UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action

Number: 101070631

D3.2 – Final Report on Multimodal, Multilingual Pre-trained XR Models

Nature	Report	**Work Package**	WP3
Due Date	30/09/2025	**Submission Date**	30/09/2025
Main authors	Marcely Zanon Boito (NAV)		
Co-authors	Barry Haddow (UEDIN), Alexandra Birch (UEDIN), Ben Peters (IT)		
Reviewers	Vlad Niculae (UVA)		
Keywords	pre-trained models, XR models, efficient training		
Version Control			
v0.1	**Status**	Draft	12/09/2025
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

In this report, we present the progress of WP3 over the course of the project, with a particular focus on RP2 (for a detailed account of RP1, see D 3.1). The primary objective of this WP is to develop pretrained models that can be effectively leveraged by other WPs. Our investigation addresses not only the development of such models, but also advances in training and adaptation strategies, pursuing approaches that are more efficient, multilingual, and robust.

During RP1, we achieved substantial progress with the release of the Tower family of models, highly competitive LLMs for translation, as well as mHuBERT-147, a foundational speech model trained on 147 languages. In RP2, we continued this trajectory, introducing the general-purpose EuroLLM model family and Spire, a speech-to-text multimodal LLM designed for transcription and translation from English speech.

Beyond model releases, we present 5 different outputs related to efficient training for the adaptation of LLMs and machine translation systems. Taken together, the contributions across RP1 and RP2 demonstrate the UTTER consortium's strong contributions to open science in the pretraining of both speech and text models.

1 Introduction

Proposal

"The purpose of this package is to develop pretrained models that can be effectively leveraged for multi-modal (written and spoken language) translation. The objective can be structured into two components:

- **Model pretraining and availability:** Collect speech and language models on selected and large datasets. A large language model developed for the BigScience project will be adapted and integrated for usage in the project. An analysis and evaluation of existing pretrained speech/textual models and their relevance for the target task will be done. All the selected speech and text models will be integrated and provided in a unified interface for usage by the other working groups.
- **Architectural investigations:** Investigate how speech and text pretrained models can be efficiently combined together for translation. We will analyze the performances of several combinations of the pretrained speech and text architectures made available in the previous tasks for end-to-end spoken and/or written translation. We will formulate recommendations for optimally combining pretrained models."

Work Presented in the First Deliverable

In the first deliverable, we discussed that since the project's proposal and the BigScience project, more competitive pre-trained models for text were released. We thus shifted focus towards the adaptation of those newer competitive models, proposing Tower, a derived model from the popular open-source LLaMA-2 (Touvron et al., 2023a). For speech, observations from the self-supervised learning (SSL) SUPERB benchmark (Yang et al., 2021) convinced us of the potential of building a multilingual speech model based on the Hidden Units BERT (HuBERT, Hsu et al. (2021)) architecture. We trained and released the first general-purpose multilingual HuBERT: mHuBERT-147. Both text (Tower) and speech (mHuBERT-147) models were shown to be very competitive on their domains.

Regarding efficient training, we proposed a parameter efficient distillation approach for the speech architecture Whisper, which allows us to bridge the gap between large and small speech-to-text Whisper models without much cost for inference. We also proposed a CTC regularization approach via coarse labels for improving the performance of speech translation models.

Work Presented in this Deliverable

Manuscripts:

- **3 conference papers:** Ambilduke et al. (2025), Baziotis et al. (2024), Wang et al. (2025b)
- **1 workshop paper:** Iyer et al. (2024)
- **3 arXiV papers:** Martins et al. (2024, 2025), Zhu et al. (2025)

Code and data:

- **EuroLLM models:** HuggingFace
- **SPIRE models:** HuggingFace and GitHub

2 Task 3.1: Investigating and pretraining text and speech models (NAV*, UEDIN, IT)

Proposal highlights

In this task, we gather our work on pretraining multilingual speech, text and multimodal models. These backbones are used to support the work of our other WPs, and are also shared with the scientific community to advance research in the topic.

Summary of completed work

We present two key contributions related to this task. The first one is the EuroLLM collection of models, covering many languages and model sizes (Section 2.1). The second contribution is the Spire collection of multimodal (speech and text) models, which allowed us to investigate speech integration in text-only LLMs (Section 2.2).

2.1 Output 1: `EuroLLM` models

This output is described in Martins et al. (2024, 2025). The EuroLLM project has been an effort led by members of the UTTER consortium to develop an LLM that addresses the needs of European citizens by training it from scratch with a significant proportion of non-English data. This work was made possible by a successful bid for resources from a EuroHPC Extreme Call where we got approved 420k node hours (4xH100) on October 2023 for Barcelona Super Computer on project EHPC-EXT-2023E01-042. We have also been selected as one of the best 15 Extreme call projects for JUREAP and assigned 220k node hours on JUPITER running from May-October 2025.

We have worked together to develop a series of LLMs trained from scratch to support all 24 official European Union languages and 11 additional languages. The website for the project is linked here: https://eurollm.io/, and the repository with the models and data sets are linked at HuggingFace. The project consisted of a phase of scaling experiments. This led to the release of a 1.7B base and instruct model on the 6 August 2024. This model has been downloaded 215k times from Huggingface to date (1 September 2025). Then, a few months later on 2 December 2024 we released the 9B base and instruct (280k downloads to date). We have also released a 22B preview model on 11 June 2025.

Data and models. In two technical reports we describe the development process and report our main results. The EuroLLM-1.7B and EuroLLM-1.7B-Instruct models were described in Martins et al. (2024). The EuroLLM-9B models were described in Martins et al. (2025). EuroLLM addresses the issue of European languages being underrepresented and underserved in existing open LLMs. In the reports we provide a comprehensive overview of EuroLLM's development, including tokenizer design, architectural specifications, data filtering, and training procedures. We describe the pre-training data collection and filtering pipeline, including the creation of EuroFilter, an AI-based multilingual filter, as well as the design of EuroBlocks-Synthetic, a novel synthetic dataset for post-training that enhances language coverage for European languages (both detailed in D 2.2). Evaluation results demonstrate EuroLLM's competitive performance on multilingual benchmarks and machine translation tasks, establishing it as the leading open European-made LLM of its size.

Pre-trained	**Arc** (25-shot)	**Hellaswag** (10-shot)	**MMLU** (5-shot)	**TruthfulQA** (0-shot)	**MMLU-pro** (5-shot)	**MUSR** (0-shot)	**Borda C ↓**
Non-European							
Gemma-2-9B	**67.89**	**67.73**	**66.19**	50.63	29.75	**9.70**	**1.3**
LLaMa-3.1-8B	55.46	58.86	55.54	49.49	19.94	5.44	3.0
Granite-3-8B	47.42	51.73	47.10	49.34	20.38	7.07	4.0
Qwen-2.5-7B	50.68	52.17	62.44	**54.06**	**31.63**	8.04	2.2
OLMo-2-7B	38.25	42.23	41.32	45.24	13.91	4.53	5.8
Aya-23-8B	47.53	53.48	45.44	47.64	14.04	3.64	4.7
European							
Mistral-7B	51.31	53.38	50.09	47.15	17.36	8.69	2.3
Occiglot-7B-eu5	46.90	51.12	42.52	44.81	11.98	3.83	3.7
Salamandra-7B	61.15	64.73	42.75	46.06	5.25	2.63	3.0
EuroLLM-9B	**66.48**	**67.00**	**55.68**	**51.84**	**17.60**	**10.97**	**1.0**

Table 1: Comparison of the pre-trained versions of open-weight LLMs on multilingual benchmarks, averaged across EU official languages. For Arc, Hellaswag, MMLU, and TruthfulQA we use EU20 benchmark. For MMLU-Pro and MUSR we translate the English version with Tower to 7 EU languages (German, French, Spanish, Portuguese, Italian, Dutch, and Czech). Scores of MMLU-PRO, and MUSR are normalized between random baseline and maximum possible score, following the methodology used in *Open LLM Leaderboard*.

Post-trained	**Arc** (25-shot)	**Hellaswag** (10-shot)	**MMLU** (5-shot)	**TruthfulQA** (0-shot)	**MMLU-pro** (5-shot)	**MUSR** (0-shot)	**WMT24++** en→xx (0-shot)	**WMT24++** xx→en (0-shot)	**Borda C ↓**
Non-European									
Gemma-2-9B-IT	**64.60**	**64.28**	**65.13**	**60.65**	27.42	**8.38**	**80.47**	**80.39**	**1.1**
LLaMa-3.1-8B-IT	51.39	56.92	57.07	55.05	24.22	4.01	77.43	77.70	3.0
Granite-3-8B-IT	46.98	52.86	49.36	56.04	20.10	7.90	66.59	67.24	4.1
Qwen-2.5-7B-IT	47.91	51.61	62.27	57.88	**29.68**	7.62	69.32	69.54	3.0
OLMo-2-7B-IT	39.00	43.30	41.86	48.57	12.38	4.02	62.43	63.20	5.9
Aya-Expanse-8B	47.78	54.94	51.33	53.57	19.77	5.52	72.02	73.90	3.9
European									
Mistral-7B-IT	52.63	53.40	48.29	**58.01**	**18.19**	6.94	70.08	70.01	4.0
Ministral-8B-IT	51.71	55.36	51.22	52.53	17.41	6.17	73.52	73.77	4.1
Occiglot-7B-eu5-IT	42.39	49.52	39.75	48.10	11.77	4.17	61.14	59.59	6.4
Salamandra-7B-IT	55.16	63.46	47.30	51.15	7.01	7.17	81.38	78.21	3.8
Pharia-1-LLM-7B-C	38.58	43.13	34.68	45.80	10.10	**9.83**	51.71	49.44	6.8
Teuken-7B-IT-R-v0.4	55.42	60.96	37.65	54.75	9.29	2.25	75.19	70.50	4.8
Teuken-7B-IT-C-v0.4	53.77	60.25	40.05	52.68	9.79	2.94	76.44	71.65	4.6
EuroLLM-9B-IT	**60.67**	**64.94**	**55.37**	53.99	17.04	9.02	**84.19**	**83.94**	**1.6**

Table 2: Comparison of the post-trained versions of open-weight LLMs on multilingual benchmarks, averaged across EU official languages. For WMT24++ we average the Comet-22 scores on all 21 language pairs which include English as source (for en→xx) or as the target language (for xx→en). Scores of MMLU-PRO, and MUSR are normalized between random baseline and maximum possible score, following the methodology used in *Open LLM Leaderboard*.

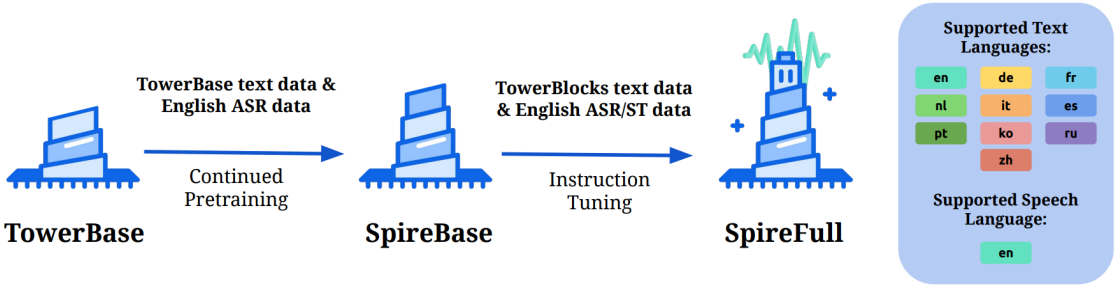

Figure 1: The two-step training approach to build Spire from Tower.

To support open research and adoption, we release all major components of this work, including the base and instruction-tuned models, the EuroFilter classifier, and the synthetic post-training dataset.

Findings. As we can see in Tables 1 and 2, in 13 languages, our pre-trained model outperforms the other European open-weight models in all the tasks. In the case of post-trained models, in 14 languages, EuroLLM-9B-IT outperforms all the other open-weight models in all the tasks, except TruthfulQA. For this specific task, it ranks fourth on average. Our initial investigation suggests that it is mainly due to the challenging characteristics of this benchmark. Finally, in the machine translation task, EuroLLM-9B-IT outperforms all other European models across all language pairs and translation directions, with the sole exception of Greek→English.

2.2 Output 2: From Tower to Spire: Adding the Speech Modality to a Translation-Specialist LLM

This output is described in Ambilduke et al. (2025). The Spire family of models was an effort to adapt the text-only LLM Tower (Alves et al., 2024) to also handle speech. In its final form, the model excels at automatic speech recognition (ASR) and speech translation (ST) while maintaining TOWER's abilities on textual tasks such as machine translation. The work, to be presented at EMNLP 2025 (Findings), included researchers from four of UTTER's partners: Instituto de Telecomunicações, Unbabel, the University of Edinburgh, and NAVER LABS Europe; as well as several collaborators from other institutions. We have released[1] five models developed during the project. The best-performing version of the model, SpireFull,[2] has been downloaded 3.4k times as of 8 September 2025. Spire, like Tower, uses a two-stage training approach, shown in Figure 1, which combines a continued pretraining (CPT) stage whose role is to imbue the model with new knowledge and abilities with an instruction tuning (IT) stage whose role is to finetune the model to follow instructions related to specific tasks.

Speech as discrete units. The core of the Spire approach is speech *discretization*: whereas previous models have used an ASR pipeline or various techniques to project speech representations into the space of a text LLM, Spire incorporates speech by making use of a cluster-based speech tokenizer based on HuBERT. This tokenizer converts a continuous speech signal into a sequence of discrete units (DSUs); these units can then be processed by the LLM as though they were text.

[1] https://huggingface.co/collections/utter-project/spire-67d4253d6af8d6a0308527e0
[2] https://huggingface.co/utter-project/SpireFull

One advantage of this approach is its architectural simplicity; while other models require special modules to map between modalities, the only architectural change between Tower and Spire is the addition of the DSUs to the model's vocabulary.

Continued pretraining. After adding 5000 new types for DSUs to TowerBase-7B[3]'s embedding matrices, we trained the model on a total of 6 billion tokens. Of these, 5 billion consisted of discretized ASR examples from publicly available datasets, while the remaining billion consisted of monolingual and parallel text from various public sources. Notably, the 5 billion DSU tokens amounts to only 35k hours of English speech, far less than many speech foundation models.

Instruction tuning. Following continued pretraining, we finetune the model on a mixture of text and speech instructions for textual machine translation, speech recognition, and speech-to-text translation. The training set combines TowerBlocks (Alves et al., 2024) data, which includes high quality instructions for translation-related tasks, with speech recognition examples, gold standard speech translation examples, and pseudo-labeled speech translation examples in which speech recognition transcripts are machine-translated into other languages.

Findings. Experiments show that Spire can effectively handle ASR, MT, and ST tasks. For ASR, the fully trained SpireFull model outperforms Whisper-base and matches or surpasses several multimodal baselines trained on far larger datasets, though it lags behind large-scale systems like Whisper-large-v3 and SeamlessM4T. On MT benchmarks, Spire maintains Tower's strong translation performance across 10 languages, confirming that integrating speech does not compromise text-based capabilities. In ST, SpireFull demonstrates robustness, performing competitively in both direct and cascaded settings, especially when cascading ASR and MT. However, its direct ST performance remains dataset-dependent and is weaker than models trained on larger and more diverse speech corpora.

[3] https://huggingface.co/Unbabel/TowerBase-7B-v0.1

3 Task T3.2: Efficient Training (UEDIN*, NAV, IT)

Proposal highlights

In this task, we report our research on efficient training. This includes work that improve and/or investigate efficiency and training performance. It also includes fine-tuning approaches leveraging pre-trained models.

Summary of completed work

We present the following contributions related to the task of efficient training.

Training and Adaptation of LLMs:

- Data balancing optimization (Section 3.1.1);
- Long-context instruction-tuning (Section 3.1.2);
- Language impact for chain-of-thought in LLMs (Section 3.1.3).

Improving machine translation training:

- Low-resource translation with LLMs (Section 3.2.1);
- Data augmentation for encoder-decoder models (Section 3.2.2).

Overlapping contributions are mentioned in Section 3.3.

3.1 Training and Adaptation of LLMs:

3.1.1 Output 3: HBO: Hierarchical Balancing Optimization for Fine-Tuning LLMs

This output is described in Wang et al. (2025a). Fine-tuning LLMs on a mixture of diverse datasets poses challenges due to data imbalance and heterogeneity. Existing methods often address these issues across datasets (globally) but overlook the imbalance and heterogeneity within individual datasets (locally), which limits their effectiveness.

We introduce Hierarchical Balancing Optimization (HBO), a novel method that enables LLMs to autonomously adjust data allocation during fine-tuning both across datasets (globally) and within each individual dataset (locally). *HBO* employs a bi-level optimization strategy with two types of actors: a *Global Actor*, which balances data sampling across different subsets of the training mixture, and several *Local Actors*, which optimizes data usage within each subset based on difficulty levels. These actors are guided by reward functions derived from the LLM's training state, which measure learning progress and relative performance improvement.

Data and models. To validate the effectiveness of HBO, we conduct extensive experiments using three model backbones: Llama-3.1-8B, Qwen2.5-7B, and EuroLLM-9B. These experiments span two setups, covering a total of nine tasks: a Multilingual setting (`MMMLU`, `XCOPA`, `XStoryCloze`, `XNLI`, and `MGSM`) and a Multitask setting (`MMLU`, `MultiFin-EN`, `GSM8K`, and `MedMCQA`).

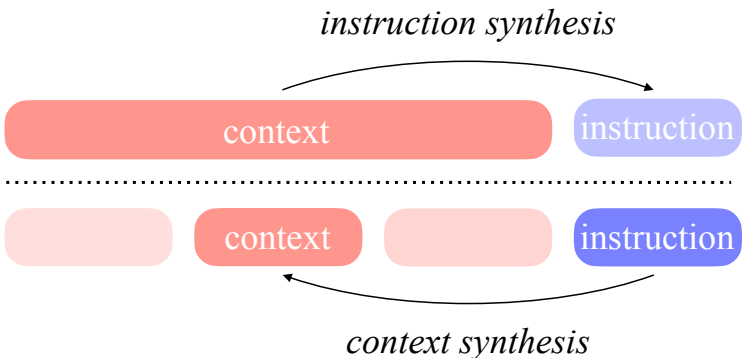

Figure 2: Illustration of two long-context instruction data synthesis frameworks: *instruction synthesis* and *context synthesis* (ours). The light-colored blocks indicate potentially lower-quality components in the synthesized data samples.

Methodology. HBO enables LLMs to autonomously adjust their data allocation both globally and locally based on their current training state. Our method employs a bi-level optimization framework, where the outer problem minimizes the objective function of training the LLM over a mixture of training datasets, and the inner problem adjusts the sampling probabilities both globally (across datasets) and locally (within datasets). To achieve this optimization, we introduce two types of actors: global actor and local actor. The global actor is responsible for balancing data allocation across the subsets of the training data mixture, while the local actor for each individual subset optimizes data usage within the subset. Specifically, we categorize the examples in each subset into four groups based on their difficulty levels. To guide the global and local actors, we define two reward functions based on the training state of the LLM. The global reward is computed as the L_2 norm of the gradients, which reflects the current learning progress of the model. The local reward is defined as the ratio of the perplexities given by the fine-tuned LLM and the original LLM, capturing the relative improvement in model performance on specific groups. By integrating these components, our *HBO* framework effectively optimizes the training process of LLMs.

Findings. Our results demonstrate that *HBO* consistently outperforms existing sampling strategies, achieving substantial improvements in model performance. Additionally, we perform a detailed analysis to investigate the contributions of the global and local actors, the robustness to varying sampling priors, the impact of data volume, and other factors. These findings underscore the effectiveness of *HBO* in addressing data imbalance and heterogeneity during fine-tuning.

3.1.2 Output 4: Generalizing from short to long: Effective data synthesis for long-context instruction tuning

This output is described in Zhu et al. (2025). Long-context modeling for LLMs has been a key area of recent research because many real world use cases require reasoning over longer inputs such as documents. The focus of research into modeling long context has been on how to model position and there has been little investigation into other important aspects of language modeling such as instruction tuning. Long context training examples are challenging and expensive to create and use. In this work, we investigate how to design instruction data for the post-training phase of a long context pre-trained model: how much and what type of context is needed for optimal and efficient post-training.

Data and models. We conduct experiments on real-world tasks from LongBench (Bai et al., 2024b) with two base models `LLaMA2-7B-64K` (Bai et al., 2024a) and `LLaMA3.1-8B-128K` (Dubey et al., 2024a).

Findings. We first presents a pilot study on artificial needle-in-a-haystack tests which allows for rigorous control of different aspects of the instruction data. This study yields three key findings: (1) instruction quality plays a crucial role in model performance; (2) models instruction-tuned on short contexts can generalize to much longer ones; (3) training with evidence embedded in distracting content helps models develop robust information extraction abilities. We leverage these findings to design a novel instruction data synthesis approach called "context synthesis" (see Figure 2) and test it on naturally occurring tasks. Experimental results demonstrate that our context synthesis approach significantly outperforms the instruction synthesis methods and comes close to the performance of fine-tuning with oracle human-annotated long-context instruction data.

3.1.3 Output 5: Demystifying multilingual chain-of-thought in process reward modeling

This output is described in Wang et al. (2025b). In order to fine-tune LLMs for reasoning, we can employ a process reward model (PRM) to assess the reasoning chain produced by the LLM. There are datasets that can be used to train PRMs for English, but for most other languages there are no datasets available, making it hard to create strong multilingual reasoning models. In this paper, we compare strategies for creating multilingual PRMs. We compare the performance of a PRM trained only on the English data (crosslingual), with PRMs trained on translated data (monolingual and multilingual).

Data and models. As process reward training data we use PRM800K (Lightman et al., 2024) and Math-Shepherd (Wang et al., 2024), translating into other languages using NLLB (NLLB team et al., 2022). We test our models on MGSM (Shi et al., 2022) and MATH500 (Wang et al., 2024). We train our multilingual PRM (***verifier***) based on the Qwen2.5-7B (Yang et al., 2024), and use three diverse LLMs as the ***generator***: *MetaMath-Mistral-7B* (Yu et al., 2023), *Llama-3.1-8B-math* (fine-tuned with the MetaMath dataset (Dubey et al., 2024b)), and *DeepSeekMath-7B-Instruct* (Shao et al., 2024).

Methodology. The PRM is trained to evaluate reasoning chains of an LLM using a binary loss across all steps. In order to evaluate a PRM, we sample reasoning chains from the LLM, score them with the PRM, and then check whether the highest scoring chain achieves a correct answer. To provide an extrinsic evaluation, we use the PRM in a reinforcement learning setup in order to fine-tune the LLM. We can then assess to what extent the performance of the LLM on reasoning tasks has improved.

Findings. Our main findings are the following:

- **Multilingual PRM consistently outperforms monolingual and cross-lingual PRMs across all three LLMs.**

- **Multilingual PRM is sensitive to both the number of languages and the amount of English training data.** Our experiment shows that training an optimal multilingual PRM requires careful consideration of how many languages to include and how much English data to use.
- **Multilingual PRM produces fewer errors in the early steps.**
- **Multilingual PRM can benefit more from more candidate responses and trainable parameters.** Our analysis demonstrates that multilingual becomes more advantageous with a larger number of candidate responses and when more trainable parameters are introduced.

3.2 Improving Machine Translation Training:

3.2.1 Output 6: Exploring very low-resource translation with LLMs: The University of Edinburgh's submission to AmericasNLP 2024 translation task

This output is described in Iyer et al. (2024). It describes the University of Edinburgh's submission to the AmericasNLP 2024 shared task on the translation of Spanish into 11 indigenous American languages. We explore the ability of multilingual LLMs to model low-resource languages by continued pre-training with LoRA, and conduct instruction fine-tuning using a variety of datasets, demonstrating that this improves LLM performance. Furthermore, we demonstrate the efficacy of checkpoint averaging alongside decoding techniques like beam search and sampling, resulting in further improvements. We participate in all 11 translation directions.

Data and Models. We build multilingual LLMs for these indigenous American languages by fine-tuning Llama-2 7B (Touvron et al., 2023b), Mistral 7B (Jiang et al., 2023) and MaLA-500 (Lin et al., 2024). We curate data from MADLAD-400, Glot500, Wikipedia, Helsinki'23 datasets from the AmericasNLP 2023 Shared Task.

Methodology. To adapt LLMs for the task of translating indigenous American languages, we follow the 2-stage training paradigm: Stage 1 Continued Pre-training with LoRA, and Stage 2 Instruction Tuning.

Findings. We conduct instruction tuning using a variety of tasks and language pairs, and show this contributes to performance improvements in MT. We also demonstrat how familiar techniques such as checkpoint averaging, beam search, and sampling help boost LLM performance for low-resourced translation.

3.2.2 Output 7: When Does Monolingual Data Help Multilingual Translation: The Role of Domain and Model Scale

This output is described in Baziotis et al. (2024). In this work we investigated the ways in which back-translation (BT) and denoising auto-encoding (DAE) pre-training could be used in encoder-decoder models for neural machine translation (MT). Both of of these are methods for using monolingual data in MT: the former uses another MT system to create synthetic data whilst the second

adds a pre-training step to train an a base models to remove artificially inserted noise from monolingual data. Several previous works had applied these methods to massively multilingual MT, but the experimental results were incomplete, and appeared contradictory at times.

Data and models. All our experiments were based on the encoder-decoder transformer model (Vaswani et al., 2017), as it was standard for MT at that time. We compared to different denoising DAE methods, mBART (Liu et al., 2020) and MASS (Song et al., 2019), which differ in the types of noise that they introduce, and the pre-training objective. We used the ML50 dataset for training (Tang et al., 2021) and for evaluation, we used FLORES-200 (NLLB team et al., 2022), NTREX-128 (Federmann et al., 2022) and TICO-19 (Anastasopoulos et al., 2020).

Methodology. In our experiments we compared the use of monolingual data via BT and DAE, as well as using only parallel data for training. We controlled both the domain of the monolingual data (with respect to the test data) and the scale of the models.

Findings. We show that BT and DAE are both sensitive to domain mismatches between the monolingual and test data, particularly on small scales. BT is best in most settings. Also, prior works have overestimated the effectiveness of DAE, and when comparing the two methods, MASS out-performs mBART. We also showed that model capacity is key for the effectiveness of both methods, especially DAE. When the scale is small, DAE can even harm multilingual MT, but it quickly improves with scale, and at larger scales it becomes competitive with BT.

3.3 Overlapping Contributions

UTTER@IWSLT The UTTER consortium participated to the IWSLT 2025 challenge on speech LLMs with two submissions. The *IWSLT Instruction-following Speech Processing Track* focused on the leveraging of LLMs and speech foundation models (SFM) to build solutions capable to perform multilingual tasks from English speech input and textual multilingual instructions (Abdulmumin et al., 2025). The consortium submitted two systems to the *short track*, where the tasks proposed were automatic speech recognition (ASR), speech translation (ST) and multilingual spoken question answering (SQA). The target languages for ST and multilingual SQA were Chinese, Italian and German. The submission from IT (Attanasio et al., 2025) focused on developing a compact solution (3B model) capable of competitive ASR. The submission from NAV (Lee et al., 2025) focused on multimodal alignment and a three-stage training process to reach optimal performance (best system in the short track). These works are described in more details in D 4.2, Task 4.1: *Adaptable, multimodal generation and translation.*

Method for cross-lingual open-ended generation instruction-tuning This contribution (Iyer et al., 2025) is presented in D 2.2 Task 2.1: *Identifying, collecting and evaluating monolingual and bilingual written and spoken language resources*, since it also includes two novel datasets `XL-AlpacaEval` and `XL-Instruct`.

4 Impact

Across RP1 and RP2, the consortium has made significant contributions related to pre-trained text and speech models and efficient training, resulting in 11 publications. Beyond academic dissemination, our models have achieved strong community adoption and visibility, as demonstrated by their widespread use, citations, and downloads.

On HuggingFace, the most recent iteration of Tower+ (Rei et al., 2025) has already surpassed 11,000 downloads as of August 20, 2025. The EuroLLM family of models (Martins et al., 2024, 2025) has reached more than 485,000 downloads in the same period, highlighting its role as a cornerstone resource for the European and international NLP community. Likewise, our speech model mHuBERT-147 has been downloaded over 330,000 times as of September 20, 2025, confirming its strong impact in advancing multilingual and low-resource speech processing.

These figures demonstrate not only the scientific relevance of our work but also its broad adoption by the research and developer communities, positioning our consortium as a key driver of innovation in Europe's AI ecosystem.

5 Conclusion

WP3 has made significant contributions across both RP1 and RP2, delivering state-of-the-art pre-trained models for speech and text, as well as methodological advances in their training and adaptation. Through the release of resources such as Tower, mHuBERT-147, EuroLLM, and SPIRE, alongside targeted research on efficiency, multilinguality, and robustness, WP3 has strengthened the foundations for the entire project. These outcomes underscore the UTTER consortium's central role in advancing open, high-quality resources for the research community.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Ashwin, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połeć, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. Findings of the iwslt 2025 evaluation campaign. In Elizabeth Salesky, Marcello Federico, and Antonios Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austria (in-person and online), 2025. Association for Computational Linguistics. To appear.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José GC Souza, and André FT Martins. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*, 2024.

Kshitij Ambilduke, Ben Peters, Sonal Sannigrahi, Anil Keshwani, Tsz Kin Lam, Bruno Martins, Marcely Zanon Boito, and André FT Martins. From tower to spire: Adding the speech modality to a text-only llm. *arXiv preprint arXiv:2503.10620*, 2025.

Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Franscisco Guzmán, Junjie Hu, Macduff Hughes, Philipp Koehn, Rosie Lazar, Will Lewis, Graham Neubig, Mengmeng Niu, Alp Öktem, Eric Paquin, Grace Tang, and Sylwia Tur. TICO-19: the translation initiative for COvid-19. In Karin Verspoor, Kevin Bretonnel Cohen, Michael Conway, Berry de Bruijn, Mark Dredze, Rada Mihalcea, and Byron Wallace, editors, *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.nlpcovid19-2.5. URL https://aclanthology.org/2020.nlpcovid19-2.5/.

Giuseppe Attanasio, Sonal Sannigrahi, Ben Peters, and André Filipe Torres Martins. Instituto de telecomunicações at IWSLT 2025: Aligning small-scale speech and language models for speech-to-text learning. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 347–353, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.36. URL https://aclanthology.org/2025.iwslt-1.36/.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. LongAlign: A recipe for long context alignment of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024a. URL https://aclanthology.org/2024.findings-emnlp.74.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilin-

gual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024b. URL https://aclanthology.org/2024.acl-long.172.

Christos Baziotis, Biao Zhang, Alexandra Birch, and Barry Haddow. When does monolingual data help multilingual translation: The role of domain and model scale. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6297–6324, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.349. URL https://aclanthology.org/2024.naacl-long.349/.

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*, 2024a. URL https://arxiv.org/pdf/2407.21783.

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*, 2024b.

Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT evaluation of 128 languages. In Kabir Ahuja, Antonios Anastasopoulos, Barun Patra, Graham Neubig, Monojit Choudhury, Sandipan Dandapat, Sunayana Sitaram, and Vishrav Chaudhary, editors, *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online, November 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.sumeval-1.4. URL https://aclanthology.org/2022.sumeval-1.4/.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021.

Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. Exploring very low-resource translation with LLMs: The University of Edinburgh's submission to AmericasNLP 2024 translation task. In Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense, editors, *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 209–220, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.americasnlp-1.25. URL https://aclanthology.org/2024.americasnlp-1.25/.

Vivek Iyer, Ricardo Rei, Pinzhen Chen, and Alexandra Birch. XL-Instruct: Synthetic Data for Cross-Lingual Open-Ended Generation. In *Findings of the 2025 Conference on Empirical Methods in Natural Language Processing*, China, November 2025. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Beomseok Lee, Marcely Zanon Boito, Laurent Besacier, and Ioan Calapodescu. NAVER LABS Europe submission to the instruction-following track. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 186–200, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.17. URL https://aclanthology.org/2025.iwslt-1.17/.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*, 2024.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual Denoising Pre-training for Neural Machine Translation. *arXiv:2001.08210 [cs]*, January 2020. URL http://arxiv.org/abs/2001.08210. arXiv: 2001.08210.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. EuroLLM: Multilingual Language Models for Europe. In *Proceedings of the EuroHPC User Day*, 2024. URL https://arxiv.org/abs/2409.16235.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. EuroLLM-9B: Technical Report, 2025. URL https://arxiv.org/abs/2506.04079.

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, July 2022. URL http://arxiv.org/abs/2207.04672. Number: arXiv:2207.04672 arXiv:2207.04672 [cs].

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. Tower+: Bridging generality and translation specialization in multilingual llms, 2025. URL https://arxiv.org/abs/2506.17080.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *CoRR*, abs/2402.03300, 2024. doi: 10.48550/ARXIV.2402.03300. URL https://doi.org/10.48550/arXiv.2402.03300.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. Language models are multilingual chain-of-thought reasoners, 2022.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: Masked Sequence to Sequence Pre-training for Language Generation. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5926–5936. PMLR, May 2019. URL https://proceedings.mlr.press/v97/song19d.html. ISSN: 2640-3498.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation from denoising pre-training. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.304. URL https://aclanthology.org/2021.findings-acl.304/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023a. URL https://arxiv.org/abs/2307.09288.

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 arXiv:2307.09288*, 2023b.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems*, volume abs/1706.03762, 2017. URL http://arxiv.org/abs/1706.03762.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-shepherd: Verify and reinforce llms step-by-step without human annotations. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 9426–9439. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.510. URL https://doi.org/10.18653/v1/2024.acl-long.510.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. HBO: Hierarchical Balancing Optimization for Fine-Tuning Large Language Models, 2025a. URL https://arxiv.org/abs/2505.12300.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. Demystifying multilingual reasoning in process reward modeling. In *Findings of EMNLP*, 2025b.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, Keming Lu, Mingfeng Xue, Runji Lin, Tianyu Liu, Xingzhang Ren, and Zhenru Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement. *CoRR*, abs/2409.12122, 2024. doi: 10.48550/ARXIV.2409.12122. URL https://doi.org/10.48550/arXiv.2409.12122.

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198, 2021. doi: 10.21437/Interspeech.2021-1775.

Longhui Yu, Weisen Jiang, Han Shi, Jincheng Yu, Zhengying Liu, Yu Zhang, James T Kwok, Zhenguo Li, Adrian Weller, and Weiyang Liu. Metamath: Bootstrap your own mathematical questions for large language models. *arXiv preprint arXiv:2309.12284*, 2023.

Wenhao Zhu, Pinzhen Chen, Hanxu Hu, Shujian Huang, Fei Yuan, Jiajun Chen, and Alexandra Birch. Generalizing from short to long: Effective data synthesis for long-context instruction tuning. *CoRR*, abs/2502.15592, February 2025. URL https://doi.org/10.48550/arXiv.2502.15592.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D3.2 Final Report on Multimodal, Multilingual Pre-trained XR Models