UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D2.2/D14 – Final Report on Data and Resources

Nature	Report	**Work Package**	WP2
Due Date	30/09/2025	**Submission Date**	dd/mm/2025
Main authors	Marcely Zanon Boito (NAV)		
Co-authors	Barry Haddow (UEDIN), Tsz Kin Lam (UEDIN), André Martins (IT), Alexandra Birch (UEDIN)		
Reviewers	Wilker Aziz (UVA)		
Keywords	data filters, data for large language models		
Version Control			
v0.1	**Status**	Draft	10/09/2025
v1.0	**Status**	Final	29/09/2025

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 WP2's progress during the second half of the UTTER project. The goal of WP2 is to gather, annotate, and release language data resources that support UTTER's overarching objectives. While the first report covered our efforts across all three tasks in this work package, the present report focuses specifically on Task 2.1. We describe 11 outputs, including datasets for speech and text machine translation, filtering approaches for synthetic data, and resources for training multilingual LLMs. We believe WP2 has successfully achieved its objectives, publishing and distributing a substantial volume of text and speech data in many languages. These resources not only enable progress within UTTER's other work packages but also stimulate broader research in diverse areas of speech and language technology.

1 Introduction

WP2 Proposal

"Gather, annotate and collect language data to achieve UTTER's objectives."

Work Presented in the First Deliverable

In the initial WP2 deliverable (D 2.1), we presented the majority of the resources gathered during the UTTER project, as most of the person-hours for this WP were invested in the project's first half. For speech, we gathered a massive amount of speech in 147 languages for self-supervised training. We also released a multilingual speech datasets for spoken language understanding in 12 languages (Speech-MASSIVE).

For text we released two datasets related to WP7, one for each use-case. We also released datasets for summarization (PMIndiaSum), natural language understanding (MULTI3NLU++), machine translation test data (WMT 2023), translation accuracy challenge data (ACES and SpanACES), and language identification (OpenLID).

Work Presented in this Deliverable

For this final report on data and resources, we do not present contributions for T 2.2 (Data for dialogue) and T 2.3 (Data for minuting and summarization), as that output was already presented in the first half of our project (see D 2.1). For T 2.1, this report presents work related to machine translation (MT) topics (5 outputs), synthetic data for LLM training (3 outputs), data filtering approaches (2 outputs) and data for NLP studies (1 output).

Manuscripts:

- **2 journal papers:** Moghe et al. (2025), Cilleruelo et al. (2025)
- **3 conference papers:** Pal et al. (2024), Waldendorf et al. (2025), Iyer et al. (2025)
- **5 workshop papers:** Zhang et al. (2024), WMT (Kocmi et al., 2024, 2025), IWSLT (Ahmad et al., 2024; Abdulmumin et al., 2025)
- **2 arXiV papers:** Martins et al. (2025), Ji et al. (2025)

Code and data:

- **WMT 2024 and 2025 test data:** WMT24 and WMT25
- **Document-level MT dataset:** HuggingFace
- **ChineseMenuCSI dataset:** GitHub
- **ACES dataset:** HuggingFace
- **EuroLLM EuroBlocks-synthetic dataset:** HuggingFace
- **EuroLLM EuroFilter:** HuggingFace
- **EMMA-500 resource:** HuggingFace
- **XL-AlpacaEval dataset:** HuggingFace
- **XL-Instruct dataset:** HuggingFace
- **MGEN dataset:** Website
- **MFDS dataset:** HuggingFace

2 Task 2.1: Identifying, collecting and evaluating monolingual and bilingual written and spoken language resources (NAV*, UEDIN)

Proposal highlights

This task focuses on the gathering of data and resources that enable the consortium to train and evaluate monolingual and multilingual models that cover text and/or speech.

Summary of completed work

In this second half of the UTTER project we produced eleven different contributions related to data and resources. We focus mostly on different topics of machine translation (MT), and data and filters for LLM tuning. The contributions are the following:

- **Machine translation (MT) data:**
  - Challenge test sets for MT competitions (Section 2.1.1);
  - Document-level data for MT (Section 2.1.2);
  - Data for Literary MT (Section 2.1.3);
  - Restaurant menu MT dataset (Section 2.1.4);
  - MT meta-evaluation data (Section 2.1.5).
- **Data for LLM training:**
  - EuroLLM synthetic data (Section 2.2.1);
  - Multilingual data covering 939 languages (Section 2.2.2);
  - Synthetic data generation for cross-lingual open-ended generation (Section 2.3.3).
- **Data Filtering approaches:**
  - EuroLLM quality filters (Section 2.3.1);
  - Multilingual filtering for data quality and domain (Section 2.3.2).
- **Data for NLP studies:**
  - Corpus generics (Section 2.4.1).

2.1 MT data

In this section we present the contributions related to speech and text MT.

2.1.1 Output 1: Challenge data for translation competitions

The consortium have been collaborating with the most well-known competitions for machine translation (WMT) and speech translation (IWSLT). These outputs are described in Kocmi et al. (2024, 2025) for WMT; and in Ahmad et al. (2024); Abdulmumin et al. (2025) for IWSLT.

WMT 2024 and 2025 challenge data The WMT General MT shared task is an annual effort organized by a large team of people from diverse institutions. It focuses on creating new challenge sets for MT systems, collecting translations for a broad selection of systems (some submitted by the authors, others gathered from APIs), and then performing extensive human and automatic evaluation of the systems. All training, test and evaluation data is published on the WMT website[1] and the results are described in the WMT conference papers (Kocmi et al., 2024, 2025). UTTER participated in the selection of the data for the test sets, which for 2024 and 2025 consisted of news, literary text, social media text and speech. For 2025 we focused on selecting more challenging source texts by initially choosing likely sources of complex language (e.g. news analysis instead of news) and then applying estimates of MT difficulty (Proietti et al., 2025) for further filtering.

IWSLT 2024 and 2025 challenge data: accented speech translation The *offline speech translation task* at IWSLT is the longest-standing task at the conference. It provides a stable evaluation framework for tracking technological advancements in spoken language translation, and it offers a good estimate to the current upper limit of our speech technologies. In 2025, an evaluation server and leaderboards[2] have been established to track the development of speech translation. The offline track is organized by a team of speech translation experts from the industry and the academic institutions, such as Zoom, FBK and KIT. UTTER participated as a task organizer and contributed with a valuable test set that focuses on the German translation of accented English speech conversations. Since 2024, this test set has been shown to be the top-2 most challenging set in both automatic and human evaluations (Ahmad et al., 2024; Abdulmumin et al., 2025). In 2025, our test set has also been used for evaluating simultaneous speech translation systems. This test set is not publicly available, in order to avoid model contamination.

2.1.2 Output 2: Document-Level MT with Large-Scale Public Parallel Corpora

This output is described in (Pal et al., 2024). Previous document-level MT studies have often relied on small, domain-specific, or proprietary datasets, which limited the reproducibility and scalability of research. In contrast, the dataset released in our paper (Pal et al., 2024) provides a large, multilingual, and extensible foundation for both training context-aware systems and benchmarking them on phenomena that sentence-level resources cannot capture.

Our contributions center around the release of a large-scale document-level parallel corpus derived from ParaCrawl, designed to overcome the long-standing scarcity of open, document-aligned resources for MT. The corpus covers five language pairs with English: Czech, Polish, German, French, and Russian. Each corpus is built by reconstructing document boundaries within ParaCrawl through URL-based alignment, yielding document-level structures that preserve natural discourse organization rather than isolated sentence pairs.

The resulting corpora are large in scale, comparable in size to widely used sentence-level resources, while uniquely preserving cross-sentence context. This makes them particularly suitable for training and evaluating context-aware NMT systems. Experiments conducted with the released data confirm its utility: models trained on the document-level corpora achieve consistent improvements in BLEU and COMET scores over sentence-level baselines, and qualitative analyses highlight

[1] See http://www.statmt.org/wmt24 and http://www.statmt.org/wmt25
[2] https://iwslt2025.speechm.cloud.cyfronet.pl/

gains in discourse-sensitive phenomena such as pronoun resolution, lexical cohesion, and consistency of named entities across sentences.

2.1.3 Output 3: *Liaozhai* through the Looking-Glass: On Explicitation via Genettean Paratexts in Literary MT

This output is described in (Shen et al., 2025). The faithful transfer of contextually-embedded meaning remains a central challenge to MT, especially in regards to culture-bound terms, expressions or concepts deeply rooted in specific languages or cultures, resisting direct linguistic transfer. Existing computational approaches to explicitating such terms have focused exclusively on in-text solutions, overlooking paratextual apparatus (e.g., footnotes and endnotes) employed by professional translators. This paper formalizes Genette (1997)'s theory of paratexts from literary and translation studies to introduce the task of *paratextual explicitation for MT*.

A dataset of 560 expert-aligned paratexts is compiled from four English translations across 150 stories of the classical Chinese short story collection *Liaozhai*. The source text, all English paratexts, and classical Chinese annotations corresponding to each paratext are released, notably avoiding release of the English translations of the stories themselves. LLMs with and without reasoning traces are then evaluated on choice and content of explicitation across various prompting and retrieval strategies. Human evaluation shows that LLM-generated paratexts significantly improve target audience comprehension, though with considerably less effectiveness than translator-authored ones.

Beyond model performance, analysis reveals that even professional translators exhibit consistent variation in explicitation choice, suggesting that paratextual mediation is inherently open-ended rather than prescriptive. These findings demonstrate the potential of paratextual explicitation for cultural mediation and advancing MT beyond literal equivalence, with promising extensions to monolingual explanation and personalized adaptation. This work was accepted to EMNLP 2025 and we will soon make the preprint and data available to the community.

2.1.4 Output 4: Cultural Adaptation of Menus: A Fine-Grained Approach

This output is described in Zhang et al. (2024). We introduce `ChineseMenuCSI`, a carefully curated bilingual corpus of 4,275 human-verified Chinese–English restaurant menu entries extracted from UK-based Chinese restaurants via a Selenium-based web crawler. Each entry is annotated with classification into CSI (Culture-Specific Item) or Non-CSI, achieving a Cohen's kappa agreement of 0.91 across annotators. The dataset enables fine-grained analysis through a CSI taxonomy that distinguishes between concrete, creative, and abstract CSIs, facilitating nuanced exploration of translation phenomena that go beyond entity-level adaptation.

Building upon this foundation, a focused test subset of 480 entries is selected through random sampling – 120 items each from concrete, creative, abstract, and Non-CSI categories – and annotated by a broader group of native Chinese speakers proficient in English. This subset includes span annotations, where specific segments within dish names are marked to indicate the CSI, yielding substantial inter-annotator agreement (Fleiss' kappa: 0.63 for CSI categorization; 0.70 for span identification). This dual-level annotation strategy supports precise evaluations of models' abilities to detect and localize nuanced cultural content in translation tasks.

Our paper underlies the development of novel CSI identification techniques based on translation

theory and linguistically inspired criteria, such as Round-Trip Translation, Cultural Uniqueness, Historical Significance. Prompts incorporating recipe-based knowledge and translation theory (e.g. cultural, functional, and descriptive equivalence) furthermore match or outperform baseline prompts in translating CSIs without relying on parallel corpora or knowledge graphs by up to 7 points on COMET.

2.1.5 Output 5: MT meta evaluation through translation accuracy challenge sets

This output is described in Moghe et al. (2025). We introduce ACES and its span-annotated extension, SPAN-ACES, both designed to advance fine-grained evaluation of MT systems and the metrics used to assess them. ACES is a large-scale challenge set comprising 36,476 examples distributed across 146 language pairs and organized into 68 distinct translation accuracy phenomena. These phenomena cover a wide spectrum of error types, ranging from simple token-level manipulations such as word addition, omission, or reordering, to more linguistically complex cases including mistranslation of named entities, agreement errors, discourse-level inconsistencies, hallucinations, untranslated material, wrong-language output, and errors requiring real-world knowledge.

The dataset was constructed through a multi-pronged approach: synthetic adversarial perturbations were automatically generated to capture specific error patterns; contrastive pairs were extracted and repurposed from existing multilingual datasets such as FLORES-101, PAWS-X, XNLI, WinoMT, and MuCoW; and human annotations supplemented areas where automated methods were insufficient. To ensure consistency and interpretability, all examples were systematically organized under the Multidimensional Quality Metrics (MQM) framework, which was extended with novel categories to better capture error types not addressed by existing ontologies, such as real-world knowledge violations and wrong-language outputs.

SPAN-ACES extends this foundation by introducing explicit error-span annotations that pinpoint the exact location of the error within each incorrect translation. These annotations are expressed in an MQM-inspired format, embedding error spans within the text through tagged markers. The annotation process was primarily automated, with rule –based methods applied successfully to 34,514 examples. For phenomena that resisted reliable automation—such as pragmatic errors, subtle mistranslations, or cases requiring fine semantic interpretation – manual annotation was performed by trained annotators, yielding an additional 2,006 high – quality, human-verified span labels. Annotation quality was maintained through a piloting phase to refine guidelines, systematic agreement checks, and targeted validation of span boundaries, achieving strong agreement rates across annotators.

Together, ACES and SPAN-ACES contribute a comprehensive, multilingual, and phenomenon-rich resource that supports both holistic and fine-grained analyses of MT evaluation metrics. Their breadth of language coverage, diversity of error phenomena, and inclusion of error-span annotations distinguish them from prior datasets, which have typically been limited in scope, error taxonomy, or annotation depth. Beyond serving as static benchmarks, these datasets open a path for designing and testing metrics that move beyond single-number scores, toward evaluations that are source-aware, error-specific, and interpretable: key properties for trustworthy assessment of modern translation systems.

2.2 Data for LLM training:

In this section we present our contributions related to LLM training with synthetic data.

2.2.1 Output 6: EuroLLM resources: `EuroBlocks-Synthetic`

This output is described in Martins et al. (2025). As part of the EuroLLM initiative, we released `EuroLLM-9B`, an LLM trained from scratch covering all 24 official European Union languages and 11 additional languages. We published a detailed technical report where we provide a comprehensive overview of `EuroLLM-9B`'s development, including tokenizer design, architectural specifications, data filtering, and training procedures. Along with the model, we released `EuroFilter`, an AI-based multilingual filter, as well as `EuroBlocks-Synthetic`, a novel synthetic dataset for post-training that enhances language coverage for European languages. All these resources are available in HuggingFace as part of the EuroLLM Collection. Our `EuroFilter` is presented in Section 2.3.1. We now focus on `EuroBlocks-Synthetic`.

To enable `EuroLLM-9B` to follow natural language instructions, we constructed `EuroBlocks`, a multilingual dataset that combines both human-written and synthetic instruction-following conversations. The human-written portion draws from several publicly available sources, including Magpie (Xu et al., 2024)[3], Aya (Singh et al., 2024), lmsys-chat-1m (Zheng et al., 2023), OpenMath-2 (Toshniwal et al., 2024), and smol-magpie-ultra (Allal et al., 2024).

To ensure data quality, we applied filtering based on complexity and readability scores. Prompts from OpenMath-2 and smol-magpie-ultra falling below a score of 4 on either dimension were removed, while low-scoring prompts from Magpie, Aya, and lmsys-chat-1m were downsampled rather than discarded entirely. We further filtered conversations using ArmoRM-v0.1 (Wang et al., 2024), removing responses with scores below 0.08 – except in cases where low readability in the prompt skewed the score.[4]

To broaden language coverage and support less-represented languages, we generated synthetic data. This involved prompting an LLM – either Llama 3 (AI@Meta, 2024) or an earlier EuroLLM checkpoint – with a monolingual document, a target language, and a category, asking it to create an instruction relevant to the document. The same model then produced an answer in a RAG-style setup, using both the document and the generated instruction. Additionally, we synthesized further supervised fine-tuning (SFT) data by translating prompt–answer pairs,[5] and incorporating high-quality examples from multilingual translation benchmarks such as NTREX-128 (Federmann et al., 2022), FLORES-200-DEV (Team et al., 2022), WMT-21 (Farhad et al., 2021), and WMT-22 (Kocmi et al., 2022), leaving WMT-23 and later editions for evaluation purposes.

Altogether, we collected approximately 4.5 million instructions. After applying filtering and deduplication, the final `EuroBlocks` dataset contains 1.95 million high-quality examples. To support further research and development of European-centric LLMs, we publicly release the synthetic portion of the dataset: `utter-project/EuroBlocks-SFT-Synthetic-1124`.

[3] Magpie datasets are generated with several models with different licenses. We used only the data from Qwen 2, Llama 3 and Phi 3 models which have commercially permissive licenses.

[4] ArmoRM-v0.1 was found to be robust on multilingual data, with a Pearson correlation above 0.7 between English and translated examples. While reward models yield uncalibrated scores, the 0.08 threshold provided a good balance between quality and data retention.

[5] Translations were produced using Tower v2 (Rei et al., 2024) or earlier EuroLLM-9B models.

2.2.2 Output 7: EMMA-500: Enhancing Massively Multilingual Adaptation of LLMs

This output is described in Ji et al. (2025). This collaboration was initiated during the sponsoring of one of the FSTP projects funded during the first cycle. The idea of the project was to create a highly multilingual corpus and use this to train a multilingual LLM. The *MaLa* corpus[6] was created by curating, cleaning and deduplicating data from diverse sources. This resulted in a corpus of 74B whitespace delimited tokens, covering 939 languages. We used 546 languages for training, finetuning Llama-2 7B to produce the EMMA model. Evaluation of the performance across such a wide range of languages is of course difficult, but using existing evaluation benchmarks we are able to show improved performance across 9 tasks, for the languages that they cover. A perplexity based evaluation shows improvement across all languages.

2.3 Data filtering approaches:

2.3.1 Output 8: EuroLLM resources: `EuroFilter`

This output is described in Martins et al. (2025). To filter high-quality English data, we use FineWeb-Edu (Lozhkov et al., 2024). For the other languages, we reuse the FineWeb-Edu annotation; however, we translate all data using Tower v2-supported languages (Rei et al., 2024). This is done to create multilingual texts paired with educational scores. Then we train a multilingual classifier on top of mDeBERTa (He et al., 2023) which we use to annotate the rest of the languages. Our filter is publicly available at `utter-project/EuroFilter-v1`. More details about this process are provided in Martins et al. (2025).

2.3.2 Output 9: Multilingual Data Filtering using Synthetic Data from LLMs

This output is described in Waldendorf et al. (2025). We approach multilingual data filtering, focusing on the MT task. We test a recently proposed method that consists of first labeling a set of texts using an LLM, then training a smaller model on that set of texts, and finally using this model to label the full training corpus. This method has been previously only applied to English, but we test it in multilingual settings.

We apply our method to the filtering of parallel data in two different scenarios: filtering for quality and filtering for domain. In the former scenario we show that our method has comparable results to COMET-KIWI, and in the latter case we show that filtering separately either the English side or the non-English side of the data results in similar performance. As for the architectural choice of the smaller model required for training on the LLM-labeled data, our work illustrates that a very efficient n-gram model can be effective. This resource available to the research community.[7]

2.3.3 Output 10: XL-Instruct: Synthetic Data for Cross-Lingual Open-Ended Generation

This output is described in Iyer et al. (2025). The task of cross-lingual open-ended generation – producing responses in a language different from that of the user's query – is an important yet

[6] https://huggingface.co/collections/MaLA-LM/mala-corpus-66e05127641a51de34d39529
[7] https://huggingface.co/datasets/waretupper/mdfs

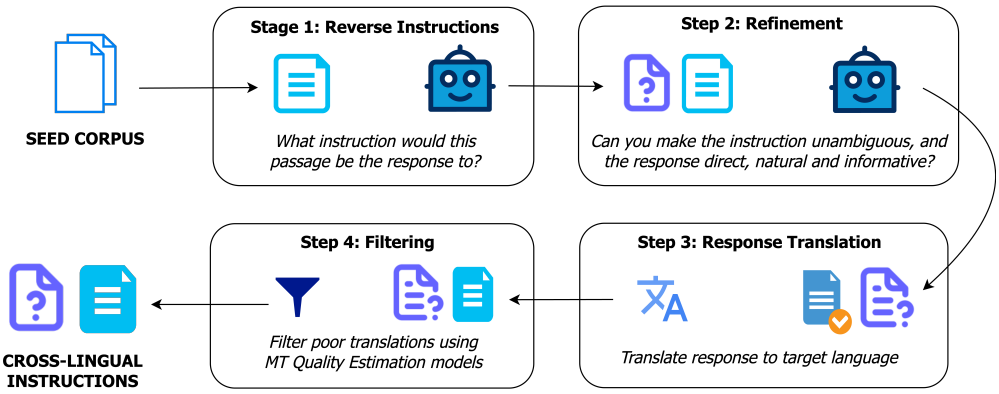

Figure 1: The `XL-Instruct` pipeline: 1) instruction generation from seed English data; 2) data refinement; 3) response translation into non-English; 4) data filtering.

understudied problem. We introduce `XL-AlpacaEval`, a new benchmark for evaluating cross-lingual generation capabilities of LLM, and propose `XL-Instruct`, a high-quality synthetic data generation technique. Fine-tuning with just 8K `XL-Instruct`-generated instructions significantly improves model performance, increasing the win rate against GPT-4o-Mini from 7.4% to 21.5%, and improving on several fine-grained quality metrics. Additionally, base LLMs fine-tuned on `XL-Instruct` yield strong zero-shot improvements in both English-only and multilingual generation tasks. Given these consistent gains, we strongly recommend incorporating `XL-Instruct` in the post-training pipeline of future multilingual LLMs. To facilitate further research, we publicly release the `XL-Instruct` and `XL-AlpacaEval` datasets, which constitute two of the scarce cross-lingual ones currently available.

2.4 Data for NLP studies:

In this section we data related to general NLP topics.

2.4.1 Output 11: MGEN: Millions of Naturally Occurring Generics in Context

This output is described in Cilleruelo et al. (2025). We present a new corpus of generics, complete with their document context. Generics are sentences that express generalizations without making use of explicit quantifiers. Examples of generics are *ravens are black* or *ticks carry lyme disease*. Generics are an interesting object of study in computational linguistics because they still lack an agreed account of their semantics. In many examples they can be paraphrased by quantified statements but the actual quantifier varies greatly, and generics admit exceptions.

The MGEN[8] corpus described in this paper was extracted using a custom filtering pipeline applied to a cleaned web-scale dataset (Zyda).[9] The pipeline consists of a syntactic analyzer and a classifier trained on hand-labeled data. Our corpus contains 4.1 million examples of generics and other quantified sentences, together with the full document context. The motivation for including the context is to enable the testing of computational theories of generics, some of which suggests that generics need to be interpreted with respect to their context.

[8] https://gustavocilleruelo.com/mgen
[9] https://huggingface.co/datasets/Zyphra/Zyda

3 Impact

For RP1, we released 10 different speech and text resources spanning all three WP tasks. For RP2, we contributed 11 resources specifically to Task 2.1. In total, these efforts resulted in 22 publications on language resources, most of which are already available, or will soon be available, to the research community. Notably, Speech-MASSIVE (Lee et al., 2024), the multilingual spoken language understanding dataset introduced in RP1, has gained significant traction, with over 19K downloads on HuggingFace as of September 2025.

4 Conclusion

In this report we presented our contributions related to data and resources produced by the UTTER consortium during the second half of the project. Our contributions mainly focus on (i) diverse data for evaluating speech and text machine translation models; (ii) data filtering approaches for better selection of synthetic data; and (iii) data for training multilingual LLMs. In total, we present 11 contributions, most of them already or soon-to-be available on sharing platforms, and 12 publications. We believe the resources produced by UTTER during the project help foster innovation and collaboration across the speech and language technology community.

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Fortuné Kponou, Mateusz Krubiński, 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, Ashwin Sankar, 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 Antonis Anastasopoulos, editors, *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.44. URL https://aclanthology.org/2025.iwslt-1.44/.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.1. URL https://aclanthology.org/2024.iwslt-1.1/.

AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Lewis Tunstall, Agustín Piqueres, Andres Marafioti, Cyril Zakka, Leandro von Werra, and Thomas Wolf. Smollm2 - with great data, comes great performance, 2024.

Gustavo Cilleruelo, Emily Allaway, Barry Haddow, and Alexandra Birch. Mgen: Millions of naturally occurring generics in context. In *Society for Computation in Linguistic*, 2025. doi: 10.7275/scil.3147. URL https://openpublishing.library.umass.edu/scil/article/id/3147/.

Akhbardeh Farhad, Arkhangorodsky Arkady, Biesialska Magdalena, Bojar Ondřej, Chatterjee Rajen, Chaudhary Vishrav, Marta R Costa-jussa, España-Bonet Cristina, Fan Angela, Federmann Christian, et al. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, 2021.

Christian Federmann, Tom Kocmi, and Ying Xin. NTREX-128 – news test references for MT evaluation of 128 languages. In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online, nov 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.sumeval-1.4.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=sE7-XhLxHA.

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.

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, 2025. URL https://arxiv.org/abs/2409.17892.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, et al. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 2022.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.1. URL https://aclanthology.org/2024.wmt-1.1/.

Tom Kocmi, Katia Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Stefano Perrella, Lorenzo Proietti, Parker Riley, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. Findings of the WMT25 general machine translation shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 1–46, Souzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.1. URL https://aclanthology.org/2025.wmt-1.1/.

Beomseok Lee, Ioan Calapodescu, Marco Gaido, Matteo Negri, and Laurent Besacier. Speech-MASSIVE: A Multilingual Speech Dataset for SLU and Beyond. In *Interspeech 2024*, pages 817–821, 2024. doi: 10.21437/Interspeech.2024-957.

Anton Lozhkov, Loubna Ben Allal, Leandro von Werra, and Thomas Wolf. Fineweb-edu, 2024. URL https://huggingface.co/datasets/HuggingFaceFW/fineweb-edu.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, et al. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*, 2025.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. Machine translation meta evaluation through translation accuracy challenge sets. *Computational Linguistics*, 51(1):73–137, March 2025. doi: 10.1162/coli_a_00537. URL https://aclanthology.org/2025.cl-1.4/.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. Document-level machine translation with large-scale public parallel corpora. 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)*, pages 13185–13197, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.712. URL https://aclanthology.org/2024.acl-long.712/.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. Estimating machine translation difficulty, 2025. URL https://arxiv.org/abs/2508.10175.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.12. URL https://aclanthology.org/2024.wmt-1.12/.

Sherrie Shen, Weixuan Wang, and Alexandra Birch. Liaozhai through the Looking-Glass: On Explicitation via Genettean Paratexts in Literary Machine Translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, China, November 2025. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. Aya dataset: An open-access collection for multilingual instruction tuning, 2024.

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. 2022.

Shubham Toshniwal, Wei Du, Ivan Moshkov, Branislav Kisacanin, Alexan Ayrapetyan, and Igor Gitman. Openmathinstruct-2: Accelerating ai for math with massive open-source instruction data. *arXiv preprint arXiv:2410.01560*, 2024.

Jonas Waldendorf, Barry Haddow, Alexandra Birch, and Mateusz Klimaszewski. Multilingual data filtering using synthetic data from large language models. In *Findings of EMNLP*, 2025.

Haoxiang Wang, Wei Xiong, Tengyang Xie, Han Zhao, and Tong Zhang. Interpretable preferences via multi-objective reward modeling and mixture-of-experts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10582–10592, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.620. URL https://aclanthology.org/2024.findings-emnlp.620/.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *ArXiv*, abs/2406.08464, 2024. URL https://api.semanticscholar.org/CorpusID:270391432.

Zhonghe Zhang, Xiaoyu He, Vivek Iyer, and Alexandra Birch. Cultural adaptation of menus: A fine-grained approach. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1258–1271, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.120. URL https://aclanthology.org/2024.wmt-1.120/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric. P Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D2.2/D14 Final Report on Data and Resources