, with 1 meaning *Not smooth at all*, and 5 meaning *Very smooth*

Did you enjoy the experience of using a conversation bot like the one you used in this evaluation?

Give a score from 1 to 5, with 1 meaning *I didn't enjoy it at all*, and 5 meaning *I enjoyed it a lot*

Do you think the bot's response time to each of your interactions was too long?

Give a score from 1 to 5, with 1 meaning *Yes, it took too long to respond*, and 5 meaning *No, I think the response time was appropriate*

Do you believe the conversation chat accurately detected the mistakes you made when using English in the video conferences you participated in?

Give a score from 1 to 5, with 1 meaning *The error detection was terrible*, and 5 meaning *The error detection was very accurate*

Did you find the explanations provided for each mistake clear?

Give a score from 1 to 5, with 1 meaning *Not clear at all*, and 5 meaning *Very clear*

Did you find it difficult to make the bot understand your doubts regarding the detected errors?

Give a score from 1 to 5, with 1 meaning *It was very difficult*, and 5 meaning *It was very easy*

Do you think the time spent in your conversation with the bot to identify, explain, and help you improve your English based on the detected errors was excessive?

Give a score from 1 to 5, with 1 meaning *Yes, it spent too much time on the same errors*, and 5 meaning *No, the time spent on each error did not seem excessive at all*

Do you think the time spent in your conversation with the bot to identify, explain, and help you improve your English based on the detected errors was insufficient?

Give a score from 1 to 5, with 1 meaning *Yes, it spent too little time on some errors*, and 5 meaning *No, the time spent on each error did not seem insufficient at all*

Did you find the resources (examples, exercises, etc.) provided by the bot to help you improve your knowledge of English related to the mistakes you made useful?

Give a score from 1 to 5, with 1 meaning *No, the resources were not useful at all*, and 5 meaning *Yes, the resources were very useful*

Do you think that interacting with the conversation bot has helped you improve your spoken English in any way?

Give a score from 1 to 5, with 1 meaning *I don't think it helped me at all*, and 5 meaning *Yes, I think some of the corrections or suggestions it provided helped me improve*

Do you think more sessions with a conversation bot like this one (in future improved versions) could help you improve your English?

Give a score from 1 to 5, with 1 meaning *I don't think it would help me at all*, and 5 meaning *I think it would help me a lot*

Would you use a conversation bot like this again (in future improved versions) to improve your English after participating in English video conferences?

Give a score from 1 to 5, with 1 meaning *No, I don't think I'd like to use it again*, and 5 meaning *Yes, I'd love to use it after each video conference*

B.6 SURE-GB

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Evaluation – SURE-GB

Identifying Stereotypical, Under-representational, and Algorithmic Gender Bias in Machine Translation

Nature	Evaluation Report	**Work Package**	WP1
Project start date	dd/mm/2024	**Project end date**	dd/mm/2024
Interim meeting	dd/mm/2024	**Report submission Date**	dd/mm/2024
Main authors	Chrysoula Zerva & Ben Peters (IT)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Evaluation of Project Execution

The **SURE-GB (Stereotypical, Under-representational, and Algorithmic Gender Bias in Machine Translation)** project aimed to develop an automated system to detect and mitigate gender bias in machine translation (MT) system outputs, focusing on occupation-related language across English, French, and Greek. The project addressed gender bias by developing a curated Knowledge Graph (KG) that integrates real-world gender occupation statistics from official sources at EU and national levels as well as linguistic features from large corpora. The SURE-GB team aims to employ this KG to identify different sources of gender bias in occupational terms.

The evaluation of the project in terms of the submitted deliverables, published work, and online meetings and presentations, revealed several strengths, some challenges, and a great potential for future expansion and exploitation of the released resources.

1.1 Project Execution and Achievement of Milestones

The project successfully met key milestones, ensuring timely completion of deliverables. The core milestones, including data collection, KG development, bias detection system implementation, and dissemination, were achieved with only minor deviations.

Milestones 1 and 3: Data Collection and Knowledge Graph Development The project team constructed a **Knowledge Graph** based on standardized data, such as EU-LFS and ISCO-08, encoding occupation-related gender statistics for English, French, and Greek texts. The KG structure integrates employment data from Greece, France, and the United Kingdom, alongside language-specific textual corpora which were selected based on popularity of usage for testing and training MT models. This resource enables analysis of gender biases at different occupational levels, providing a detailed representation of how occupations are "gendered" in the labour market versus textual data used for MT. There were no deviations in the promised implementations and both collected data and the KG are available on the team's GitHub page with the provided code to reproduce the KG with different data as well.

Milestone 2: Gender Mismatch Detection Module The **gender mismatch detection tool**, developed as part of the project, successfully identified disagreements in gender assignment between source and target texts during MT. The team explored several alternatives (including some suggested in the interim meetings) and chose to use a hybrid method that relied on large language models (LLMs) for the detection of terms of interest. They evaluated on known corpora and showed their method to have high accuracy when using open-access LLM variants (Llama 2) that are on the "larger" side (70B parameters). The rest of the mismatch detection pipeline relied on traditional NLP modules (i.e. the SpaCy library) for gender detection.

Milestone 4: Gender Bias Detection System The team built a fully operational **bias detection system** that leverages the KG to classify gender biases in MT outputs. By analyzing gender representations within official statistics and textual datasets, the system was able to detect three types of bias: *under-representational*, *stereotypical*, and *algorithmic* bias. Hence, the proposed method can uncover significant flaws in existing MT systems, particularly in cases where the biases are misaligned with real-world occupation statistics (i.e. in cases of stereotypical or algorithmic

bias). The authors demonstrated their system on the UTTER user day, and the implementation is accessible on the team's GitHub page.

Milestone 5: API Integration An **API endpoint** was developed to facilitate the integration of the gender bias detection tool into other systems. Although the core functionality was delivered, it has been noted that there is no guarantee for the maintenance and continued deployment of the API upon completion of the service, due to additional costs that would need to be covered. Although this differs from the plan, it does not constitute a significant deviation because the source code for the implemented tools is available and the API is documented.

Milestone 6: Dissemination The project successfully disseminated its findings through academic publications, open-source repositories, and a public video presentation, with plans for additional publications.

1.2 Key Achievements, Strengths and Challenges

Overall, the presented work constitutes a significant interdisciplinary contribution that bridges knowledge from the social sciences and real-world occupational analytics with statistics from textual data used for training LLMs. The proposed method provides a novel way to "decompose" bias observed in the output of LLMs used for MT and could help better understand and mitigate biased text generation. As a further point for improvement, it would be interesting to consider whether there are additional resources that could be used to understand the sources of algorithmic bias, i.e. bias that cannot be attributed to imbalances found in either the occupational or the textual data.

Additionally to the automated evaluation of existing benchmarks, the authors are carrying out a human evaluation campaign to further validate their methods and obtain a relevant dataset, which could provide a useful resource for evaluating future models and methods. However, the team noted that the annotation process is demanding because it requires annotators to be familiar with EU standards for occupational classification (ISCO standards); annotators with the relevant expertise are difficult to find.

It should be noted that although French and Greek use different scripts and come from different Indo-European subfamilies (Romance and Hellenic, respectively), they provide a limited picture of the diversity of grammatical gender phenomena in the world's languages. It would be great to see expansion of this work into further languages in terms of both presented analysis and human annotations.

Overall, the project demonstrated an interesting perspective on addressing gender bias in MT and language generation systems and laid the groundwork for future research in this critical area.

1.3 Overall Recommendation by Project Sponsor

Given the clear achievements, adherence to the proposal, and the potential for future impact and expansion, The SURE-GB team has shown more than satisfactory performance and already produced several outputs. Hence, it is recommended that the SURE-GB project receives the final funding part as originally planned.

B.7 InCroMin

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – InCroMin

Interactive Crosslingual Minuting

Nature	Final Report	**Work Package**	WP1
Project start date	01/01/2024	**Project end date**	30/09/2024
Interim meeting	26/04/2024	**Report submission Date**	30/09/2024
Main authors	XXX Sponsor (PARTNER)		
Co-authors	Ondřej Bojar, Marko Čechovič, Natália Komorníková, Dominik Macháček, Peter Polák (CUNI)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	–
v1.0	**Status**	Final	30/09/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

The planned goals of InCroMin were:

1a) To further develop MinuteMan by merging it with UTTER components.
Deviation: We consulted MinuteMan integration options with the UTTER team. We agreed on not merging MinuteMan with UTTER components because at that point, there was not a single UTTER pipeline. MinuteMan can be regarded as a standalone item in the collection of UTTER tools.

1b) To expand MinuteMan for the multi-lingual meeting situation.

No deviation.

2) Continue our research on multi-source speech transcription and translation, in order to benefit from human interpretation live.

Deviation: While interpreters now regularly serve in remote calls, it would be substantially more difficult and expensive to obtain examples of such calls. We therefore shifted our research focus on a different challenge: automatic interpretation from sign language.

3) Collect a test set of naturally multilingual meetings (harder to get access to), or as a fall-back monolingual meetings (easier to get access to and also already partially available to us) with cross-language access needs to them.

No deviation.

4) Evaluate multi-lingual interactive meeting summarization in practice and on the test set.

Deviation: Early test calls in the MinuteMan platform indicated that live meeting summarization quality is insufficient. We thus focused on the first step only: evaluating the cross-lingual meetings themselves, not their summaries. We nevertheless prepared two datasets (test-set size) for cross-lingual meeting summarization: many of our test calls have manually created minutes by our annotators (Appendix A.2) and we translated ELITR-Bench into Czech (see Objective 4 below).

Planned InCroMin project outputs:

1) Interactive meeting summarization tools (based on MinuteMan) extended for multilingual use.

As decribed in Goal 4 above, we limited our attention to summarization and instead focussed on the multilingual use of MinuteMan and the effectiveness of cross-lingual meetings as such.

2) A curated corpus of meetings, usable for evaluation and testing of speech translation and meeting summarization into minutes. Similar to ELITR Minuting Corpus but with multi-lingual meetings.

No deviation.

3) **Speech transcription and translation models modified for multi-sourcing, so that they can benefit from human interpretation.**

   Deviation: Instead, we deliver progress in sign language translation research.

Planned dissemination:

- **For academic and educational sector.**

  Negligible deviation: we stayed at university level of education. We tested cross-lingual MinuteMan use for academic and education use in several calls (consultations on both content of study as well as life and organizational matters) but we did not test the tool with any high-school level institutions.

- **For citizen support sector.**

  No deviation. Our test calls include professionals from the Integration Center Prague (ICP), an NGO focused on citizen support.

- **Publish the results of our research in the relevant top-tier peer-reviewed research conferences or journals.**

  Pending. While we have gathered enough content, the 9-month project time span proved too short to be able to polish it into a paper submission.

1.2 Development

Objective 1. Facilitate cross-lingual access to meeting transcripts/translations.

Multilinguality in MinuteMan Before InCroMin, MinuteMan (Kmječ and Bojar, 2023) was an application for real-time meeting transcription and summarization that supported only one language – English. Within InCroMin, we added the multilingual support, to enable cross-lingual communication between meeting partners who do not share a common language. We carried out the following steps:

- **Software Engineering** – MinuteMan required significant software engineering improvements, such as better logging, more stable deployment, and resolving data persistence issues. Next, the user interface was simplified based on initial users' feedback. After following these upgrades, multilingual transcriptions and summaries for cross-lingual meetings could have been introduced.

- **Multilingual models** – We integrated multilingual automatic speech recognition Whisper large-v3 (Radford et al., 2022), and a multi-lingual MT model NLLB (Costa-jussà et al., 2022) that is capable of translating English into 200 other languages. Implementation of these two models went quite smoothly, but some parameter tuning needed to be done to achieve suitable results.

- **Automatic data collection** system was implemented the collection of InCroMin Test Calls corpus (ref. Objective 3).

During the development of these features, challenges were encountered with the application design, particularly regarding extensibility, as well as user-technical difficulties during testing. Most issues were resolved, but some questions remain, particularly concerning audio recording buzzing, which may be caused by inconsistencies in the online web calling platform clients and were sometimes resolved by simply reconnecting the participants.

Based on the data and experience collected, the following further improvements are proposed: (1) Better voice activity detection to avoid translating noise into non-sense sentences, or cutting off unfinished sentences too aggressively. (2) More complex machine translation (MT) pipeline for specific language combinations, such as direct Czech-Ukrainian MT (Popel et al., 2024), could help with the translation quality. (3) Users' feedback-driven development of a user interface for asking questions about meetings would be beneficial. The last mentioned suggestion is in line with the idea of ELITR-Bench test set, see page 10.

Sign Language Translation Using a sign language or signs for communication is natural for many people, including but not limited to deaf and hard of hearing and their family members. In USA, there are 9 millions of people who report using signs in any period of their lives (Mitchell and Young, 2022). In other countries, the same proportion of 2.8% of sign language users in the population is assumed. In anyway, sign language users would largely benefit of being included into cross-lingual communication thanks to machine translation support, so the sign languages are very relevant for language technology providers. The problem is that the machine translation of sign languages is relatively underdeveloped. The state of the art, such as Zhang et al. (2024), require large computational and data resources, and still report results practically unusable in real-life application. Therefore, we aimed to focus on sign language translation research.

For that, we joined an international and interdisciplinary team of researchers at the intensive research workshop JSALT 2024.[1] The team worked on the research prototype "SignLLaVA: Sign Large Language and Visual Assistant." It is based on the general concept of LLaVa (Liu et al., 2023a,b). SignLLaVa inserts sign language video features converted through projector layer into the common vector space of the language tokens of the Llama large language model (Dubey et al., 2024). SignLLaVa aims to translate American Sign Language (ASL) into English and serve as an assistant that can e.g. answer questions about the sign language video.

The unique contribution of SignLLaVa includes applying three complementary sign language representations, a demonstration tool, and an effort to create new authentic ASL-to-English MT test set, which would resolve the problems of currently common state-of-the-art How2Sign data set (Duarte et al., 2021), which is created for the opposite direction English-to-ASL. There is an unrealistic optimal segmentation into sentences, and other problems. The progress in the ASL-to-English test set includes a plan to identify suitable ASL videos that have English translations and are available on the Internet, and a small initial probe with several videos.

InCroMin team members complemented the SignLLaVa team by the necessary expertise in machine translation, including the MT evaluation and creating the test set. Moreover, they contributed by e.g. text data normalization, and application of LLMs to texts for multi-tasking and evaluating the sign language representations. The SignLLaVa team did a significant progress towards application of sign language translation into MinuteMan and other cross-lingual communication tools.

[1] https://www.clsp.jhu.edu/sign-language-translation/

Whisper-Streaming improvements Whisper-Streaming (Macháček et al., 2023) is a tool for real-time transcription and translation of 99 source languages that are supported by Whisper. It is a necessary component of cross-lingual meeting tools. In InCroMin, we focused on technical improvements of Whisper-Streaming, to enable its integration into MinuteMan and other similar tools. Within InCroMin, we added the following features:

- **Automatic language identification**, the same method that is implemented in Whisper. The language identification is applied on every update, so it allows fully automatic switching between the source languages.
- **Voice Activity Controller** using Silero VAD Iterator. Whenever there is e.g. 0.04-second audio chunk, we run an iterative Voice Activity Detection (VAD) model to detect beginning or end of speech that has to be processed. It improved quality by avoiding Whisper to process non-voiced segments, which often lead to hallucinations, and improved latency because the end of speech (a significant pause, 0.5 seconds) triggers immediate update, not waiting for the next chunk for confirmation.
- **OpenAI API backend**. A new alternative backend that does not require local hardware for deploying the Whisper model was proposed and implemented by one GitHub contributor in our cooperation. Our tests showed that processing trough API achieves the same quality as local processing, but the latency is much larger and unpredictable. There is also a significant cost for the API, with 1 second audio chunk approximately 10-times higher than processing the offline audio once. Moreover, the API seemed to be changed in the newer version and following code maintenance is needed. The faster-whisper backend with local deployment of Whisper model is still the most recommendable option, but the API showed an alternative way that may be useful for some applications.

There is relatively large and active community of users and developers of Whisper-Streaming on the open-source code repository GitHub, documented by nearly 1 800 stars and 200 forks in September 2024.[2] Within InCroMin, we cooperated with them, responding to their issues and pull requests. Within InCroMin period—between 01/01/2024 and 24/09/2024—we responded to 77 issues or pull requests. Although many of them were relatively simple clarification questions, issues with installation, usage or quality, several issues or pull requests led to large valuable improvements, such as Voice Activity Controller, and some others to small but useful improvements such as removing duplicated variable, improved debug logging, adding warmup file to make Whisper-Streaming server to process the very first audio chunk faster, etc.

In summary, we were managing and supporting the open source community around Whisper-Streaming, and we gained lots of benefits that contributed to the InCroMin goals.

Objective 2: Facilitate cross-lingual access to meeting minutes.

The original goal of InCroMin was to provide participants of the meetings with live summary of the meeting in their language. This goal was unfortunately too ambitious for the short time span of the project. We nevertheless conducted initial experiments with LLMs for summarization. We used LLMs with 7 billion parameters (specifically this mlabonne/NeuralBeagle14-7B)[3] to obtain

[2] https://github.com/ufal/whisper_streaming
[3] https://huggingface.co/mlabonne/NeuralBeagle14-7B

summarizations, showing promising results that could be tailored to user's preference by adjusting the initial prompt. However, due to the complexity of the MinuteMan's architecture, these enhancements were not integrated. The LLMs also demonstrated the ability to reliably answer questions about the meetings, which could be particularly useful for future references. For the time being, we did not evaluate this step yet and only managed to prepare the relevant test set, see ELITR-Bench Czech on page 10.

Objective 3: Prepare test sets for cross-lingual meetings.

While low-latency speech translation systems are publicly available and sometimes even built to remote conferencing platforms (e.g. Zoom) or to cell phones (Samsung Live Translate),[4] we believe that the true level of "penetration" of the language barrier when making a cross-lingual call has not been properly assessed yet.

To this end, we organized calls between participants with no mutual language understanding, so they have to rely on speech-to-text translation provided by our MinuteMan. The calls are topically diverse, spanning from travel or living-abroad experience up to regular examples of project meetings or technical consultations. We collected the sound and transcripts of the calls, we manually revised them, and organized them in the **InCroMin Test Calls** corpus. We also collected experience of the participants in a feedback questionnaire. To complement the subjective assessment, and to start a semi-formal analysis of misunderstandings in communication, we equipped the corpus with an annotation of misunderstandings. Through this we hope to get a better idea of how frequent misunderstanding are, how often they can be attributed to speech translation errors, what needs to be fixed first to reduce misunderstandings count, and also if large language models are capable of identifying misunderstandings in meeting transcripts.

Details on this activity are provided in the report in Appendix A

Objective 4: Rigorous evaluation of underlying models.

Speech translation support for cross-lingual calls can be realized using a considerable number of architectures and technical components. Some setups can be end-to-end, with one model achieving the full needed process, some setups can first transcribe the sound and then translate the transcribed text to the target language.

Thanks to InCroMin, we contribute to rigorous evaluation of the necessary components on two fronts:

- **Analysis of automatic evaluation of speech translation latency**. A crucial aspect of a speech translation system for cross-lingual meetings is its latency, i.e., the duration required to provide the translation to the users. Unlike in speech recognition, latency measurement in translation is complex due to reordering. In recent years, researchers have proposed various latency measures. Currently, the IWSLT campaign (Ahmad et al., 2024) employs five different latency metrics to evaluate submitted systems. However, our detailed analysis in Appendix F reveals that these metrics do not correlate strongly with each other. To improve

[4] https://www.samsung.com/latin_en/support/mobile-devices/how-to-use-live-translate-for-phone-calls-on-the-galaxy-s24/

the reliability of InCroMin system evaluations, we examine which metrics reflect actual latency with the highest precision.

Results show that the DAL (Arivazhagan et al., 2019) is the most robust metric and should be used for comparing different systems. Alternative latency metrics can be employed in system development, as they generally show a good correlation with true latency when the compared systems are similar.

More details are in Appendix F.

- **Dialogue Translation Test Sets** Dialogue and conversational content in general are not yet well covered in established evaluation campaigns such as WMT or IWSLT, although they bring specific challenges not seen e.g. in news text or monologues. We used InCroMin funds to prepare test sets geared towards InCroMin goals for recent as well as future use: WMT24 evaluations, MultiWOZ Czech and German Small Test Set, and ELITR-Bench Czech.

WMT24 Translation Evaluation The Czech-to-Ukrainian WMT24 test set source was collected through the Charles Translator for Ukraine[5] (Popel et al., 2024). With users' consent, the service can log their inputs for the purpose of creating a dataset of real use cases. The data includes cross-lingual dialogues of Czechs communicating with Ukrainians (mostly refugees). There is also a subset of originally spoken data from users using the Charles Translator mobile app. The Czech source translations were translated to Ukrainian by professional translators (funded from other sources) in order to create the gold reference for WMT 2024[6] and InCroMin funds were used for the manual evaluation of this language pair and also English-to-Czech. The results will appear in WMT24 Findings.

MultiWOZ Czech and German Small Test Set For the purposes of evaluation of automatic translation of meeting transcripts, we had professionally translated a small portion of the MultiWOZ (Ye et al., 2022)[7] dialogues dataset from English into Czech, and as a second step from Czech into German. This two-step procedure was adopted to maintain consistency in important linguistic features which are often not explicit in English but which are expressed in Czech and German, primarily the gender and level of politeness. We instructed the translators to attribute gender to the speakers arbitrarily and ensure consistency within each dialogue.

The actual use of these test sets for MT system development or selection was not carried out in the limited time of InCroMin. Instead, we will include these test sets into WMT25 General Translation Task and possibly also WMT25 Test Suites, where we would be assessing phenomena critical for dialogue fluency such as gender and politeness preservation.

Interleaving our translated dialogues with each other or with the original English version allows us to construct also simulated cross-lingual setting and develop multilingual access methods for this content. This is however left for future work.

MultiWOZ Czech and German Small Test Set is available upon request from Ondřej Bojar, until it will have served in WMT evaluations.

[5] https://translator.cuni.cz/
[6] https://www.statmt.org/wmt24
[7] https://github.com/smartyfh/MultiWOZ2.4

Q: Who were the participants of the meeting?
A: [PERSON14], [PERSON10], [PERSON5], [PERSON9], [PERSON1], [PERSON11]
Q: What was the main purpose of this meeting?
A: Discuss and finalize the technical setup for a demo
Q: How many scenarios were discussed?
CONTEXT-FREE Q: How many scenarios were discussed for the demo setup?
A: 3 (plans A, B and C)
Q: Which scenario was chosen eventually?
CONTEXT-FREE Q: Which scenario was chosen eventually for the demo setup?
A: Plan C

Figure 1: An illustration of context-sensitive and context-free variants of ELITR-Bench questions as provided to translators.

ELITR-Bench Czech ELITR-Bench (Thonet et al., 2024) is a collection of questions and answers in English created to complement English meetings from the ELITR Minuting Corpus (Nedoluzhko et al., 2022) with questions and golden-truth answers. This test set is meant to evaluate the accessibility of information in meeting minutes with QA systems or LLMs.

In InCroMin, we had ELITR-Bench questions professionally translated from English into Czech. This variant of the test set allows anyone to evaluate *cross-lingual* access to meeting content: asking questions in Czech, locating answers in English meeting minutes, reporting answers in Czech and comparing them to the golden-truth Czech answer.

ELITR-Bench contains questions in two settings: as a continuous dialogue where the formulation of the question may need the previous context of the dialogue, and as independently formulated questions. When preparing the translation batches for translators, we noticed that these two settings overlap greatly. Of the total of 271 question we were translating, only 33 have a separate context-free formulation. We substantially saved the translation costs by providing translators with the 271+33 questions in a single sequence, with context-free variant coming right after the context-dependent one, as illustrated in Figure 1.

ELITR-Bench Czech is available upon request from Ondřej Bojar, to prevent LLMs learning from it accidentally.

1.3 Dissemination

Research seminars InCroMin was twice briefly presented as an ongoing project at research seminars within one-hour lecture of Dominik Macháček who presented his PhD. research "Multi-Source Simultaneous Speech Translation." First, at the Linguistics Mondays at ÚFAL MFF CUNI (Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics) on 04/03/2024.[8] There were around 20 participants on-site and online. The lecture recording on the institute's website[9] has 387 views, YouTube[10] reports 25 views and 318 subscribers.

Second, at the research seminar of the CSTR and ILCC groups at the University of Edinburgh, School of Informatics, on 22/04/2024. There were approximately 20 attendees.

[8] https://ufal.mff.cuni.cz/events/multi-source-simultaneous-speech-translation
[9] https://lectures.ms.mff.cuni.cz/view.php?rec=534
[10] https://www.youtube.com/watch?v=FpumkKjCJO0

JSALT 2024 Closing Day The SignLLaVa team presented their results at JSALT 2024 Closing Day on 02/08/2024. There were approximately 65 attendees on-site, the YouTube recording[11] has 566 views and 2 380 followers.

Machine Translation Marathon is a week-long gathering of machine translation researchers, developers, students and users. In 9/2024, it was organized at ÚFAL MFF CUNI[12] in Prague, Czech Republic, and there were approximately 50 participants. InCroMin was presented to the participants at several points:

- Project proposals,
- Project midweek and final reports,
- Whisper-Streaming demo at the open poster session.[13]

The collection of slides presented at MT Marathon is provided in Appendix C.

UTTER Users Days We presentated InCroMin project at UTTER Users Days online conference, highlighting the goals, some technical details, and progress of this project. Hopefully we were inspiring enough for other attendees to consider FSTP project funding as well. The slides are provided in Appendix D.

ELF Conference Ondřej Bojar presented InCroMin at the English as Lingua Franca international conference[14] organized by the Prague City University. The presentation raised attention because relying on speech translation goes against the spirit of the universal and inclusive use of English. It is however clear that the best approach is to combine InCroMin-like tools (in early stages, when mutual understanding is not possible) with gradual adoption of a common language such as English. As a follow-up of this dissemination, one test call was acquired and there are good connections established between us and some ELF participants for future joint research where the new colleagues would cover primarily "soft" aspects of the task, e.g. running InCroMin test calls with non-technical people, or evaluating the (mis-)understandings more rigorously.
The slides are provided in Appendix E.

In-house presentations We made use of the opportunity of ÚFAL offsite seminar to informally introduce InCroMin to our colleagues and students from our department we don't often get in touch with. As a result, several InCroMin test calls were again acquired. We also showcased the project to our colleagues in Gen Digital Inc. which sparked their interest and also motivated them to voluntarily provide us more test calls and valuable feedback. They were happy to assist with any further testing of similar applications.

1.4 Ethics

In InCroMin, we followed our established practice of consent collection. The practice was defined during the EU project ELITR and conforms to the standards of Charles University.

The consent form was updated to reflect InCroMin test calls corpus collection. See Appendix B for the full text.

[11] https://www.youtube.com/watch?v=65L7tklQbyc
[12] https://ufal.mff.cuni.cz/mtm24/
[13] https://ufal.mff.cuni.cz/mtm24/abstracts.html
[14] https://www.praguecityuniversity.cz/elf

The data collected into the InCroMin corpus were deidentified by removing all personal names from the transcripts and the subsequent documents (translations and summaries). The occurrences of participants' names were also silenced in the provided recordings.

Depending on the circumstances, the participants were paid or unpaid volunteers. We paid participants from the Integration Center Prague (IC Praha) and several other participants with no direct interest in language and speech technologies. Another set of volunteers was solicited from students of relevant subjects at Charles University and from participants of the JSALT 2024 workshop; these participants donated their time for free because they were interested in testing out the state-of-the-art system from their field of study. The last group were participants from MT Marathon. It was technically impossible to pay these participants (our university would require to prepare short-term contracts with them, which is simply infeasible during the one week of MT Marathon), so we provided them with a souvenir from Prague.

2 Summary of Results and Plans

2.1 Results

This is the summary of tangible outputs of InCroMin.

- Cross-lingual meetings tools:
  - **MinuteMan**: The project is fully open-sourced and well documented at GitHub page https://github.com/fkmjec/minuteman.
  - **Whisper-Streaming**: The commits and author's activity in the public repository https://github.com/ufal/whisper_streaming are outputs of InCroMin.
- Research results:
  - **Sign language translation**: The output is research progress documented in JSALT 2024 Sign LLM, Large Sign Language Model team final report.
  - **Analysis of latency measures for speech translation**: The output is technical report in Appendix F.
- Datasets:
  - **InCroMin Test Calls** is described here in Appendix A and the deidentified data are publicly available at https://github.com/ELITR/incromin-test-calls
  - **MultiWOZ Czech and German Small Test Set** is available upon request from Ondřej Bojar, until it will have served in WMT 2025 evaluations.
  - **ELITR-Bench Czech**, the translation of ELITR-Bench into Czech, is available upon request from Ondřej Bojar; aimed for future evaluations of LLM applicability in cross-lingual access to meetings.

2.2 Business plan

We do not have any business plan that would directly exploit InCroMin results.

2.3 Future plans

- **Research publications**. We plan to publish research publications with the following content:
  - Sign language translation research results.
  - Analysis of lantecy measures in speech translation.
  - InCroMin Test Calls, including misunderstanding annotation.
  - ELITR-Bench Czech.
  - MultiWOZ Czech and German Small Test Set.
- **Research projects**: We used InCroMin findings and preliminary results for prioritization and motivation for proposing future research projects, including but not limited to Horizon Europe MSCA postdoctoral fellowship project on live credible translation.
- **Student projects**: We propose projects and theses to Charles University students, such as cross-lingual communication tool with synchronization, multi-lingual post-editing, etc. A reimplementation of MinuteMan is proposed as a team software project.
- **MinuteMan**: Future plans with MinuteMan are: (1) Integrate Whisper-Streaming with enhanced voice activity detection as a module. (2) Finish the implementation of propagation of user edits of transcripts to other languages. (3) Optional automatic transcript scrolling. (4) Explore the idea of generating subtitles from the transcripts in real time so it could be embedded into a remote call software removing the need of external application for users that only need to watch and not interact with the transcripts.

2.4 Blurb for public dissemination on UTTER's website

In InCroMin, we examined and carefully evaluated the applicability of recent state-of-the-art speech-to-text translation tools in real cross-lingual calls, i.e. calls between parties that do not have a common language. We adapted MinuteMan (https://github.com/fkmjec/minuteman) for this purpose and collected a corpus of such calls. The deidentified part of the corpus is available here: https://github.com/ELITR/incromin-test-calls. Additional results of InCroMin include an evaluation of latency metrics for speech translation, translation of ELITR-Bench (https://github.com/utter-project/ELITR-Bench) into Czech to allow evaluation of cross-lingual access to past meeting content or translation of a part of MultiWOZ dialogues into Czech and German to assess translation quality of dialog-critical features such as participants' gender preservation. All the outputs are detailed in InCroMin Final Report.

3 Recommendation by Project Sponsor

From:	Laurent BESACIER
To:	Maryam Hashemi Shabestari
Cc:	Barry Haddow
Subject:	FW: InCroMin Final Report
Date:	woensdag 2 oktober 2024 18:04:45
Attachments:	SUBMITTED.pdf InCroMin Wrap-up Meeting Slides.pdf

---

Dear Maryam

Here is the InCroMin final report they submitted
We also had today the final wrap up call during which they shared some slides (also attached to this message)

Based on these two documents and on our discussions, here is our final assessment (which is positive for unlocking the 2d part of the project money) - they did not share the overleaf so we could not include it directly into a single report, sorry for that.

=======
We had a very productive review meeting for InCroMin. Overall, the project exceeded expectations. In summary, they:
- Adapted MinuteMan to support cross-lingual calls.
- Collected a new and potentially valuable corpus of simulated cross-lingual meetings.
- Conducted practical tests to assess the usability of the extended MinuteMan and identified areas for improvement.
For well-supported languages, MinuteMan appears close to being fully operational. Additionally, the potential founding of a spin-off for MinuteMan is under consideration, with FSTP funding playing a crucial role in bringing the system closer to production-ready.
Finally, even though it wasn't initially planned, InCroMin developed a Czech version of the ELITR-Bench meeting, which will soon be added to the UTTER/ELITR-Bench repository. This could also spark future collaboration between Naver and Charles University on cross-lingual Q&A on long documents (meeting transcripts).
======

Best
Laurent & Barry

-----Original Message-----
From: "Ondrej Bojar"<bojar@ufal.mff.cuni.cz>
To: "Laurent BESACIER"<laurent.besacier@naverlabs.com>; "Barry Haddow" <bhaddow@staffmail.ed.ac.uk>;
Cc: "Dominik Machacek"<machacek@ufal.mff.cuni.cz>; "Marko Cechovic" <markocechovic@gmail.com>; "naty komorka"<naty.komorka@gmail.com>; "Peter Polak, FW" <polak@ufal.mff.cuni.cz>;
Sent: Mon, Sep 30, 2024 22:29 (GMT+02:00)
Subject: InCroMin Final Report

Dear Laurent, Barry,

attached please find our final report for InCroMin.

I am not sure what the "Sponsor (PARTNER)" cell should be, so I put all of us authors into Co-authors.

Talk to you on Wednesday at 14.00 Prague time.

Thanks,

Ondrej.

--
Ondrej Bojar (mailto:obo@cuni.cz / bojar@ufal.mff.cuni.cz)
http://www.cuni.cz/~obo

B.8 pyannote.mobile

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP1 Final – pyannote.mobile

On-device streaming speaker diarization

Nature	Final Report	**Work Package**	WP1
Project start date	15/01/2024	**Project end date**	14/10/2024
Interim meeting	24/06/2024	**Report submission Date**	30/09/2024
Main authors	Marcely Zanon Boito, Laurent Besacier (NAVER LABS)		
Co-authors	Hervé Bredin (CNRS, IRIT)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v1.0	**Status**	Final	30/09/2024

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

No major deviation from the original plan.

1.2 Development

Objective 1. Streaming extension of `pyannote.audio` open-source toolkit

Support for streaming speaker diarization has been added to `pyannote` toolkit (Bredin (2023)).

The offline speaker segmentation model architecture (Bredin and Laurent (2021); Plaquet and Bredin (2023)) has been adapted to work in causal manner with support for variable latency (between 0ms to 1s). Changes include removing instance normalization step, switching from bi-directional to unidirectional internal recurrent neural networks, and adding a look-ahead mechanism. The inference pipeline has also been adapted to support this new type of causal segmentation model. This allowed us to dive deeper and more efficiently into the study of the latency/accuracy trade-off. The work achieved in this part of the project has been published at Interspeech 2024 (Rahou and Bredin (2024)). Paper abstract is repeated here for convenience:

> *We address the task of streaming speaker diarization and propose several contributions to achieve a better trade-off between latency and accuracy. First, computational latency is reduced to its bare minimum by switching to a causal frame-wise speaker segmentation architecture. Then, a multi-latency look-ahead mechanism is used during training to support adaptive latency during inference at no additional computational cost. Finally, we detail the method used during inference to achieve the final frame-wise segmentation. We evaluate the impact of these contributions on the AMI meeting dataset with a focus on the speaker segmentation step, seen through the prism of voice activity detection, overlapped speech detection and speaker change detection.*

Objective 2. Proof-of-concept of on-device streaming speaker diarization

A `pyannote` SDK targeted at iOS and macOS platforms has been developed and provides a fully end-to-end streaming speaker diarization API to be used in iOS and macOS apps.

Two demo apps depicted in Figure 1 (one on iOS and one on macOS) have also been developed to showcase how developers can use this new streaming SDK. Main features include live speaker diarization from the microphone (smartphone or laptop), batch speaker diarization from an existing recording, replay/visualization of the results, and export to CSV file format.

1.3 Dissemination

Research results obtained in *Objective 1* have been presented to the speech processing community at Interspeech 2024 in September 2024 in Kos (Greece), as a poster repeated in Figure 2 for convenience.

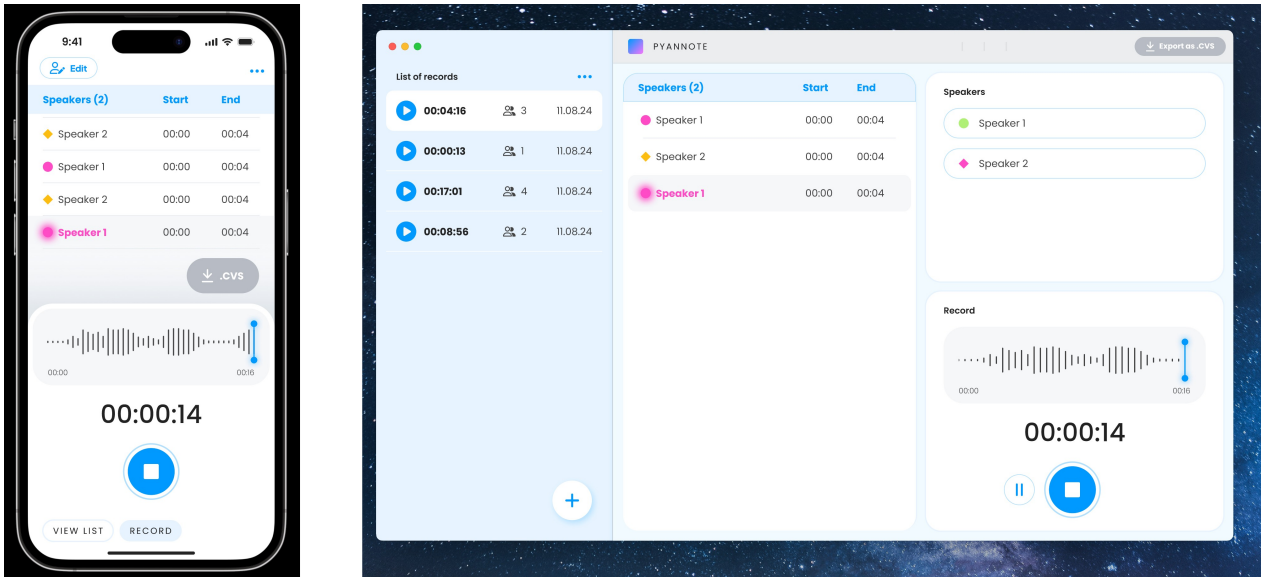


Figure 1: Screenshot of iOS and macOS demo apps

Streaming support has been added to `pyannote.audio` open-source toolkit. MIT-licensed code is currently being reviewed, already functional and available for anyone to try at the following address: `github.com/pyannote/pyannote-audio/pull/1544`.

Finally, beta builds of iOS and macOS demo apps are being uploaded to `Apple TestFlight` and we started granting access to a first batch of beta testers (mostly iOS app independent developers) for feedback.

iOS and macOS streaming speaker diarization SDKs are implemented in `Swift` and are not meant to be open-sourced in the short term as there are plans to commercialize them (see next section).

1.4 Ethics

This project dealt with the automatic processing of live recordings of conversations between several people and relies on deep learning models, that relies themselves on large collection of data that are known to contain societal biases.

We did not record any new personal data for the purpose of this project. We relied exclusively on existing and well-established academic speaker diarization benchmarks (such as AMI, VoxConverse, DIHARD, or AliMeeting).

Though the proposed technology can be used to help people better communicate (e.g. for deaf or hard-of-hearing people), we are also well aware that such speaker recognition technologies, if used by the wrong people, may lead to less desirable applications such as mass-surveillance for instance. We do believe, however, that open-sourcing the technology allows anyone to study and understand it, and therefore raise awareness of the general public.

2 Summary of Results and Plans

2.1 Results

We extended `pyannote.audio` open-source speaker diarization toolkit by adding support for the streaming scenario (where audio streams are processed in real-time instead of in batch after they completed).

We also developed a proof-of-concept of on-device (iOS/macOS) streaming speaker diarization, including an SDK implemented in Swift that we plan to distribute to interested actors in the field, through the local university tech transfer office.

2.2 Business plan

As stated above, we will work hand-in-hand with the local university tech transfer office (Toulouse Tech Transfer) in order to look for potential industrial partners interested in the iOS and macOS SDKs. In particular, `pyannoteAI` (a company building on top of `pyannote` open-source toolkit, co-founded by Hervé Bredin) will likely become one of the first user of this new piece of technology.

2.3 Future plans

Future plans include a collaboration with `pyannoteAI` university spin-off company, for them to distribute the real-time speaker diarization SDK to interested companies building iOS and/or macOS apps. Extension to other platforms such as Android phones, Raspberry Pi, or even edge device is also envisioned.

2.4 Blurb for public dissemination on UTTER's website

`pyannote.mobile` project led to the extension of the `pyannote.audio` open-source speaker diarization toolkit to perform speaker diarization in real-time while controlling the trade-off between latency and accuracy. It also led to the creation of an iOS/macOS streaming speaker diarization SDK which will be handed over to interested parties through the local university tech transfer office.

3 Recommendation by Project Sponsor

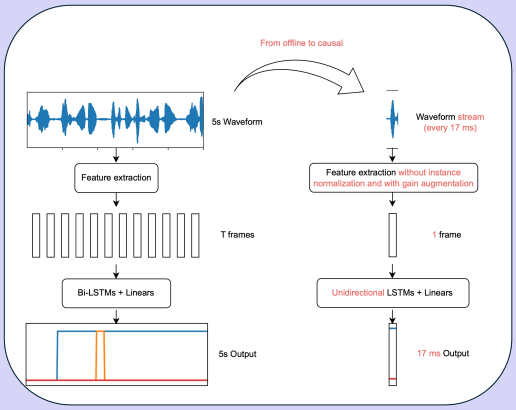
As the sponsor of this project, we confirm that the project successfully delivered its planned results. The dissemination efforts were effective, including an iOS application soon to be available, and a scientific paper published at Interspeech 2024. The project lead also provided comprehensive documentation and effective communication throughout the duration of the project, which were appropriate and well-aligned with the project's objectives. Overall, we recommend this project positively, as it has met its key objectives and demonstrated potential for future impact.

Multi-latency look-ahead for streaming speaker segmentation

Bilal Rahou Hervé Bredin

github.com/pyannote/pyannote-audio

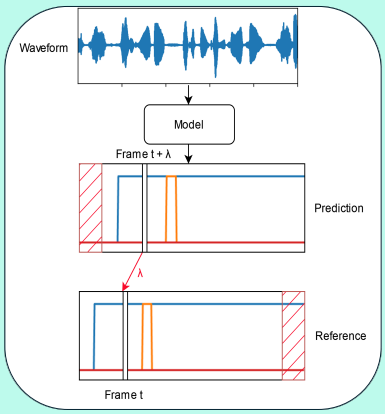
1 From offline to causal frame-wise segmentation model



The few changes (removing the instance normalization and the bidirectionality of the LSTMs) significantly worsen the performance of the model. The gain augmentation is added to compensate the lack of normalization. The table below summarizes the impact of these changes.

LSTM direction	Instance norm.	Gain augm.	5s chunk DER%
⟷	✓		17.3
⟶	✓		20.3
⟶			22.2
⟶		✓	21.1

2 Training with look-ahead



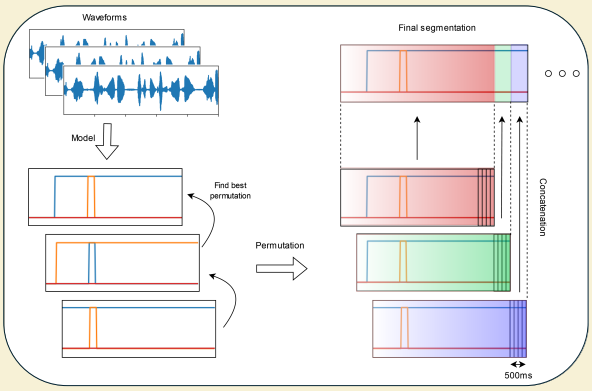
To improve the model's deteriorated quality, we introduce a look-ahead mechanism. The structure of the model does not change, the only change occurs during training, where the loss is calculated between shifted predictions and references. The shift λ corresponds to the added latency to the system. Below is the formula of the loss:

$$\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \min_{p \in \mathbb{P}} \mathcal{L}_{\text{CE}}\left(p(\mathbf{y}_{0 \to T-\lambda}), \hat{\mathbf{y}}_{\lambda \to T}\right)$$

The approach can easily be generalized to multiple latencies, though that needs a slight modification of the final classifier layer. We duplicate the final classifier layer K times, so that the model now outputs K predictions, one for each latency. The rest of the model is shared by every latency. The training loss is computed as the sum of the aforementioned look-ahead training loss over each latency:

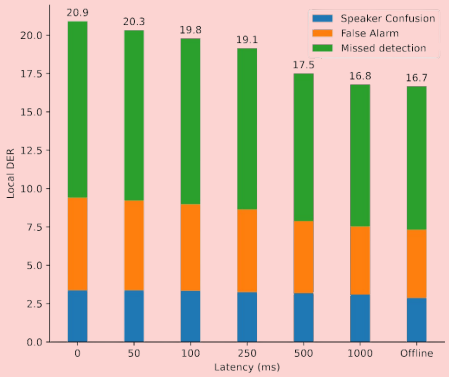
$$\mathcal{L}\left(\mathbf{y}, \hat{\mathbf{y}}\right) = \sum_{k=1}^{K} \min_{p \in \mathbb{P}} \mathcal{L}_{\text{CE}}\left(p(\mathbf{y}_{0 \to T-\lambda_k}), \hat{\mathbf{y}}^k_{\lambda_k \to T}\right)$$

3 Inference



To keep the advantages of hybrid speaker diarization approaches, we stick with an approach based on sliding windows (5s chunks with a 500ms stride). With the exception of the very first chunk that is used entirely, only the final 500ms of each subsequent chunk is used in the final concatenated output.

4 Results



As expected, the performance of the streaming system increases with the latency, almost closing the gap with its offline counterpart for a 1s latency (AMI).

Figure 2: Poster presented at Interspeech 2024

References

Hervé Bredin. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH 2023*, pages 1983–1987, 2023. doi: 10.21437/Interspeech.2023-105.

Hervé Bredin and Antoine Laurent. End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech 2021*, pages 3111–3115, 2021. doi: 10.21437/Interspeech.2021-560.

Alexis Plaquet and Hervé Bredin. Powerset multi-class cross entropy loss for neural speaker diarization. In *INTERSPEECH 2023*, pages 3222–3226, 2023. doi: 10.21437/Interspeech.2023-205.

Bilal Rahou and Hervé Bredin. Multi-latency look-ahead for streaming speaker segmentation. In *Interspeech 2024*, pages 1610–1614, 2024. doi: 10.21437/Interspeech.2024-923.

C Reports from Project Teams from Second Call

C.1 DETOEX

Unified Transcription and Translation for Extended Reality

(UTTER)

Horizon Europe Research and Innovation Action

Number: 101070631

D6/D1.2 – FSTP2 Final Report – DETOEX

Nature	Interim report		
Project start date	15/01/2025	**Project end date**	15/07/2025
Interim meeting	12/05/2025	**Report submission Date**	09/05/2025
Main authors	Sponsor (PARTNER)		
Co-authors	Awardees (ORG)		
Reviewers	Chryssa Zerva		
Version Control			
v0.1	**Status**	Draft	
v1.0	**Status**	Final	

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innova- tion (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

It should be mentioned that due to miscommunication with the organisers, the Grant Agreement was signed on Jan 20, 2025. As a result, the project actually started on that date, with the kickoff meeting with the assigned reviewer/"sponsor" taking place in Mar 2025 and the interim review on May 12, 2025. The project was completed on Jul 20 (after the foreseen duration of six months), with all milestones successfully achieved.

No significant deviations have taken place, except for the delay of the delivery of certain outputs. By the time of the project's completion, all milestones foreseen in the Third Party Agreement have been successfully achieved. A minor deviation concerns the delay in the submission of the foreseen scientific paper: a preprint has been prepared but not yet submitted, since we are still considering different target scientific venues that are most appropriate for the publication of our work.

Another minor deviation concerns the means used to conduct the evaluation. According to our original proposal, the plan was to use the CrowdHeritage platform to support the evaluation by human participants. However, due to the need to measure multiple evaluation indicators (e.g., decision agreement, reasons for disagreement, explanation content and fluency rating, shortcomings definition, etc), we chose the more modular option of online dedicated spreadsheets that were distributed to the human participants. This approach allowed us to readily incorporate multiple evaluator input types and effectively conduct a more in-depth evaluation process in due time.

1.2 Development

DETOEX aims to fulfill the following three main objectives:

a) *Creation of a multilingual vocabulary of politically-charged toxic terms:* The vocabulary includes terms that are considered offensive or derogatory towards groups and members in connection with their beliefs or characteristics. The vocabulary focuses on terms that convey i) denigration, including attacks on the character or reputation of one or more persons in connection with their beliefs and membership; ii) negative stereotyping, meaning any reference to negative traits assigned to a group and its members in relation to protected characteristics and ideologies (nationalism, religious beliefs, etc). The vocabulary curates terms from existing lexicons and extends them with additional information about their meaning in a contentious context and the dimensions based on which they are considered toxic (e.g. based on ethnicity, religion, political affiliation etc). The vocabulary is exploited by the toxic language detection system (see below). It covers terms in Greek, French, and English and is made openly available under a CC0 licence.

b) *Creation of a politically-charged toxic speech detection system*: Leveraging traditional NLP techniques and LLMs, we aim to create a hybrid system that utilises the curated vocabulary to detect toxic expressions that are derogatory towards a group or its members due to their characteristics of beliefs. The tool also provides an explanation as to why an expression is considered toxic with political valence within the specific context. In order to create a high-precision system, we first deploy a string-matching module that utilises NLP techniques such as lemmatization to analyse free text and match it with entries from the vocabulary. In order to disambiguate whether the use of a term is used in a derogatory way or not within a certain context, the explanations provided by the expert vocabulary are exploited to ground an LLM. The LLM is also instructed to produce an enhanced explanation that aligns the expert information with the input text. Given the high variance

of politically-charged toxic language, we also decided to employ a complementary pipeline that is not restricted to detecting vocabulary terms but is instructed to identify any toxic expression directed towards a group or their members due to their identity characteristics.

c) *Human evaluation and refinement*: The performance of the tool, with respect to its decisions about the presence of politically charged toxic language and the respective generated explanations, is reviewed and evaluated by humans. The evaluation feedback is analysed to draw useful conclusions about the tool's accuracy and possible biases.

1.2.1 Problem framing

Currently, there is no universally agreed-upon toxic or hate speech definition[1]. For the purposes of our project, we used the definition of the UN Strategy and Action Plan on hate speech[2] and the research conducted for the European Commission. During this research, we noticed that toxic language can be divided into different types, depending on the intensity of the expression, but also on whether it is illegal or not[3].

For our project, we decided to focus on expressions that reflect denigration or negative stereotyping, including attacks on the character or reputation of one or more persons in connection with their beliefs and membership, or references to negative traits assigned to a group and its members. We will not address the issue of illegality, i.e., we are not investigating whether the expressions are illegal or not.

We base our work on the definitions and categories provided by the European Commission's publication on online hate speech[4], in particular:

Definitions

- Toxic language: is a broad term that captures various forms of offensive or harmful language, an umbrella term that includes several different types of languages, including offensive language, abusive language, and hateful language

- Offensive language: a sub-category of toxic language and is closely connected with linguistic and societal phenomena, such as abusive and aggressive language, cyberbullying, racism, extremism, radicalisation, toxicity, profanity, flaming, discrimination, hate and hate speech. Or else, any form of non-acceptable language (profanity) or a targeted offense, which can be veiled or direct.

- Abusive language: can be seen as a sub-category of offensive language, used to refer to any impolite, rude or hurtful language using profanity, that can show a debasement of someone or something, or show intense emotion.

- Hate speech: is more restricted than abusive language that incites, promotes or justifies violence or hatred against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin.

For the purposes of this project, we define *politically-charged toxic language* as any spoken or written communication that uses derogatory or offensive language toward an individual or group based on who they are — that is, on the basis of religion, ethnicity, nationality, race, political affiliation, color, descent, gender, or other identity-related characteristics, beliefs, or factors. This definition aims to capture all forms of toxic language, including subtle or indirect expressions, and to determine whether such language is offensive or

[1] https://www.coe.int/en/web/freedom-expression/hate-speech
[2] https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf
[3] https://www.echr.coe.int/documents/d/echr/fs_hate_speech_eng
[4] The European Online Hate Lab, 2023

derogatory either explicitly or implicitly due to an identity-related feature. Our definition is, on the one hand, broader than the narrower concept of *hate speech*, as it covers a wider range of identity characteristics and is not limited to overt expressions of hate. On the other hand, it is narrower than the more general category of *toxic language*, as it excludes offensive content that is not linked to identity-related factors.

It should be emphasised that the project does not consider "political" as the language spoken by or directed towards politicians or other people who deal with public affairs. We adopt a broad definition of the concept of "political": language has a political valence, and it is pejorative towards a group identified based on their characteristics or beliefs.

This offensive or derogatory language that is harmful towards a group or its members can compromise democratic processes and hinder social inclusion, and can be manifested in various ways by:

- Portraying politicians, institutions, or minority/marginalised groups and their members as subhuman, dangerous, or undeserving of basic rights
- Stereotyping via derogatory language
- Blaming societal problems on specific groups
- Using euphemisms that appear innocent but signal prejudice to specific audiences
- Using slurs to disgrace or humiliate specific groups of people
- Denying the right of certain groups to participate in political processes

Characteristics of groups towards which politically-charged toxic language is directed

For this project, our definition of politically-charged toxic language hinges on the types of groups towards which the use of certain language is harmful. To this end, to guide our analysis, we developed a classification of these groups, identifying key characteristics that define them. These characteristics include:

1. **Age**: The time period that a person has lived.
2. **Disability**: A physical, mental, cognitive, or developmental condition that substantially limits one or more major life activities, including physical disabilities affecting mobility or bodily functions, sensory disabilities, cognitive, psychological or mental health conditions.
3. **Ethnicity**: A Social group that shares a common cultural heritage, ancestry, language, religion, traditions, or national origin
4. **Gender**: Gender identity refers to a person's internal, deeply felt sense of their own gender. It's about who you are, not who you're attracted to**.**
5. **Public Institutions:** Refers to an organisation established and funded by government authorities to serve the public interest, provide essential services, and implement public policies at local, regional, national, or EU level.
6. **Political Affiliation:** A person or a group's connection with a political party
7. **Race:** A social categorization based primarily on physical characteristics, particularly skin color and other visible traits
8. **Religion:** The belief in the existence of a god or gods, and the activities that are connected with the worship of them; one of the systems of faith that are based on the belief in the existence of a particular god or gods
9. **Sexual Orientation:** Refers to an individual's pattern of emotional, romantic, and/or sexual attraction to others.
10. **Socioeconomic:** Refers to expressions of hatred of, contempt for, or prejudice against people that belong to a particular group of economic affluence

11. Addiction: Addiction refers to persistent, difficult-to-control behaviors that typically provide short-term reward but lead to substantial psychological distress, functional impairment, or risk of harm when continued.

12. **Physical Appearance**: Physical appearance refers to the observable external characteristics of a person, such as height and build, body shape, posture, and gait.

Ways in which politically-charged toxic language is manifested

Based on the above, we distinguish between two main ways in which politically-charged toxic language is manifested:

a) *Use of stand-alone politically-charged toxic terms*: This refers to words or expressions that are inherently offensive or degrading towards a group of people. This case mainly covers slurs and insulting words that are harmful towards a group of people with specific characteristics or ideology. For example, the Greek term "αδελφή", when used with the meaning of "faggot", is by itself derogatory towards gay people, even when it is being used as a broader reaching insult (not directly addressed to the queer community). It should be emphasised that the meaning of the term and whether it is used in a contentious way or not highly depends on the context. For example, the Greek term "αδελφή" is also used with the neutral/acceptable meaning of "sister" and "nun".

b) *Toxic language directed towards a group and their members*: This aspect refers to toxic expressions that are explicitly directed against a group, as defined by our categorisation. This aspect covers any type of toxic expressions —including common slurs that are not included in the curated vocabularies of politically-charged toxic terms— that are used against a group identified by certain characteristics or a common ideology. For example, the Greek word "μαλάκας", which is commonly used with the meaning of "asshole", although toxic, is not considered by itself politically-charged if it is used as a personal insult, with no politically-charged connotations ("You're an asshole!"). However, it is considered part of a politically-loaded toxic expression if it is addressed against a person or group due to their ethnicity or religion ("All [ethnicity] are assholes!").

1.2.2 Creation of vocabularies with politically charged toxic terms

In accordance with the aforementioned interpretation of politically-charged toxic language, our objective is to curate three vocabularies (one per the considered languages) that contain derogatory and offensive terms, which have at least one meaning that is politically charged as such, i.e. terms that are per se derogatory and offensive towards a group and their members, based on their characteristics or beliefs. In this respect, the vocabularies are not meant to include common slurs, which may be directed against certain groups considered by our categorisation in certain contexts, but are not by themselves derogatory or offensive towards such a group.

1.2.2.1 Overview of the vocabulary structure

The vocabulary should contain the following pieces of information:

- Term: the politically-charged toxic term
- Description: Free text description of all the meanings of the term (including both offensive and non-offensive ones). The description should mention why and under which circumstances the term is used in an offensive way.
- Category: The group or groups towards which the term is by itself offensive or derogatory, based on the groups defined in Section 1.1.
- Source: A link to the repository from which the term was sourced.

Below we provide an example of a possible vocabulary entry:

- Term: Αδελφή
- Source: https://el.wiktionary.org/wiki/αδελφή

- Category: Gender; Sexual Orientation
- Description: Ο όρος "αδελφή" στην κύρια του σημασία αναφέρεται στη γυναίκα που έχει γεννηθεί από τον ίδιο γονέα με κάποιον άλλον, στη μοναχή στο χριστιανικό πλαίσιο, ή στη νοσοκόμα/νοσηλεύτρια στην ιατρική ορολογία. Ωστόσο, όταν χρησιμοποιείται με τη μορφή "αδελφή" (ή συχνότερα "αδερφή") για να αναφερθεί σε άντρα ομοφυλόφιλο, αποκτά έντονα μειωτικό και υποτιμητικό χαρακτήρα. Σε αυτή την περίπτωση, ο όρος λειτουργεί ως προσβλητικός χαρακτηρισμός που στηρίζεται σε στερεότυπα για τους ομοφυλόφιλους άνδρες, υπονοώντας θηλυπρεπή συμπεριφορά. Η χρήση του όρου με αυτή την έννοια θεωρείται ομοφοβική και κοινωνικά απαράδεκτη στο σύγχρονο λόγο.

1.2.2.2 Methodology for constructing the vocabularies

We conducted extensive research on existing vocabularies in all three languages. From the onset, we noticed significant limitations on existing Greek vocabularies[5], while for English and French we found vocabularies that had a specific focus, i.e., groups targeted based on gender, race[6], sexual orientation, or religion. In other cases, we tried to contact directly the organisations or researchers that have worked on similar vocabularies, using the European Online Hate Lab website, where most of the EU research on hate speech is collected. However, it was difficult to obtain these vocabularies, mainly for proprietary purposes[7]. For instance, the team reached out to weaponizedworld.com, an organisation that seemed to have existing vocabularies in all three languages, including most of the categories of our interest, but after several attempts and repeated communication, we did not manage to get access to the resources.

By inspecting the aforementioned vocabularies, we found out that they suffer from one or more of the following limitations:

a) Most vocabularies contained a rather small number of terms, something that was particularly true for the vocabularies we found for Greek;
b) They did not include sufficient definitions about the meaning of the terms, information that is particularly important for providing trustworthy input to the toxic speech detection system.
c) In several vocabularies, many terms were not associated with any tags indicating the perspective (e.g., ethnicity, religion) under which they are considered derogatory or offensive.

Given these limitations, we decided to base our approach on the Wiktionary multilingual, web-based free content dictionary of terms. Wiktionary entries are associated with rich and reliable information, including definitions and usage examples. Wiktionaries contain an extensive crowd-sourced list of offensive and derogatory terms, as well as slurs, and other terms that could be useful for our vocabulary, along with contextual information and metadata for each term, like descriptions, synonyms, related terms, etc. Most importantly, abusive words are marked with tags such as pejorative, derogatory, or vulgar. We identified tags/categories that are of interest to us:

- Greek:
  - Μειωτικοί όροι (νέα ελληνικά) | μειωτικός | μειωτική | μειωτικό | μειωτικά
  - Κατηγορία:Υβριστικοί όροι (νέα ελληνικά) | υβριστικός | υβριστική | υβριστικό | υβρισιτκά
  - Κατηγορία: Χυδαιολογίες (νέα ελληνικά) | χυδαίος | χυδαία | χυδαίο
  - βρισιά | βρισιές
- English:
  - Category:English derogatory terms
  - Category:English vulgarities
  - Category:English offensive terms
- French:

[5] https://hatespeechdata.com/#Greek-header (last accessed 08/02/2025, page is no longer available)
[6] http://www.rsdb.org/ (last access 08/05/2025)
[7] https://weapbbsaaonizedword.org/

- Catégorie:Termes péjoratifs en français
- Catégorie:Insultes en français

The methodology we followed consists of the following steps, which may vary depending on the characteristics of the Wiktionary in the specific language:

1) *Collection of initial list of terms and respective descriptions from Wiktionary*: As a first step, we collect all Wiktionary terms that are tagged with one of the aforementioned labels. All definitions (concerning both derogatory/offensive uses and neutral ones) are collected. It should be mentioned that in all languages other than English, the Wiktionary API can only retrieve the page titles/ids of terms, not the definitions. In those cases, the Wiktionary definitions are only available in plain HTML, and thus special parsing is necessary to extract the relevant information. Given that the HTML does not have a consistent structure and that there is "garbage" within some definitions, such as URLs, special cleaning needs to be applied to extract the relevant definitions in a consistent format. The process led to the collection of 965 terms for Greek, 11310 for English, and 3749 for French.
2) *Filtering to keep only politically-charged toxic terms*: Human experts inspected the initial list of Greek terms and excluded terms, mainly common slurs, that are not, by themselves, derogatory or offensive towards a certain group. For English and French, due to the high number of terms tagged as vulgar, derogatory, or offensive, and given the limited resources of the project, an LLM is instructed to assess which terms are politically charged to speed up the filtering process. In this case, humans validate a sample of the automatic outputs.
3) *Indication of categories*: Depending on the size of the terms ultimately included in the vocabulary, this step is conducted either solely by a human or assisted by an LLM. For Greek, the categories towards which the term is considered offensive or derogatory are added by a human. For English and French, an LLM is also instructed to indicate the categories, which are later sampled and validated by a human.
4) *Production of enriched descriptions*: The Wiktionary definitions corresponding to the filtered terms along with appropriate instructions are provided as input to an LLM, so as to produce a continuous text that describes the meaning(s) of the term with an emphasis on why and under which circumstances the term is used in an offensive way. Claude Sonnet 3.7 has been used with a dedicated system prompt[8] that describes the task, emphasizing the inclusion of contextual, linguistic, cultural, and historical information. A user prompt[9] is given that

[8] "You have an expert understanding of the [Greek] language and slang, and how it can be used in a derogatory manner to target individuals or groups through stereotypes, negative generalizations, or the use of identity-related markers (e.g., ethnicity, origin, profession) as a negative trait. This derogatory nature may be evident in the etymology or structure of the word (e.g., compound words using a component metaphorically to evoke a stereotype).
Your task is to help the user create a vocabulary of derogatory [Greek] terms. You provide vocabulary entries with clear, concise descriptions for each term that explain:
1. In which context(s) the term is considered offensive or inappropriate.
2. If and when the term can be used in a neutral or acceptable way.
3. Why or how the term came to acquire its derogatory meaning, if such information is available. Any relevant linguistic, cultural, or historical background that helps understand its offensive nature.
The user will provide you with a [Greek] term and a set of definitions extracted from the Wiktionary entry of this term.
Your task is to write a short, free-text explanation in [Greek], based on this information. The output should begin with a reference to the term itself (e.g., ["Ο όρος..."]), and describe its derogatory or inappropriate use clearly and concisely. If relevant, also mention in which cases the term may be acceptable. The explanation should sound like a neutral usage note or definition — not a chat response — and should contain no introductory phrases or closing remarks. Do not explicitly refer to the Wiktionary definitions."

[9] Example of user prompt in [Greek]: Term: αδελφή
Wiktionary definitions:
1. (οικογένεια) αυτή που έχει γεννηθεί από τον ίδιο γονέα με κάποιον άλλον
2. (χριστιανισμός) η μοναχή

includes the term and its Wiktionary entries.

5) *Human inspection and correction*: Human experts inspect the list of filtered terms and produce descriptions and correct them, if needed. Depending on the number of terms, a selected sample is validated by humans.

1.2.2.3 Creation of vocabulary for Greek

Below, we describe how the vocabulary creation methodology was applied for Greek.

1) *Collection of initial list of terms and respective descriptions from Wiktionary*: The terms were collected using the Greek Wiktionary API. The pages of some of the terms we were interested in had been directly tagged with the respective category (e.g., the term "αδερφή" is marked with the category Μειωτικοί όροι (νέα ελληνικά)), these were collected using the API parameters "action: query" and "list: categorymembers". Other pages were only tagged with a link to the Wiktionary page of the tag literal (e.g. the term "πούστης" is not marked with any category but one of the definitions contains a tag linking to the page for "μειωτικά"), these were collected using the API parameters "action: query" and "prop: linkshere". The process led to the collection of 965 terms using the tags and the methodology described in the previous section. Definitions were extracted from the term pages by parsing the HTML and detecting the sections containing definitions. Since the Wiktionary pages are only designed to be human-readable there is no consistent way to isolate the sections containing definitions, so some assumptions and work-arounds had to be made. Sections containing definitions almost universally start with a heading describing the part-of-speech of the word. A page may contain many such sections containing definitions (e.g. the term "αδερφή" contains definitions both under section "Ουσιαστικό" and "Κλιτικός τύπος επιθέτου"). All heading tags for all pages were collected and the ones referring to parts-of-speech were isolated. Other sections containing irrelevant information, such as the pronunciation of a term, were discarded. Within the sections kept, the term definitions are almost exclusively structured as a list (even if the section contains a single definition), so the HTML tags <ul> and <ol> were used to detect the lists of term definitions. For each definition, only plain text was kept, discarding any links or tags contained in the HTML. For each term, all definitions collected were unified in a single numbered list and stored in a CSV file.
2) *Filtering to keep only politically-charged toxic terms*: The filtering was conducted manually by members of the HomoDigitalis team. **288** terms were maintained as politically-charged toxic terms.
3) *Indication of categories*: Members of the HomoDigitalis team indicated manually the categories associated with each term.
4) *Production of enriched descriptions*: Anthropic's Claude Sonnet 3.7 was used to produce the descriptions based on the lists of Wiktionary definitions and appropriate instructions (see Section 1.2.2.2 for the exact prompts used).
5) *Human inspection and correction*: Members of HomoDigitalis reviewed the descriptions produced under the previous step and corrected them, where needed. The vast majority of the descriptions were considered of sufficient quality in terms of information accuracy, coverage of required aspects/meanings, and language use. A limitation that we inspected in some descriptions is that Wiktionary does not always sufficiently reflect probable recent shifts in the usage of terms in certain social environments, especially regarding the extent to which certain terms are used in a reclaimed manner (e.g. the reclaiming of "τραβέλι" by trans people). As a result, the descriptions generated by Claude based on the Wiktionary definitions often treat such reclaimed usages as being derogatory. Where possible, this limitation has been addressed by adapting the automatic descriptions accordingly. Furthermore, the LLM to

---

3. (ιατρική) η νοσοκόμα, η νοσηλεύτρια
※ Η αδελφή είπε ότι θα μου έβγαζαν πλάκες. (Θανάσης Βαλτινός, Ο γύψος. Συλλογή διηγημάτων Δεκαοχτώ κείμενα. Αθήνα: Κέδρος, 1970)
4. (μειωτικό) άντρας ομοφυλόφιλος → δείτε τη λέξη αδερφή"

be used for the disambiguation part (see below) is also prompted to be cautious of such usages by the affected community itself, so as to avoid over-flagging.

The resulting vocabulary of politically-charged derogatory terms in Greek can be found here.

1.2.2.4 Creation of vocabulary for French

The same steps were followed to create the French vocabulary. Using the relevant categories and flag terms identified in Section 1.2.2.2, we collected 3749 terms from the French Wiktionary. The main difference compared to the creation of the Greek vocabulary was that filtering and categorization (Steps 2 and 3) were performed by Claude, followed by human validation of a sample, rather than manual processing of the full set. This approach was necessary due to the large size of the initial term collection. Steps 4 and 5 were carried out as previously described. A validation was conducted on a sample of the vocabulary with respect to the decisions, generated categories, and descriptions f. Due to some variations of the category names produced by the model, a normalisation to the categories as defined above was also conducted. In all the steps the human validators found the results satisfactory in more than 90% of the cases. As a result, we compiled a vocabulary of politically-charged derogatory terms in French, consisting of **1644** terms, available here.

1.2.2.5 Creation of vocabulary for English

For the English vocabulary, the same procedure as with the French was followed. The appropriate categories and flag terms identified in Section 1.2.2.2 were used to collect 11310 terms from the English Wiktionary. The only difference was that the descriptions of the terms were retrieved directly via the English Wiktionary API, which supports access to term definitions, unlike the Wiktionaries in other languages, where this was not possible. Regarding human validations, in all the steps the human validators found the results satisfactory in more than 90% of the cases. The resulting vocabulary of **3904** politically-charged derogatory terms in English can be found here.

1.2.3 Design of politically-charged toxic language detection system with explanations

The politically-charged toxic language detection system aims to perform two complementary tasks:

1. *Term-based detection of politically-charged toxic terms*: The objective of this pipeline is to analyse a given text and detect the presence of terms as defined in the curated vocabularies. The pipeline should be able to disambiguate whether an identified term is used in a contentious way or not in a certain context. The pipeline should also be able to provide an explanation on why the term is considered derogatory in the given context, based on the input text and the term description.
2. *Detection of toxic language directed against certain groups and their members*: The objective of this pipeline is to analyse a given text and detect whether some toxic expression is explicitly directed against a group defined by certain traits or beliefs, as defined by our categorisation. The main aim is to capture cases where any type of toxic expression, including common slurs that are not included in the curated vocabularies of politically-charged toxic terms, is used against such groups (see example in Section 1.1).

It should be mentioned that the proposal only describes the first objective-pipeline among its objectives. We decided to also consider the second objective-pipeline, at an experimental level, so as to capture more cases of politically-motivated language that the first vocabulary-based approach may miss.

To fulfill the aforementioned complementary objectives, two distinct technical approaches are set up.

a) A pipeline for the detection of politically-charged toxic terms combines a vocabulary-based string matching approach with the use of LLMs. As a first step, through the use of a lemmatiser and string matching rules, occurrences of the vocabulary terms are detected in the analysed text. This term-based approach parses the given text, applies a hybrid neural and rule-based lemmatizer that makes use of syntactical information, and detects matches with an ingested form of the vocabulary. As a second step, an LLM is used to resolve semantic ambiguity and determine whether the term is used in a contentious way or not. To this end, the LLM is given as input, besides the analysed text, the description of the detected term as included in the curated vocabulary. The LLM is then prompted to produce an explanation as to why the use of the term is acceptable or unacceptable within the given context, based on the vocabulary description.
b) The pipeline concerning the detection of toxic language directed against groups and their members relies on an LLM-based approach. In this case, the LLM is provided as input, besides the analysed text, the categories on which our definition of politically-charged toxic language is based on, along with dedicated instructions. The LLM is prompted to determine whether the given text contains toxic speech that is explicitly addressed to a person, group or entity determined based on the characteristics following from the identified categories (e.g. a person or group characterised based on their religion, ethnicity, sexual orientation etc). If toxic language is detected, then the LLM is instructed to produce an explanation of why the language is offensive towards a certain group in the given context.

The two aforementioned complementary pipelines are applied in parallel:

- If only one of the pipelines detects the presence of toxic speech, then its output is provided as the result of the overall detection system.
- If both pipelines detect the use of toxic speech, then the two explanations are combined by fusing them into a common output through an LLM. The prompt used for the fusion is provided below:

```
# Task: Merge Toxicity Analysis Texts

You will be given two separate texts in [Greek] that describe the politically-charged toxic content found in another piece of [Greek] text. Your task is to merge these two analyses into a single, coherent description in [Greek] that eliminates redundancy and flows naturally.

## Instructions:

1. **Combine the information** from both texts into a unified analysis
2. **Reuse the existing text** - the output should consist of a re-ordering and combination of the input texts, not original phrasing
3. **Remove duplicate information** - if both texts mention the same politically-charged toxic elements, include them only once
4. **Reorganize for better flow** - arrange the information in a logical order that reads smoothly
5. **Maintain accuracy** - preserve all unique politically-charged toxic content identified in either text
6.  **Keep it focused** - focus on explaining why parts of the text are politically toxic without explaining why they do not fall under acceptable use cases of this term or expression
7. **Keep it brief** - one or two sentences for each politically toxic expression or term are enough

## Input Format:
**Text 1:** [First toxicity analysis]
**Text 2:** [Second toxicity analysis]

## Output Format:
Provide a single, well-structured paragraph that comprehensively describes the politically-charged toxic content without repetition. Do not include any opening or closing remarks in the paragraph.
```

Example:

Text 1: [Ο όρος "καριόλα" στο συγκεκριμένο tweet χρησιμοποιείται ως πολιτικά-φορτισμένη τοξική γλώσσα καθώς αποτελεί μέρος μιας προσβλητικής επίθεσης με έμφυλα χαρακτηριστικά. Η φράση στοχεύει στη μείωση και τον εξευτελισμό μιας γυναίκας μέσω σεξιστικής γλώσσας, συνδέοντας την με υποτιμητικές αναφορές σε σεξουαλική συμπεριφορά και φυλετικά στερεότυπα. Η χρήση του όρου σε αυτό το πλαίσιο παραβιάζει βασικές αρχές σεβασμού και ισότητας των φύλων]

Text 2: [Το κείμενο περιέχει πολιτικά φορτισμένη τοξική γλώσσα που στοχεύει άτομα με βάση τη θρησκεία τους. Συγκεκριμένα, ο όρος "ισλαμιπιθηκους" αποτελεί υποτιμητικό και απανθρωποποιητικό χαρακτηρισμό για τους μουσουλμάνους, συνδυάζοντας τη λέξη "Ισλάμ" με τη λέξη "πίθηκος", υπονοώντας κατωτερότητα. Επιπλέον, το κείμενο περιέχει σεξιστική γλώσσα ("μωρη καριολα") που στοχεύει το φύλο του παραλήπτη και περιλαμβάνει σεξουαλικά υπονοούμενα βίας ("περιμένεις να σε γαμησουν"). Η γλώσσα είναι άμεσα επιθετική και στοχευμένη, χωρίς να αποτελεί παράθεση ή έμμεσο λόγο.]

Merged Output: [Το συγκεκριμένο κείμενο παρουσιάζει πολλαπλά επίπεδα τοξικής γλώσσας. Ο όρος "καριόλα" στο συγκεκριμένο tweet αποτελεί μέρος μιας προσβλητικής επίθεσης με έμφυλα χαρακτηριστικά. Η φράση στοχεύει στη μείωση και τον εξευτελισμό μιας γυναίκας μέσω σεξιστικής γλώσσας και περιλαμβάνει σεξουαλικά υπονοούμενα βίας ("περιμένεις να σε γαμησουν"). Επίσης, αποτελεί υποτιμητικό και απανθρωποποιητικό χαρακτηρισμό για τους μουσουλμάνους, συνδυάζοντας τη λέξη "Ισλάμ" με τη λέξη "πίθηκος", υπονοώντας κατωτερότητα.]

Figure 1 provides an overview of the two complementary pipelines employed by the toxic speech detection system.

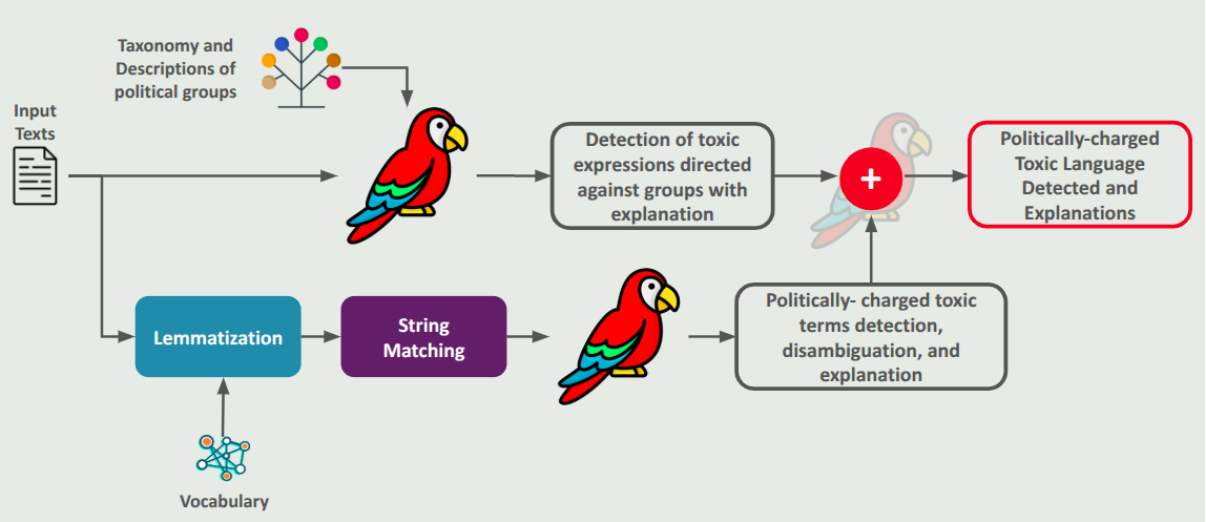


Figure 1: Overview of the approach to detecting politically charged and motivated toxic language.

Considering pipeline (a), which relies on the detection of vocabulary terms, the initial text processing, including lemmatization is done using existing lemmarisers, such as the Stanford Stanza NLP package[10]. The string matching module of the pipeline matches vocabulary terms in the input texts, by comparing their lemmatized forms. We have also added some extra checks in order to be sure that we match with the proper vocabulary terms each time. For example, we examine whether two vocabulary terms share the same lemma, e.g., considering grammatical gender variations. In such cases, we match the one with the shortest Levenshtein distance[11] between the un-lemmatized term and the text span,

[10] https://stanfordnlp.github.io/stanza/
[11] https://en.wikipedia.org/wiki/Levenshtein_distance

since the contentious nature of the term may have subtle differences based on gender, so the different vocabulary entries with the same lemma may differ semantically. It also considers matches that overlap, always keeping the longest match, e.g., whenever the term corresponding to the Wiktionary entry "nigger chaser" produces a match, the entry "nigger" will also produce a match, but only the longer one will be kept. After the string matching module, an LLM is prompted to disambiguate the meaning of the term within context and provide a brief analysis of the text, given the vocabulary descriptions. It is then tasked to decide whether the usage of the term is politically-charged toxic or not and to provide an explanation as to why. Different prompt variations (e.g. breaking down the task to simpler steps, asking the model to assign a toxicity rating, dedicated instructions to avoid over-detection etc) have been tested on sample data from the evaluation datasets (see Section 1.2.4) until achieving satisfactory results in terms of decision and explanation quality. For the final prompts used, we refer to 1.2.3.1-1.2.3.2.

For pipeline (b), the text is provided as part of a user prompt, along with a system prompt that explains the task to the model. Different prompt variations have been tested that refine how the task is described to the model to better align its behavior to our definition of the task. More details on which aspects of the prompt were refined are in section 1.2.3.1.

1.2.3.1 Pipelines implementation

For comparison reasons, we deployed two different LLMs for both pipelines. A proprietary larger model, and a smaller open-weight model was used for each language, in order to compare their performance and provide alternative solutions for implementation depending in the budget and available resources. We used Claude Sonnet 3.7 by Anthropic as the large proprietary LLM for all three languages, and we selected the smaller open-weight LLMs to be fine-tuned versions of the Llama family. For the latter, different versions have been used depending on the language, as described below.

The instruction-following version of Llama-3.1-8B was our original choice for the Llama-based LLM for English. Since there are no major recent releases of French LLMs around 7B parameters and considering that Llama-3.1-8B has been trained on extensive amounts of French text[12], we decided to use it for French as well. However, we found that Llama refused to fulfill the task due to guardrails about producing toxic speech, even after severe prompt modifications. We therefore chose to use Hermes 3 Llama 3.1 8B[13], which is a fine-tuned Llama that "aligns LLMs to the user" and didn't refuse to fulfill the task. Given that French outputs of Hermes were considered satisfactorily fluent by French speakers we consulted, we opted to use Hermes for both English and French.

For Greek, we selected Llama-Krikri-8B-Instruct[14], which has been developed by ILSP as a fine-tuned version of Llama-3.1-8B. Its finetuning included instruction-following in both English and Greek, as well as translating between the two languages.

Another point of examination is the use of the Stanza lemmatizers. While for English and French the Stanza lemmatizers were deemed good enough, initial tests with the Stanza lemmatisers for Greek proved disappointing, so we also examined the ILSP Greek lemmatizer[15] which performed considerably better. However, the ILSP lemmatizer is not offered as part of a deployable software package, but is only exposed via an API endpoint[16] that processes Greek text. It should be mentioned that the endpoint is considerably slower than Stanza. Due to these restrictions, the finalized online version of the tool as a service makes use of the lower-quality Stanza lemmatiser However, we have deployed a custom pipeline that uses it for the offline processing of the Greek evaluation dataset (see below). Interested stakeholders who wish to locally deploy the tool are advised to have it interoperate with the ILSP lemmatiser's API.

---

[12] French is officially supported in the official Hugging Face release while Greek is not.
[13] https://huggingface.co/NousResearch/Hermes-3-Llama-3.1-8B
[14] https://huggingface.co/ilsp/Llama-Krikri-8B-Instruct
[15] https://inventory.clarin.gr/tool-service/694
[16] https://nlp.ilsp.gr/nws/apidoc/

Initial prompts for both tasks were created with a simple description of the task and our definition of politically-charged toxic speech. The prompts were iteratively refined by manual qualitative inspections of the model outputs. Dedicated instructions were included to deal with reclaimed uses of originally discriminatory language as well as with references to opinions stated by a third person, which should not be flagged as toxic.
For the term-based pipeline, we also found that the model was quick to flag most usages of ambiguous terms (such as "αδερφή") as toxic, even when used within an acceptable context. In order to avoid over-flagging, we adjusted the prompt to a chain-of-thought format that allowed the model to first carefully disambiguate the meaning of the term and then make a decision on its toxic usage.
Regarding the language of the prompts, extensive manual qualitative evaluations were conducted on sample data with prompts provided in Greek and English. No considerable variations in the decisions and the explanation quality of the model were observed between the two languages.

Below we provide the final system and user prompts used for Greek. For French and English we followed the same pattern, translating the Greek parts of the prompt into English and French respectively.

Pipeline (a) prompt (term-based):

```
system_prompt = '"You are an expert content moderator specializing in detecting
politically-charged toxic language in Greek text. Your task is to analyze Greek text and distinguish
politically-charged toxic from neutral uses of a specific term based on the following definition:

**Politically-charged toxic language** refers to spoken or written communication that attacks or
uses pejorative or discriminatory language with reference to a person or a group based on
identity-related characteristics, including:
- Religion (θρησκεία)
- Ethnicity (εθνότητα)
- Nationality (εθνικότητα)
- Race (φυλή)
- Political affiliation (πολιτικές προτιμήσεις)
- Color (χρώμα δέρματος)
- Descent (καταγωγή)
- Gender (φύλο)
- Sexual orientation (σεξουαλικός προσανατολισμός)
- Socioeconomic status (κοινωνικοοικονομική κατάσταση)
- Age (ηλικία)
- Disability (αναπηρία)
- Addiction (εξάρτηση)
- Physical appearance (εξωτερική εμφάνιση)
- Association with public institutions (σχέση με δημόσιους οργανισμούς)

**You will be given:**
- A Greek term
- A description in Greek of how that term can be used in toxic and non-toxic/neutral ways
- A piece of Greek text containing this term
- One or more target characteristics the term may be offensive toward (e.g., "Sexual Orientation",
"Ethnicity")

Your goal is to analyze the text and then decide if the term is used in a politically-charged toxic
way. **All your output must be in Greek.** Use the following reasoning steps:
- **Step 1**: If the description includes multiple possible meanings of the term, identify which
meaning is used in the text. If disambiguation is particularly difficult, rely on non-toxic uses of the
term. If it has only one clear meaning, write "Μη αμφίσημος όρος" Do not evaluate toxicity yet.
```

- **Step 2**: Based on the meaning you identified, consider whether the term corresponds to the toxic usage described earlier. Consider both the possibility of it being used in a politically-charged toxic way and the possibility of it being used in a neutral/non-politically-charged toxic way.
- **Step 3**: Decide whether the use of the term in the text is politically-charged toxic or not and simply write "Τοξική" or "Μη τοξική".
- **Step 4**: Provide a clear, concise explanation (under 100 words) of your judgment. In your explanation use the phrasing provided in the term description you will be given. Do not include, or refer to any previous Step.

Important considerations for analysis:
- **Indirect speech**: Any toxic speech contained in the text as part of a quote or paraphrased from a different source should influence your decision significantly less or not at all.
- **Reclaimed language (Επανοικειοποιημένη γλώσσα)**: Some terms which are usually derogatory can be used with a reclaimed, empowering way by members of the same community they target. In these cases, the level of toxicity should be significantly lower or non-existent.
- **Self-identity versus targeting others**: Distinguish between someone who describes themself or their own community versus targeting others with the same language.

Format your output using XML tags as follows:
<STEP_1>
[Βήμα 1]
</STEP_1>

<STEP_2>
[Βήμα 2]
</STEP_2>

<STEP_3>
[Βήμα 3]
</STEP_3>

<STEP_4>
[Βήμα 4]
</STEP_4>"'

user_prompt = "'Όρος: {}
Περιγραφή: {}
Κείμενο: {}
Χαρακτηριστικά: {}"'

Pipeline (b) prompt (term-free):

system_prompt = "'You are an expert content moderator specializing in detecting politically-charged toxic language in Greek text. Your task is to analyze Greek text and determine if it contains politically-charged toxic language based on the following definition:

Politically-charged toxic language refers to spoken or written communication that attacks or uses pejorative or discriminatory language with reference to a person or a group based on identity-related characteristics, including:
- Religion (θρησκεία)
- Ethnicity (εθνότητα)
- Nationality (εθνικότητα)

- Race (φυλή)
- Political affiliation (πολιτικές προτιμήσεις)
- Color (χρώμα δέρματος)
- Descent (καταγωγή)
- Gender (φύλο)
- Sexual orientation (σεξουαλικός προσανατολισμός)
- Socioeconomic status (κοινωνικοοικονομική κατάσταση)
- Age (ηλικία)
- Disability (αναπηρία)
- Addiction (εξάρτηση)
- Physical appearance (εξωτερική εμφάνιση)
- Association with public institutions (σχέση με δημόσιους οργανισμούς)

Important considerations for analysis:
- **Non-targeted speech**: Do not consider slurs as politically-charged toxic language, unless they are directed towards an individual or group defined by their identity characteristics as listed above.
- **Indirect speech**: Any toxic language that the text contains as part of a quote or paraphrased from a different source should affect your decision significantly less or not at all.
- **Self-identity versus targeting others**: Distinguish between someone who describes themself or their own community versus targeting others with the same language.
- **Greek cultural context**: Consider Greek cultural references, historical context, slang, colloquialisms, regional expressions, sarcasm, irony, and political terminology specific to Greek society.

Output Requirements:
For each text, provide a decision of whether it contains politically-charged toxic language and a a clear, concise explanation (under 100 words) in Greek that covers:
- Which elements influenced your decision
- If politically-charged toxic language was found, which identity characteristics of the ones defined above are targeted
- Any ambiguities or nuances you considered
- Specific quotes from the text when necessary for your argument

Always use the exact XML format specified in the user prompt.
"

user_prompt = '''Ανάλυσε το παρακάτω ελληνικό κείμενο και αξιολόγησε το αν περιέχει πολιτικά φορτισμένη τοξική γλώσσα:

Κείμενο προς ανάλυση:
{}

Παρακαλώ δώσε την απάντησή σου ακριβώς στην παρακάτω μορφή:

<DECISION>
["Τοξικό" ή "Μη τοξικό"]
</DECISION>

<EXPLANATION>
[Αναλυτική εξήγηση στα ελληνικά της αξιολόγησης σου]
</EXPLANATION>'''

1.2.4 Evaluation

1.2.4.1 Datasets preparation

Because of the specific focus of our task and the requirement to evaluate explanations, our evaluation relies primarily on human annotators rather than automated metrics, as the available ground truth data does not fully meet the criteria following from the definition of "politically-charged" toxic language. Nevertheless, the identified datasets provide a valuable resource to generate system outputs, which human evaluators can then use to assess the detection quality and the relevance and clarity of the system's explanations.

We conducted a thorough search for existing datasets suitable for evaluating our system. Since we introduce a rather novel task of detecting and explaining politically-charged toxic speech, there is no existing dataset appropriately annotated to fully meet the requirements for evaluating both detection performance and explanation quality. To address this gap, we expanded our search to include general toxic and hate speech datasets, which contain terms and expressions that align with our approach. While there is a rich body of resources available for English, datasets in French — and especially in Greek — are notably scarce. We have identified several key datasets that we assessed and curated as needed, to be used for the detection system's evaluation. Specifically, we have considered the SemEval-2020 Task 12 (OffensEval2020) datasets in Greek and English as well as the MLMA hate speech dataset, which includes both French and English texts. We also explored additional datasets such as HateXplain and FRENK, and other supersets[17][18] that combine multiple hate speech datasets available in open repositories like Huggingface, in order to compile a diverse evaluation set that adequately covers the range of toxic speech phenomena targeted by our system.

Our selection choices were guided by the need to ensure (a) the use of datasets with similar characteristics across the three languages; (b) a sufficient number of examples that include politically-charged references (and not just toxic expressions); (c) and that pertain multiple group characteristics (e.g. race, ethnicity, gender , political affiliation). Based on these criteria, we made a random selection of of 2000 positive (annotated as toxic/hateful) and 2000 negative (annotated as non-toxic) tweets from the following datasets,:

- Greek: The Greek set of the OffensEval2020 dataset (this was the only dataset we were able to discover for Greek).
- English: Examples representing tweets from the English Hate Speech Superset. The examples were selected after shuffling to ensure a fair representation from various subsets that covered different types of toxicity such as sexism, racism etc).
- French: Examples representing tweets from the French Hate Speech Superset following the same approach as for English.

1.2.4.1 Application of pipelines on the datasets

The pipelines comprising the politically-charged toxic language detection system with explanations have been applied on the three aforementioned datasets. Table 1 provides an overview of the decisions made by (i) the term-based pipeline (a) ; (ii) the term-free-based pipeline (b); and the final fused model when using the Claude and Llama-based LLMs respectively. "Yes" means that the model detected politically-charged toxic language in the analysed tweet and "No" means that it did not. In cases where the term-based approach did not detect any term, its output is indicated as "Blank".

[17] https://huggingface.co/datasets/manueltonneau/french-hate-speech-superset
[18] https://huggingface.co/datasets/manueltonneau/english-hate-speech-superset

		Fused_Yes	Fused_No	TermBased_Yes	TermBased_No	TermBased_Blank	TermFree_Yes	TermFree_No	BOTH_Yes
EN	Claude	1901	2099	1226	495	2279	1338	2662	662
	Llama-based	2171	1829	1460	262	2279	2178	1822	1110
FR	Claude	1262	2738	670	246	3085	2886	1114	522
	Llama-based	1322	2678	829	86	3085	764	3204	526
GR	Claude	847	3153	198	59	3743	780	3220	130
	Llama-based	823	3177	223	34	3743	719	3281	118

Table 1: Results of the detection tool applied on a selection of tweets collected from datasets on toxic/hateful language.

It becomes evident that the datasets are not balanced with respect the presence of *politically-charged* toxic language (while they were selected to contain 50% negatively and 50% positively annotated examples with respect to hateful/toxic language): significantly fewer incidents of politically-charged toxic language were detected in the Greek dataset in comparison with the French and English ones.

For English and French, the Llama-based model (Hermes 3 - Llama-3.1 8B) has the tendency to detect more positive cases than the Claude model. This is particularly evident in the disambiguation step of the term-based pipeline, where Claude considers significantly more examples as negative. For Greek, the decisions made between the two models do not differ significantly. Regarding the presence of vocabulary terms that were identified, 43% tweets were found to include at least one term in the English dataset, 23% in French, and 7% in the Greek dataset (the fact that the dictionaries contain a decreasing number of terms for the respective languages should also be considered in this respect). The majority of these occurrences (with the percentage varying depending on the model and/or language) are disambiguated as being politically-charged toxic language.

We can also draw some useful insights about the complementary role of each pipeline. 34-51% of the positive cases for English, 40-41% for French and 14-15% for Greek are detected as positives by both pipelines. From the positive cases, around 32-35% for English, 37-43% for French, and 72-82% for Greek are only detected by the terms-free pipeline. The vast majority of those cases are due to the variety of toxic expressions not captured by the vocabularies (i.e. the term-based pipeline returns “blank”). The percentage of positive cases that are captured by the term-based pipeline and not by the terms-free one are 16-28% for English, 19-36% for French, and 8-12% for Greek. These are mostly cases where the tweets do not explicitly attack an identity-related group or its members, but use a biased term that is by itself derogatory based on the defined characteristics. For example, the term-based pipeline explains that the tweet “Females think dating a pussy is cute now?” is considered positive due to “the term being used to question a man's masculinity in a derogatory way”, but the term-free pipeline does not consider it to be directed against a protected characteristic or identity group.

1.2.4.3 Setup of human evaluation process

The results produced by the automatic detection system - including both the decisions about the existence of toxic terms and the corresponding explanations - have been inspected and evaluated by human participants. A subset of the datasets described in the previous Section was selected for evaluation by human participants. A selection of 1000 tweets per language and model was made, so that 50% of them represented examples of automatically detected politically-charged toxic language and 50% examples that did not contain such language (based on Claude’s decisions, which are more reliable). Each tweet was included twice in the evaluation sample, with the objective of having the same tweet evaluated by two different evaluators.

The evaluators were presented with the analysed text, the system’s final (fused) decision (Yes or No)

and the respective explanation in case the text was found to contain politically-charged toxic language. Evaluators were mainly members of the HomoDigitalis network and were contacted via the organisation's mailing list. They were asked to fill in a participation form indicating, among others, their language skills. They should be either native speakers or have full professional proficiency of the language they were invited to inspect. 16 participants contributed to the evaluation process (with varying levels of contribution in terms of the number of evaluated tweets). Detailed instructions that explain the objective of the task and provide concrete examples were provided. A dedicated online event was organised to inform participants about the project's objectives and engage them in the evaluation activities. The event also included a hands-on session on a sample dataset, so that participants get familiarised with the task and their questions are answered.

The participants were invited to provide the following pieces of information with respective instructions:

- *Agree with the detection decision*: Select Yes / No / Unsure based on whether you agree with the system's detection.
- *Reason for disagreement*: If you chose, you can select one or more reasons from the following list. "Not toxic" (Language is not toxic at all); "Not politically-charged" (Toxic, but not politically-charged, according to the definition); "Politically-charged toxic speech undetected" (politically-charged toxic language is present but was missed) ; "Unclear" (for example, there is no sufficient context to make a clear decision);"Other" (Specify further in the comments)
- *Decision comments*: Briefly explain your decision (optional but helpful, especially in case of disagreement).
- *Explanation content quality (1-5)*: Rate how well the explanation captures the reasons the text is politically-charged and toxic. Consider relevance, completeness, and correctness. 1 indicates the lowest quality, and 5 the highest .
- *Explanation fluency (1-5):* Rate how well-written the explanation is. Focus on grammar, clarity, and readability. 1 indicates the lowest fluency, and 5 the highest.
- *Feedback on explanation*: Select from a dropdown why you gave a certain score. Options: Too vague; Repetitive; Incorrect Details; Too Verbose; Other.
- *Explanation comments*: Add more thoughts about the explanation or anything that didn't fit elsewhere.

Each evaluator worked on a dedicated online spreadsheet with a certain structure that reflects the aforementioned elements (see example here). This also allowed us to periodically monitor the evaluators' contributions, perform quality checks, and intervene on-time in cases where the evaluator did not follow the right instructions.

1.2.4.4 Human evaluation results

Table 2 provides an overview about the extent of contributions received by evaluators with respect to different dimensions. The number of "decision evaluations" reflects the number of collected judgments about the tool's decision (i.e. agree, disagree or unsure). The number of "explanation evaluations" represents the number of explanations (produced only for tweets that have been found positive by the tool) for which at least one rating has been added. We see that evaluators who assessed a decision, often did not provide any feedback about the respective explanation - since the latter task is more demanding. Similarly, evaluators who disagreed with the tool's decision, did not always state their cause of disagreement.

	Claude		Llama-based		All			
	#decisionEvaluations	#ExplanationEvaluations	#decisionEvaluations	#ExplanationEvaluations	#decisionEvaluations	#disagreementCauseStated	#tweets evaluated	#tweets evaluated by 2 participants
EN	794	369	786	347	1580	304	995	585
FR	338	150	371	125	709	47	562	147
GR	538	227	543	185	1081	163	602	479

Table 2: Contributions by evaluators.

We calculated three variants of precision and recall, in order to account for the borderline cases, defined as those in which at least one human evaluator indicated their decision to be “Unsure” or where there was a disagreement between evaluators:

- A version that disregards all borderline cases.
- A version that promotes a strict approach, meaning it favors flagging borderline cases as positive. In this version, if no evaluator has stated “No” (i.e. all responses correspond to “Unsure” or “Yes”), then the tweet is considered positive (thereby promoting the stricter decision as the correct one).
- A version that promotes a permissive approach, meaning it favors treating borderline cases as negative. In this version, if no evaluator has stated “Yes” (i.e. all responses correspond to “Unsure” or “No”), then the tweet is considered negative (promoting the more lenient decision).

Table 3 provides an overview of the precision, recall, and F1 score metrics for the different models and languages.

		Safe			Permissive			Strict		
		Precision	Recall	F Score	Precision	Recall	F Score	Precision	Recall	F Score
EN	Claude	0.92	0.91	0.92	0.57	0.91	0.70	0.95	0.82	0.88
	Llama-based	0.80	0.86	0.83	0.5	0.86	0.63	0.88	0.77	0.82
FR	Claude	0.99	0.97	0.98	0.85	0.97	0.91	0.99	0.91	0.95
	Llama-based	0.97	0.79	0.87	0.84	0.79	0.82	0.97	0.73	0.83
GR	Claude	0.75	0.96	0.84	0.46	0.96	0.62	0.85	0.85	0.85
	Llama-based	0.86	0.88	0.87	0.53	0.88	0.66	0.91	0.74	0.82

Table 3: Versions of precision, recall and F score considering different handling of borderline cases.

In terms of comparison between the LLM models, for English, Claude outperforms Llamma in terms of both precision and recall. For Greek KriKri seems to perform better than Claude. We notice a disproportionate high precision in case of French for both models. However, we remain cautious regarding the quality of the evaluation feedback collected for French; additional input from native or proficient speakers is needed to ensure reliable results.

For the tweets that do not fall under the borderline cases (considering the "safe" model that disregards them), both the precision and recall are quite high for English and French and somewhat lower for Greek. When also considering the borderline cases, we see that all LLM models for English and Greek have a clear tendency to flag borderline cases as positives: the "strict" metric version receives a higher score than the permissive one, with the recall in the latter remaining high while the precision dropping significantly. As already mentioned, we are mindful of reaching safe conclusions about the tool's behaviour for French.

Table 4 provides some more insights about the nature and reason of disagreements between the tool's and the evaluators' decisions. As already seen, for Greek and English disagreement due to false positives is more frequent than disagreement due to false negatives. In the former case, the most common reason for disagreement is "Not-political" for Greek and either "Not Toxic" or "Not political" for English.

Percentages		Model Yes_EvalNo	Model No_EvalYes	Model Yes_Unsure	Model No_Unsure	Agree	DisagreeCause1	DisagreeCause 2
EN	Claude	0.04	0.02	0.02	0.009	0.82	Not Toxic (47)	Not Political (31)
	Llama-based	0.06	0.03	0.02	0.01	0.76	Not Toxic (53)	Politically-charged Toxic Speech Undetected (53)
FR	Claude	0.003	0.01	0.004	0.01	0.94	Politically-charged Toxic Speech Undetected (5)	Not Political (1) Not Toxic (1)
	Llama-based	0.01	0.05	0.02	0.01	0.82	Politically-charged Toxic Speech Undetected (29)	Not Toxic(6)
GR	Claude	0.06	0.01	0.02	0.006	0.80	Not Political (49)	Politically-charged Toxic Speech Undetected (16)
	Llama-based	0.04	0.03	0.02	0.009	0.81	Politically-charged Toxic Speech Undetected (34)	Not Political (28)

Table 4: Percentages and nature of disagreements between the model and the evaluators.

Another interesting dimension to investigate concerns inter-rater agreement, which we measure by calculating the Krippendorff's alpha (with 1 indicating perfect agreement and 0 agreement no better than chance). For its calculation, the decisions are interpreted as an interval with scale values 0.0 for No, 0.5 for Unsure, and 1.0 for Yes. Table 5 provides an overview of the agreement between the evaluators. The low inter-rater agreement observed across all languages underscores the subjective nature of the task.

	Krippendorff's alpha	Percentage of tweets with disagreement between the raters (among tweets rated by 2)
EN	0.41	0.33
FR	0.67	0.24
GR	0.55	0.30

Table 5: Inter-rater agreement.

With respect to the evaluation of the provided explanations, Table 6 presents an overview of the collected ratings. The results indicate that the different LLM models produce explanations with comparable content and fluency. The most frequently identified issue is the inclusion of irrelevant information within the explanations. A closer examination of the corresponding examples reveals that this issue pertains primarily to false positives.

	Model	Content	Fluency	Main issue 1	Main issue 2
EN	Claude	4.24	4.66	Irrelevant Information (31)	Too Verbose (14)
	Llama-based	3.92	4.59	Irrelevant Information (49)	Other (10)
FR	Claude	4.45	4.65	Too Verbose (13)	Irrelevant Information (3)
	Llama-based	4.40	4.66	Too Verbose (4)	Irrelevant Information (2)
GR	Claude	3.35	4.12	Irrelevant Information (31)	Too Verbose (10)
	Llama-based	3.19	4.04	Irrelevant Information (31)	Too Verbose (7)

Table 6: Explanations evaluation.

A more careful look at the free text comments made by evaluators, allow us to draw some more useful insights. The cases where evaluators indicated that they were “unsure” were mainly either due to the lack of sufficient context to make a trustworthy judgement (which is often the case for short tweets) as

well as uncertainty as to whether some expressions should be considered as politically-charged toxic or not. The most common example concerns the use of the terms "#chinavirus", "#wuhanvirus", "wuflu" and similar variations. Many evaluators were unsure about when this reflected a neutral fact (that the coronavirus first appeared in China) and to what extent it was pejorative towards Chinese people. It should be mentioned that the model itself did not have a consistent treatment of such terms, although it flagged their use most of the time. This was also a frequent case of disagreement. Similarly, many users felt that the use of established vulgar expressions that include mainly sexist terms, but are broadly used in not politically-charged contexts, should be acceptable.
Regarding the explanations assessment, in cases of low-quality explanations, besides the predefined reasons indicated in Table 3, it is also often the case for all languages that the level of toxicity is exaggerated, e.g. phrases such as "the language used is extremely toxic, the text contains intensively toxic language etc".

1.3 Dissemination

The means of dissemination of the project results include:

1) Scientific dissemination: A scientific paper has been written and will be soon submitted to a peer-reviewed venue.

2) Dissemination to the IT and research communities:
a) The source code of the detection system has been documented and made available on GitHub. It is also offered as a Docker container with the tool's functionalities exposed via an API. The tool's repository with the aforementioned results is available here.
b) The multilingual semantic vocabularies are published on Zenodo, to facilitate its further reuse.

3) Communication to broader audiences: The HomoDigitalis mailing list has been used to inform its network's members about the project results. A virtual event (in Greek) to inform members of the HomoDigitalis network about the project's objectives and engage participants into the evaluation activities has been organised (the recording is available here). The online evaluation process has also acted as a means of outreach, by mobilising more than 15 participants in the process of detecting and understanding different types of hate speech. A LinkedIn post informing about the project's paper has also been made.

1.4 Ethics

DETOEX made use of established pretrained LLMs -Claude by Anthropic and LLM versions based on Llama 3.1-8B, which are open source . Both models align with the European AI Act and are compliant with privacy regulations, while adopting mechanisms that filter out and correct known sources of misinformation and strategies to avoid biases in trained data. Claude is used by the European Parliament and adopts the "Constitutional AI" approach developed by Anthropic for training AI systems, so that they are harmless and helpful without relying on extensive human feedback". Moreover, all data used to ground and guide the LLM were collected in accordance with fairness and data minimisation principles, guaranteeing that all data processed throughout DETOEX is relevant and limited to the purposes of the research project.
During the data collection activities, we made sure to verify the presence of a lawful basis for the further processing of data provided by a third party, which pertains to the analysis of any raw data to be selected from existing open data repositories. The creation of the semantic vocabulary was the responsibility of experts with a strong legal background and expertise in the protection of human rights in the digital age. Particular attention was paid to the delicate task of harmonising the use of derogatory terms with the cultural intricacies characteristic of the different languages considered by the project. Finally, the FAIR principles for making the data Findable, Accessible, Interoperable, and

Reusable will be followed.
With respect to the human evaluation campaigns, participation was voluntary. Concerning the protection of personal data, we ensured that appropriate GDPR-compliant procedures were in place. Among others, the following were observed: informed consent; measures ensuring confidentiality, concerning the collection, storage and management of data; and the right to withdraw.

2 Summary of Results and Plans

2.1 Results

The project has led to the following main outputs:

- Three vocabulary with politically-charged toxic terms and accompanying descriptions in Greek, English and French.

- The tool for detecting politically-charged toxic language in Greek, English and French, made available as source code, Docker container and API.

- The evaluation results, providing information about the perception of the tool's results by humans and comparative insights about the use of different LLMs across the three considered languages.

- A scientific paper describing the overall methodology and achieved results.

2.2 Business plan

The vocabularies and software components that have been developed in the context of the project have been made openly available in appropriate open repositories - Zenodo and GitHub respectively - in accordance with open access principles and with the aim to promote their further takeup and exploitation. Moreover, the tool is offered as a free online service exposed via an API, to facilitate its reuse by applications developed by interested stakeholders.
Datoptron will be responsible for maintaining the online service for at least 2 years after the end of the project. Datoptron and HomoDigitalis plan to further disseminate the project results, and particularly the developed tool for detecting politically-charged languages, so as to broaden its use by interested stakeholders to analyse various types of textual data.

2.3 Future plans

There are several directions for future work towards improving the performance of the detection tool as well as expanding its applicability to more languages. As new, more powerful versions of LLMs are being released, their incorporation to the modular setup of the tool is expected to lead to higher-quality results. Moreover, possible adaptations of the detection tool for new application areas, so that it can deal with other types of toxic language or biases, will also be considered.

2.4 Blurb for public dissemination on UTTER's website

DETOEX presents a novel approach to hate speech detection that combines large language models (LLMs) with a curated vocabulary of derogatory language and traditional NLP techniques. The method focuses on identifying language that is offensive or derogatory toward groups or their members based on identity-related characteristics or beliefs. It also provides contextualized explanations for why certain expressions are considered offensive. The system integrates outputs from

two complementary pipelines. The first is a term-based pipeline designed to detect terms that are inherently offensive towards certain groups of people. To support this, a vocabulary of toxic terms and accompanying usage descriptions is developed and used to guide and ground an LLM that disambiguates the use of the term within a specific context. The second pipeline focuses on identifying expressions explicitly directed against groups or individuals defined by particular traits or beliefs. DETOEX has been implemented to analyze text in Greek, French, and English. Its outputs - both the detection decisions and the accompanying explanations - are evaluated by human participants. The feedback is further analyzed to draw insights into the tool's accuracy, potential biases, and the inherently subjective nature of the task.

3 Recommendation by Project Sponsor

The DETOEX project set out to build a multilingual resource and system for detecting *politically-charged toxic language* in three languages, Greek, French, and English. All contractual milestones were reached within the six-month window, accounting for the switch in the human evaluation campaign format, i.e. the decision to switch from the CrowdHeritage platform to the use of formatted spreadsheets.

A short overview of achieved milestones is summarised below.

Multilingual vocabularies of politically-charged toxic terms
The team delivered rich lexicons covering 965 Greek, 1644 French and 3904 English terms, collected after a a combination of automated and manual validation. The vocabularies have been made available on Zenodo under a CC0 licence, supporting the community and future work in this field.

Hybrid detection pipeline (term-based + term-free) & explanations
An end-to-end pipeline combining a curated term matcher with an LLM-based "term-free" component was released as Docker, open-source code and a public REST API. They demonstrated the comparative advantage of the dual approach, since each of the two pipeline components are better tailored to capture different types of politically charged toxic speech, thus contributing to better performance besides optimising the system efficiency.

Maintaining the API beyond the project's lifetime will be key to sustaining impact.

Evaluation campaign
The evaluation process confirmed the performance and suitability of the proposed architecture. Sixteen trained annotators rated more than 2K tweets in total, including assessments for decision correctness and explanation quality. The study confirms strong precision but also exposes challenging cases where inter-rater agreement drops. Publishing the guidelines and annotation template—currently available only internally—would let others replicate the and extend the protocol.

Dissemination & exploitation
All core artefacts (code, data, vocabularies) are public; a webinar has taken place and an article is in preparation for submission. These actions demonstrate commitment to open science, yet continued outreach—especially to civil-society groups in the newly covered languages—will determine real-world uptake.

In summary, every contractual milestone has been met, and most were exceeded in terms of openness and transparency. Future iterations should broaden language coverage, benchmark against generic toxicity detectors, and release the annotation materials to maximise community benefit.

Key Achievements, Strengths and Challenges

The project has met—and in several respects exceeded—its objectives. By coupling a carefully curated vocabulary of politically charged terms with a large-language-model component, the team produced a genuinely novel, hybrid detection pipeline. Crucially, the system does more than flag content: it generates concise, human-readable explanations, boosting transparency and user trust. All code, Docker images, vocabularies and evaluation datasets were released under permissive licences, underscoring a strong commitment to open science and making it easy for NGOs, researchers and industry practitioners to adopt or extend the work. A rigorous human-in-the-loop study further confirmed that the pipeline achieves high precision and recall across English, French and Greek, especially under the stricter definitions of safety for

AI, while the public REST API and detailed documentation lower the barrier to real-world deployment.

A few challenges emerged that pave the way for potential future work and related projects. First, political toxicity is inherently subjective, and the task of automated detection is particularly challenging, as reflected in inter-annotator agreement during the human evaluation—especially on borderline cases of reclaimed slurs or disease-related hashtags. In addition, language coverage is potentially skewed: the Greek vocabulary and positive-case pool are noticeably smaller than their English and French counterparts, hinting at geographic and cultural biases that future work could investigate. Finally, some false positives revealed that the models still struggle with context—particularly when irony or reclaimed language is involved—suggesting promising paths for future improvements.

Overall, the achievements clearly outweigh the limitations. With targeted follow-up—benchmarking against generic toxicity detectors, expanding language-specific resources and publishing the annotation template—the project can set a new standard for transparent, multilingual detection of politically charged toxic speech.

Recommendations for Future Work

1. **Benchmark against generic hate-speech / toxicity models.** A systematic comparison with well-known detectors that are *not* politically specialised would help understand potential overlaps, attributes of hate-speech in general and reveal complementary strengths.

2. **Extend language-specific analysis.** Expand evaluation to analyse and better understand the specific attributes of each language that affect performance, and expand to additional EU languages to improve representativeness. In addition, specific linguistic aspects pertaining to challenges highlighted above (irony and sarcasm detection, disambiguation of reclaimed slurs, etc) could be further investigated.

3. **Publish the annotation guidelines & spreadsheet template.** Making these resources citable will help other projects replicate or extend the human-evaluation protocol and foster best practices.

Overall recommendation

Given the clear achievements, adherence to the proposal, and the potential for future impact and expansion, the DETOEX team has shown more than satisfactory performance and already produced several outputs. Hence, it is recommended that the project receives the final funding part as originally planned.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D6/D1.2 FSTP2 Final – DETOEX

C.2 Cognifit Harmony

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D13/D1.3 – FSTP2 Final – *Cognifit Harmony*

Cognifit Harmony: Home-based Mixed Reality Therapy for Dementia

Nature	Final Report	**Work Package**	WP1
Project start date	02/01/2025	**Project end date**	02/07/2025
Interim meeting	dd/mm/2025	**Report submission Date**	dd/mm/2025
Main authors	Valeria Villani (UNIMORE)		
Co-authors	Awardees (ORG)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	Valeria Villani
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

The aim of *Cognifit Harmony* is to develop and test a MR application for older adults with cognitive decline, aiming to stimulate cognitive and motor functioning and promote active ageing.

The activities were carried out according to the planned schedule. In Month 1 (M1), the focus was on the definition of the system architecture, including the hardware and software components. The chosen technologies included Unity, Mixed Reality Toolkit (MRTK), World Locking Tool (WLT), and Microsoft HoloLens 2. The system architecture also incorporated the API Groq, along with a Groq client and LangChain. Following this, Months 2 and 3 (M2-M3) were dedicated to the implementation of physical training. This involved developing three different games with increasing difficulty, where the user would move around a room, collect virtual objects, and categorize them. Cognitive training was developed in Months 3 to 5 (M3-M5). An LLM-based dialogue system was integrated in the MR application, with the aim of stimulating conversation with the user. The approach consists in MR-mediated storytelling utilizing LLMs to support remembering, telling, and sharing events prompted by the domestic environment and detected pictures. Conversation topics are identified from the surrounding environment, e.g., pictures and objects in the room. Finally, the MR systems was validated in Month 6 (M6) with young and elderly healthy subjects, to assess the overall feasibility of the proposed physical and cognitive approach.

1.1 Deviations from original plan

Nothing to be reported.

1.2 Development

This section summarizes the activities carried out in the project. The first two activities (Definition of system architecture and Implementation of physical training) were described in details in the interim report and are summarized here for the sake of completeness.

Definition of the system architecture (M1)

The MR application was developed using Unity[1], a widely used cross-platform game engine known for its intuitive interface and component-based architecture. Unity uses C#, allowing developers to create custom scripts and features.

To enhance the MR experience, the project integrated Microsoft's Mixed Reality Toolkit (MRTK) and World Locking Tools (WLT). MRTK[2] provides essential components for MR development, including gesture recognition, real-world interaction, and scene management. WLT[3] offers a stable world-locked coordinate system, ensuring holograms remain aligned with the physical world. Its Space Pins[4] feature addresses scale errors and arbitrary coordinates by anchoring virtual objects relative to real-world elements.

[1] Unity Development Platform, https://unity.com/
[2] Mixed Reality Toolkit, https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/
[3] World Locking Tools, https://learn.microsoft.com/en-us/mixed-reality/world-locking-tools/documentation/concepts
[4] Space Pins, https://learn.microsoft.com/en-us/mixed-reality/world-locking-tools/documentation/concepts/advanced/spacepins

After development, the application is deployed to Microsoft HoloLens 2[5], a wearable MR headset with see-through lenses, spatial mapping, and advanced sensors. These features allow accurate blending of holograms with the environment, adapting easily to various physical spaces.

Implementation of physical training (M2-M3)

The application includes three mini-games with increasing difficulty, all based on the same core mechanic: the user moves through the room, collects virtual objects, and categorizes them by placing them on specific platforms.

In the first game, red and green cubes appear in the room. The user picks them up one by one and places them on the platform of the corresponding color. A counter updates and a sound confirms correct placement. The game ends when all cubes are sorted.

The second game follows the same logic, but the cubes are semi-transparent, making them harder to detect and increasing the challenge.

The third game is inspired by the Wisconsin Card Sorting Test (WCST). Various objects with different colors and shapes appear, and the user must place them on one of four platforms. The correct sorting rule (by color, shape, etc.) is initially unknown and must be inferred. The rule changes periodically without notice, requiring the user to adapt.

The application source code is available at https://github.com/ARSControl/app_HoloLens.

Implementation of cognitive training (M3-M5)

Between months 3 and 5, a cognitive training module was developed and integrated into the MR app. This module leverages a LLM to create interactive dialogues aimed at stimulating memory recall and conversational engagement, particularly through autobiographical storytelling.

The cognitive training is triggered through interaction with familiar physical items in the environment, referred to as "cognitive proxies", for example, a photo, painting, or object marked with a QR code. When the MR headset detects and continuously tracks such a QR code for a predefined duration, a virtual avatar appears and initiates a conversation with the user.

The avatar's speech is generated using a text-to-speech system via HoloLens speakers. Dialogue content is dynamically produced by the LLM, which receives structured prompts including the user's speech transcription (captured via a speech-to-text module) and contextual information drawn from a precompiled database. This database contains detailed semantic knowledge about each cognitive proxy, such as the people or events depicted in a photo or the origin of a particular object, allowing for highly personalized and meaningful interactions.

The avatar then engages the user in a multi-turn conversation, encouraging them to recall and elaborate on personal memories related to the identified item. Depending on the system's configuration, the avatar can adopt two different interaction strategies: a supportive mode, where it assists the user in retrieving memories; or a challenging mode, where it introduces deliberate factual errors (e.g., misidentifying a person in a photo) to assess the user's memory accuracy and capacity for error detection.

[5] Microsoft HoloLens 2, https://www.microsoft.com/en-us/hololens

The interaction concludes either when the user moves away from the object, causing the QR code to disappear from view, or when the dialogue naturally comes to an end. As users explore their environment, the system continuously scans for new cognitive proxies, prompting new conversations as they are detected.

From a technical standpoint, the working elements are: the augmented reality environment accessible through HoloLens, a speech-to-text model, a text-to-speech model, a LLM instance to actively converse with the user. The LLM client is instantiated via the GROQ-API. Each call to the LLM is provided with a structure that contains context information and user's transcription. The user transcription is provided by a speech-to-text model based on the recordings from the microphone on the MR headset. The context information is transparent to the user (i.e., the user has no access and does not know of its existence) but is passed to the model to contextualize either what the user is talking about or what the known facts about a specific cognitive proxy are. The user interface is produced by HoloLens, which provide an augmented reality world. The augmented reality world is composed of the actual environment plus a digital assistant that interacts with the user in a verbal conversation.

Validation (M6)

After developing the MR application, the next step of this research was to test it with elderly people, to assess both the feasibility and usability of the approach. A user study was set up, involving volunteer elderly subjects using the system at their home. The user study focused on physical training. This was initially accomplished by introducing healthy elderly subjects to Microsoft HoloLens 2 and seeing how they interact with the device. To this end, the MR application was deployed in real domestic environment, as shown in figures from Fig. 1 to Fig. 3. Afterwards, two distinct phases of testing were carried out in the participant's home, allowing the person to move around in a familiar environment. All participants were healthy volunteers, with the only requirement being a minimum age of 65 years.

1) First phase: During this phase, the user was asked to play a simple game, in which the user is asked to move around their home environment, collect virtual objects, and place them on a single platform. Although similar to the activities proposed for physical training in the MR application, this task is simplified because it does not require the categorization of the collected objects. The goal of this first user study is not to stimulate cognitive functioning and movement, but rather to determine whether elderly individuals can use the device and interact with it. The first user study included 5 participants (3 men and 2 women), with an average age of 76.6 years. During each test, several data were collected: the user's positions recorded with 1 Hz sampling rate and the exact moment when each cube was placed on the platform. These data were used to calculate the total distance traveled and the time taken to complete the task, providing insights into participant's movement and task duration. From the total traveled distance, we were able to determine how many steps the user walked during the exercises and compare this value with the recommended daily step count, providing therapists with more accurate information on the user's progress and the actual impact of the application.

2) Second phase: For the second phase, participants were asked to complete the three tasks developed for physical training with the MR application. They were instructed to complete the games consecutively, using the menu to transition from one game to the next. The participants recruited for this second phase of tests were 5 (3 men and 2 women, different from those in the first phase), with an average age of 68.8 years. During these tests, each participant was timed to understand

Figure 1: Deployment of the first interaction task for physical training. Red and green cubes are spread across the room, blending in with the physical environment.

Figure 2: Deployment of the second interaction task for physical training. Red and green semi-transparent cubes are spread across the environment.

how long they took to complete the three tasks.

At the end of the user studies for both phases, each participant was asked to fill out a questionnaire, obtained by combining the questions from the NASA Task Load Index (NASA TLX) and the System Usability Scale (SUS). The NASA TLX is a widely used tool for assessing the mental workload experienced while completing a task (Hart and Staveland, 1988). It is a subjective measure, since users rate their own performance. The SUS questionnaire is an efficient tool for the evaluation of usability of any kind of system (Lewis, 2018). It is in a 5-point Likert scale format, meaning that participants had to assign a value between 1 to 5 to each statement, with 1 being "I completely disagree" and 5 being "I completely agree". To ensure that the questionnaire followed a 5-point Likert scale format, the questions taken from the NASA TLX were reworded. Participants were also asked to give personal feedback on the experience.

Concerning cognitive training through the use of the LLM-driven storytelling app, its effectiveness was confirmed via an in-lab study.

Figure 3: Deployment of the third interaction task for physical training. Several objects with different colors and shapes are found in the environment.

1.3 Dissemination

The MR app for physical training has been described in the following scientific paper

Marta Gabbi, Valeria Villani, and Lorenzo Sabattini, "Towards User-Friendly MR Solutions for Cognitive and Motor Stimulation in Active Ageing," *2025 IEEE International Conference on Robot and Human Interactive Communication (RO-MAN2025)*.

The conference will take place in Eindhoven (NL) in the days August 25th-29th, 2025 (https://www.ro-man2025.org/). The paper will be published in the conference proceedings.

1.4 Ethics

The study was approved by the Ethics Committee of the University of Modena and Reggio Emilia (UNIMORE), specifically the Comitato Etico di Ateneo per la Ricerca (CEAR). All participants were enrolled on a voluntary basis and provided written informed consent prior to their inclusion in the study.

To minimize ethical risks, participants were informed in advance about the nature of the study, the procedures involved, and their right to withdraw at any time without penalty. The system was tested in a controlled environment, and researchers were present during all sessions to ensure participant safety and intervene if necessary, particularly considering the elderly population involved.

Data collection focused on assessing system functionality and perceived usability through physical and cognitive tasks. No sensitive personal or medical data were collected. All collected data were anonymized at the time of storage: any identifying information was removed or replaced with random codes, ensuring that individual participants could not be identified.

Project data management complied with the relevant data protection regulations, including the General Data Protection Regulation (GDPR). Data were stored securely on encrypted drives accessible only to authorized research personnel and will be retained only for the duration necessary to fulfill the research objectives.

2 Summary of Results and Plans

2.1 Results

This section describes the results of the user study focused on the physical training system. Once all questionnaires were completed, the average for each statement was calculated. Specifically, averages were determined based on the questionnaires from participants in the first phase and those from participants in the second phase. Furthermore, an overall average was calculated for all participants. The results are shown in Table 1.

STATEMENTS	**RATINGS**		
	I phase	**II phase**	**I and II phase**
I think that I would like to use this product frequently.	2.8	3.6	3.2
I found the product unnecessarily complex.	1.6	1.4	1.5
I thought the system was easy to use.	4.4	4	4.2
I think that I would need the support of a technical person to be able to use this product.	2	3.2	2.6
I found the various functions in the product were well integrated.	4.8	3.8	4.3
I thought there was too much inconsistency in this product.	1.4	2	1.7
I imagine that most people would learn to use this product very quickly.	4.2	3.6	3.9
I found the product very awkward to use.	1	2	1.5
I felt very confident using the product.	4.8	4.8	4.8
I needed to learn a lot of things before I could get going with this product.	1.4	1.8	1.6
The task was mentally demanding.	1.4	2.2	1.8
The task was physically demanding.	1.2	1.6	1.4
The pace of the task was hurried or rushed.	1	1.6	1.3
I was successful in accomplishing what I was asked to do.	4.4	3.8	4.1
I had to work hard to accomplish my level of performance.	2.6	3	2.8
I was annoyed, discouraged, stressed while performing the tasks.	1	1.2	1.1

Table 1: Average ratings for each questionnaire statement: first column presents data from the first testing phase, second column displays data from the second testing phase, and third column shows the overall mean across all participants.

During the first phase of the user study, users' positions throughout the activity and the time taken to complete it were recorded. From these data, the average number of steps walked by the user while completing the single game is approximately 200 steps, with an average duration of 7 minutes and 6 seconds. Moreover, the number of collected cubes was monitored to check for potential excessive fatigue that might have caused the person to stop earlier, but this metric was found unnecessary as all participants completed the task successfully. During the second testing phase, these metrics were not recorded, but we can hypothesize that participants took approximately 3 times longer than the first group, since participants in the second group performed the three games in sequence. Consequently, the number of steps taken while completing the entire activity is expected to be higher.

Moreover, the recorded coordinates were plotted into a graph, in order to create a visual 2D rep-

resentation of the user's path. Fig. 4 shows an example of these graphs, with the red point being the starting location and the blue one indicating the finishing location.

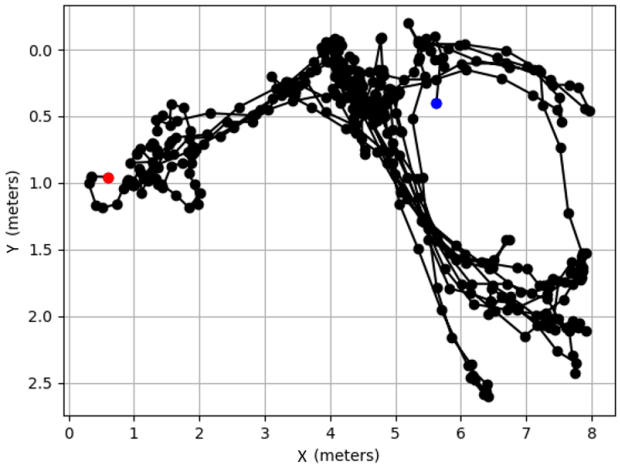


Figure 4: 2D representation of a user's path while performing the task.

The findings from the two phases evaluating the feasibility and usability of the approach have shown promising outcomes. As reported in the third column of Table 1, participants gave an overall positive evaluation of the approach. Participants in both the first and second groups stated that the task was easy to learn and perform, as reflected by the high results for ease of use and well-integrated functions, along with low ratings for unnecessary complexity and inconsistency. Users felt confident while performing the tasks and found the device neither awkward nor uncomfortable to use. The pace of the activity was not rushed, and participants felt successful in achieving their objectives, without experiencing significant stress or discouragement. The approach has a user-friendly design and a cohesive functionality that likely contributed to its overall usability. However, the moderate scores for frequent use of the approach and the occasional need for technical support suggest areas for improvements.

Analyzing and comparing the results reported in the first and second columns of Table 1, the ratings are generally consistent with each other. However, a few statements show a notable change in the ratings, with participants in the second phase giving more strict evaluations. These differences are most likely attributable to the fact that the second group was asked to play all three games consecutively. This means they faced a more complex and structured activity compared to the first group. This is reflected in lower ratings for the integration of features, higher scores for perceived inconsistency, and a greater need for technical support. Despite these differences, the overall results remain positive, indicating that participants perceived the approach as user-friendly, easy to learn, and comfortable to use.

2.2 Business plan

Not applicable.

2.3 Future plans

As this work is still in its initial phase, it presents several limitations that should be acknowledged. These will be addressed in future developments of the *Cognifit Harmony* project.

To begin with, the usability and feasibility assessment was conducted with a small number of participants, which may limit the generalizability of the findings. Additionally, only healthy individuals were involved during the preliminary studies, meaning that the approach has yet to be evaluated on elderly people with cognitive impairments. Furthermore, accessibility considerations for color-blind individuals were not addressed during either the development or the testing phases, highlighting the need for further refinements to enhance inclusivity. Lastly, an additional user study is set to be conducted to assess the feasibility and effectiveness of combining physical and cognitive training using the *Cognifit Harmony* MR application.

2.4 Blurb for public dissemination on UTTER's website

Cognifit Harmony is a Mixed Reality (MR) solution designed to promote active ageing by combining physical and cognitive training for elderly individuals, including those with early cognitive decline. Developed using Microsoft HoloLens 2 and tools like Unity, MRTK, and World Locking Tools, the system includes three interactive MR games for physical engagement, along with an LLM-powered storytelling module for cognitive stimulation.

The project included a home-based user study involving elderly volunteers to assess usability, comfort, and feasibility. Results were highly promising: participants found the system intuitive, non-intrusive, and engaging, even when used in domestic environments. Usability and workload assessments using NASA-TLX and SUS scales confirmed the system's accessibility and user-friendliness.

Dissemination efforts include a peer-reviewed publication accepted at the IEEE RO-MAN 2025 conference (https://www.ro-man2025.org/): Gabbi, M., Villani, V., Sabattini, L. "Towards User-Friendly MR Solutions for Cognitive and Motor Stimulation in Active Ageing".

The MR application source code is publicly available at: https://github.com/ARSControl/app_HoloLens

Although no immediate commercialization is planned, future work includes extended testing with users affected by cognitive impairment and the integration of adaptive features for broader accessibility. The project highlights the potential of MR and AI-powered systems for home-based digital therapies and age-friendly innovation.

3 Recommendation by Project Sponsor

This space is intended for the assessment by the project Sponsor. Did the project deliver its planned results? Was it successfully disseminated? Did the project team document business plans and future plans, and do these seem appropriate? Are there opportunities for exploitation of results within UTTER?

References

Sandra G. Hart and Lowell E. Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. In Peter A. Hancock and Najmedin Meshkati, editors, *Human Mental Workload*, volume 52 of *Advances in Psychology*, pages 139–183. North-Holland, 1988. doi: https://doi.org/10.1016/S0166-4115(08)62386-9.

James R. Lewis. The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction*, 34(7):577–590, 2018. doi: 10.1080/10447318.2018.1455307.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D13/D1.3 FSTP2 Final – *Cognifit Harmony*

C.3 TEASE

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP2 Final – TEASE

TExt And Schematic for Education

Nature	Final Report	**Work Package**	WP1
Project start date	02/01/2025	**Project end date**	02/07/2025
Final meeting	30/06/2025	**Report submission Date**	27/06/2025
Main authors	Thomas Gerald		
Co-authors	CNRS, Université Paris Saclay, LISN		
Reviewers	Bryan Eikema		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

1 Project Execution

During this project we experimented different solutions for annotation of text-image documents related to education. We focused on question-answering annotation that rely upon textual information and image, especially documents that contain graphical content such as maps or diagrams. To this end, we already gathered documents from wikipedia and the openstax platform. Today we collected about 2000 pairs of texts and images. We have started to work on the automatic annotation using Visual Language Models (VLM) by generating questions and answers from data pairs. Nonetheless this study and the quality of the annotation is not yet sufficient to publish the dataset, but it is an objective to do so as soon as possible. In the meantime, we are currently working on an automatic evaluation process using VLM to validate different criteria. Still, some additional work is required to validate the evaluation of the data generated, especially considering the alignment of the results obtained by the methods with human judgement.

The planned objective in the next few months following the end of the project will be to work on improving automatic annotation, comparing different models, pipelines or input provided. A second objective that we would complete will be to leverage automatic annotation methods to ensure the relevance of the generated annotation, particularly suiting educational competencies. In addition, those methods will allow to benchmark the different models.

Deviations from original plan

In the current state, the project objective has not been completely achieved yet. Particularly, we continue to advance on the two objectives: (a) the creation of a high-quality question-answering dataset and (b) methods for automatic evaluation. We review in the next section the advancement of the project and discuss its future development.

2 Development

At the current advancement of the project, we collected resources from Wikipedia focusing on schematics or infographic resources. Today, we have recruited two human resources to work full-time on the project:

- A research engineer (April 1st to July 2nd 2025)
- A trainee (April 1st to September 1st 2025)

Notice that the trainee is not recruited from the UTTER founding.

2.1 Related Works

Recent studies in large language models (LLMs), such as GPT Radford and Narasimhan (2018), LLaMA Touvron et al. (2023), and Bloom Workshop et al. (2023), have significantly improved performance in natural language understanding and generation tasks, particularly

in question answering (QA) over purely textual content. However, extending QA systems to complex documents, comprising interrelated text, tables, charts, and diagrams, remains a relatively underexplored research direction. Traditional Visual Question Answering (VQA) datasets, such as VQA Agrawal et al. (2016) and CLEVR Johnson et al. (2016), predominantly rely on photographic imagery and often address shallow factual questions, limiting their applicability in domains that require multimodal reasoning.

To address these limitations, several recent efforts have proposed integrating textual and visual modalities. For example, ScienceQA Lu et al. (2022) incorporates multimodal reasoning through diagrams and textbook-style questions, while ChartQA Masry et al. (2022) blends human-annotated and model-generated question-answer pairs for chart understanding. However, these corpora either employ templated question formats or focus on narrow media types, thus failing to capture the diversity and structural complexity of educational materials. DiagramQG Zhang et al. (2025) targets diagram-based QA but does not leverage textual context, which is essential for semantic interpretation in instructional settings.

In parallel, visual language models (VLMs) Bordes et al. (2024), such as Flamingo Alayrac et al. (2022), LLaVA Liu et al. (2023b), and Pixtral Agrawal et al. (2024), have demonstrated promising capabilities in multimodal generation and understanding. Nonetheless, these models are typically pre-trained on web-scale, open-domain corpora and thus exhibit limited robustness when applied to domain-specific documents. Domain adaptation techniques — including parameter-efficient fine-tuning approaches such as LoRA Hu et al. (2021), prefix tuning Li and Liang (2021), and prompt-based methods Wei et al. (2023) — have shown potential in aligning pre-trained models with downstream multimodal tasks, yet their comparative efficacy in complex educational QA remains an open question.

This work aims to bridge these research gaps by constructing a multimodal corpus tailored to educational content and developing interpretable, robust QA systems capable of reasoning across heterogeneous document structures.

2.2 Objective 1: Collecting resources and Question-Answer generation

At the project's current point of development, we collected different articles containing images (maps, schematics and diagrams) and text in both languages (French and English). We are currently focusing our preliminary collection on specific domains (history and biology). This first step aims to help us draw the objectives, characterize the annotations (kinds of questions and answers) we would create, and highlight the difficulties. Today, the engineer (Julie Lascar) focuses on these topics with the help of the person responsible for the project (Thomas Gerald). The collected documents come from two sources, Wikipedia (in French and English) and the Openstax platform.

Wikipedia For Wikipedia, we manually selected articles based on their relevance for high-school education. We then splitted the documents based on the sections of the document. We only kept sections, where there is an image with a tag related to diagrams, maps, or timelines. We only have a small subset of Wikipedia articles (20 articles and 110 relevant sections) to experiment with the automatic annotation process.

The use of the cleaned Wikipedia corpus WIT (Srinivasan et al., 2021) is envisioned later to complete the collection.

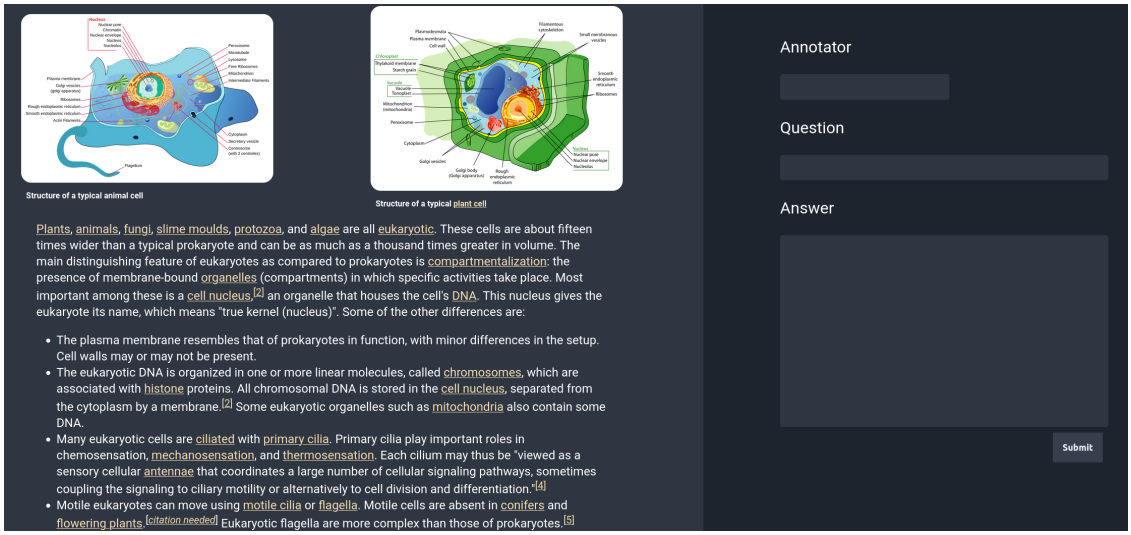


Figure 1: Screenshot of the current version of the annotation interface

OpenStax We began to collect information from the OpenStack[1] platform, providing educational material (in English) in HTML format. This resource contains questions (MCQ format) that will be integrated into the annotated dataset (with specific metadata). Notice that answers to the questions on the content of the course are not always available on the platform.

We collected all sections that contained images. Moreover, on openstax, some of the section have a question associated[2]. We collected those questions but, we do no use them in our pipeline yet.

At the time, we collected 1144 text and image pairs related to Biology and 726 related to history.

Other resources Other resources will be considered later, such as the free resources of the French government[3]. A part of the data will be annotated manually. To this end, an annotation platform has been developed (pictured in figure 1).

Automatic annotation (question and answer) To automate annotation, we would rely on a simple mechanism, we selected different VLMs models and input text and images. At the time we kept the question generation pipeline simple, experimenting the two different pipelines (see figure2) :

Pipeline 1

1. Generation of the question considering the context (text, image, captions)
2. Generation of the answer considering context and the question

[1] https://openstax.org/books/biology-2e/pages/18-review-questions
[2] for instance https://openstax.org/books/biology-2e/pages/2-3-carbon
[3] https://docs.forge.apps.education.fr/cartographie/manuels-libres/

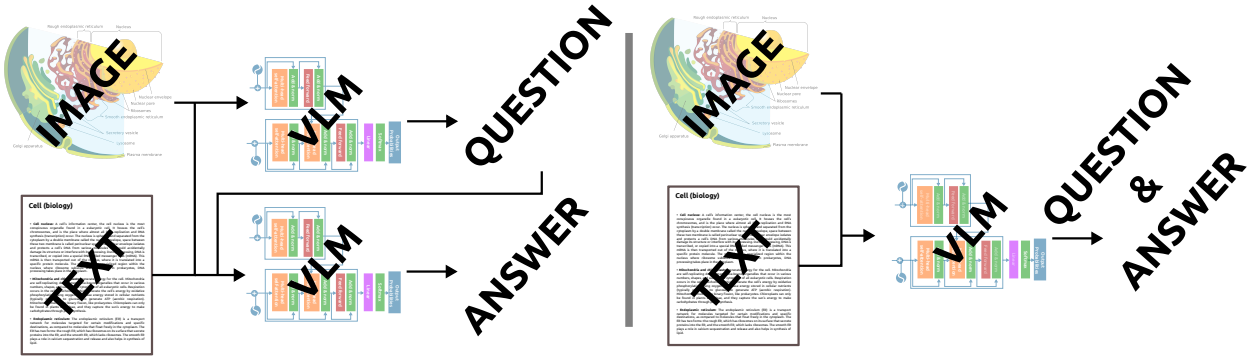


Figure 2: Two pipelines for question and answer generation

Pipeline 2

1. Generation of the question and the answer considering the context (text, image, captions)

To generate and evaluate (see next section), we experimented with different models and prompts. Today we generated questions and answers using the models Qwen2.5, llava 1.5 and Gemma 3 (4B). We also actively compare different prompting approaches, based on modifying instruction in the prompt. Below you will find an example of prompt used for the generation (pipeline 2):

> Image caption : [legend] Image info : [hidden_legend] Course excerpt: [context] Based on the provided image and the textual information, generate a question and its corresponding answer. - Write a question whose answer relies on the image and/or the text. - The question should be designed for a student aged 12 to 18. - The answer should rely mostly on the image and/or the text. Return the response structured as follows:{"question": "...", "answer": "..."}"[a]
>
> [a] words between square brackets are replaced with textual values

Notice that different prompting approaches are envisioned, such as few-shot prompting approaches (providing examples of questions/answers) or leveraging questions extracted from Openstax.

The validation of the generated examples, and hence of the capacity of the models, will result from the evaluation pipeline.

While we generated more than 2000 annotations, they do not have the quality requested for the publication of the dataset. In the meantime, we develop techniques to evaluate/filter generated questions and answers to validate the generation process. Moreover, different generation processes such as using few-shot learning are envisioned, e.g integrating validated course questions in the context for the generation.

2.3 Objective 2: Automatic evaluation Pipeline

The second objective in this project is the creation of an automatic evaluation pipeline. An intern is dedicated to this part of the project. Today, the first experiments using

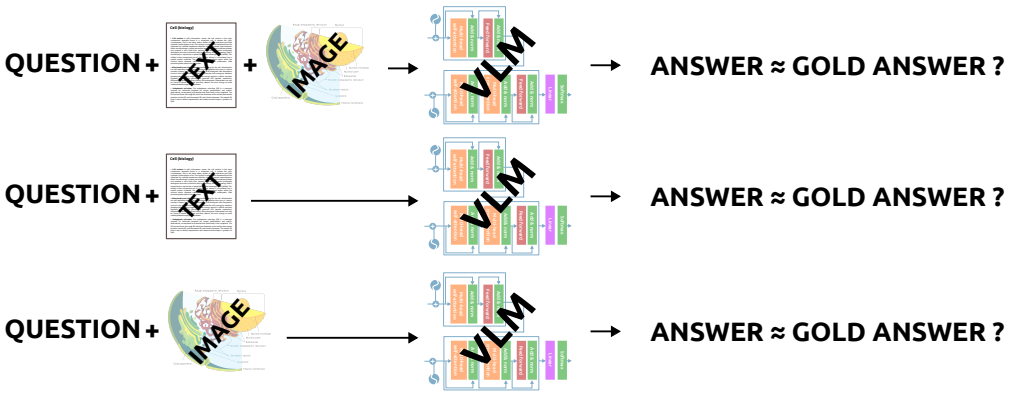


Figure 3: The pipeline tested to locate which part of the document allow to produce a correct answer. We generate answer to the question considering as context: image-only, text-only and text and image. Then we check the correctness from a gold answer.

few-shot/zero-shot prompting approaches focus on prompting open-source VLMs models (LLaVA Liu et al. (2023a), Falcon-11B Malartic et al. (2024) and PaliGemma Beyer et al. (2024)). At the time, the intern focuses on testing and running the models using different configurations. This step is crucial to determine what are the pros and the cons of models. And it will allow us to gather insights on the frequency and the cause of VLM errors.

Location of the answer (ongoing experiment)

A first step beyond verifying the relevancy of generated annotation was to locate if the information to answer the question was in the textual content, the image or both.

We designed a simple two steps protocol to predict what part of the information is necessary to answer the question:

- **Step-1** Generating an answer using partial information using a VLM (llava1.5)
  - The question and the images $answer_{image-only}$
  - The question and the textual content $answer_{text-only}$
  - The question, the textual contents and the images $answer_{text-image}$
- **Step-2** Classify the correctness of the previously generated answers. We used a VLM (llava 1.5) with the following context: the gold answer (ground truth), the question, and one of the previously generated answers. The model is attended to generate one of the two labels correct/incorrect

Once the results are obtained, the objective is to determine the relevant part of the information to answer the question (text, image, or both). For instance, getting the correct answer by providing only the text and the question would mean that the text contains sufficient information to answer it. To this end, we designed a pipeline, where the objective was to decide if to answer a question you need to use solely the text (text-only), the image only (image-only) and both (image-text). For this experiment we used the LLaVa 1.5 model.

We considered the following combination depending on the correctness of the different generated answers:

- Only image is necessary when correct for image-only and image-text but not text-only
- Only text is necessary when correct for text-only and image-text but not image-only
- Both information are required when solely correct for image-text
- Image or text contains enough information each to answer the question when the label predicted is correct for all configurations
- None: every other cases

To assess the relevance of the method, we manually annotated the question regarding the documents and asked annotators to label the question with one of the previous cases. However, at the time, we did not get an explicit agreement between annotators (considering 36 documents); further evaluation should be required to assess the relevancy of the method.

While locating the type of resource needed is crucial for our corpus (as we would retain question/answer that at least need the image to answer the question), we are also interested in the question's relevancy for an audience of students. The following section describes a preliminary framework to assess the relevance of the automatic annotations for students.

Does question fit the scope (education) ?

As the aim is to produce a corpus with high-quality questions for education (students), we proposed to evaluate different criteria for assessing student competencies. Similar to previous methods, we leverage automatic annotation on different criteria. At the time, we designed 12 criteria that we report in the table1.

name	objective	labels
correctness-1	decide if the answer is correct or not	binary
correctness-2	decide if the answer is logically coherent ?	binary
relevance-1	decide if the question is aligned with the document content	4-classes
clarity-1	decide if the question is clearly worded and easy to understand ?	3-classes
appropriateness-1	is suitable for the target (students aged 12–18) ?	3-classes
visual-interpretation-1	analyzing and understanding diagrams, charts, or tables	binary
visual-interpretation-2	understanding and using chronological references	binary
visual-interpretation-3	understanding and using spatial references	binary
language-1	use disciplinary language and representations	binary
reasoning-1	apply reasoning and methods	binary
reasoning-2	cognitive complexity of the task	$\in [0, 1]$
creating-1	create, represent, and model information	binary

Table 1: Summary of the criteria evaluated on the corpus

Those criteria are, in their vast majority, inspired by the target competencies a student must acquire. To study the relevance of the proposed criterion, we evaluate the automatic prediction of models for each criterion. Models are prompted with instructions requiring the model to associate the generated annotation with a label and an explanation. For instance, for the criterion *visual-interpretation-1*, we prompted the model with the following context:

Criterion	Model	Percent yes	Kappa Cohen	Interpretation
correctness-1	GEMMA-3 QWEN-2.5	91.0% 84.7%	0.50	Moderate agreement
correctness-2	GEMMA-3 QWEN-2.5	91.0% 85.7%	0.71	Substential agreement
visual-interpretation-1	GEMMA-3 QWEN-2.5	88.9% 57.7%	0.26	Fair agreement
visual-interpretation-2	GEMMA-3 QWEN-2.5	83.6% 3.7%	0.01	No agreement
visual-interpretation-3	GEMMA-3 QWEN-2.5	76.2% 0.0%	0.00	No agreement
language-1	GEMMA-3 QWEN-2.5	89.9% 83.6%	0.64	Substential agreement
reasoning-1	GEMMA-3 QWEN-2.5	89.4% 65.6%	0.33	Fair agreement
creating-1	GEMMA-3 QWEN-2.5	88.4% 50.3%	0.19	Slight agreement

Table 2: Preliminary results for automatic evaluation comparing two Visual Language Models (Gemma and Qwen 2.5) for binary criterion (Yes/No) on a subset (189 annotations) of OpenStax Biology (generated with LLaVa model)

> Considering the image(s) and the associated caption(s):
>
> [captions]
>
> The document:
>
> [document_text]
>
> evaluate the pedagogical content by determining whether the question "[question]" and the answer "[answer]" assess the competency of analyzing and understanding diagrams, charts, or tables (extract relevant information, compare sources). Answer by label "yes" if it asses the competency and by "no" if not. You should write the output in JSON format with the field "label" containing one of the previous labels and a field "explanation" containing the justification of the label chosen.

We conducted experiments by prompting two recent multi-modal models[4]. We report in the table 2 the results on the different binary criteria and how the two models agree on the evaluation (using the Kappa Cohen score).

Notice that while there is at least slight agreement for both model for the different criteria, some of them especially visual interpretation criterion 1 and 2 are having a kappa score close to 0. While we did not investigated this yet, the explanations provided by the model Gemma 4B have inconsistencies with the label chosen. For instance, we got the following evaluation for the model for the criterion *visual-interpretation-3*:

[4] Gemma 4b (https://huggingface.co/google/gemma-3-4b-it and Qwen2.5 (https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct)

{
"label": "**yes**",
"explanation": "[...] The competency of Spatial References **is not directly assessed in this case**, as the question and answer are about biological classification, not map reading or spatial understanding.
}

We can easily observe that the label chosen is inconsistent with the explanation. This behavior could be caused by many factors such as bias in the instruction data of the model or incomplete or imprecise instruction. We thus cannot yet get spare of human evaluation. Notice that this evaluation does not aim to replace human judgment, it will work as a filter for human annotators. In the meantime we will investigate other approaches to strengthen and compare the evaluation system, especially we can evaluate the consistency of a model using many generations (Lee et al., 2025), evaluate the perplexity of the model according to the generated response (or observing the likelihood of the label tokens) or perform pairwise comparison on different answer generation as it has been stated that "comparative assessment outperforms absolute scoring" (Liusie et al., 2024).

Notice if we want to validate the relevancy of the generated data we would need to compare the LLM judgment to human one and that the current results cannot be properly interpreted to judge the quality of the annotation. An evaluation step was planned using human annotators to evaluate the content (annotations) created. However we did not have time to organize it during the project duration.

2.4 Dissemination

We did not yet publish research paper including results of the project. However, publication using the materials produced during the project duration are still planned. It will particularly fit the LREC conference scope (Language Resources and Evaluation). A submission is planned for the beginning of october 2025 (october 17^{th}).

Nonetheless, the project have been selected for the innovation session taking place during the CORIA-TALN conference (a French national conference) the 4^{th} of july. The objective will be to present the current state of the project to the NLP french-speaking community.

Additionally, dataset and code will be shared on a github platform after the first publication.

2.5 Ethics

The project's primary ethical concern is the data's license. To guarantee the right to disseminate the resources, we only collected data under permissive licenses. OpenStax's content license is CC BY 4.0 "Attribution 4.0 International License", meaning we are free to share and adapt the data if the appropriate credit and a link to the license are provided. We also contacted an Openstax representative to inform them of the use of the data in our project. The Wikipedia license is also permissive concerning textual content. For images, the license may vary from one image to another; therefore, we have only included images with permissive licenses (public domain and Creative Commons licenses).

Concerning later human annotation (and evaluation), the shared data will be anonymised.

3 Summary of Results and Plans

Describe the results (their relevance to project goals, impact, etc.), business plans (if applicable), and any future plans relevant to exploitation of results, their sustainability and future operation of the project.

3.1 Results

Currently, we do not have exploitable results; we cannot yet share the question-answering corpus. At the time, we selected the different resources we planned to use for the corpus; we produced the first annotation subset by leveraging generative visual language models, and we started experimenting with automatic evaluation approaches. Nonetheless, we will present the project and the work already completed to the conference TALN[5].

Notice that the current code and data are available at https://nextcloud.lisn.upsaclay.fr/index.php/s/rtWCFkXJKT7ABrQ (password: ikR4txqGcr).

3.2 Business plan

As the project members are not part of a company (but from a public research lab), we do not have a business plan involving selling a product. However, we plan to share the resources we developed under the research-complying license. In addition, we planned to publish scientific articles (on the methodologies of automatic annotation) based on the works we did and the ones we will do.

3.3 Future plans

As some parts are not completed yet, the next two months will be dedicated to the completion of the two main objectives of the project. The code of the project and the data will be shared through the GitLab page[6] once the dataset is published (planned submission in October 2025). Additionally, the interns recruited in TEASE will continue to work on the project until September.

3.4 Blurb for public dissemination on UTTER's website

The project TEASE aims to produce a new corpus of text-image for education. Especially the project leverages automatic annotation mechanisms and LLMs as judge methods. We particularly focus our work on the evaluation and how models can align to human judgment.

The corpus and the different methods will be published soon (https://gitlab.lisn.upsaclay.fr/nlp/corpora/tease).

Draft a short summary of the project's results, dissemination activities, business plans and future plans. This will serve as a starting point for public dissemination on UTTER's website. Where possible, share links to publicly available results and/or dissemination.

[5] https://coria-taln-2025.lis-lab.fr/
[6] https://gitlab.lisn.upsaclay.fr/nlp/corpora/tease

4 Recommendation by Project Sponsor

This space is intended for the assessment by the project Sponsor. Did the project deliver its planned results? Was it successfully disseminated? Did the project team document business plans and future plans, and do these seem appropriate? Are there opportunities for exploitation of results within UTTER?

This project broadly delivered on its intended outcomes. The dissemination results (a dataset, an automatic evaluation procedure and research paper) are available in a preliminary unpublished format in the form of this report and a privately shared download link to the present version of the dataset. The current project results will also be presented on July 4th at the CORIA-TALN conference. The project team also intends to make the dataset available to the broader public and a submit a research paper in the near future, continuing development with a trainee hired outside of UTTER funding. The results of this project could be useful for the training and evaluation of multimodal systems in UTTER. The project sponsor recommends this project for payment.

References

Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. Vqa: Visual question answering, 2016. URL https://arxiv.org/abs/1505.00468.

Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, Albert Q. Jiang, Kartik Khandelwal, Timothée Lacroix, Guillaume Lample, Diego Las Casas, Thibaut Lavril, Teven Le Scao, Andy Lo, William Marshall, Louis Martin, Arthur Mensch, Pavankumar Muddireddy, Valera Nemychnikova, Marie Pellat, Patrick Von Platen, Nikhil Raghuraman, Baptiste Rozière, Alexandre Sablayrolles, Lucile Saulnier, Romain Sauvestre, Wendy Shang, Roman Soletskyi, Lawrence Stewart, Pierre Stock, Joachim Studnia, Sandeep Subramanian, Sagar Vaze, Thomas Wang, and Sophia Yang. Pixtral 12b, 2024. URL https://arxiv.org/abs/2410.07073.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning, 2022. URL https://arxiv.org/abs/2204.14198.

Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C. Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, Mark Ibrahim, Melissa Hall, Yunyang Xiong, Jonathan Lebensold, Candace Ross, Srihari Jayakumar, Chuan Guo, Diane Bouchacourt, Haider Al-Tahan, Karthik Padthe, Vasu Sharma, Hu Xu, Xiaoqing Ellen Tan, Megan Richards, Samuel Lavoie, Pietro Astolfi, Reyhane Askari Hemmat, Jun Chen, Kushal Tirumala, Rim Assouel, Mazda Moayeri, Arjang Talattof, Kamalika Chaudhuri, Zechun Liu, Xilun Chen, Quentin Garrido, Karen Ullrich, Aishwarya Agrawal, Kate Saenko, Asli Celikyilmaz, and Vikas Chandra. An introduction to vision-language modeling, 2024. URL https://arxiv.org/abs/2405.17247.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning, 2016. URL https://arxiv.org/abs/1612.06890.

Noah Lee, Jiwoo Hong, and James Thorne. Evaluating the consistency of LLM evaluators. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10650–10659, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.710/.

Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021. URL https://arxiv.org/abs/2101.00190.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023b. URL https://arxiv.org/abs/2304.08485.

Adian Liusie, Potsawee Manakul, and Mark Gales. LLM comparative assessment: Zero-shot NLG evaluation through pairwise comparisons using large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.8/.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

Quentin Malartic, Nilabhra Roy Chowdhury, Ruxandra Cojocaru, Mugariya Farooq, Giulia Campesan, Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Ankit Singh, Maksim Velikanov, Basma El Amel Boussaha, et al. Falcon2-11b technical report. *arXiv preprint arXiv:2407.14885*, 2024.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning, 2022. URL https://arxiv.org/abs/2203.10244.

Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018. URL https://api.semanticscholar.org/CorpusID:49313245.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, pages 2443–2449, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. URL https://arxiv.org/abs/2302.13971.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan

Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL https://arxiv.org/abs/2211.05100.

Xinyu Zhang, Lingling Zhang, Yanrui Wu, Muye Huang, Wenjun Wu, Bo Li, Shaowei Wang, Basura Fernando, and Jun Liu. Diagram-driven course questions generation, 2025. URL https://arxiv.org/abs/2411.17771.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D6/D1.2 FSTP2 Final – TEASE

C.4 INFINITY

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP2 Interim/Final – INFINITY

Inclusive Networking Framework for Immersive XR Technology Integration

Nature	Final Report	**Work Package**	WP1
Project start date	01/01/2025	**Project end date**	30/06/2025
Interim meeting	28/04/2025	**Report submission Date**	27/06/2025
Main authors	Nikos Tantaroudas (DASKALOS APPS)		
Co-authors	Andrew McCracken (DASKALOS APPS)		
Reviewers	Maryam Hashemi (UVA) — Amin Farajian (UNBABEL)		
Version Control			
v0.1	**Status**	Draft	28/04/2025
v1.0	**Status**	Final	27/06/2025

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

Goal for Interim meeting. *For the Sponsor to i) check that the Project Team has clear plans and are on track, and ii) get clarity/confirmation on the next steps and deadlines.*

Goal for Final meeting. *The final evaluation of a project will be performed by the Project Sponsor after the dissemination activities took place. The project team is required to report their results, business plans, secured venture capital for further development and future plans. The Pilot Board will assess the finished projects and evaluate the immediate results. It will also formulate recommendations for sustainability and future operation of the project. The Project Sponsor will then prepare a short report (to be made public) and recommend to the Pilot Board to approve (or not) the final payment to the project Awardee.*

How to complete this report. *The Sponsor asks the Project Team to fill in Sections 1 and 2 prior to the meeting (this entire report is likely no longer than 3–5 pages). After the meeting, the Sponsor writes a recommendation to the Pilot Board (Section 5). Some parts of this report are possibly not relevant to the interim meeting (e.g., subsections 2.2–2.4), they can be addressed in the final version of this report.*

1 Project Execution

Provide an overview of the project execution: development (Section 1.2) and dissemination (Section 1.3). If major deviations from the plan took place, document those first (Section 1.1).

The INFINITY project focused on developing advanced AI-based multilingual communication tools within an immersive XR environment. The execution phase included key technological advancements, such as speech-to-text transcription, text-to-text translation, text-to-speech translation, and text-to-sign-language translation, with a particular focus on International Sign Language (ISL).

During the research phase, an extensive literature review was conducted on sign language translation, revealing that each country has its own distinct sign language with unique linguistic structures. However, International Sign Language (ISL), though widely understood by deaf individuals across the globe, lacked large-scale datasets required for developing robust translation models. To address this gap, the consortium prioritized ISL, aiming to achieve the widest possible accessibility impact. Unlike national sign languages (such as American Sign Language or British Sign Language), ISL is a visual communication system used by deaf individuals from diverse linguistic backgrounds in international contexts, such as conferences, sporting events, online platforms, educational institutes and cultural events. It draws on signs from multiple sign languages, iconic gestures, and universal visual cues to enable understanding across national boundaries. As highlighted by the European Union of the Deaf, ISL has emerged as an effective auxiliary language used when participants do not share a common sign language, allowing for broader accessibility in diverse international settings . By implementing ISL in the avatar's communication, INFINITY ensures that deaf users from various countries, regardless of their national sign language, can engage with and benefit from its features, supporting inclusive design and global reach (European Union of the Deaf, 2018).

To develop an ISL translation model, the team analyzed approximately 750 videos of ISL gestures using Google MediaPipe, extracting key movement coordinates and hand position data. This dataset served as the basis for creating an ISL translation model. Subsequently, an API was developed to map hand positions and gestures to avatar animations within a virtual reality (VR) environment,

allowing for real-time ISL interpretation.

One of the core developments was the implementation of speech-to-text transcription and text-to-text translation, leveraging Facebook's No Language Left Behind (NLLB) model. This enabled real-time, high-accuracy transcription and translation between multiple languages, laying the foundation for inclusive communication in educational and professional XR settings. Furthermore the well-established open source model of Piper (Rhasspy contributors, 2023) was explored for text-to-speech delivery through the 3D avatars within the VR environment. However, because the package depended on several not so well maintained python packages and libraries, a security vulnerability issue was raised for future use and the team decided to switch to a production grade service provided by AWS Polly which is an excellent TTS service.

The project explored several VR-based educational scenarios, where users select their preferred language through an interactive questionnaire. In response, a 3D avatar provides real-time translations, delivering both spoken language and ISL-based sign translations of selected words and phrases. This immersive approach fosters language learning by allowing individuals to simultaneously develop spoken and signed language proficiency.

The final application was built using the Unity engine and deployed on Meta Quest 3 headsets, providing an engaging, fully immersive learning experience. By integrating AI-driven language models and avatar-based ISL translations, the INFINITY project successfully advances accessibility in XR environments, making multilingual communication more inclusive for diverse user groups.

1.1 Deviations from original plan

If relevant, document and justify changes in objectives, changes to the project schedule, as well as changes in the development and dissemination plan. For most projects, this is not applicable.

No deviations occurred during the project. All objectives, including speech-to-text transcription, text-to-text translation, text to speech translation, and ISL translation, were completed as planned. Development and dissemination activities remained on schedule, with no modifications required to the project scope or timeline.

1.2 Development

Provide a brief overview of the progress towards achieving each of the project's objectives. If the work carried out deviates from the work planned, use this section to document the work carried out (list and justify changes in Section 1.1).

The INFINITY project successfully progresses towards its core objectives, focusing on enhancing AI-driven multilingual communication in immersive XR environments. Below is an overview of the key objectives and achievements:

Objective 1. Development of Speech-to-Text and Text-to-Text Translation

- Implemented speech-to-text transcription using AI-based models.
- Integrated Facebook's No Language Left Behind (NLLB) model for high-accuracy text-to-text translation between multiple languages.

- Indicators of Success: Achieved real-time transcription and translation with high accuracy, ensuring accessibility for multilingual users.

Objective 2. Research and Development of International Sign Language (ISL) Translation

- Conducted a literature review on sign language translation, identifying ISL as a global standard despite limited datasets.
- Processed 750 ISL videos using Google MediaPipe and OpenCV to extract hand movements and sign coordinates.
- Constructed an ISL model to facilitate real-time ISL translation.
- Indicators of Success: Created an accurate ISL translation model applicable to a broad audience of deaf individuals worldwide.

Objective 3. Integration of ISL and Speech-Based Translation into a VR Environment

- Develop an API to map hand positions and gestures to 3D avatar animations.
- Design an interactive VR learning scenario where users select a language, and the avatar provides both sign language and spoken translations through text to speech translations.
- Implement the solution in Unity, optimised for Meta Quest 3 headsets.
- Indicators of Success: Preliminary VR-based language learning prototype, enabling users to learn spoken languages and their corresponding ISL signs interactively.

1.3 Dissemination

Describe activities carried out to disseminate the results. Discuss the relevance of the activity, the audience it reaches, etc. Document as much publicly visible information as possible. If the dissemination carried out deviates from what was planned, use this section to document the dissemination carried out (list and justify changes in Section 1.1)

The INFINITY project actively disseminates its results through multiple channels to reach a diverse audience, including researchers, educators, language institutions, and XR industry stakeholders. Below is a structured overview of the key dissemination activities carried out:

- **Partnership and Presentation at TEDx Patras (May 17, 2025)**
  - INFINITY has become a partner of TEDx Patras TEDx Patras, a widely recognised event promoting innovative ideas.
  - The presentation of INFINITY will highlight the role of AI and XR in breaking language barriers through real-time multilingual translation and sign language integration.
  - Audience: General public, technology enthusiasts, and accessibility advocates.
- **Scientific Paper at Salento XR 2025 (June 14–17, 2025)**

  - A conference paper has been accepted for publication paper detailing the development and impact of INFINITY was accepted for presentation at Salento XR 2025. (title: "AI-based Services to Support Language-Learning for Deaf and Hearing Individuals in Immersive XR Settings")
  - The paper covers speech-to-text, text-to-text, and ISL translation, including AI-driven avatar interactions in XR environments.
  - Audience: Researchers, XR developers, and language accessibility experts.

- **Startup Summits in Paris and London (June 2025)**
  - INFINITY was pitched at two leading startup summits in Paris and London, engaging with investors, tech companies, and language institutions.
  - The focus will be showcasing the solution to industry partners and explore partnerships.
  - Audience: Industry leaders, investors, and EdTech companies.

- **Dedicated Project Website and Online Presence**
  - A dedicated project website INFINITY XR was created to share updates, research findings, and interactive demos.
  - Regular updates and discussions are maintained on LinkedIn, fostering engagement with researchers and industry professionals.
  - Audience: Educators, students, XR developers, and language accessibility stakeholders.

- **Collaboration with Deaf Community and Sign Language Educators**
  - Partnered with KENG, a nationally recognized institution for Greek and International Sign Language education based in Thessaloniki, Greece.
  - A 1.5-hour online pilot demonstration was conducted on June 20, 2025, featuring INFINITY's AI-driven pipeline: speech-to-text (Whisper), multilingual translation (NLLB), text-to-speech (AWS Polly), and ISL avatar-based sign rendering.
  - Participants included professional ISL interpreters, deaf educators, and members of the deaf community. The session focused on evaluating the clarity of avatar gestures, timing of translations, emotional cues from sentiment analysis, and overall inclusivity of the XR experience.
  - Feedback emphasized the need for more flexible sign language dialect options, improved motion transitions, and extended vocabulary. Participants highlighted the platform's potential for use in deaf education, language learning, and inclusive digital communication training.

All dissemination activities aligned with the original plan, ensuring maximum outreach and engagement with relevant communities. Future efforts will focus on expanding industry collaborations and increasing adoption within the language education sector.

1.4 Ethics

Document ethical implications and risks relevant to the project, as well as mitigation strategies. How did you minimise ethical risks? How did you ensure the project data management's compliance with all relevant regulations?

The INFINITY project adheres to strict ethical guidelines to ensure responsible development and deployment of AI-driven multilingual translation technologies in XR environments. The primary ethical considerations include privacy and data security, inclusivity and non-discrimination, AI bias mitigation, and transparency.

To minimize ethical risks, all collected data, including speech and sign language datasets, are anonymised and processed securely in compliance with GDPR regulations. The project uses privacy-preserving AI models as API endpoints and do not stores or personally identifies user data, ensuring that sensitive information remains protected. The inclusivity principle is central to the project, with a focus on International Sign Language (ISL) to provide accessibility for a global audience of deaf and hard-of-hearing individuals.

To ensure compliance with ethical and regulatory frameworks, the project follows FAIR (Findable, Accessible, Interoperable, Reusable) data principles and conducted regular audits of AI models to prevent unintended biases. An informed consent process is implemented for all study participants, ensuring transparency about how their data was used.

2 Summary of Results and Plans

Describe the results (their relevance to project goals, impact, etc.), business plans (if applicable), and any future plans relevant to exploitation of results, their sustainability and future operation of the project.

2.1 Results

The INFINITY project successfully develops a multimodal AI-driven translation system integrated within an immersive XR environment, making multilingual communication more accessible for deaf, hard-of-hearing, and linguistically diverse individuals. The results include:

Visual Description of the Envisioned Scenario

The following flowchart in Fig: 1 illustrates the overall workflow of the INFINITY system, showing the integration of AI models for speech recognition, text translation, ISL conversion, and real-time avatar-based learning in XR.

Speech-to-Text and Text-to-Text Translation

- Achieved real-time multilingual transcription using Whisper AI model (Radford et al., 2023) as shown in Fig: 2.
- Ensured high-accuracy text conversion (Facebook, 2022), supporting multiple languages in live interactions as shown in Fig: 3.

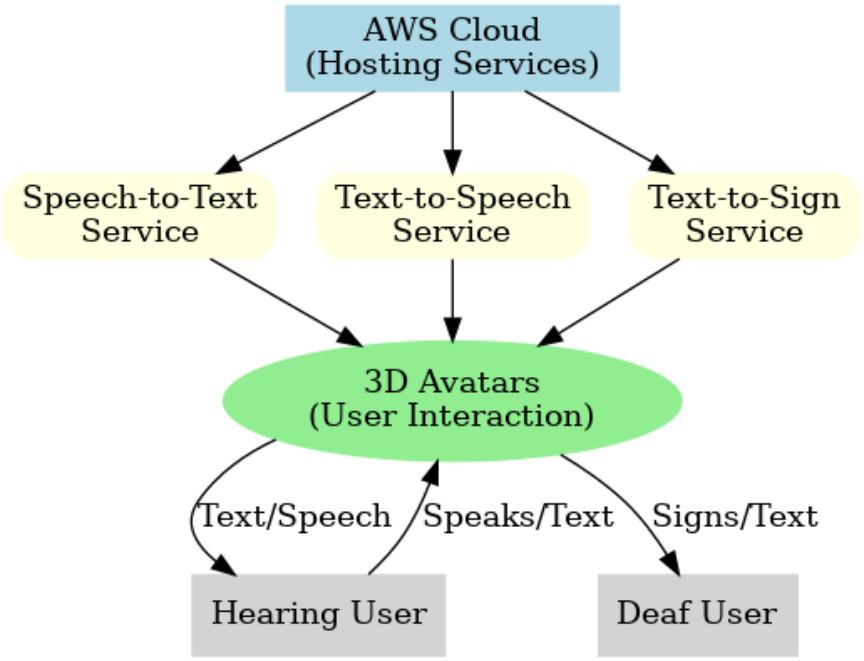


Figure 1: High level description of envisioned scenario and architecture of Infinity.

- Developed a sequential API workflow to ensure automated processing. User speech is transcribed into text, which is then automatically translated into the selected language.

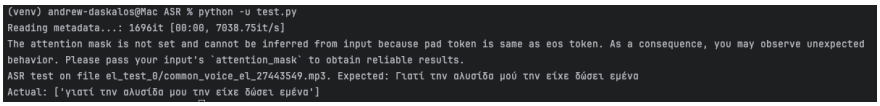


Figure 2: Speech-to-Text and AI-Powered Translation Workflow using Whisper AI. The system transcribes and translates speech in real time across multiple languages.

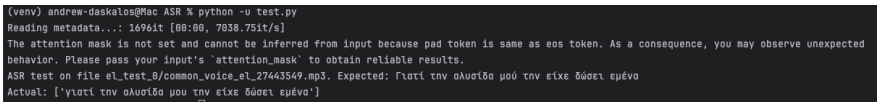


Figure 3: Text-to-Text Translation Workflow using Facebook NLLB. The system translates text in real time across multiple languages.

Text-to-Speech Synthesis

- Integrated AWS Polly to generate natural, real-time audio output in the selected language.
- Used in tandem with translated text to provide multilingual spoken feedback through the avatar system.
- Ensures accessibility for hearing users and enhances multimodal interaction.

Text-to-Speech Benchmarking

Virtual Reality applications demand ultra-low latency text-to-speech (TTS) performance with specialised requirements for spatial audio, avatar lip-sync, and conversational realism. With these requirements in mind, a comprehensive benchmarking of four common TTS services: AWS Polly Standard, Google Cloud TTS Standard, Microsoft Azure Speech, and ElevenLabs has been carried out.

Table 1: Latency Performance Metrics

Service	First Byte Latency	FTTS	Total Response Time
AWS Polly Standard	50–100ms	450ms	780ms
Google Cloud Standard	300–2,000ms	600ms	1,200ms
Microsoft Azure	150–800ms	1,140ms	1,500ms
ElevenLabs	300–500ms	840ms	1,250ms

Performance Testing The data in Table 1 comes from the Picovoice TTS Latency Benchmark (2024) study (Picovoice, 2024). Our testing has also reflected these timings, at least for the first byte latency. The FTTS (First Token to Speech) and total reponse time are affected by the testing pipeline employed. In the benchmark study, there were some other processing layers which introduced their own latency. In our VR application, there is further processing too, so these metrics provide an insight into performance when chained with other steps. These models are under constant development, so the performance is likely to only improve over time.

Table 2: Voice Quality Metrics (Reported by Service Providers)

Service	Mean Opinion Score (MOS)	Word Error Rate (%)
AWS Polly Standard	3.5–3.8	4.2
Google Cloud Standard	3.2–3.5	Variable
Microsoft Azure	3.8–4.0	~3.0
ElevenLabs	3.83–4.2	2.83

Voice Quality Metrics Table 2 shows some key TTS metrics in terms of quality. The higher the Mean Opinion Score (MOS) and the lower the Word Error Rate (WER) the better the quality. However, the MOS and WER values here are taken directly from the service documentation, so will likely have some bias. These metrics have not been tested by ourselves due to time constraints, but have been included to show that they are all roughly the same in terms of the quality needed for our application.

Service Capabilities and Pricing Table 3 summarises the offering and pricing for each service. These are important considerations for our application that needs to be as accessible as possible. Having a service that is going to cost $1000s per month means that we would have to pass the costs onto the end user. This reduces the reach that the application would finally have.

Table 3: Service Capabilities and Pricing (Standard Models)

Service	Languages	Voices	Max Length	Price/1M chars
AWS Polly Standard	34	66	3,000	$4
Google Cloud Standard	40+	220+	5,000	$4
Microsoft Azure	140+	110+	5,000	$4–16
ElevenLabs	29	5,000+	5,000	$22–1,320/mo

Key Findings and Recommendation Based on the benchmarking:

1. **Latency:** AWS Polly Standard provides the lowest and most consistent first byte latency (50–100ms), essential for real-time VR interactions. While not achieving the ideal sub-50ms threshold for perfectly natural conversation, it significantly outperforms other services and remains within acceptable limits for most interactions.
2. **Cost-Effectiveness:** At $4 per million characters, AWS Polly Standard enables broad accessibility without compromising the application's reach. This is critical given the goal of maximising user accessibility.
3. **Consistency:** AWS Polly demonstrates the most consistent performance across our testing, avoiding the high variability seen with Google Cloud TTS (300-2,000ms range). This consistency is crucial for applications where unpredictable latency spikes can break immersion and cause user discomfort.
4. **Feature Coverage:** With 34 languages and 66 voices, AWS Polly provides sufficient variety for most applications and the 3,000 character limit is sufficient for conversational interactions when broken into 1 second chunks

Sentiment Analysis with Emoticon Mapping

- Implemented a sentiment analysis module using a RoBERTa-based transformer model, specifically the `twitter-roberta-base-sentiment` model from CardiffNLP (Barbieri et al., 2020).
- Developed an API endpoint that processes translated text input and returns classified sentiment labels such as "happy", "neutral", or "sad".
- Mapped each sentiment to corresponding emoticons displayed in real time within the VR scene next to the avatar, enhancing affective expressiveness.
- Enabled additional emotional context in multilingual communication, making avatar interaction more natural and socially relevant.

Meeting Summarization for XR Dialogue Context

- Integrated the `flan-t5-base-samsum` model (Schmid, 2022) that is available on Hugging Face for real-time meeting summarization of dialogues in multilingual, XR-based educational scenarios.

Figure 4: Emoticon-based sentiment feedback in the XR environment based on RoBERTa sentiment classification.

- Deployed the summarization API to AWS Lambda as part of the backend services for INFINITY.
- Designed for deaf or hard-of-hearing users, interpreters, and educators to receive concise, natural language summaries of verbal exchanges within the XR scene.
- Planned future refinement includes embedding a ”Summarize” button within the VR interface to generate and display summaries post-session or at key milestones.

International Sign Language (ISL) Translation

- Created an ISL model after processing of 750 videos using Google MediaPipe (Lugaresi et al., 2019) and OpenCV (Bradski and Kaehler, 2008).
- Developed an API to map hand positions to avatar gestures for ISL representation in VR as shown in Fig:7.

Immersive VR Learning Environment

Built using Unity and optimized for Meta Quest 3 headsets as shown in Fig: 8.

Minimum Viable Product (MVP) Demonstration Video

To showcase the integration of AI services into the immersive XR experience, we produced a short MVP video demonstrating real-time interaction with the INFINITY platform. The demo features

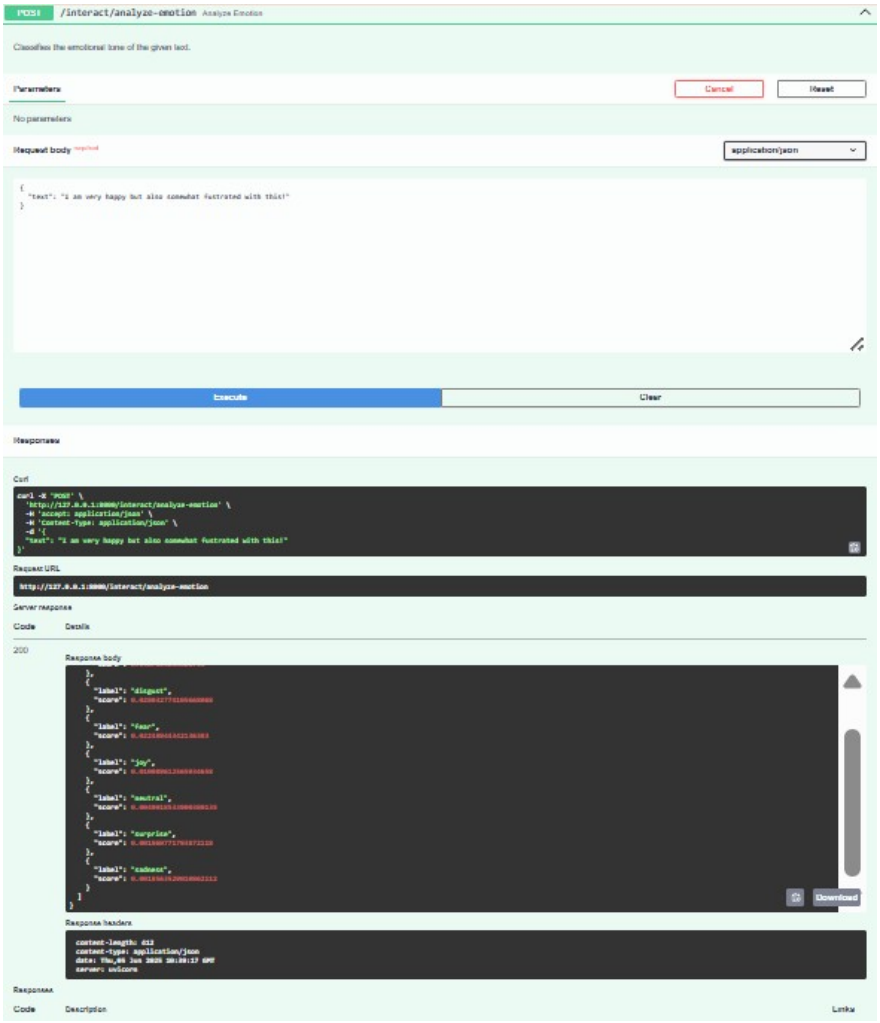


Figure 5: API request and response format for the RoBERTa-based sentiment analysis module.

speech-to-text, multilingual translation, sign language rendering via the avatar, and sentiment feedback through visual emoticons. The video also highlights the user flow. It provides a glimpse into the full pipeline from voice input to sign and spoken output in a Meta Quest 3 environment.

Watch the MVP video here: https://www.youtube.com/watch?v=l4AKG7jbHBE

2.2 Workshops and Pilot Preparations/ Preliminary Feedback Collection

On **May 30, 2025**, the INFINITY team organized a hands-on workshop to demonstrate the current XR prototype and gather targeted feedback from early adopters. We invited **six key stakeholders** from the education, accessibility, and technology sectors to test the platform, interact with the AI-powered avatar, and assess the usability and potential of the system.

Participants experienced the full pipeline—from speech input to real-time text translation and ISL sign output—within the immersive virtual learning scenes. Their feedback guided the next development sprint.

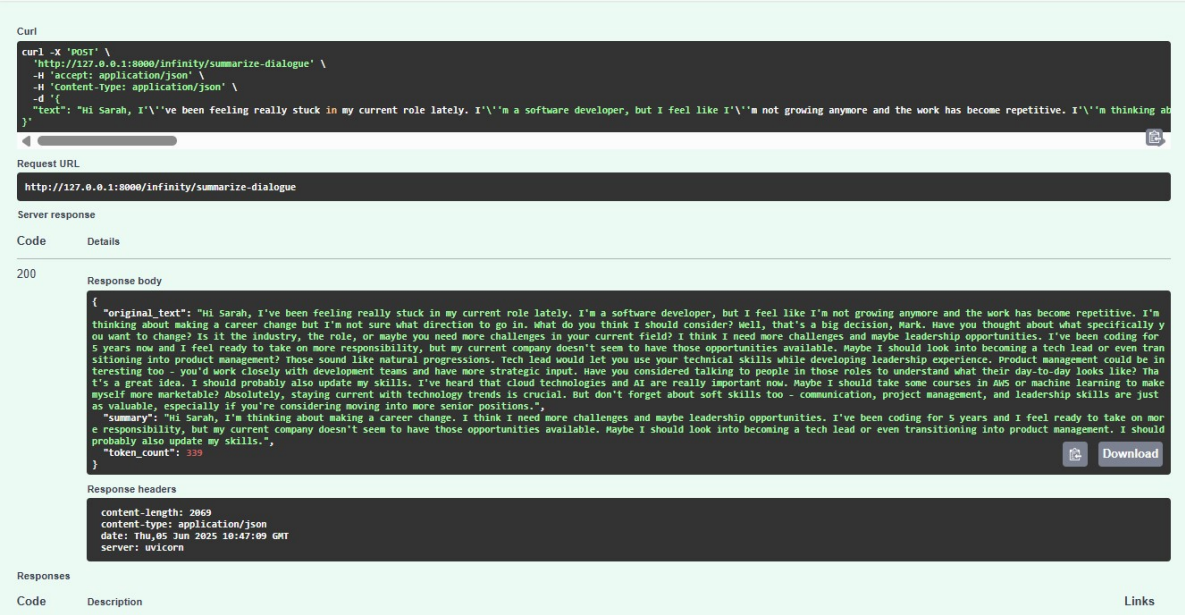


Figure 6: API request and response format for the meeting summarization module based on flan T5 base samsum model.

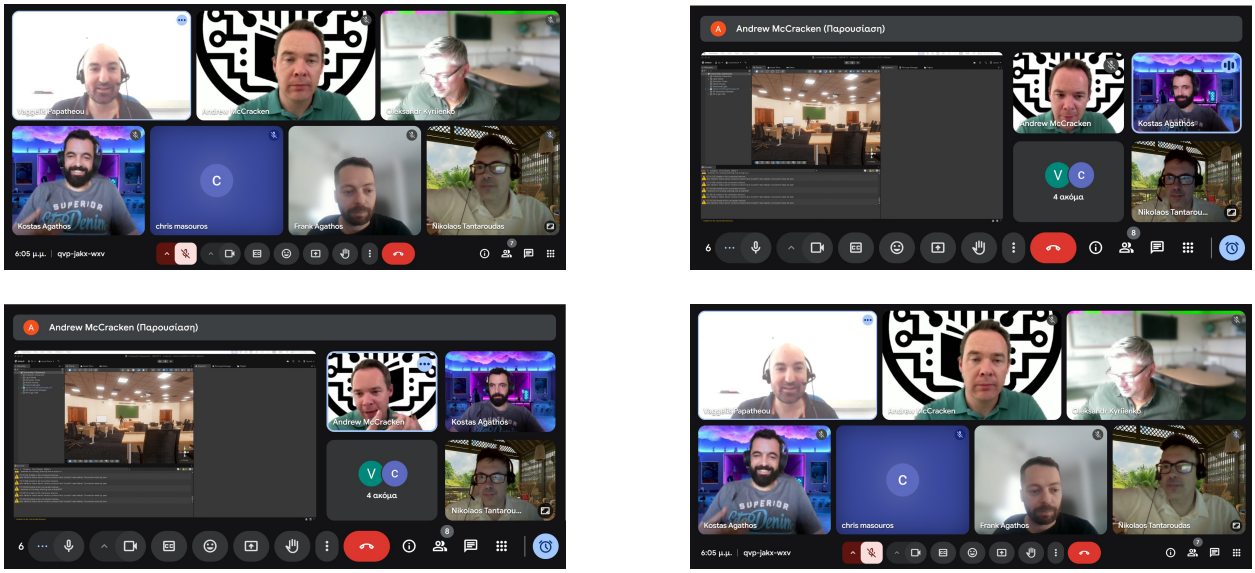


Figure 10: INFINITY stakeholder workshop on May 30, 2025

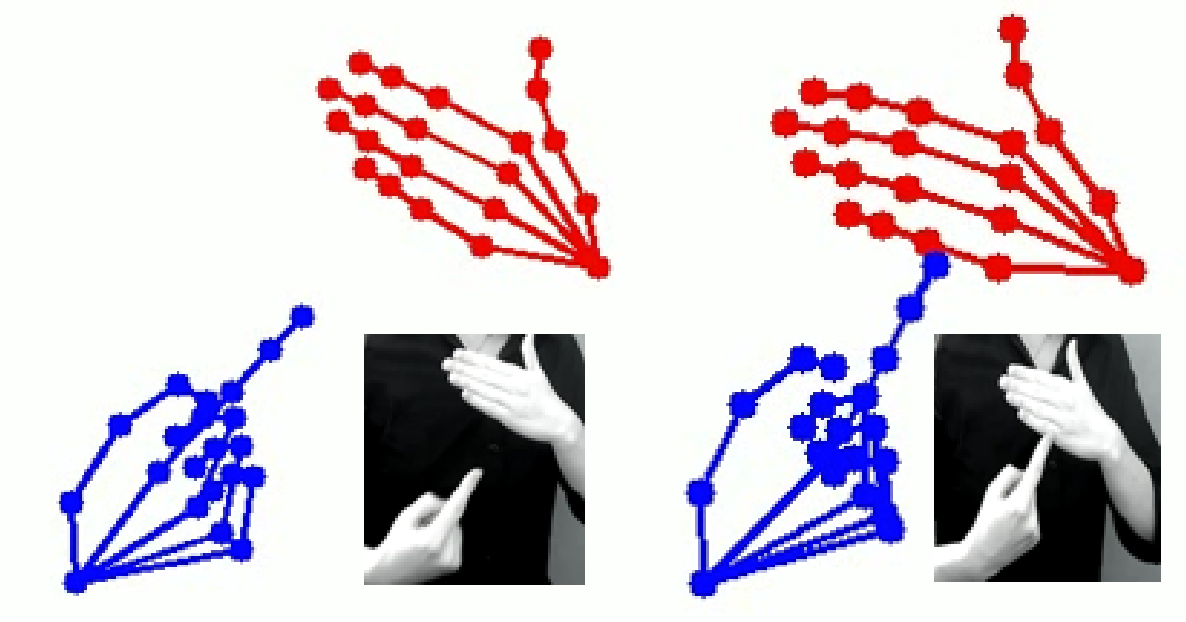


Figure 7: A visual sequence of real-time avatar animation driven by extracted gesture landmarks from the ISL dataset.

Table 4: Key Outcomes from INFINITY Workshop (May 30, 2025)

Workshop Theme	**Key Findings**
Ease of Use	Users found the XR interface intuitive, especially the in-VR language selector.
Avatar Responsiveness	The ISL avatar was perceived as engaging and responsive but would benefit from smoother gesture transitions.
AI Translation Accuracy	Speech-to-text and text-to-text outputs were considered high quality, especially for English, Greek, and Spanish.
Suggestions for Improvement	Requests included more language options, support for deafblind users, and custom vocabulary training.
Potential for Integration	All stakeholders expressed interest in piloting INFINITY in formal or informal educational settings.

2.3 Pilot Demonstration

In collaboration with KENG, one of Greece's leading institutions for sign language education based in Thessaloniki, the INFINITY team organized a dedicated pilot demonstration on **June 20, 2025**. The 1.5-hour online session brought together professional sign language educators and accessibility experts to evaluate the INFINITY platform and provide structured feedback.

The demonstration included a presentation of each AI component integrated into the platform:

- **Speech-to-text transcription** using Whisper AI.

Figure 8: Immersive VR-Based Language Learning System. The AI-powered avatar translates spoken language into International Sign Language (ISL) within a multilingual VR environment.

Figure 9: Screenshot from the MVP demo video showing multilingual avatar interaction in VR.

- **Text-to-text translation** using Facebook's No Language Left Behind (NLLB) model.
- **Text-to-speech synthesis** powered by AWS Polly.
- **Sign language gesture modeling** for International Sign Language (ISL), using Google MediaPipe and AI-driven motion classification.
- **Sentiment analysis** using a pretrained RoBERTa transformer model, which classifies user input into emotional categories (e.g., happy, neutral, sad). The results are mapped to visual emoticons displayed in real time within the XR scene, enriching communication and emotional clarity for learners.

Participants observed a live walkthrough of the XR learning activity, which showcased avatar-driven International Sign Language (ISL) and real-time multilingual speech translation. Following the demonstration, participants were invited to complete feedback questionnaires assessing usability, accessibility, and emotional resonance. Additionally, an open discussion was held with representatives from KENG to explore their impressions in greater depth. Their feedback will directly inform the future iterations of system improvements, including more natural gesture animation, expanded vocabulary coverage, and enhanced delivery of emotional context through the avatar.

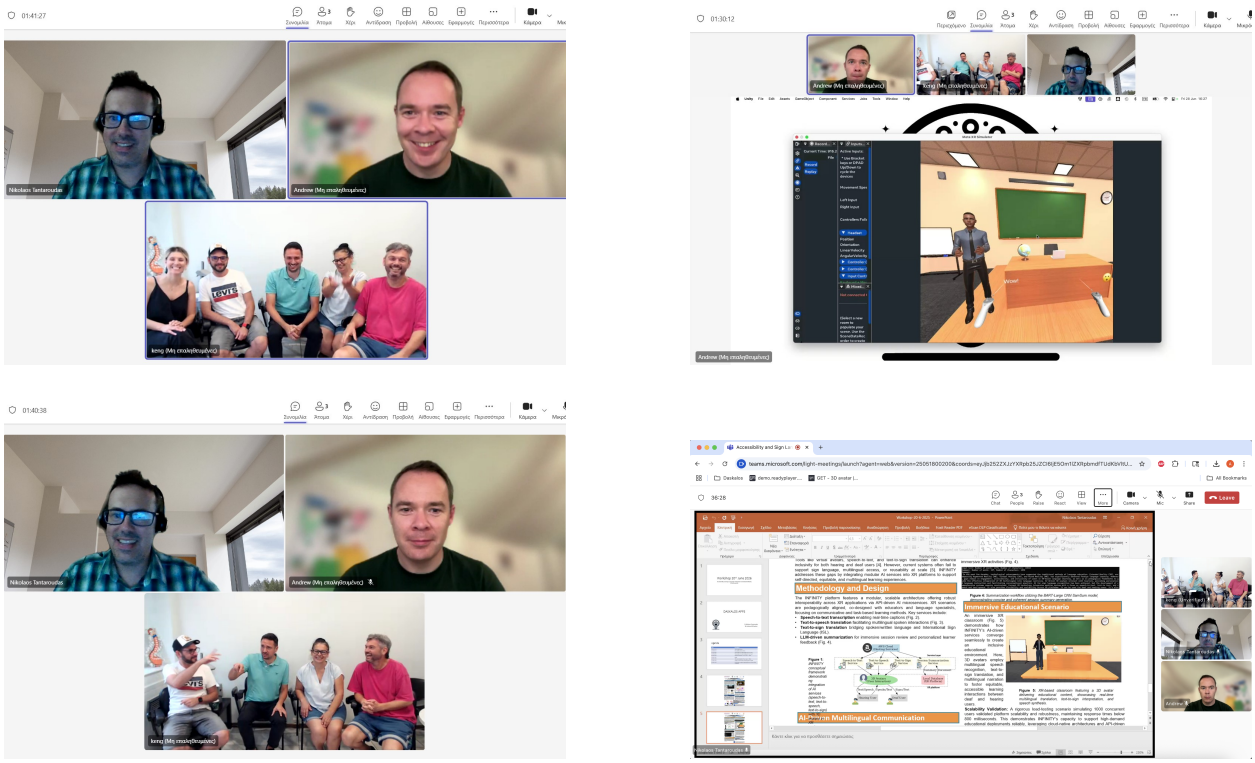


Figure 11: Online demonstration with KENG (June 19, 2025)

Table 5: Key Outcomes from KENG Online Pilot Demonstration (June 19, 2025)

Evaluation Theme	**Summary of Feedback**
Sign Language Modeling	Focus on ISL was appreciated; suggestion to expand to local variants for wider accessibility.
Avatar Performance	Gestures were described as generally natural and intuitive. However, participants highlighted the need for realistic facial expressions to provide critical emotional and contextual cues in sign language. They also noted that internal hand and arm rotations should be refined for more anatomically accurate and fluid animations.
Text-to-Speech Output	Clear and multilingual, with minor latency noted in longer texts.
Sentiment-Driven Feedback	Visual emoticons enhanced engagement; participants praised the addition of affective elements in VR.
Future Integration	High potential for use in deaf education and inclusive language training; participants recommended early school pilots.

2.4 Benchmarking NLLB against EuroLLM of UTTER

To evaluate the potential benefits of adopting the UTTER project's EuroLLM models for the machine translation component of INFINITY, we conducted a comprehensive benchmark comparison between the currently deployed Facebook NLLB-200 model and the EuroLLM 1.7B variants. This benchmarking exercise aimed to assess both translation quality and inference speed, providing data-driven insights for the UTTER team's continued development of European language models.

Benchmarking Methodology

The benchmark evaluation was conducted using a simplified test dataset comprising 10 English-to-French translations with varying complexity levels:

- **Simple sentences**: Basic conversational phrases (3 examples)
- **Medium complexity**: Short technical sentences (4 examples)
- **Complex sentences**: Longer sentences with specialized terminology (3 examples)

For each model, we measured:

1. **Translation Quality**: Using BLEU scores to compare generated translations against reference translations
2. **Inference Speed**: Recording the time taken for each translation request
3. **Resource Utilization**: Monitoring memory usage and model loading times

Technical Implementation

The benchmarking was performed using:

- NLLB-200-distilled-600M model (approximately 600M parameters)
- EuroLLM-1.7B Base model (1.7B parameters)
- EuroLLM-1.7B-Instruct model (1.7B parameters, instruction-tuned)
- Sequential model loading to manage memory constraints
- GPU-accelerated inference with float16 precision

Benchmark Results

The comparative analysis revealed significant performance differences between the models:

Table 6: NLLB vs EuroLLM Performance Comparison

Metric	NLLB-200	EuroLLM 1.7B Base	EuroLLM 1.7B Instruct
Average BLEU Score	79.25	27.58	**84.34**
Average Translation Time (s)	0.596	1.509	**0.529**
Model Load Time (s)	26.63	25.37	40.99
Memory Usage (GB)	~2.5	~3.5	~3.5
Successful Translations	10/10	10/10	10/10

Key Findings

- **Translation Quality**: EuroLLM 1.7B Instruct achieved the highest BLEU score (84.34), outperforming NLLB (79.25) by 5 points. However, the base EuroLLM model performed poorly (27.58), highlighting the importance of instruction tuning for translation tasks.
- **Inference Speed**: EuroLLM 1.7B Instruct demonstrated the fastest inference time (0.529s), marginally faster than NLLB (0.596s). The base model was significantly slower at 1.509s per translation.
- **Architecture Insights**: Purpose-built Seq2Seq models (NLLB) showed strong performance with an average BLEU of 79.25, while Causal LM models averaged 61.37 BLEU across all variants, though the instruction-tuned version significantly exceeded this average.
- **Resource Requirements**: EuroLLM models require more memory (~3.5GB) compared to NLLB's distilled version (~2.5GB), with longer initial loading times for the instruction-tuned variant.

Integration Considerations for INFINITY

The benchmark results present a compelling case for considering EuroLLM 1.7B Instruct as an alternative to NLLB in INFINITY's translation pipeline:

1. **Superior Performance**: EuroLLM 1.7B Instruct offers both better translation quality (6% higher BLEU) and faster inference (11% speed improvement), making it ideal for real-time XR applications.
2. **European Language Focus**: While NLLB supports 200+ languages, EuroLLM's specialized focus on European languages aligns well with INFINITY's target markets and the UTTER project's objectives.
3. **Memory Trade-offs**: The increased memory footprint (1GB additional) is a reasonable trade-off for the performance gains, especially in server-based deployments.

2.5 Business plan

The business plan for INFINITY focuses on leveraging only existing and open source AI services for speech-to-text, text-to-text, and text-to-speech translation, while enhancing user engagement

through immersive VR experiences. The team currently works on creating attractive and interactive VR scenarios designed to make language learning engaging and effective. These VR scenes allow students to interact with AI-powered avatars that provide both spoken and sign language translations, making the learning process more inclusive.

For sustainability, we explore different pricing models, including a pay-per-application approach or a monthly licensing model for institutions. This would allow organizations to integrate INFINITY into their existing learning environments while ensuring ongoing support for students using the platform.

To bring this solution to the market, we actively explore pitching opportunities of INFINITY to foreign language institutions, exploring partnerships for integrating the system into language schools and training centers.

Looking ahead, the focus is on expanding INFINITY's reach and adoption in the education and training sectors. We plan to continue engaging with language institutions, demonstrating the value of AI-powered multilingual translation in virtual reality (VR) learning environments. By partnering with language schools, and training centers, we aim to integrate INFINITY into their programs to support real-time spoken and sign language translation.

To increase accessibility, we will refine and optimise the VR application, ensuring ease of use and compatibility with different XR platforms. Additionally, we plan to offer customised solutions tailored to the needs of different institutions, providing ongoing technical support and training for educators and students.

2.6 Blurb for public dissemination on UTTER's website

Draft a short summary of the project's results, dissemination activities, business plans and future plans. This will serve as a starting point for public dissemination on UTTER's website. Where possible, share links to publicly available results and/or dissemination.

The INFINITY project has successfully progressed towards the development of an AI-driven multilingual translation system integrated into an Extended Reality (XR) environments, making language learning more accessible for deaf, hard-of-hearing, and linguistically diverse individuals. The system combines real-time speech-to-text transcription, text-to-text translation using Facebook's No Language Left Behind (NLLB) model, and International Sign Language (ISL) interpretation. Users are able to interact with AI-powered avatars in VR learning scenarios, where both spoken and sign language translations are provided, enhancing engagement and inclusivity.

To disseminate the project's results, INFINITY was showcased at TEDx Patras (May 17, 2025), highlighting the role of AI in breaking communication barriers. A peer-reviewed scientific paper has also been accepted and was presented as POSTER presentation at Salento XR 2025 (June 17–20, 2025), providing an in-depth look at the system's development and impact. The solution was also being pitched to technology leaders at startup summits in Paris and London (June 2025), promoting partnerships for commercial adoption. Additionally, ongoing dissemination efforts include a dedicated project website, social media engagement on LinkedIn, and regular updates on AI-driven accessibility innovations.

From a business perspective, INFINITY is exploring pay-per-use and monthly licensing models to integrate the system into language schools, and professional training centers. The goal is to provide

a scalable and sustainable solution while offering continuous support for educators and students. Moving forward, the focus will be on expanding adoption, integrating in XR environment the AI models and optimising the learning experience to further enhance accessibility in multilingual education.

3 Dissemination Highlights

TEDx Patras (May 17, 2025) The INFINITY team has partnered with TEDxPatras and was invited to present its inclusive multilingual XR learning system at TEDx Patras, one of Greece's most prominent forums for sharing visionary ideas. In our booth, we showcased how INFINITY's AI-powered avatars combine real-time speech recognition, translation, and International Sign Language (ISL) to create immersive and accessible learning environments. The talk generated significant interest from educators, accessibility advocates, and technologists. We documented the event with photos and social media outreach to amplify the project's visibility.

Salento XR 2025 (June 17–20, 2025) Our scientific paper titled *"AI-Driven Sign Language and Multilingual Translation for Inclusive XR Learning"* has been accepted for publication and presentation at Salento XR 2025. This peer-reviewed publication outlines the technical architecture of INFINITY and reports initial feedback from pilot evaluations. The poster paper is available for download on our official project website. This marks a key milestone in disseminating INFINITY to the scientific community to promote collaboration with XR researchers.

VivaTech Paris & London Tech Week (June 2025) INFINITY was presented at both VivaTech Paris (June 11-14) and London Tech Week (June 9-13), where it was pitched to startups, investors, and educational technology companies. These events served as key commercial outreach moments, enabling conversations about licensing, academic collaboration, and pilot deployments in language schools and training centers.

4 Lessons Learned and Future Work

The interim review and pilot demonstrations provided valuable insights into both the technical capabilities and practical use cases of the INFINITY platform. A key takeaway concerns the educational and social value of the XR experience. Reviewer feedback emphasized the potential for extending the single-user experience into a multiplayer VR setting where a hearing and a deaf user can communicate in real time. This would allow us to better showcase our integrated sentiment analysis system, powered by a RoBERTa-based classifier that maps detected emotions to visual emoticons—making the interactions more expressive and inclusive. Feedback from deaf participants during the pilot demonstration also highlighted the need for improved gesture landmark normalisation across diverse interpreters to ensure smoother and more consistent avatar anima-

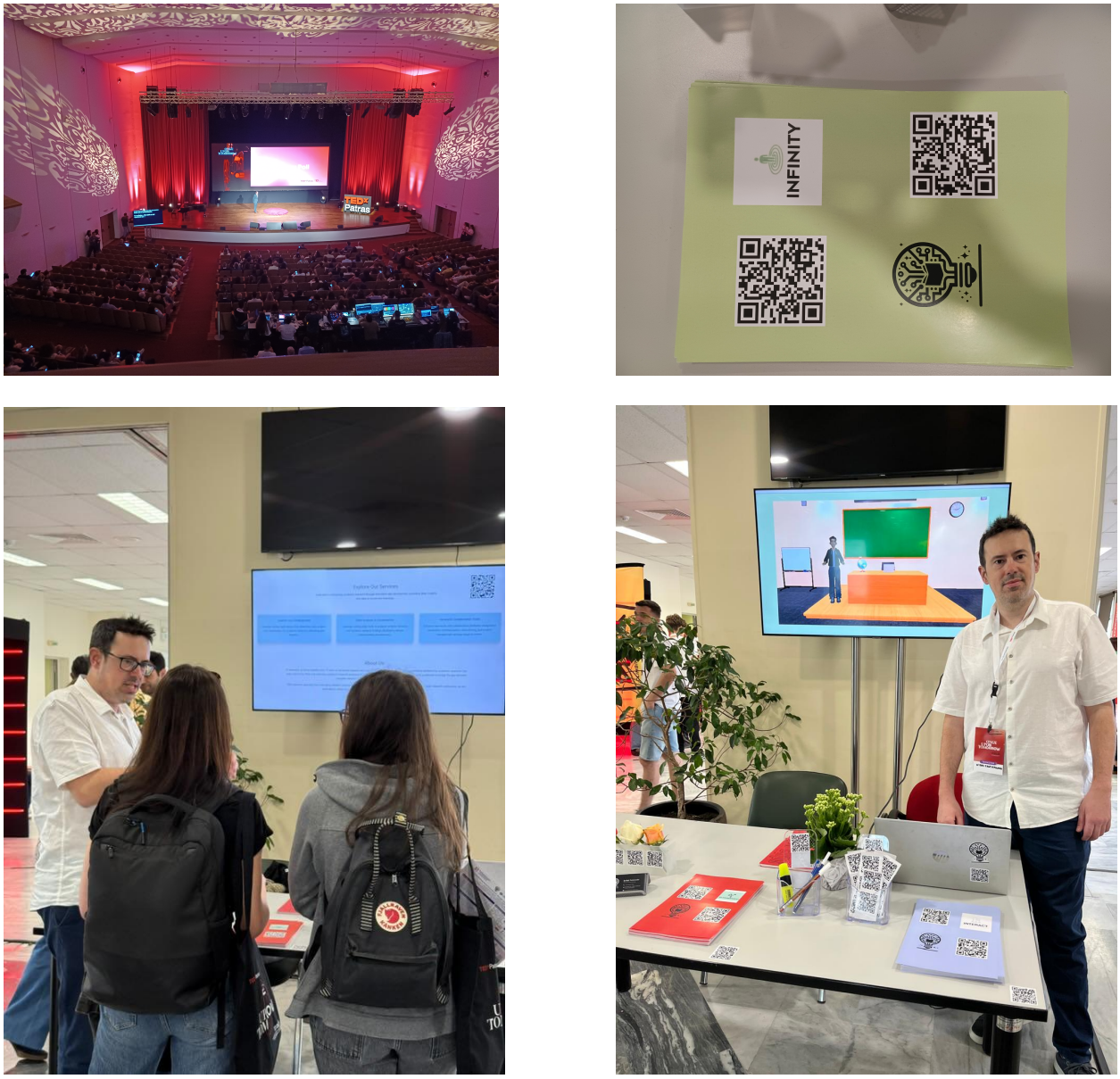


Figure 12: INFINITY showcased at TEDx Patras (May 17, 2025)

tions. In addition, internal rotations of shoulders should be further refined for more realistic avatar animations. Also, the avatar facial expressions should be explored as these provide key context and are very important in sign language communication. Based on these insights, our future work will focus on expanding multiplayer support, improving gesture mapping smoothness, normalizing input data across signers, and incorporating avatar lypsinc functionality and facial expressions, to enrich the learning and communication experience.

5 Recommendation by Project Sponsor

This space is intended for the assessment by the project Sponsor. Did the project deliver its planned results? Was it successfully disseminated? Did the project team document business plans and future plans, and do these seem appropriate? Are there opportunities for exploitation of results

Figure 13: INFINITY at Salento XR 2025 (June 17–20, 2025)

within UTTER?

Figure 14: INFINITY at VivaTech and London Tech Week (June 10 & 11 2025)

Table 7: Public Dissemination Links and Social Media Posts

Channel	**Link**
INFINITY Project Website	https://infinity-xr.netlify.app
TEDx Patras Dissemination Post	LinkedIn Post – TEDx Patras 2025
Salento XR Paper Announcement	Salento XR Paper
General INFINITY Updates	https://www.linkedin.com/company/infinity-xr-app/
INFINITY Demo Video (YouTube)	https://www.youtube.com/watch?v=Xlp-KPgWfn0

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. Tweeteval: Unified benchmark and comparative evaluation for tweet classification, 2020. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment.

Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly Media, Sebastopol, CA, 2008.

European Union of the Deaf. Eud position paper: International sign language. https://eud.eu/eud/position-papers/international-signs/, 2018. Accessed 28 April 2025.

Facebook. No Language Left Behind: Scaling Human-Centered Machine Translation, 2022. URL https://arxiv.org/abs/2207.04672. Preprint.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming-Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019. URL https://arxiv.org/abs/1906.08172. Accessed 28 April 2025.

Picovoice. Tts latency benchmark. https://picovoice.ai/docs/benchmark/tts-latency/, 2024. Accessed: 2025-06-24.

Alec Radford, Jeffrey W. Kim, Tianyi Xu, Greg Brockman, Chris McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the International Conference on Machine Learning*, pages 28492–28518. PMLR, July 2023.

Rhasspy contributors. Piper: A fast, local neural text to speech system. https://github.com/rhasspy/piper, 2023. Release 2023.11.14-2; MIT License; accessed 28 April 2025.

Philip Schmid. flan-t5-base-samsum. https://huggingface.co/philschmid/flan-t5-base-samsum, 2022. Hugging Face model repository.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D6/D1.2 FSTP2 Interim/Final – INFINITY

C.5 VISIXR

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action

Number: 101070631

D6/D1.2 – FSTP2 Final – VISIXR

Vision AI for XR

Nature	Final Report	**Work Package**	WP1
Project start date	02/01/2025	**Project end date**	30/06/2025
Interim meeting	13/06/2025	**Report submission Date**	30/06/2025
Main authors	Leonardo Ranaldi (UEDIN)		
Co-authors	Sebastian Göttert (ZAUBAR)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	**Status**	Draft	XXX
v1.0	**Status**	Final	XXX

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

This section provides an overview of the project's execution, detailing the development work undertaken, the dissemination activities planned, and the ethical framework that guided the project. The project was completed successfully, achieving its objectives as outlined in the original plan.

1.1 Deviations from original plan

There were no major deviations from the objectives, schedule, or plans originally proposed. The project successfully followed the planned development path, with technical decisions and refinements, such as the selection of specific models and platforms, made within the scope of the original plan to optimize performance and outcomes.

1.2 Development

The development phase has been successfully completed, resulting in an innovative, AI-driven XR experience for customer support and product demonstration. The work established a robust technical foundation and a functional prototype that integrates advanced image analysis, a multimodal interaction pipeline, and a real-time user interface.

Objective 1: Initial Development of Core Image Segmentation and Modification

The core engine for image analysis and interaction was successfully built. Following a thorough requirements analysis, BuildShip was chosen as the node-based backend, Grounded-SAM was selected for its context-aware image segmentation capabilities, and LiveKit was integrated as the voice pipeline to enable sophisticated voice interactions. The core segmentation algorithm was fully implemented: images uploaded to the Directus CMS trigger a workflow where an OpenAI Assistant analyzes the image, generates a prompt for Grounded-SAM, and the resulting segmented images are processed and stored for use in the application. A real-time modification pipeline was also established, featuring personality and language selection, a multimodal voice pipeline (STT, User Intent Agent, LLM, TTS), context-related AI image generation, and a custom cut-out tool for real-time user-drawn segmentation.

Objective 2: Integration with Unity3D for Real-Time Rendering

To deliver the user experience, a Unity3D environment was developed and finalized. The project utilizes a WebGL build based on Unity 6, ensuring broad accessibility across web browsers. Bidirectional communication between the web front-end and the WebGL build was established for dynamic control. An asset pipeline using pre-generated Asset-Bundles was integrated to optimize performance and reduce loading times. The final front-end supports fully 2D real-time

interaction through both touch and mouse/keyboard controls, ensuring a responsive and accessible experience across a wide range of devices.

Objective 3: Preliminary Testing and Refinement

A continuous cycle of testing and refinement was employed throughout the project to ensure system stability and performance. The project maintained a stable production-ready page for presentations while conducting new feature development on a detached branch. Performance was rigorously monitored using a live tracker for latency across the entire technology stack. Bug tracking was managed through daily reports and usage analysis. Significant effort was dedicated to optimization and fine-tuning, particularly through extensive prompt engineering for all LLM implementations and a detailed focus on minimizing latency to ensure a smooth and responsive user experience.

1.3 Dissemination

With the development phase complete, the project is now moving into its dissemination phase. The focus has shifted from technical development to sharing the project's advancements with academic, industry, and public audiences. The dissemination plan includes preparing and submitting academic papers to relevant journals and conferences in the fields of AI, XR, and customer experience technologies. Pilot implementations with select retail partners are being planned to generate case studies that demonstrate the real-world impact and tangible benefits of the technology. Further activities will involve presenting at conferences, releasing select components of the system as open-source projects to foster community engagement, and actively engaging with media to promote the project's results.

In addition to these planned measures, several concrete dissemination activities have already been carried out:

- **VIR Innovation Days (June 24–25, 2025)**
  Presentation of the system as an info guide for travelers.
  About the event: https://v-i-r.de/events/vir-online-innovationstage/

- **Open House at the Federal Ministry for Economic Affairs (August 23–24, 2025)**
  Demonstration of the general core functions to a broad public audience.
  About the event: https://www.bundesregierung.de/breg-de/schwerpunkte/tag-der-offenen-tuer/tag-der-offenen-tuer-2373830

- **Web and Social Media**
  The project has also been communicated through a dedicated blog post on the project website and a LinkedIn post, ensuring broader awareness and visibility beyond physical events.

LinkedIn post: [click here]
Blog post: [click here]

Furthermore, an additional dissemination activity is already scheduled:

- **Meetup Event at hubraum – Tech Incubator of Deutsche Telekom, Berlin Campus (September 25, 2025)**
  Presentation of the system during a meetup organized by the project team.

Through these combined measures, the dissemination activities ensure visibility within the scientific community, industry stakeholders, and the general public. By participating in large trade fairs, industry events, government outreach initiatives, and targeted online communication, the project establishes a broad foundation for future adoption and collaboration.

.

1.4 Ethics

The project was committed to addressing all relevant ethical implications and risks through proactive mitigation strategies. Key ethical concerns identified included data privacy, AI bias, transparency, accessibility, and the potential for misinformation. To minimize these risks, the project implemented robust data protection measures, including secure storage and encryption, in alignment with GDPR principles. To combat AI bias, the team applied rigorous testing to identify and mitigate biases in the AI models, using diverse training data and regular audits of AI-generated outputs. Transparency was ensured by clearly informing users that they are interacting with an AI system. The development process also prioritized inclusive design to ensure the platform is accessible to users with various abilities. To address potential misinformation, a multi-layered content moderation system combining AI filtering with human oversight was implemented. The system's design also incorporates clear policies on intellectual property for user-uploaded content and AI-generated materials. Through these measures, the project has ensured that its data management and operational procedures are compliant with all relevant regulations and uphold the highest ethical standards.

2 Summary of Results and Plans

The project successfully delivered on its goals, producing a functional prototype and a clear path forward for future development, exploitation, and dissemination.

2.1 Results

The project has yielded significant results, culminating in an integrated, end-to-end system that demonstrates the core vision. The most critical result is the successful development of a functional prototype that combines a sophisticated back-end AI pipeline with an interactive front-end user experience. This includes a fully operational workflow for image segmentation where images are uploaded, analyzed by an AI to identify key features, segmented using Grounded-SAM, and made available for interaction. A key technical achievement is the real-time multimodal interaction pipeline, which effectively processes user speech, understands user intent, queries a dynamic knowledge base with a high-performance LLM, and generates a spoken response, creating a fluid conversational experience. The successful integration with a Unity-based WebGL front-end makes the experience accessible on a wide range of devices and demonstrates the feasibility of delivering XR-like interactions through a standard web browser. The iterative testing and optimization cycles led to important improvements, such as switching to more efficient models and refining the LLM pipeline to significantly reduce response latency. These results validate the project's technical approach and confirm that a revolutionary customer support and product demonstration platform has been successfully created.

Find a live demo and tutorial video here: https://about.zaubar.com/en/utter-zaubar-assistants-for-xr

2.2 Business plan

The core business objective is to commercialize the developed technology as a sophisticated, visually-driven AI assistant platform, targeting a significant gap in the e-commerce and customer support markets. Current solutions are predominantly text or voice-based, lacking the intuitive, contextual interaction that visual exploration provides. Our platform directly addresses this by allowing users to "talk to the image," transforming static product photos into interactive, informative experiences.

Target Markets and Value Proposition:

Our primary target markets are businesses that rely on visual appeal and detailed product information to drive sales and support customers.

- **E-commerce and Retail:** For online stores, our platform offers a way to bridge the gap between online browsing and an in-store experience. It enhances customer engagement, increases conversion rates by answering questions instantly, and reduces the burden on

human support agents. The agent's personality can be customized to align perfectly with the brand's image.

- **Complex and Technical Products:** For industries like consumer electronics, automotive, or industrial machinery, where products have numerous features, our tool allows customers to self-serve and understand complex functionalities in a simple, visual way. This improves the pre-sales experience and provides a powerful post-sales support tool.
- **Training and Education:** The platform can be used to create interactive training materials where trainees can visually explore and inquire about complex diagrams, equipment, or scenarios.

Product and Service Model:

The platform will be offered as a Software-as-a-Service (SaaS) solution, allowing for easy integration into existing websites and e-commerce platforms. A tiered pricing model will be based on factors such as the volume of interactions, the number of products supported, and the level of customization required for the AI agent's knowledge base and personality.

Concrete Use Case Examples:

- **High-Fashion E-commerce:** A customer is viewing a runway look on a luxury brand's website. They can click directly on the model's shoes and ask the AI agent, "What is the heel height?" or "Are these available in a different color?". The agent, adopting a sophisticated brand voice, provides the answer instantly and might proactively suggest, "The matching handbag for this look is also available. Would you like to see it?".
- **Automotive Configurator:** A potential buyer is exploring a car's interior on the manufacturer's website. They can click on the dashboard and ask, "Show me how the driver-assist features are controlled from the steering wheel." The agent can highlight the relevant buttons on the image and provide a step-by-step explanation, creating a much more engaging experience than a static user manual.
- **DIY/Home Improvement:** A customer is looking at a power tool on a hardware store's website. They are unsure about a specific attachment. Using the custom cut-out tool, they can circle the attachment and ask, "What is this piece used for?". The AI agent can identify the part, explain its function, and even link to a tutorial video showing it in action.
- **Technical Support:** A user is having trouble with a new appliance. They can take a photo of the control panel, upload it, and ask the AI agent, "Why is this light blinking?". The agent can analyze the image, cross-reference the blinking light with its technical database, diagnose the issue, and guide the user through the solution.

Strategic Value for UTTER:

This project provides significant strategic value to the broader UTTER ecosystem. It introduces a powerful visual and spatial interaction layer that complements UTTER's existing strengths in speech and language processing. The technology can be integrated into other UTTER initiatives to create richer, more capable solutions. For instance, UTTER's meeting assistant could be enhanced to allow participants to visually query charts and diagrams within a presentation, or the multilingual customer support assistant could use our visual context to provide more accurate

and helpful responses. This project, therefore, not only stands as a viable commercial product on its own but also as a key enabling technology that expands the scope and capability of the entire UTTER framework.

2.3 Future plans

The system is designed with future expansion in mind. While the current implementation is optimized for 2D image interactions in web browsers, the underlying architecture is prepared for integration with 3D models and more immersive XR experiences. The forward-thinking approach paves the way for a seamless transition to specialized hardware interfaces, such as haptic gloves for tactile feedback or eye-tracking systems for more intuitive navigation. This scalability ensures that as XR technologies evolve and become more mainstream, our solution can easily adapt, providing businesses with a future-proof investment that can grow alongside emerging technologies and changing customer expectations. Furthermore, we plan to actively engage with the broader tech community by releasing selected components of our system as open-source projects. This approach will foster community engagement, drive further innovation in XR customer support technologies, and accelerate the adoption and development of these technologies by a wider range of businesses and developers.

2.4 Blurb for public dissemination on UTTER's website

This project introduces an innovative XR chatbot that revolutionizes customer support and product demonstrations through advanced AI and image analysis technologies. Our system combines sophisticated image segmentation, vision-based AI analysis, and a dynamic knowledge base to create an interactive, visually-driven customer experience. Users can explore product images through clickable segments, interacting with a customizable spatial agent in a web-based interface. By addressing critical gaps in current e-commerce and customer support solutions, the project introduces immersive, context-aware interactions that set new standards for engaging, informative, and personalized customer support in the digital age. Initially web-based, the system is designed for future expansion into full XR experiences with 3D models and specialized hardware, demonstrating the practical and powerful application of XR principles in everyday commercial interactions.

3 Recommendation by Project Sponsor

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D6/D1.2 FSTP2 Final – VISIXR

C.6 SwarmChat

Unified Transcription and Translation for Extended Reality

(UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D6/D1.2 – FSTP2 Final – SwarmChat

SwarmChat: Enabling Intuitive Swarm Robotics with Natural Language

Nature	Final Report	Work Package	WP1
Project start date	02/01/2025	Project end date	15/07/2025
Interim meeting	14/05/2025	Report submission Date	01/07/2025
Main authors	Sponsor (NAV)		
Co-authors	SwarmChat (ORG)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	Status	Draft	14/5/2025
v1.0	Status	Final	01/7/2025

This project has received funding from the European Union's Horizon Europe Research and Innov ation programme under Grant Agreement No 101070631 and from the UK Research and Innova tion (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

Goal for Interim meeting. For the Sponsor to i) check that the Project Team has clear plans and are on track, and ii) get clarity/confirmation on the next steps and deadlines.

Goal for Final meeting. The final evaluation of a project will be performed by the Project Sponsor after the dissemination activities took place. The project team is required to report their results, business plans, secured venture capital for further development and future plans. The Pilot Board will assess the finished projects and evaluate the immediate results. It will also formulate recom mendations for sustainability and future operation of the project. The Project Sponsor will then prepare a short report (to be made public) and recommend to the Pilot Board to approve (or not) the final payment to the project Awardee.

How to complete this report. The Sponsor asks the Project Team to fill in Sections 1 and 2 prior to the meeting (this entire report is likely no longer than 3–5 pages). After the meeting, the Sponsor writes a recommendation to the Pilot Board (Section 3). Some parts of this report are possibly not relevant to the interim meeting (e.g., subsections 2.2–2.4), they can be addressed in the final version of this report.

1 Project Execution

1.1 Deviations from the original plan

No major deviations. Two low-risk tasks were re-ordered to accelerate UX delivery without affecting scope, budget, or milestones.

1.2 Development

All core milestones have been met. An end-to-end system now converts spoken or typed commands in **nine EU languages** into validated XML behaviour trees (BTs) and executes them in a 50-agent simulator at ≈30 Hz.

Example Objective 1. System Development

ID	Objective	Key Work Performed	Status
O1	Multimodal translation pipeline	Integrated Seamless M4T-v2-Large (audio) and EuroLLM-9B (text) achieving ~80 % exact-match on 90 multilingual prompts.	✅ Completed
O2	Safety screening	Llama-Guard-3-8B filter.	✅ Completed
O3	Behaviour-tree synthesis	Prompt augmentation, dictionary injection, and strict XML validator.	✅ Completed
O4	Live swarm simulation	"Violet" Pygame simulator renders 50 agents with live BT overlay and pause/stop controls (30 FPS).	✅ Completed
O5	Synthetic training corpus	Self-instruct expansion from 35 seeds to 2062 BT examples(70/20/10 split).	✅ Completed
O6	Model fine-tuning	11 candidate LLMs benchmarked; top-3 shortlisted and Falcon-V.3-10B-Instruct currently .	✅ Completed
O7	Documentation & release engineering	Model cards, Website, Demo, and repo readme	✅ Completed

Example Objective 2. Model Selection & Evaluation

Eleven open-weight LLMs (6.7 B – 14 B parameters) were evaluated with zero-, one-, and two-shot templates. BLEU, ROUGE-L, and XML validity were computed for five swarm-control scenarios. Top three: Falcon-10B, Mistral-7B, Qwen-2.5-Coder-Instruct.

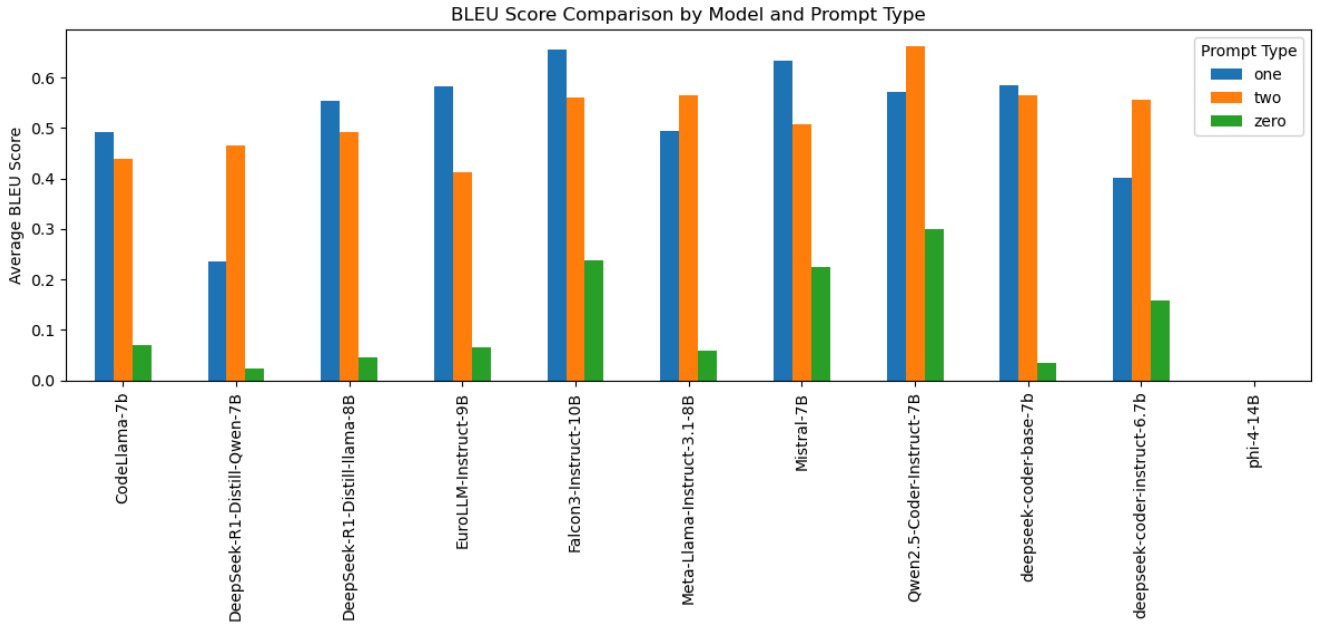


Fig.1: 11 Performance of 11 open-source LLMs on behaviour-tree synthesis

1.3 Dissemination

Our communication plan follows FAIR principles and Horizon Europe open-science mandates, ensuring every artefact is publicly accessible, traceable, and released under CC/Apache license[1].

Open-Source Hub

- GitHub: Source code, evaluation notebooks, website
- Hugging Face: Model weights, datasets, model cards, live demo
- The **Project website** is accessible online via https://swarmchat.github.io/ with direct links to demos, codes, and the paper (to be shared after publication)

Visual Identity

- Logo & slide deck– A minimalist swarmChat logo (Fig. 2) and branded slide template boost recognisability at events and online.

Fig. 2: SwarmChat Logo

Social & Community Outreach

- LinkedIn posts
  - LinkedIn | Thrilled to share our major progress on the SwarmChat journey!
  - LinkedIn | We are thrilled to announce that our proposal has been selected for funding

[1] Some of the LLM models we used are licensed under the Apache 2.0 license; consequently, portions of our work are also released under Apache 2.0.

  - under the UTTER FSTP Open Call 2!
  - Post | Dive into the SwarmChat Final Report
- Blog posts on our company website.
  - Third Post
  - Second Post
  - First Post
- Dedicated website SwarmChat Website
- Medium
- X posts
  - Second post
  - First post
- Youtube
- GitHub,
- Huggingface,

Scholarly Dissemination

Full paper targeted for the 9th International Symposium on Swarm Behaviour & Bio-Inspired Robotics 2025 (deadline 15 Jul 2025)

Community adoption (open-source)

- Freely available to academic labs, makerspaces, NGOs, and etc.
- Foster ecosystem growth via plug-in modules (new behaviours, translators, simulators).

1.4 Ethics & safety

- **Data privacy** — Speech recordings are discarded immediately after transcription; no personal data is retained.
- **Transparency & reproducibility** — Code, datasets, and evaluation scripts released under permissive licences.
- **Compliance** — Repository follows GDPR and FAIR data management guidelines.
- **Safety**
  - The LLM assembles a behavior tree from a pre-defined set of atomic behaviors that we developed manually. This guarantees that no harmful behaviors can be generated by the LLM itself.
  - The Integrated LLAMA-Guard model filters the input and blocks any harmful prompts. (e.g., “kill a person”; Fig.6).

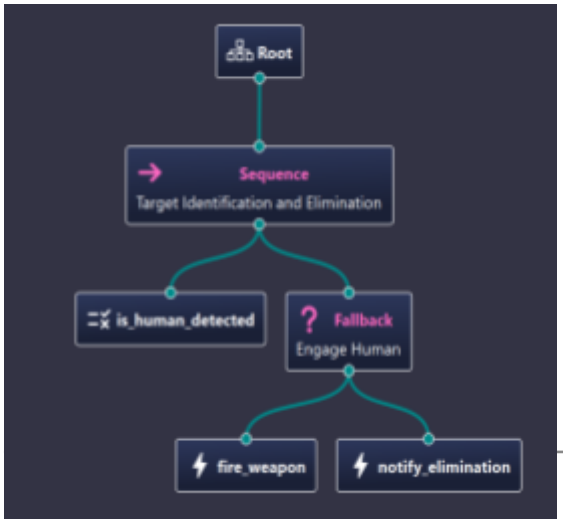


Fig. 6: Example of the model generating a harmful behavior tree.

2. Summary of Results and Plans

2.1 Results: Model benchmarking & fine-tuning:

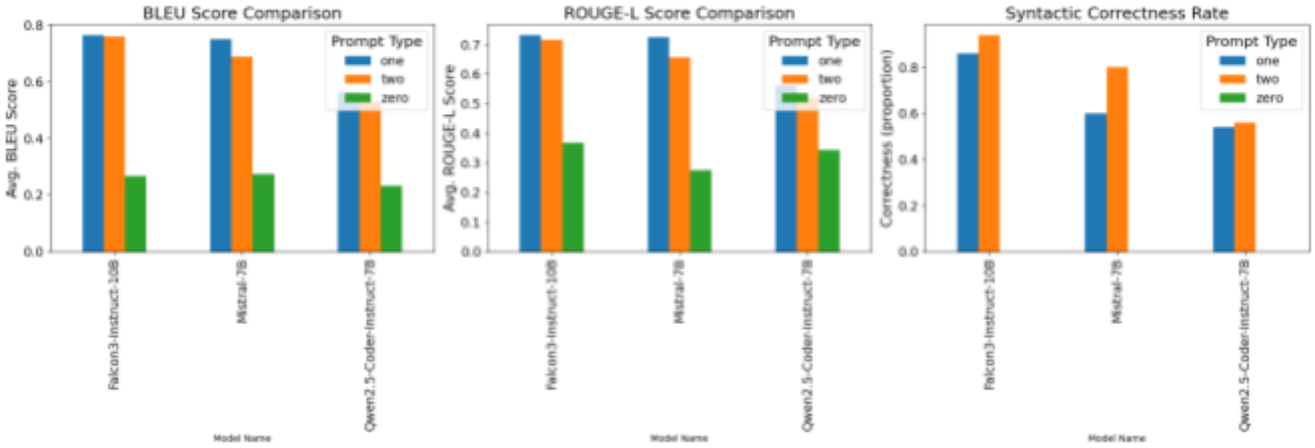


Fig.3: The three top-performing LLMs: Falcon3-Instruct-10B, Mistral-7B-v3, and Qwen2.5

Eleven open-source LLMs (ranging from 6.7 B to 14 B parameters, all 4-bit quantized) were assessed across five standard swarm scenarios using zero-, one-, and two-shot prompts (Fig.1). The top performers were Falcon3-Instruct-10B, Mistral-7B-v3, and Qwen2.5-Coder-Instruct-7B (Fig.3), each achieving BLEU and ROUGE-L scores ≥ 0.56 and syntactic validity ≥ 0.54 in the two-shot setting.

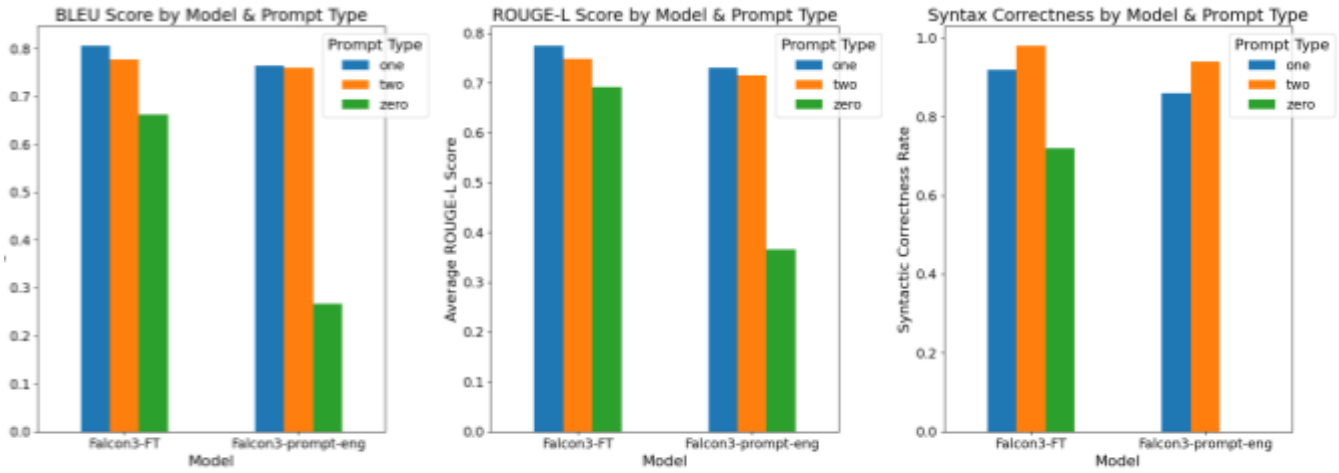


Fig.4: Falcon3-10B before and after LoRA fine-tuning

Using LoRA to fine-tune Falcon3-10B on 2,063 synthetic instruction-BT pairs significantly improved its zero-shot performance: BLEU jumped from 0.27 to 0.66, ROUGE-L from 0.37 to 0.69, and syntactic validity from 0 % up to 72 % (Fig.4).

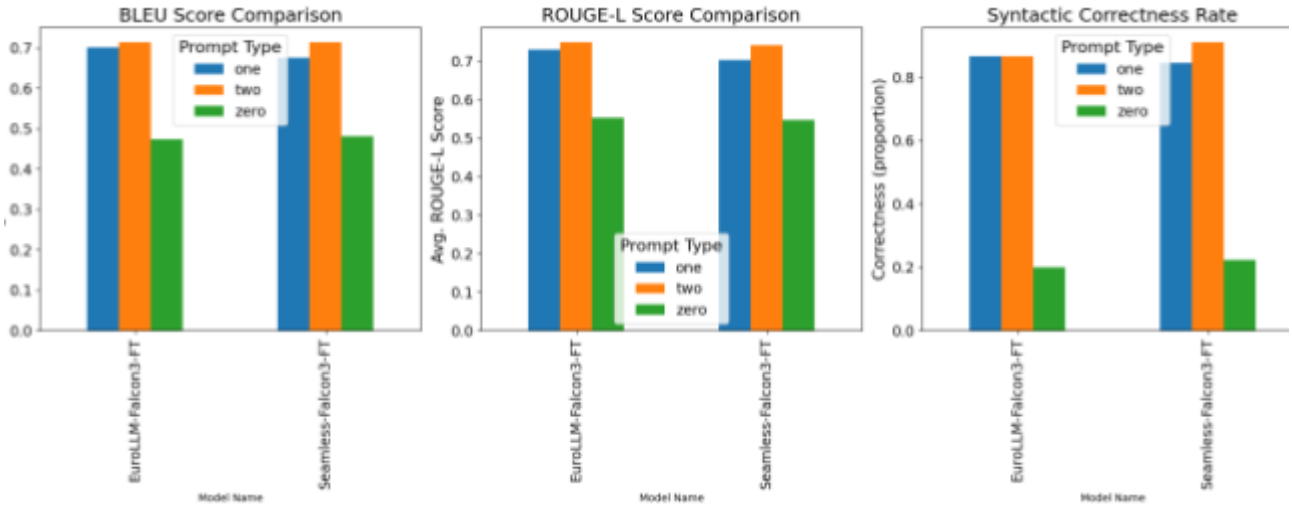


Fig.5: Comparison of EuroLLM-Falcon3-FT vs. Seamless-Falcon3-FT across BLEU, ROUGE-L, and syntactic correctness in zero-, one-, and two-shot settings.

EuroLLM-Falcon3-FT and Seamless-Falcon3-FT perform almost identically, with only minor shifts: EuroLLM-Falcon3-FT is a touch stronger in one-shot learning as indicated by BLEU (0.702 vs. 0.676) and ROUGE-L (0.729 vs. 0.702) scores. On the other hands, Seamless-Falcon3-FT holds a slight edge in zero-shot learning as shown by the corresponding BLEU score (0.481 vs. 0.473) and two-shot syntactic correctness (91.1 % vs. 86.7 %) (Fig.5).

2.2 Business plan

Short-term — SwarmChat v2.0 with hierarchical BTs and expanded swarm behaviours, pursued through a follow-up Horizon Europe project.
Mid-term — SaaS APIs and SDKs for industrial, agricultural, and search-and-rescue robotics.
Long-term — Integral voice-driven control for extraterrestrial robotic swarms (cf. OffWorld.ai) supporting astronaut-robot collaboration on the Moon and Mars.

2.3 Future plans

Building on our current successes, we plan to:

Real-world data & benchmarking

- Collect and integrate real mission transcripts and logs from field exercises to expand our training corpus.
- Benchmark our open-source models against leading enterprise AI services to quantify performance, cost, and latency trade-offs.

Hierarchical BT generation

- Extend the prompt-schema approach to emit nested and recursive subtrees, enabling multi-stage, long-horizon missions that exceed today's context-window limits.

Expanded modalities

- Add support for additional languages (e.g., Asian and Middle Eastern) and incorporate vision inputs—allowing users to upload sketches, map snapshots or camera frames so

SwarmChat can interpret visual context alongside text when generating behavior trees.

Physical swarm trials

- Port SwarmChat to small drones and ground robots; validate real-world performance and safety in indoor and outdoor environments.

2.4 Blurb for public dissemination on UTTER's website

SwarmChat is an open-source, multilingual pipeline that converts everyday speech or text into **validated XML behaviour trees (BTs)** and streams them to swarms of autonomous robots. Built under the EU-funded **UTTER** programme, it couples state-of-the-art transcription (SeamlessM4T / EuroLLM), safety filtering (Llama-Guard), and a LoRA-fine-tuned **Falcon-10B** behaviour-tree generator. The result: operators—expert or novice—can task dozens of heterogeneous machines *in real time* with a single spoken sentence.

SwarmChat Website, Inventors Hub site, Medium, GitHub, Huggingface, Youtube

3 Recommendation by Project Sponsor

This space is intended for the assessment by the project Sponsor. Did the project deliver its planned results? Was it successfully disseminated? Did the project team document business plans and future plans, and do these seem appropriate? Are there opportunities for exploitation of results within UTTER?

Mid-term: The project is very well on-track, EuroLLM9B model is used in the pipeline and a demo is planned at the end of the project.

End-of-project: All key milestones have been achieved. The full system now transforms spoken or typed commands in nine EU languages into validated XML behavior trees (BTs) and executes them within a 50-agent simulator. I watched a video of the demo prototype and also tested it myself. It worked well, although it was a bit slow due to being hosted on Hugging Face's basic (free) infrastructure. Finally, the FSTP team provided correct documentation of the ethical implications and risks associated with their project.

Overall, the report and presentation were both clear, and all aspects of the project—model, demo, and blog posts—were made readily accessible in the dissemination materials. I recommend proceeding with the payment of the remaining portion of the FSTP funding for this project.

References

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631
D6/D1.2 FSTP2 Interim/Final – SwarmChat

C.7 EOLAS

Unified Transcription and Translation for Extended Reality

(UTTER)

Horizon Europe Research and Innovation Action Number: 101070631 D6/D1.2 – FSTP2 Final – EOLAS: E-Learning Of Language Augmented Services

Nature	Final Report	Work Package	WP1
Project start date	28/03/2025	Project end date	04/07/2025
Interim meeting	12/05/2025	Report submission Date	07/07/2025
Main authors	Ben Peters (IT)		
Co-authors	Ian Mills, Ryan McCloskey, Daniel Hickey (SETU/Walton Institute)		
Reviewers	Maryam Hashemi (UVA)		
Version Control			
v0.1	Status	Draft	*Final meeting - 20250704*
v1.0	Status	Final	*Revisions & updates - 20250707*

This project has received funding from the European Union's Horizon Europe Research and Innovation programme under Grant Agreement No 101070631 and from the UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee (Grant No 10039436).

Contents

1 Project Execution

1.1 Deviations from original plan

There are no deviations to report.

1.2 Development

Phase 1:

T1.1 **High level requirements**

In this task we delivered high-level requirements for the project.

T1.2 **Application Design and Integration**

As part of the design work on EOLAS we developed an architectural diagram, data flow and a series of wireframes for the AR application. Our architecture laid out the internal structure of the platform and how each component communicated to each other. Our data flow contained the communications pathways, and the design wireframes we created in figma guided our development activities.

T1.3 **Develop AR Application and Chat Features**

As part of this task, we have developed an AR (augmented reality) application which allows the end user to observe the world around them, highlight a particular object and receive the translated recognition object, this is facilitated directly through our Eolas API to the OpenShift

cluster. The Eolas API allows the AR application to send the outputs from our object recognition model to the deployed fine-tuned model to translate and generate sentences for the end user/learner to practice. Our object recognition model is based on a vision transformer model called FastViTMA36F16 and is designed to take input from the user selection and interpret the contents of the image. The Eolas LLM can then use this input to translate and provide additional sentences/information for the recognised object. The output from the translation is visible on screen to the end user and a chat bar is available to send additional queries about the recognised output.

T1.4 **Dataset collection/Model Evaluation**

We performed evaluation of Euro LLM (1.7B, 9B, 9B-instruct) models on our local deployment to translate regular queries. As part of our development, we tried to deploy 1.7B on device but were unsuccessful. UCCIX-Llama2-13B[1]-Instruct was selected to use as our base model for fine tuning.

T1.5 **Fine tune & Deploy LLM**

We sourced Bunachar Náisiúnta Moirfeolaíochta | Irish National Morphology Database a database of 43,000 Irish words and used it during the fine tuning process through LoRA, a parameter-efficient fine tuning technique that has been shown to maintain strong performance while substantially reducing memory requirements. The LLM was deployed to our openshift cluster and is available through our Eolas API endpoints.

T1.6 **Testing and Refinement**

We have tested the main application running UCCIX-13B-instruct with our internal group in Walton Institute, the feedback aiding in the refinements of the application. Our latest iteration with fine-tuned components is being deployed to our TestFlight at present in order to further test the efficacy of AR language learning.

Phase 2:

T2.1 **Create and Distribute Promotional Materials**

We have contacted our internal marketing team to start a press/promotion campaign for the application and its benefits. We are developing the press release with regard to EOLAS with them. We have sent the promotional material/videos and images of the application to our marketing team. This will be published through our marketing portals.

T2.2 **Partner Engagement and Presentations**

During the project we had contact with several prospective partners. An initial trial has taken place with the SETU Irish language learning club gauging feedback on capabilities and impacting our refinement process in T1.6. Based on user feedback, the application output initially contained verbose responses. We have since altered this through our fine tuning process so that it provides more concise grammatical information to the end user/learner. We have also

[1] *https://huggingface.co/ReliableAI/UCCIX-Llama2-13B-Instruct/*

added the ability to request variations in modern/traditional Irish words as some learners were more familiar with the modern variants. We implemented simplified responses as a result of the feedback.

T2.3 **Outreach and Publication**

As part of the above dissemination activities, we targeted educational conferences within the XR field. An upcoming call for papers call for Digital Education Conference (DEC 2025[2]) is expected towards late summer and we will submit our findings to date. More activities are pending. We have contacted the Sneem Digital Hub in order to organise a trial in Sneem, a town in Kerry which is part of the Iveragh Gaeltacht (Irish speaking regions where the language is promoted and persevered) and are looking to organise a trial shortly with the community as well.

1.3 Dissemination

We have held an internal trial, contacted Gaeltacht organisations, and are planning future trials. A post on the Walton Institute website[3] and social media posts is currently under review and will be published shortly on the main website and social media outputs. The structure of the planned paper is to test the efficacy of our fine tuned model against the Llama 13B instruct model as mentioned in T2.3. A number of regional events are planned in the Gaeltacht and through our connection in SETU we are planning to showcase EOLAS at Sneem summer festival (https://www.sneemfestivals.ie) for further feedback and dissemination.

1.4 Ethics

As part of our requirements in T1.1, we stated that data privacy was a high priority. In order to deliver on this requirement, we designed the application to transmit any queries securely through our developed secure API. The object recognition component of the AR application only processes on device and nothing beyond the result of the recognition is transmitted to Eolas infrastructure. No user data is collected on device or on server and the application is distributed through our own TestFlight application through a shared link to end users in collaboration with our SETU group.

2 Summary of Results and Plans

Results

A series of requirements were delivered through T1.1. A suitable application data flow and architecture was developed in T.12. As part of T1.3 an augmented reality application was developed with on device object recognition capabilities. We developed a secure API that facilitates the transmission of recognition results and additional queries from end users. We evaluated local deployments of LLMs for translation in T1.4 such as EuroLLM and UCCIX locally and tested them with Irish language queries. In T1.5 an OpenShift deployment with UCC-IX was

[2] *https://www.setu.ie/events/dec25*

[3] https://waltoninstitute.ie

deployed and exposes the endpoints needed for communication with the applications. The architecture was designed in such a manner to be extensible in future with additional models for testing through addition of multiple endpoints. In T1.6 we have identified and processed the data required for fine tuning through Bunachar Náisiúnta Moirfeolaíochta | Irish National Morphology Database. This was then used to fine tune using LoRA on our current deployment. T1.6 Testing and refinement was carried on the deployed application using acceptance testing within an internal stakeholder group. Feedback regarding the output was used to provide refinements to both the application itself and updated interactions with the API. We have engaged in dissemination activities to promote the application. Marketing materials for the project site and event promotion were provided to the marketing team in T2.1, we engaged in internal trials and planned a Gaeltacht region event in Sneem in T2.2, and are planning outputs through publication in DEC25 and an outline paper structure agreed in T2.3.

2.1 Business plan

We developed our business and dissemination plan and liaising with our Walton institute/SETU marketing department and Technology Transfer office in order to commence the discussions and identify target key groups to whom this project may be of interest in order to identify an exploitation plan.

- Follow-on channels such as Enterprise Ireland commercial funding for researchers will also be considered and investigated in collaboration with SETU technology transfer office as a possible future project option
- Invention disclosures will be completed, and licencing options will also be investigated with our SETU technology transfer office.

We see great potential in the expansion of EOLAS features. As such we would like to progress development either through additional funding opportunities in the EU or as stated about a commercialisation output. We have prepared a business plan containing our future feature, target audiences, revenue models and route to market.

2.2 Future plans

While the EOLAS project has come to end we see a broader future with enhanced features as mentioned in our business plan and BMC (Business Model Canvas):

- On device LLM models
- Conversational chat mode with 3D dynamic avatar

- Local scene recognition models
- Synthetised voice input and output

With the successful integration of XR with LLMs, the virtual 3D assistant can act a novel means of means of interaction with the end-user audience, making the conversational experience with EOLAS even more personable. AR is the medium of choice for this project and seeing initial results we can expand this to include MR headsets with passthrough enabled.

2.3 Blurb for public dissemination on UTTER's website

E-Learning Of Language Augmented Services (EOLAS) delivered teaching of the Irish language through an XR enhanced platform. Through a fusion of LLM interaction and object recognition technologies to identify and translate everyday objects into centre of the conversation, EOLAS has delivered a novel means of language learning open to all learning levels. EOLAS is bolstered by the addition of an open source grammatical database, enabling the LLM to become your own personal language tutor. With EOLAS, which in Gaeilge (Irish) means knowledge, we leveraged an Irish Large Language Model (LLM) to facilitate learning through interaction. EOLAS creates a dynamic and interactive language learning environment, making Irish language education more engaging and effective for learners of all levels. More information is available through https://waltoninstitute.ie and through the EOLAS landing page at https://github.com/Ian-Mills/EOLAS.

3 Recommendation by Project Sponsor

Given the release of the language learning app leveraging an XR-based LLM and the planned activities for dissemination in both academic and non-academic contexts, the project has delivered the expected results in all aspects. Therefore, it is recommended that EOLAS receive the final funding part as originally planned.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D6/D1.2 FSTP2 Final – EOLAS

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D11/D1.3 Report on second set of FSTP projects