UTTER

Unified Transcription and Translation for Extended Reality (UTTER)

Horizon Europe Research and Innovation Action Number: 101070631

D4.2 – Final Report on adaptable and context-aware models

Nature	Report	**Work Package**	WP4
Due Date	30/09/2025	**Submission Date**	30/09/2025
Main authors	Chrysoula Zerva (IT)		
Co-authors	Marcely Zanon Boito (NAV), Pierre Erbacher (NAV), Ben Peters (IT), André Martins (IT), Marcos Treviso (IT), Vlad Niculae (UvA), Barry Haddow (UEDIN), TszKin Lam (UEDIN), Leonardo Ranaldi (UEDIN), Radina Dobreva (UEDIN)		
Reviewers	Lexi Birch (UEDIN)		
Keywords	adaptation, human-aligment, long-context, speech translation		
Version Control			
v0.1	**Status**	Draft	18/09/2025
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 document WP4's progress throughout the project, with a focus on RP2 (for a detailed view of progress in RP1, see D4.1).

Work Package 4 (WP4) of UTTER set out to advance adaptable, context-aware, and multimodal generation models with a focus on translation and transcription. The effort was structured across three interrelated tasks: Task 4.1 explored adaptation strategies across modalities and domains, Task 4.2 developed context- and discourse-aware methods for translation and multilingual interaction, and Task 4.3 initially targeted simultaneous translation. In RP1, our emphasis was on establishing adaptation techniques and foundational datasets for context and emotion; in RP2, progress shifted toward alignment with user feedback, cross-lingual and multimodal adaptation, long-context modelling, and conversational translation, while Task 4.3 redirected its scope toward integrating speech capabilities in LLMs and full-duplex conversational models. At the end of the report, we also reflect briefly on impact.

Across both reporting periods, WP4 has delivered substantial outputs in the form of peer-reviewed publications, open-source models and datasets, and co-organisation of shared tasks. Progress was steady and proceeded according to plan, with no significant risks encountered. Collectively, the results advance the state of the art in adaptation, contextualisation, and speech translation, while providing the community with resources and methodologies that will support ongoing research beyond UTTER.

1 Introduction

Work Package 4 (WP4) of UTTER focuses on advancing adaptable, context-aware, and multimodal generation models, with a strong emphasis on translation and transcription for spoken and textual dialogue. The objectives of WP4 are to design adaptation strategies that allow large language models (LLMs) to operate effectively across modalities and domains, to incorporate context, discourse, and affective signals into translation and multilingual interaction, and to improve speech and real-time translation capabilities. The work is structured into three tasks: Task 4.1 on adaptable and multimodal generation and translation, Task 4.2 on contextualisation and emotion tracking, and Task 4.3 on simultaneous translation.

In the first reporting period (RP1), as documented in D4.1, emphasis was placed on establishing initial adaptation methodologies for steering LLMs towards better translation, both in terms of instruction tuning and in-context learning. Further we investigated methods to optimise multilingual training and retrieval-augmented approaches, and laid the foundations for context- and emotion-aware translation. Early efforts were also dedicated to simultaneous translation, with a focus on benchmarking and identifying modelling challenges.

In the second reporting period (RP2), progress expanded towards broader adaptation strategies, including alignment with human feedback, inference-time and cross-lingual adaptation, and exploration of alternative architectures such as state-space models. Contextualisation was strengthened through long-context modelling, document-level and conversational translation, and the extension of methods to multimodal speech translation in the IWSLT shared task. Task 4.3 was redirected from simultaneous translation to the development of speech-enabled LLMs and full-duplex conversational models, ensuring that work remained aligned with project needs and opportunities.

WP4 has also contributed extensively to dissemination, with outputs including peer-reviewed publications, open-source models and datasets, and the co-organisation of shared tasks and benchmarking campaigns. This report (D4.2) documents the results of RP2, complementing the foundations laid in D4.1. It is structured as follows: Section 2 details contributions to Task 4.1, organised thematically across model architectures, human alignment, decoding strategies, cross-lingual transfer, and speech adaptation. Section 3 presents Task 4.2, covering context usage, document-level and conversational MT, and long-context methods. Section 4 summarises Task 4.3, including the associated change of focus and current progress. Finally, Section 5 provides the overall conclusions of WP4.

Summary of Output

The work described in this report has contributed to several publications, models, and code repositories, as well as co-organised events, summarised below.

Manuscripts: 1 journal article (TACL), 19 conference papers (4 EMNLP'24, 6 WMT'24, 1 WMT'25, 4 ACL'25, 1 EAMT'24, 1 NeurIPS'24, 2 EACL'24, 3 NAACL'25, 1 COLM'25, 1 MT Summit'25), 2 workshop papers (IWSLT), and 5 arXiv pre-prints.

Code and data:

- https://github.com/deep-spin/ssm-mt
- https://github.com/deep-spin/latim
- https://github.com/deep-spin/mt-pref-alignment

- https://github.com/deep-spin/reranking-laws
- https://github.com/cisnlp/mPLM-Sim
- https://huggingface.co/sardinelab/xTower13B
- https://github.com/Henry8772/ChineseMenuCSI
- https://github.com/deep-spin/gender-bias-qe-metrics
- https://github.com/cisnlp/XAMPLER
- https://github.com/deep-spin/adasplash
- https://github.com/sweta20/chat-qe
- https://github.com/deep-spin/interp_llm
- https://github.com/WMT-QE-Task/wmt-qe-2024-data
- https://github.com/wmt-conference/wmt25-mteval/tree/main

Events:

- WMT Chat Translation Shared Task 2024
- WMT Metrics Shared Task 2024
- WMT Quality Estimation Shared Task 2024
- WMT Shared task on Automated Translation Quality Evaluation Systems 2025

2 Task 4.1: Adaptable, multimodal generation and translation (IT*, UEDIN, UVA, NAV)

Proposal highlights

In this task we focus on adaptation strategies for multiple generation tasks: machine translation (MT), transcription, and summarization. The key proposal highlights are listed below.

- Dynamic adaptation techniques that combine in-context learning and adaptors as an alternative to expensive fine-tuning.
- Investigating new adaptation techniques that can handle multiple modalities.
- Direct and cascaded approaches to speech translation, using large language models.

Summary of completed work

This task focuses on adaptable, multimodal generation and translation, building directly on the foundations laid in D4.1. During RP1 we concentrated on establishing methodologies for steering Large Language Models (LLMs) towards translation (D4.1, Section 2.1) and multilingual adaptation (D4.1, Section 2.2), themes still highly relevant to UTTER and currently presented contributions. However, an increasing amount of work in the last part of UTTER also considered alignment to user preferences, and cross-lingual adaptation to specific domains, tasks, and modalities, including speech. Moreover, we have been considering alternatives to transformer architectures, such as State Space Models (SSMs), exploring their potential in MT and multilingual generation.

Thus, our current contributions for this task can be briefly summarised as follows:

- **State-Space Models:** Exploration of SSMs, assessing their effectiveness compared to transformers for translation, and proposing a method for token attribution (Section 2.1).
- **Aligning with Human Feedback:** Building on the Quality Estimation section of RP1 (D4.1, Section 4.5), we continue exploring related research, with an emphasis on aligning with humans to improve translation quality and its assessment (Section 2.2).
- **Adaptation at the Decoding Stage:** We present works that focus on steering LLMs and further improving the output quality of generation models at the decoding stage (Section 2.3 is a continuation of D4.1 Section 4.1).
- **Cross-lingual Adaptation:** This section, building on D4.1 Section 4.1, 4.2, and 4.4, addresses cross-lingual adaptation, encompassing both pairwise transfer (low-resource to high-resource) and multilingual model adaptation (Section 2.4).
- **Speech adaptation:** Adapted multilingual and multimodal LLMs for speech translation in the IWSLT 2025 shared task, highlighting challenges of variability, noise, and cross-modal alignment (Section 2.5).

We summarise in Table 1 the contributions (publications and repositories) reported throughout UTTER in this task. We note that we do not include publications for which the main description is provided in a different report to avoid duplication of results.

Period	Venue	Paper	ACK	Code	ACK	citations
RP1	WMT	Bogoychev and Chen (2023)	✔ (UKRI)	✗	-	13
	ICML	Zhang et al. (2023)	✔ (UKRI)	✗	-	394
	EMNLP	Alves et al. (2023)	✔	https://github.com/deep-spin/translation_llm	R	39
	COLM	Alves et al. (2024)	✔	https://huggingface.co/collections/Unbabel/tower-659eaedfe36e6dd29eb1805c	R	145
	EMNLP	Farinhas et al. (2023)	✔	https://github.com/deep-spin/translation-hypothesis-ensembling	✔	34
	TACL	Fernandes et al. (2023b)	✔	✗	-	94
	NAACL	Baziotis et al. (2023)	✔ (UKRI)	✗	-	3
	ACL	Zhu et al. (2024)	✔ (UKRI)	https://github.com/NJUNLP/QAlign	✗	31
	EMNLP	Iyer et al. (2023a)	✔ (UKRI)	✗	-	7
	LREC-COLING	Klimaszewski et al. (2024)	✔	https://github.com/mklimasz/transferable-modularity	✔	2
	ACL	Wang et al. (2023)	✔(UKRI)	https://github.com/weixuan-wang123/ReMaKE	✔	36
	EAMT	Martins et al. (2023)	✔	✗	-	4
	EACL	Pal and Heafield (2023)	✔ (UKRI)	✗	-	2
	WMT	Iyer et al. (2023b)	✔ (UKRI)	https://data.statmt.org/ambiguous-europarl/	NA	42
	WMT	Fernandes et al. (2023a)	✔	https://github.com/google-research/google-research/tree/master/palm2_automqm	✗	95
	WMT	Rei et al. (2022c)	✗	https://huggingface.co/Unbabel/wmt22-cometkiwi-da	NA	214
	WMT	Rei et al. (2022a)	✔	https://github.com/Unbabel/COMET	✔	335
	WMT	Rei et al. (2023)	✔	✗	-	47
	WMT	Zerva et al. (2022)	✔	https://github.com/WMT-QE-Task/wmt-qe-2022-data	✔	88
	WMT	Blain et al. (2023)	✔	https://github.com/WMT-QE-Task/wmt-qe-2023-data	✔	47
	WMT	Freitag et al. (2023)	✔	✗	-	93
RP2						
	WMT	Pitorro et al. (2024)	✔	https://github.com/deep-spin/ssm-mt	✔	3
	arXiv	Pitorro and Treviso (2025)	✔	https://github.com/deep-spin/latim	✔	1
	ACL	Treviso et al. (2024)	✔	https://huggingface.co/sardinelab/xTower13B	R	11
	EAMT	Ramos et al. (2024)	✔	https://github.com/deep-spin/mt-pref-alignment	✔	15
	EMNLP	Agrawal et al. (2024a)	✔	-	-	6
	WMT	Zerva et al. (2024)	✔	https://github.com/WMT-QE-Task/wmt-qe-2024-data	✔	28
	WMT	Freitag et al. (2024)	✔	✗	-	41
	WMT	Lavie et al. (2025)	✔	https://github.com/wmt-conference/wmt25-mteval/	✔	0
	NeurIPS	Farinhas et al. (2024)	✔	https://github.com/deep-spin/reranking-laws	✔	1
	EACL	Waldendorf et al. (2024)	✔ (UKRI)	✗	-	16
	EACL	Lin et al. (2024)	✔	https://github.com/cisnlp/mPLM-Sim	R	3
	WMT	Iyer et al. (2024)	✔(UKRI)	✗	-	8
	EMNLP	Ranaldi et al. (2024)	✔(UKRI)	✗	- 10	
	NAACL	Gomes et al. (2025)	✔	✗	-	1
	IWSLT	Attanasio et al. (2025)	✔	https://github.com/deep-spin/it-iwslt-2025	✔	0
	IWSLT	Lee et al. (2025)	✔	✗		0

Table 1: Research outputs (manuscripts and code) from T4.1. Note that for ACK{nowledgements} *R* signifies requested/under process acknowledgement and *NA* signifies 'not applicable'. Citations refer to publications and are obtained from Google Scholar as of September 30, 2025.

2.1 State-Space Models

State Space Models (SSMs) have emerged as a promising alternative to transformer architectures, motivated by the need to overcome the quadratic computational and memory complexity of self-attention. While transformers remain the standard for neural Machine Translation (MT), their scalability issues hinder efficient adaptation to long-context and multilingual generation tasks. SSMs such as Mamba provide linear-time inference and reduced memory usage, raising the question of whether such architectural efficiency can be achieved without sacrificing translation quality or discourse competence. We explored this trade-off through two complementary contributions: (i) a systematic evaluation of SSMs for machine translation, comparing their effectiveness against transformer baselines across sentence-level, document-level, and multilingual settings (§2.1.1), and (ii) the development of LATIM, a novel interpretability framework enabling token-level decomposition in Mamba models, thereby uncovering how SSMs process contextual dependencies (§2.1.2). Together, these studies advance our understanding of the strengths and limitations of SSMs for translation, highlighting both their potential efficiency advantages and their current shortcomings in discourse-sensitive tasks.

2.1.1 How Effective are State Space Models for Machine Translation?

Transformer architectures dominate neural machine translation (NMT), but their quadratic self-attention complexity makes scaling to long sequences costly in both computation and memory. State Space Models (SSMs), such as Mamba, have recently been introduced as an alternative, offering linear-time inference and improved efficiency. Yet their effectiveness for MT remains underexplored, particularly regarding their ability to capture long-range dependencies and adapt to multilingual, discourse-sensitive translation.

In this study, we carried out a **systematic evaluation of SSMs for MT**. We trained and tested Mamba-based models under multiple conditions and compared them against strong transformer baselines of equivalent size. Experiments covered high- and low-resource language pairs from the IWSLT and WMT benchmarks, with both sentence-level and document-level settings. To evaluate adaptability, we also tested bilingual vs. multilingual training and analysed cross-lingual transfer performance.

Our methodology combined three evaluation axes:

1. **Sentence-level translation quality**, measured with BLEU, CHRF, and COMET.
2. **Contextual and discourse modelling**, assessed through pronoun resolution accuracy and consistency in document-level MT.
3. **Efficiency**, quantified via training throughput, inference speed, and memory consumption for long sequences.

Findings indicate that SSMs achieve translation quality competitive with transformers at the sentence level, particularly in high-resource scenarios. They also deliver substantial efficiency gains, with faster inference and up to 40% lower memory usage on long inputs. However, their limitations became evident in discourse-sensitive tasks: SSMs exhibited weaker propagation of contextual information across sentences, resulting in lower performance on pronoun resolution and reduced consistency in long documents. Moreover, in multilingual settings, transfer to morphologically rich or distant languages was less effective than with transformer-based models.

Overall, this work demonstrates that SSMs are a promising architecture for efficient MT but are not yet a full replacement for transformers in tasks requiring deep contextual reasoning. Their strengths in scalability and efficiency complement transformers, while their weaknesses motivate future adaptations aimed at improving discourse-level modelling. This contribution aligns with UTTER's objective of exploring adaptable architectures by providing the first detailed assessment of SSMs in machine translation.

This work is reported in more detail in (Pitorro et al., 2024).

2.1.2 LaTIM: Measuring Latent Token-to-Token Interactions in Mamba Models

Beyond performance, interpretability is a key desideratum for many generative tasks, especially as LLMs are expanding across several user-facing applications. Unlike transformers, which expose attention scores that can be directly inspected for interpretability, SSMs lack explicit mechanisms for visualizing token-to-token interactions. Existing interpretability methods for Mamba, such as

MambaAttention and MambaLRP, only approximate these interactions and fail to provide fine-grained decompositions across layers, limiting our ability to understand how Mamba processes sequences.

To address this limitation, we introduced **LATIM**, a novel token-level decomposition framework for Mamba-1 and Mamba-2. LATIM reformulates SSM computations into a form that supports token-by-token analysis, enabling us to adapt transformer interpretability techniques such as ALTI and ALTI-Logit (Ferrando et al., 2022, 2023b). Specifically, LATIM unrolls Mamba's recurrence relations and defines contribution vectors that capture the influence of one token on another across the sequence. We developed several variants depending on the aggregation function, including ℓ_2 norms, ALTI-style contextual mixing, and ALTI-Logit attribution. While the non-linear SiLU activation complicates exact decomposition, we proposed both approximation-based and exact strategies, the latter obtained by removing the non-linearity.

We evaluated LATIM on three tasks. On a synthetic copying benchmark (Jelassi et al., 2023), we observed that LATIM produces sharper and more faithful diagonal interaction patterns compared to MambaAttention and MambaLRP, with quantitative improvements in area under the curve, average precision, and recall-at-K. On machine translation, using the IWSLT17 EN↔DE dataset (Cettolo et al., 2017), we showed that LATIM produces token alignment maps that align more closely with gold references, improving alignment error rate (AER) relative to previous methods (see also Figure 1). Finally, on retrieval-based generation with the RULER benchmark (Hsieh et al., 2024), we uncovered systematic weaknesses in Mamba models, including difficulty with multi-key retrieval and frequency counting, which often led to misattributed focus or degraded recall over repeated tokens. These insights were not visible with previous interpretability tools.

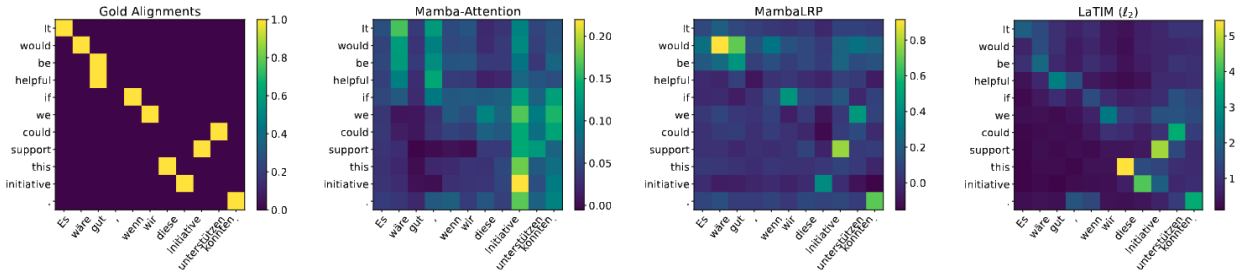

Figure 1: Interpretability heatmaps for Mamba-1 (370M) fine-tuned on DE→EN data from the IWSLT17 dataset. LATIM ($\ell 2$) produces alignments that more closely match the ground truth.

Through ablation, we found that removing the SiLU activation yields exact decompositions with zero approximation error and maintains competitive downstream performance. This suggests a path toward more interpretable SSM architectures without major trade-offs. Overall, LATIM provides the first token-level decomposition method for Mamba models, successfully extends transformer-style attribution to recurrent SSMs, and delivers new insights into both the interpretability and limitations of state space models for long-context reasoning.

This work is reported in more detail in (Pitorro and Treviso, 2025).

2.2 Aligning with human feedback

Steering LLMs and specifically neural machine translation models, often requires optimising surrogate objectives (e.g., via fine-tuning) that, however, have been found to correlate poorly with human judgments (Koehn and Knowles, 2017; Ott et al., 2018). To bridge this gap, we pursued two complementary directions. First, we compared the use of established MT quality metrics towards aligning MT models using different strategies as described in §2.2.1. Second, we developed MT-PREF, a large-scale preference dataset built from expert ratings and metric ensembles, enabling effective preference optimization of translation-specific LLMs as we show in §2.2.2. Both approaches demonstrate the potential of incorporating feedback signals towards the improvement of translation quality. Beyond these contributions described below, we investigated sociolect-related biases that are learned at the alignment stage through reward modelling and preference learning and are reflected in model preferences (§2.2.3), as well as gender biases that are learned and revealed in MT quality estimation models (§2.2.4). Finally, following the same motivation described in RP1, we continue to put emphasis on automated metrics for the evaluation of translations, co-organising three related shared tasks in the WMT conference (2024 and 2025) as described in §2.2.5.

2.2.1 Aligning Neural Machine Translation Models: Human Feedback in Training and Inference

This work investigates how human-aligned reward models can be integrated into the neural machine translation (NMT) pipeline to improve translation quality. The motivation stems from limitations of maximum likelihood estimation (MLE) training, which suffers from exposure bias (Bengio et al., 2015; Ranzato et al., 2016; Wiseman and Rush, 2016), and from the poor correlation between model likelihood and actual translation quality (Koehn and Knowles, 2017; Ott et al., 2018). Inspired by reinforcement learning from human feedback (RLHF)(Stiennon et al., 2022), we leverage existing MT evaluation metrics as reward models, avoiding the need to train reward functions from scratch.

We compare three integration strategies: (i) **data filtering**, where parallel corpora are curated using estimated quality scores from COMET-QE (Rei et al., 2020); (ii) **reinforcement learning**, where models are trained with policy optimization (PPO; Schulman et al., 2017) using COMET (Rei et al., 2022a) and COMET-QE as reward signals; and (iii) **inference-time reranking**, using N-best reranking (Ng et al., 2019; Bhattacharyya et al., 2021) and minimum Bayes risk (MBR) decoding (Eikema and Aziz, 2022). The authors also explore combinations of these techniques.

Experiments use both large-scale noisy data (WMT15 EN→FR (Bojar et al., 2015), WMT16 EN→DE (Bojar et al., 2016)) and smaller clean data (IWSLT2017 EN→DE and EN→FR (Cettolo et al., 2012b, 2017)). Models are based on T5-Large (Raffel et al., 2020), fine-tuned with MLE and further optimized with RL. Evaluation covers BLEU (Papineni et al., 2002), chrF (Popović, 2015), METEOR (Banerjee and Lavie, 2005), COMET, COMET-QE, and BLEURT (Sellam et al., 2020).

Results show that quality-aware data filtering is crucial for stabilizing RL training, with COMET-QE-based filtering yielding the best improvements. RL training with COMET or COMET-QE rewards consistently improves over MLE baselines, often surpassing reranking approaches. Combining RL and MBR decoding further boosts consistency across metrics, although with a higher computational cost. Notably, using reference-free COMET-QE as a reward signal proved highly competitive, suggesting potential for unsupervised NMT training without human references. Overall, we demonstrate that integrating neural quality metrics as reward models at multiple stages of

the MT pipeline provides substantial and reliable gains in translation quality.

This work is presented in more details in Ramos et al. (2024).

2.2.2 Modeling User Preferences with Automatic Metrics: Creating a High-Quality Preference Dataset for Machine Translation.

MT evaluation and training typically rely on single reference translations or small sets of human annotations, which fail to capture the diversity of valid outputs and are costly to scale. To overcome this, we collected human preference judgments from professional linguists for English→German and Chinese→English outputs produced by five state-of-the-art MT systems. Using Direct Assessment (DA) with Scalar Quality Metrics (SQM), we obtained fine-grained ratings that allowed us to systematically compare the quality of system outputs.

We then analysed how well a wide set of automatic quality estimation metrics correlate with these human judgments and found that an ensemble of XCOMET-XL and XCOMET-XXL provided the most reliable proxy (Guerreiro et al., 2023). Based on this ensemble, we induced preference triplets by selecting the best and worst hypotheses per source sentence across multiple systems. This process allowed us to construct MT-PREF, a dataset of around 18,000 instances covering 18 language directions and diverse domains.

Using MT-PREF, we fine-tuned translation-specialised LLMs, namely TOWERINSTRUCT-7B and TOWERINSTRUCT-13B, with preference learning methods: direct preference optimisation and contrastive preference optimisation. Our experiments on WMT23 and FLORES benchmarks show that models aligned with MT-PREF consistently outperform their baselines, achieving higher scores on COMET and XCOMET and narrowing the gap with larger closed-source MT systems, particularly in non-English directions.

We demonstrate that automatically induced preference datasets, when grounded in metric ensembles validated against human judgments, provide scalable and effective supervision for aligning MT models with user preferences and thus improve translation quality in multilingual and low-resource scenarios.

This work has been presented in more detail in Agrawal et al. (2024a).

2.2.3 Rejected Dialects: Biases Against African American Language in Reward Models

This work by (Mire et al., 2025) is reported in D5.2, under Task 5.2, as it closely relates to robustness and transparent evaluation.

2.2.4 Watching the watchers: Exposing gender disparities in machine translation quality estimation

This work by Zaranis et al. (2025) is described in D5.2, since it relates more to explainability and robustness.

2.2.5 Findings from WMT Quality Estimation and Metrics Shared Tasks

In addition to methodological research, following the motivation discussed in RP1, the UTTER consortium actively contributed to the organisation of the WMT Shared Tasks that relate to the automated evaluation of the machine translation models (compared with human judgements). We have been co-organising the tasks reported below, both contributing data annotations (Unbabel; WMT Quality Estimation 2024) and being involved in the task design, planning, and implementation. These initiatives provide a crucial benchmark for evaluating the reliability, robustness, and interpretability of automatic evaluation metrics, especially in the emerging context of large language model (LLM)-based translation and evaluation.

WMT24 Metrics Shared Task. The 2024 edition of this task investigated the ability of automatic metrics to assess MT model outputs across three language pairs (English–German, English–Spanish (LatAm), and Japanese–Chinese), supported by professional Multidimensional Quality Metrics (MQM) annotations. The meta-evaluation framework was revised to prioritise pairwise accuracy at both system- and segment-levels, moving beyond traditional correlation-based evaluations to better reflect the real-world deployment of metrics. Results confirmed that fine-tuned neural metrics such as COMET, XCOMET, and MetricX remain highly effective, with ensemble "metametrics" (MetricX-24-Hybrid, XCOMET) achieving the strongest reliability Rei et al. (2022b); Guerreiro et al. (2023); Juraska et al. (2023).

The task also introduced several challenge sets, allowing participants to target robustness towards diverse linguistic phenomena and domains. These revealed that even top-performing metrics remain vulnerable to failure cases in low-resource languages, domain mismatches, and irregular outputs (e.g. empty strings, mixed language). The findings demonstrated both the maturity of current neural metrics for LLM evaluation and the continuing need for robustness testing across broader language and domain contexts.

The aforementioned findings are reported in more detail in Freitag et al. (2024).

WMT24 Quality Estimation Shared Task. This edition advanced reference-free evaluation by extending sentence-level prediction, fine-grained error span detection, and, for the first time, linked quality estimation with automatic post-editing (APE). The release of new datasets was a major step forward: MQM annotations for English–German, English–Spanish, and English–Hindi, alongside extended Indic language resources, substantially increased language coverage. Systems ranged from traditional encoder-based approaches to LLM-driven methods, showing that LLMs are becoming competitive in QE.

The novelty of incorporating APE within the QE framework opened new directions, showing how QE signals can be directly exploited to correct translations and promoting multi-task quality estimation models. Span-level error detection remained challenging, particularly in classifying error severity, while challenge sets on idioms, omissions, word order, and gender bias revealed persistent weaknesses. The results highlighted a dual trend: rapid improvements through LLM integration, but also a clear need for more interpretable and computationally efficient approaches to ensure sustainable deployment of QE systems.

The aforementioned findings are reported in more detail in Zerva et al. (2024).

WMT25 Unified Metrics and QE Shared Task. The 2025 edition marked a turning point, motivated by the growing capabilities of LLMs to act as multi-tasking evaluators. For the first time, the previously separate Metrics and QE tracks were unified into a single shared task, reflecting the insight that LLM-based systems can handle diverse evaluation functions within one framework. The task comprised three subtasks: (i) segment-level quality score prediction, (ii) span-level error detection, and (iii) quality-informed post-editing. Evaluation covered sixteen language pairs, including low-resource and morphologically rich languages, and domains such as speech transcripts and social media text. Results indicated that large LLMs are highly competitive in score prediction, but span-level detection and balanced error correction remain unsolved problems.

The novelty of this unified setup lies in its ability to leverage cross-task information. By treating metrics, QE, and post-editing as facets of the same evaluation ecosystem, LLM-based multi-tasking models can learn from overlapping signals, such as using error span predictions to guide post-editing or integrating reference-based and reference-free judgments for more stable scoring. Challenge sets revealed major weaknesses in current systems—susceptibility to fluent but semantically irrelevant outputs, systematic gender bias, and low performance in under-represented languages—but the unified format has opened a new research direction where integrated LLM evaluators could benefit from richer, cross-task supervision. This paradigm shift highlights the potential for building holistic, adaptive evaluation pipelines that mirror real-world translation workflows.

The aforementioned findings will be reported in more detail in Lavie et al. (2025), to be presented at WMT 2025.

Key Insights.

- Fine-tuned neural metrics remain reliable for LLM-based MT, but robustness to low-resource and domain-specific phenomena requires attention.
- QE has expanded beyond sentence-level prediction to include span-level error detection and downstream post-editing, with LLMs playing an increasingly central role.
- The unification of Metrics and QE in WMT25 reflects the field's move toward integrated evaluation ecosystems, where LLM-based models can benefit from cross-task information to provide more robust judgments.
- Challenge sets across all years revealed persistent weaknesses: gender bias, poor handling of idiomatic language, over-reliance on fluency, and instability in low-resource scenarios.

Overall, these shared tasks trace a trajectory from validating metrics for LLM-based MT (WMT24 Metrics), through expanding QE to downstream correction (WMT24 QE), to a unified evaluation ecosystem (WMT25). The insights that we obtain and discuss directly support UTTER's objectives of developing robust, explainable, and user-aligned evaluation methodologies for adaptable multilingual generation.

2.3 Adaptation at the Decoding Stage

Adaptation can also be achieved at inference time, where LLMs are steered to better outputs, without retraining, through modifications and decisions at the decoding stage. We investigated

two complementary directions. First, we focused on reranking strategies, and introduced a novel communication-theoretic analysis of reranking, formalising how error probabilities decay as additional candidate hypotheses are considered, and deriving general "reranking laws" that govern the trade-off between candidate set size, scoring functions, and expected quality improvements (§2.3.1). These theoretical results were validated empirically in both machine translation and code generation, showing how inference-time reranking can systematically improve performance. Second, we extended contrastive decoding to multilingual machine translation, where "expert" model and "amateur" models are jointly combined: the expert generates fluent outputs, while the amateur—trained or constrained to suppress hallucinations—provides a corrective signal. We further introduced adaptive weighting, dynamically adjusting the influence of the amateur based on the expert's confidence, thereby reducing hallucinations without degrading fluency (§2.3.2). Together, these contributions illustrate how inference-time adaptation, through principled reranking and contrastive decoding strategies, can enhance the robustness of multilingual LLMs beyond what is achievable with standard greedy search decoding.

2.3.1 Reranking Laws for Language Generation: A Communication-Theoretic Perspective

This work introduces a communication-theoretic framework for understanding and improving reranking strategies in LLMs generation. The central idea is to interpret the generator-reranker pipeline as analogous to a noisy communication system: the generator acts as a sender, producing multiple candidate hypotheses in response to a query, while the reranker serves as a receiver, tasked with selecting the most reliable hypothesis. Each hypothesis is viewed as a message transmitted through a noisy channel, where noise corresponds to hallucinations or errors in generation.

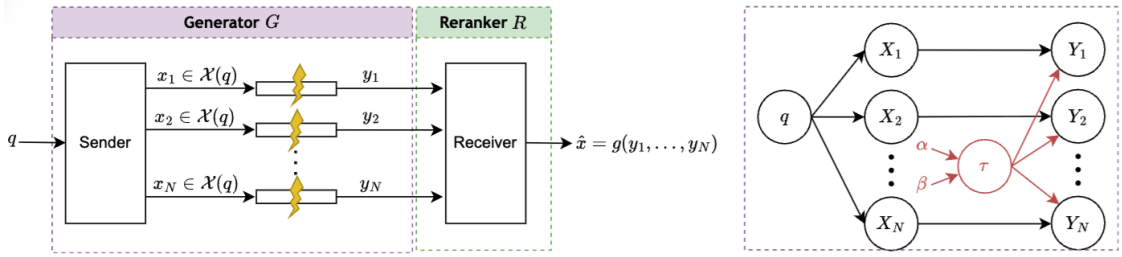

Figure 2: **Left**: A generator-reranker system (G, R) depicted as a communication system. Given a query q with acceptance set X (q), the sender sends N descriptions through noisy channels. The receiver's goal is to decode an acceptable answer through reranking. **Right**: Graphical model of the generator G. We consider two different models: a simplified version with N independent hypotheses, represented in black, and a scenario with exchangeable hypotheses, represented in red.

We propose *reranking laws* that describe how the error probability of the combined generator–reranker system decays with the number of generated hypotheses N. Several scenarios are analyzed (see also Figure 2).

- **Independent hypotheses:** if hypotheses are independent, a perfect reranker yields exponentially decaying error probability (ϵ^N).

- **Imperfect rerankers:** using Mallows or Zipf–Mandelbrot models of noisy ranking, the authors prove that even imperfect rerankers can achieve asymptotically error-free protocols under mild conditions.
- **Dependent hypotheses:** for correlated hypotheses (e.g., exchangeable via a Beta prior), error probability decays with a power law rather than exponentially.

The analysis leverages tools from order statistics and information theory, including the Hurwitz zeta function and dominated convergence theorem.

Experiments validate the theoretical predictions on two tasks:

- **Text-to-code generation:** using DeepSeek-Coder 7B on the MBPP dataset (Austin et al., 2021).
- **Medical machine translation:** using TowerInstruct 13B Alves et al. (2024) on the TICO-19 dataset Anastasopoulos et al. (2020).

Evaluated reranking methods included majority voting, minimum Bayes risk decoding with COMET, and quality-estimation rerankers (CometKiwi) Rei et al. (2022a,c). Results confirmed that empirical error decay closely matches the predicted reranking laws.

The study demonstrates that reranking systematically increases robustness: even with imperfect rerankers, generating multiple candidates and reranking them can asymptotically eliminate unacceptable outputs as N grows. Independent hypotheses guarantee exponential convergence to error-free decoding, while dependent hypotheses lead to slower, power-law decay. Our results provide quantitative guidance on how many candidates are needed to achieve a desired reliability. The work further highlights connections to forward error correction in communication systems, suggesting that future error-correcting protocols for LLMs could be designed by borrowing from coding theory.

This work has been **distinguished as a spotlight paper at NeurIPS** 2024 and has been presented in detail in Farinhas et al. (2024).

2.3.2 Contrastive Decoding Reduces Hallucinations in Large Multilingual Machine Translation Models

In Neural Machine Translation (NMT), models will sometimes generate repetitive or fluent output that is not grounded in the source sentence. This phenomenon is known as hallucination and is a problem even in large-scale multilingual translation models. Contrastive decoding (CD) is a technique that has been used to increase diversity when LLMs are used generatively, but here we adapt it to reduce hallucinations. Contrastive decoding makes use of a strong *expert* model and a weaker *amateur* model, where the latter is specifically designed as a model likely to hallucinate. The key idea is to modify the scoring used in beam search to disprefer predictions where the expert agrees with the amateur. We test several different possible amateurs, as well as variants on the original CD scoring, with the aim of reducing hallucination in NMT.

We use the M2M (Fan et al., 2020) set of models in our experiments, with the FLORES-101 (Goyal et al., 2021) test sets for evaluation. In the original formulation of CD the score assigned to a given token during inference is the difference between the log-probabilities of the expert and amateur

models, incorporating a weighting hyperparamter. We also apply a threshold so that only tokens with a sufficiently high probability under the expert are considered. This ensures that if the expert is confident (only one token reaches threshold) then we ignore the amateur, and also that tokens which are considered unlikely by the expert are not used. We experiment with having an amateur that is simply a smaller model, as well as amateurs that are versions of the expert but "damaged" in some way that introduces hallucinations, for instance by reducing or removing the effect of attention, or by removing the encoder.

To evaluate CD we apply a hallucination detection pipeline to the output of the baseline model to divide the test set into hallucinatory and non-hallucinatory examples. We then measure the COMET score of the CD variants on each of the subsets to see whether CD can increase the score on hallucinatory examples without significantly affecting the non-hallucinatory examples.

We show that using CD in conjunction with amateur models that have reduced source contributions mitigates hallucinations. We extend the CD algorithm, dynamically setting the weight given to the amateur to limit the effect of CD when the expert is confident. Overall, we evaluate across 21 language pairs using the M2M family of models on the FLORES-101 dataset, reporting a mean increase of 14.6 ± 0.5 and 11.0 ± 0.6 COMET on sentences causing hallucinations for the M2M Small model and M2M Medium models respectively.

This work is reported in detail in Waldendorf et al. (2024).

2.4 Cross-lingual Adaptation

Adaptation across languages is central to enabling multilingual LLMs to generalise beyond high-resource language settings. Building on the work of the previous reporting period we introduce further work that addresses multilingual adaptation challenges across MT and relevant generation tasks. We thus introduce mPLM-Sim in §2.4.1, a metric derived from multilingual pretrained model representations that predicts transferability between source and target languages. We further consider the challenges of adapting to low-resource languages: we systematically studied the effects of data scale and diversity on adapting LLMs for low-resource translation, demonstrating how both quantity and variety of training data critically determine performance (§2.4.3). We also showed how program-aided LLMs can improve multilingual reasoning by planning in a high-resource pivot language and executing in the target language, particularly benefiting low-resource languages (§2.4.4). We further discuss xTower (Treviso et al., 2024), a multilingual LLM designed to explain and correct translation errors across languages (§2.4.2. Finally, we consider multimodal evaluation, proposing multilingual finetuning strategies for CLIP-based metrics that enable fairer and more accurate assessment of multilingual image captioning as described in §2.4.5. In addition to these methodological advances, we considered the scope of cross-lingual adaptation to explicitly address cultural and literary dimensions. We introduced a resource and methodology for systematically evaluating how culture-specific items are translated across languages, advancing research on cultural sensitivity in MT and supporting systematic evaluation of cultural adaptation strategies(§2.4.6) Along the same lines, we also provided a novel dataset and framework that highlight how MT systems can mediate cultural meaning beyond the text itself (§2.4.7). Together, these contributions demonstrate how multilingual and multimodal signals can be leveraged to enhance the adaptability and fairness of LLMs across diverse linguistic and cultural contexts.

2.4.1 mPLM-Sim: Unveiling Better Cross-Lingual Similarity and Transfer in Multilingual Pretrained Language Models

Multilingual pretrained language models (mPLMs) encode strong language-specific signals even when these are not explicitly provided during pretraining. However, it remains unclear whether such internal representations can be systematically exploited to measure language similarity and whether these similarity signals can improve source-language selection for zero-shot cross-lingual transfer. To address this gap, we proposed **mPLM-Sim**, a similarity measure that relies on model-internal representations rather than external linguistic resources.

Our **evaluation** considered a diverse set of eleven mPLMs covering different architectures (encoder-only, encoder–decoder, and decoder-only), model sizes, tokenization strategies, and pretraining corpora. To compute similarity, we relied on multi-parallel corpora, including FLORES, the Parallel Bible Corpus (PBC), and the speech corpus Fleurs. For each layer of an mPLM, we obtained sentence embeddings for all languages using mean pooling (or position-weighted pooling for decoder models) and computed pairwise cosine similarity across languages. Averaging these values across 500 multi-parallel sentences per language yielded similarity matrices for each layer of every model. These model-based measures were then compared against seven traditional linguistic similarity measures, including lexical, genealogical, geographical, syntactic, and phonological distances, as well as inventory and feature-based measures.

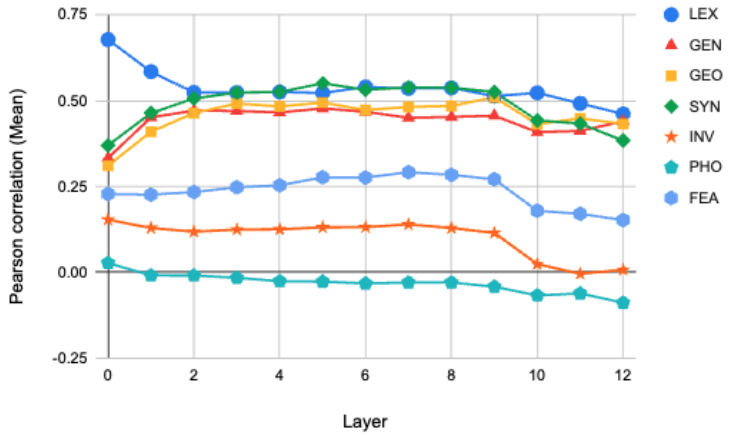

Figure 3: Pearson correlation (MEAN) between mPLM-Sim and linguistic similarity measures across layers for Glot500 and Flores on 32 languages. Correlation between mPLM-Sim and LEX peaks in the first layer and decreases, while the correlation with GEN, GEO, and SYN slightly increases in the low layers before reaching its peak.

Our analysis revealed that mPLM-Sim correlates strongly with lexical similarity and moderately with genealogical, geographical, and syntactic measures, while phonological and inventory-based measures were less well reflected. The layer at which similarity is measured proved critical: lower layers tend to capture lexical overlap, middle layers reflect structural similarities such as genealogy and syntax, and upper layers lose fine-grained distinctions as they encode more task-specific semantics. Architectural differences also played a role: encoder-only models such as XLM-R and mBERT yielded more robust similarity patterns than decoder-only models, while larger models more consistently encoded high-level similarity.

Across tasks and corpora, using mPLM-Sim to select source languages led to improvements of

one to two percentage points on average compared to traditional linguistic similarity measures, with particularly large gains for low-resource and typologically distant languages. Nevertheless, performance gains were uneven, as pretraining coverage and model type remained limiting factors.

Overall, this work provides both methodological and empirical insights into how language similarity is encoded within mPLMs. By demonstrating that model-internal similarity signals can serve as effective predictors of cross-lingual transfer performance, mPLM-Sim advances the analysis of multilingual models and provides a practical tool for more informed source-language selection.

This work is reported in detail in Lin et al. (2024).

2.4.2 xTower: A Multilingual LLM for Explaining and Correcting Translation Errors

We introduce **xTower**, a translation-oriented LLM built on top of TowerBase-13B (Alves et al., 2024) to **explain** span-marked translation errors and **propose** a corrected translation. The model is obtained by distilling GPT-4 explanations on MQM-annotated WMT22 data (EN→DE, EN→RU, ZH→EN), followed by multilingual finetuning mixed with TowerBlocks MT data. Prompts present an "annotated translation" (error spans with severities) and adopt an explanation-then-correction format, which xTower handles in referenceless or reference-based modes. Crucially, xTower is agnostic to the source of spans—human or automatic (e.g., xCOMET)—and can thus plug into existing QE pipelines. WE illustrate our approach in 4.

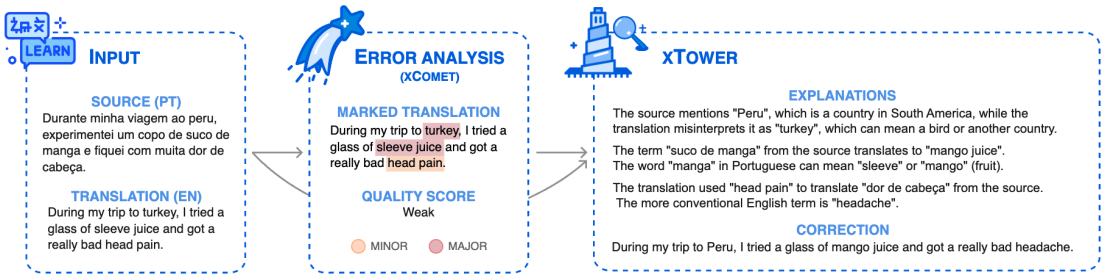

Figure 4: Illustration of xTower approach. In this example, the input consisting of a source and a translation is passed to xCOMET, which annotates the translation with error spans and produces a (discretized) quality score. The full input, marked translation, and quality score are passed to xTower, which, in turn, produces an explanation for each error span along with a final suggestion for a new, corrected translation.

We evaluate on WMT23 MQM test sets for EN→DE, HE→EN, and ZH→EN, measuring (i) *relatedness* of explanations to the marked spans and (ii) *helpfulness* for understanding the error and guiding a fix, via expert human annotation. Relatedness is higher when spans are human-labeled (about 4.3 on a 0–6 scale) than when predicted by xCOMET (about 3.2), confirming the impact of span quality. Helpfulness scores average 4.4–4.6 for error understanding and 3.3–3.9 for guidance, indicating that explanations are generally informative and often suggest the path to a better translation.

Conditioning on spans and the generated explanations, xTower refines the translation and is assessed with COMET (primary), BLEURT, and COMET-Kiwi. Across language pairs, xTower improves over the original MT by roughly +1 to +3 COMET in referenceless setups; a hybrid strategy that keeps the original translation when COMET-Kiwi is high and otherwise adopts the xTower correction yields further gains (up to ∼+2 COMET on HE→EN). Improvements concen-

trate on lower-quality inputs (original COMET $\leq$ 80). Moreover, xTower fixes the majority of span-marked errors and compares favorably to strong LLM baselines when used for post-editing.

Overall, xTower shows that free-text, span-grounded explanations can be both **useful to humans** and **actionable for models**, leading to measurable MT quality gains in multilingual settings and enabling practical hybrid protocols for cost-effective post-editing.

This work is presented in more detail in (Treviso et al., 2024).

2.4.3 Quality or quantity? On data scale and diversity in adapting large language models for low-resource translation

Many recent works have shown strong results for LLMs in MT, but few have dealt with very low-resource languages (which may only have 1000s of parallel sentences available, or less). The idea of this paper was to test recently proposed LLM-based methods on very low-resource translation. In particular, we applied the "two-stage" approach (Xu et al., 2023) where continued pre-training (CPT) on monolingual (and maybe parallel) data is followed by supervised fine-tuning (SFT) on a much smaller amount of high-quality parallel data (see also Figure 5). Work on the Tower models (Alves et al., 2024) had further shown that task and language diversity within the two-stage approach could also increase MT quality.

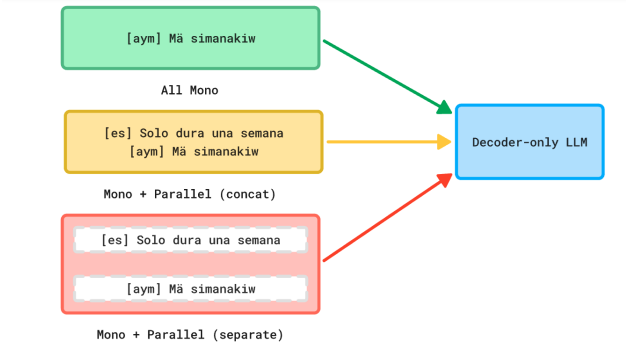

Figure 5: Strategies explored for incorporating parallel data during Continued Pre-Training. We show a Spanish (es) to Aymara (aym) example from our parallel data.

We used data from two recent shared tasks on low-resource MT: AmericasNLP (Ebrahimi et al., 2024) and the WMT23 shared task on low-resource Indic translation (Pal et al., 2023). In some experiments we supplemented this with training daya from OPUS (Tiedemann, 2012). We experiment with 3 base LLMs of varying sizes: Gemma 2B (Gemma Team et al., 2024), Mistral 7B (Jiang et al., 2023) and Llama 3 8B (Dubey et al., 2024).

We explore the question *What would it take to adapt LLMs for low-resource MT?* by varying the two-step training approach and evaluating on the low-resource tasks. For the first step, we test whether training only on monolingual data (as recommended by Xu et al. (2023)) is possible, or whether parallel data is necessary. For the SFT we test whether including diverse tasks and languages can help, and examine the effect of scaling parallel data.

Our contributions are as follows:

1. In contrast to findings for high-resource LLM-MT (Xu et al., 2023), we observe that for low-resource languages, LLMs benefit hugely from scale of parallel data, during both CPT and SFT stages

2. Linguistic and task diversity during SFT leads to negative interference for low-resource languages LLM-MT, with focused multilingual MT fine-tuning for more epochs being the most effective recipe.

3. Choosing larger base LLMs is crucial. For under-represented (zero-resource) languages, it is more effective to train larger LLMs with smaller vocabularies rather than vice-versa.

4. Fine-tuning embeddings is critical. Across the board, we observe that fine-tuning embeddings along with LoRA modules yields huge gains.

5. SFT alone is effective, but CPT+SFT (i.e. the two stage approach) yields best results.

6. The gains of CPT on parallel data carry over to SFT scaling, particularly for the higher-resourced languages in our experiments, and the very low-resourced languages benefit more from SFT scaling.

For full details please refer to the paper (Iyer et al., 2024).

2.4.4 Empowering multi-step reasoning across languages via program-aided language models

The paper proposes Cross-PAL, a cross-lingual extension of Program-Aided Language Models and structure reasoning trajectories operating via two modules: an understander that plans a solution in a pivot language (typically English) using program-like steps, and a solver that executes and answers in the original query language. Then, we introduce a self-consistent extension, SCross-PAL, that ensembles multiple cross-lingual reasoning paths to select the most consistent outcome. Figure 6 summarises our contribution. We show that this method targets an ongoing limitation of multilingual reasoning, such as instability and drop-offs in low-resource languages, by aligning planning in a high-resource language with language-specific execution.

The core evaluation centres on multilingual arithmetic reasoning: MGSM (250 GSM8K items translated and double-checked into 10 languages) and MSVAMP (the SVAMP suite in 9 languages). To probe generality beyond arithmetic, we also study results on XCOPA for multilingual causal commonsense reasoning. To deliver broader results, we operate using both close-weight (GPT-based) and open-weight models (Llamas and Phi) of different scales. We compute the final results using the accuracy score computed by exact match against the gold numeral/text, with answers normalised to the target language format.

We demonstrate that Cross-PAL improves multilingual reasoning performance across tasks and model sizes. On MGSM and MSVAMP, it achieves higher accuracy than existing prompting methods, with self-consistent Cross-PAL (SCross-PAL) providing further gains. The modular approach is critical, as both single-step prompting and first-step-only variants underperform, highlighting the value of explicit planning followed by execution. Smaller open-weight models such as Llama-3-8B and Phi-3 benefit substantially, limiting the gap with larger models when led by program-aided demonstrations. We also observe that using English as the pivot language strengthens reasoning, particularly for low-resource languages like Telugu, Swahili, Bangla and Thai, whereas native-language planning is less effective. Self-consistency mode stabilises performance by ensembling across languages, with English playing a pivotal role in low-resource settings. Finally, we prove that the approach generalises beyond arithmetic: on XCOPA, Cross-PAL and its self-consistent

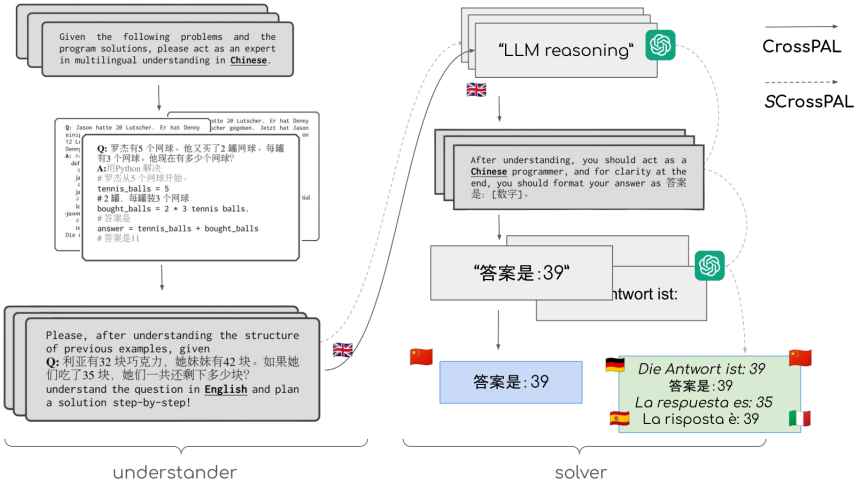

Figure 6: Cross-PAL elicits the LLM to generate reasoning programs across different languages. In this example, given separate problems in language LS (Chinese), the conducted steps for solving it are: (1) generate a structured planning strategy in English (using in-context demonstrations in LS), (2) collect the planned strategy and finalize the solution in LS (that is the language of the original problem).

variant remain competitive for causal commonsense reasoning, underscoring the applicability of the method.

This work is presented in more detail in Ranaldi et al. (2024).

2.4.5 Evaluation of Multilingual Image Captioning: How far can we get with CLIP models?

Beyond MT and moving towards multi-modal generation we considered the task of image captioning evaluation, which has seen great advancements with CLIP-based models, which can encode and compare both image and text modalities. However, such models so far have been largey English-centric, thus restricting their applicability to multilingual and multicultural contexts, where captions must not only be semantically accurate but also culturally appropriate. Without proper adaptation, models may fail to capture the diversity of meanings and cultural references across languages, leading to biased or unreliable evaluations. Recognising this, we sought to extend CLIPScore to multilingual scenarios and to investigate whether multilingual finetuning can align such metrics with human judgments across diverse linguistic and cultural settings.

We relied on a wide range of multilingual multimodal datasets, including CrossModal-3600, VICR (translated), VALSE multilingual variants, XVNLI, and MaRVL, which together cover different languages, domains, and cultural perspectives. Using these resources, we evaluated zero-shot transfer of English CLIP models and proposed adapted variants through multilingual finetuning on preference data.

Our experiments revealed that multilingual CLIPScore, when finetuned, matches the performance of English-only models on general benchmarks while significantly improving correlation with human judgments in multilingual and multicultural settings. In particular, the adapted models better capture semantic adequacy and cultural nuances, demonstrating that cross-lingual adaptation of multimodal metrics is both feasible and necessary for fair evaluation of multilingual captioning

systems.

This work is described in Gomes et al. (2025).

2.4.6 Cultural Adaptation of Menus: A Fine-Grained Approach

As the main contribution of this paper includes a dataset, we describe it in more detail in D2.2.

2.4.7 Through the Looking-Glass: On Explicitation via Genettean Paratexts in Literary Machine Translation

As the main contribution of this paper includes a dataset, we describe it in more detail in D2.2.

2.5 Speech Adaptation

Extending multilingual generation models from text to speech poses additional challenges, as systems must handle variability in acoustic input, robustness to noise, and alignment between spoken and textual modalities. In this section, we report on contributions to the IWSLT 2025 shared task (Salesky et al., 2025), where multilingual and multimodal LLMs were adapted for speech translation. The focus was on designing instruction-following mechanisms, leveraging pretraining on speech–text pairs, and evaluating generalisation across languages and domains. These works demonstrate how methods developed for text-based translation within UTTER can be transferred to spoken language, highlighting both the feasibility and the remaining challenges of multimodal adaptation.

2.5.1 IWSLT 2025 Instruction Following - NAV submission

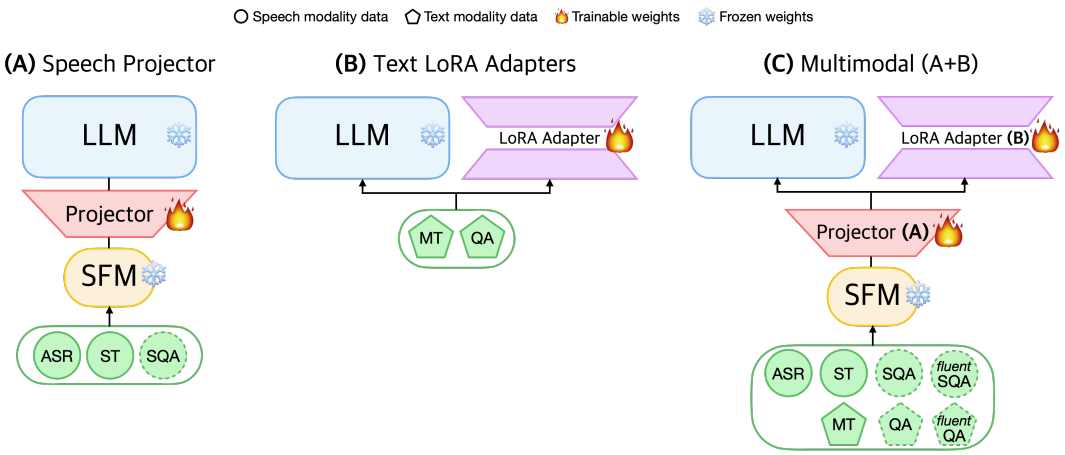

Figure 7: The training pipeline. A speech projector (A) and text LoRA adapters (B) are trained in parallel using speech-to-text and text-to-text data, respectively. These modules are then integrated during a brief multimodal adaptation step (C).

The UTTER consortium participated to the IWSLT 2025 challenge on speech LLMs. The *IWSLT Instruction-following Speech Processing Track* focused on leveraging LLMs and speech foundation models (SFM) to build solutions capable of performing multilingual tasks from English speech

input and textual multilingual instructions (Abdulmumin et al., 2025). NAV participated in the constrained setting of the IF track of IWSLT 2025, meaning that data and models were controlled and defined by the organizers. Their submission focused on the modality-specific training for both text and speech, followed by multimodal alignment. They overall achieved the best results for the short track. Notably, their scores for multilingual SQA were even superior to the commercial system Microsoft-Phi. The system report is available at Lee et al. (2025). The track results and discussion is available at Abdulmumin et al. (2025).

Participants were allowed to use the speech backbone `SeamlessM4T-v2-large` (Barrault et al., 2023) and the text LLM `Llama-3.1-8B-Instruct` (Grattafiori et al., 2024) for both training and data generation. For training the models, they leveraged the constrained setting datasets: CoVoST2 (Wang et al., 2020), EuroParlST (Iranzo-Sánchez et al., 2020), and SpokenSQuAD (Lee et al., 2018). With the agreement of the organizers, they also chose to take advantage of the `SeamlessM4T-v2-large` to produce extra synthetic speech data (Seamless TTS) and multilingual text data (Seamless MT). `Llama-3.1-8B-Instruct` is used to rephrase SQA answers. ACL 60-60 (Salesky et al., 2023) is used for validation and evaluation only.

The approach is presented in Figure 7. Two systems are trained in parallel: (1) speech-to-text ASR/ST/SQA projectors that project the averaged speech representation from the SFM encoder to the embedding space of a frozen LLM; (2) text-only LoRA adapters (Hu et al., 2022), plugged on top of the same frozen LLM and trained on MT/QA. Once both systems are separately trained, they show that it is possible to merge those learned components, increasing overall speech performance, by fine-tuning for only 1K steps on multimodal multilingual data.

We observed that for translation tasks, such as ASR and ST, projector-only approaches for Speech LLMs are well-suited. However, for some complex generation tasks, such as SQA, adapters are required to better control the output of the model. They extensively experimented with synthetic data and data mixture regimes, submitting their best multimodal multilingual system to all the challenge languages (en-en, en-de, en-it, en-zh). For further information, please refer to the system paper Lee et al. (2025).

2.5.2 IWSLT 2025 Instruction Following - IT submission

The consortium submitted a second entry to the IWSLT 2025 Shared Task on Instruction-following Speech Processing (*unconstrained* track). In particular, this work addresses the challenge of extending small-scale language models with speech processing capabilities, a direction of growing relevance given the increasing interest in efficient multimodal models. While recent advances in speech-language modeling have largely focused on large-scale architectures (7B+ parameters), such models are computationally costly and limit reproducibility. Motivated by efficiency and accessibility, we investigated whether **models under 2B parameters can be effectively aligned with speech encoders to perform automatic speech recognition (ASR), speech translation (ST), and spoken question answering (SQA)**. This line of research is particularly relevant for broadening participation in speech-to-text learning, ensuring systems can be trained and deployed in resource-constrained settings.

The proposed system integrates a pretrained continuous speech encoder (w2v-BERT 2.0) with compact Qwen 2.5 models (0.5B and 1.5B parameters) via a two-stage training curriculum. The first phase, modality alignment, adapts speech representations into the LM's embedding space, equipping the model with general ASR capabilities. The second phase, instruction fine-tuning,

expands to multi-task instruction-following behavior across ASR, ST, and SQA. Training relies on high-quality, open-licensed CC-BY datasets such as LibriSpeech, CommonVoice, FLEURS, VoxPopuli, and CoVoST2, complemented with synthetic corpora generated through pseudolabeling and large model filtering for ST and SQA.

Our best model achieves solid performance on English ASR, despite challenging test conditions (spontaneous, technical-domain speech). Performance on ST and SQA was more limited, with frequent task misalignment and reduced output quality, largely due to training data imbalance and the short audio length cutoff (120s). Nonetheless, findings highlight that **compact speech-language models can already attain competitive ASR performance at a fraction of the cost of larger systems**, showing promise for efficient multimodal modeling. The study further emphasizes the importance of carefully balanced multi-task training and enhanced synthetic data generation for scaling small-scale models to more complex speech-to-text tasks.

In sum, this work demonstrates the feasibility of adapting small-scale language models into multimodal systems with speech capabilities, advancing efficiency and reproducibility in speech-to-text research.

This work was presented in Attanasio et al. (2025).

3 Task 4.2: Contextualisation and emotion tracking (IT*, UEDIN, UVA, UNB, NAV)

Summary of completed work

In RP1, our work on this task established the foundations for context-aware dialogue and translation. We investigated the role of prosody in speech-to-text translation (D4.1, Section 3.1), proposed annotation frameworks for customer support chat (D4.1, Section 3.2), and introduced the MAIA dataset with dialogue quality and emotion labels (D4.1, Section 3.3). These efforts underlined the importance of context and emotion for conversational systems and provided resources and evaluation methods to measure their impact.

Rapid advances in LLMs have emphasised the need to further elaborate on the use of context. Specifically, LLMs are now capable of processing tens of thousands of tokens, enabling applications such as document-level translation and multi-turn dialogue modeling. However, as highlighted both in our earlier work and in recent publications, simply enlarging context windows does not ensure that models exploit contextual information effectively. LLMs frequently display shallow context usage, positional biases, and persistent errors in discourse-level phenomena such as pronoun resolution, lexical cohesion, and register maintenance. This evolution of capabilities and challenges has therefore made **context-aware generation, discourse reasoning, as well as reasoning over long-context** the primary scientific focus of RP2.

Thus, this reporting period, our contributions concentrated on analysing how LLMs use context, developing methods to understand and enhance their discourse competence, and benchmarking their performance in document- and chat-level translation. Our contibutions in this sense are more closely related to D4.1 Section 4.2, but generalise beyond translation to context usage for large multilingual generation models. We group these contributions into four themes:

1. **Context Usage and Challenges of In-Context Learning.** We analysed how LLMs exploit few-shot prompts in multilingual MT and reasoning, exposing positional biases and limits of natural language demonstrations (Section 3.1).
2. **Contextualisation for document-level Machine Translation** We benchmarked document-level MT, introduced multilingual datasets such as *DocBlocks*, and proposed evaluation frameworks targeting relevant discourse phenomena (Section 3.2).
3. **Long-Context Modelling.** We developed sparse attention mechanisms (ASEntmax) and assessed memory mechanisms for sustaining cross-turn consistency (Section 3.3).
4. **Conversation-Aware Translation Frameworks.** We designed frameworks for real-time bilingual conversations, combining context-augmented fine-tuning and decoding, and co-organised the WMT 2024 Shared Task on Chat Translation (Section 3.4).

Taken together, these advances consolidate UTTER's trajectory from RP1 to the final reporting period: moving from laying the groundwork with context and emotion resources to addressing the central challenge of how LLMs can exploit extended context meaningfully for discourse-level translation and multilingual interaction.

We summarise in Table 2 the contributions (publications and repositories) reported throughout UTTER in this task. We note that we do not include publications for which the main description is provided in a different report to avoid duplication of results.

Period	Venue	Paper	ACK	Code	ACK	citations
RP1						
	EACL	Zhou et al. (2024)	✔ (UKRI)	✗	-	7
	ACL	Fernandes et al. (2023c)	✔	https://github.com/CoderPat/MuDA	✔	39
	EACL	Mohammed and Niculae (2024)	✔	✗	-	6
	MT Summit	Honda et al. (2023)	✔	https://github.com/su0315/discourse_context_mt	✗	0
	MT Summit	Menezes et al. (2023)	✔	✗	-	4
	GEM	Mendonça et al. (2023)	✔	github.com/johndmendonca/MAIA-DQE	✗	3
	WMT	Farinha et al. (2022)	✔	https://github.com/Unbabel/MAIA	✗	15
RP2						
	EMNLP	Zaranis et al. (2024)	✔	https://github.com/deep-spin/interp_llm	✔	4
	ACL	Ranaldi et al. (2025a)	✔(UKRI)	✗	-	2
	NAACL	Lin et al. (2025)	✔	https://github.com/cisnlp/XAMPLER	R	3
	arXiv	Ramos et al. (2025)	✔	✗	-	2
	MT Summit	Mohammed and Niculae (2025)	✔	✗	-	1
	arXiv	Vasylenko et al. (2025)	✔	✗	-	1
	ICML	Gonçalves et al. (2025)	✔	https://github.com/deep-spin/adasplash	✔	1
	arXiv	Ranaldi et al. (2025b)	R	✗	-	2
	WMT	Pombal et al. (2024b)	✔	✗	-	2
	arXiv	Pombal et al. (2024a)	✔	✗	-	1
	WMT	Mohammed et al. (2024)	✔	✗	-	4
	TACL	Agrawal et al. (2024b)	✔	https://github.com/sweta20/chat-qe	R	3

Table 2: Research outputs (manuscripts and code) from T4.2. Note that for ACK{nowledgements} *R* signifies requested/under process acknowledgement and *NA* signifies 'not applicable'. Citations refer to publications and are obtained from Google Scholar as of September 30, 2025.

3.1 Context Usage and Challenges of In-Context Learning

In-context learning has become a defining feature of large language models, enabling them to adapt to new tasks from demonstrations embedded directly in the prompt. However, despite their apparent flexibility, it remains unclear whether LLMs truly exploit context or merely learn superficial patterns. This section investigates how models attend to demonstrations (§3.1.1), what limitations arise when natural language alone is used to encode examples (§3.1.2), and how targeted retrieval strategies can improve effectiveness (§3.1.3). The works presented here move from diagnosing shallow or biased use of context to showing that natural-language demonstrations are sometimes insufficient for multilingual reasoning and finally to introducing learned retrieval mechanisms that provide more informative examples.

3.1.1 Analyzing context contributions in LLM-based machine translation

In this work we focus specifically on MT, since few-shot prompting has been shown to boost the performance of LLMs in translation tasks. Yet, it remains unclear how different parts of the provided context in the examples, i.e., the sources, their translations, or target prefixes, actually influence the generated outputs. Understanding these contributions is crucial both for improving translation reliability and for efficiently diagnosing hallucinations. Thus we evaluated English–German and English–Russian translation tasks, using five-shot prompts constructed from high-quality parallel examples. We relied on contribution attribution techniques to decompose how different context segments affect translation outputs. We first partition the prompt into the current source, few-shot example sources and targets, and target prefixes, and then quantify average contributions at the token and sequence levels. Comparative analyses are conducted across base, pre-

trained, and fine-tuned models, specifically using LLaMA-2 7B, Tower, and TowerInstruct) (see also Figure 8.

E1 | SRC	Es gibt auch zwei schöne Parks in der Nähe, den Espanya Industrial Park und den Parc de Joan Miró.
E1 | TGT	There are also two beautiful parks nearby, the Espanya Industrial Park and the Parc de Joan Miró.
E2 | SRC	Das Frühstück ist im Preis (10 €) enthalten, es ist aber optional.
E2 | TGT	Breakfast is included in the price (10 €), but it is optional.
E3 | SRC	Es gibt auch kostenlose Internet 24/7 and WiFi in allen Zimmern.
E3 | TGT	There is also free internet 24/7and wifi in all rooms.
E4 | SRC	Bisher gibt es noch keine Bewertungen für S-Plus Company!
E4 | TGT	There are no reviews for S-Plus Company yet!
E5 | SRC	Die Größe der Wohnung ist 15 m2, es ist klein, aber sehr gemütlich.
E5 | TGT	The size of the apartment is 15 m2, it's small but very cosy.
SRC	Die gibt es zwar auch (anscheinend?) bei den MarathonPlus Reifen, aber der Großteil ist schon breiter.
LLAMA-2 ✓	
MT	There are also (apparently?) at Marathon Plus Tyres, but the majority is wider.
TOWER ✗	
MT	There are also two beautiful parks nearby, the Espanya Industrial Park and the Parc de Joan Miró.

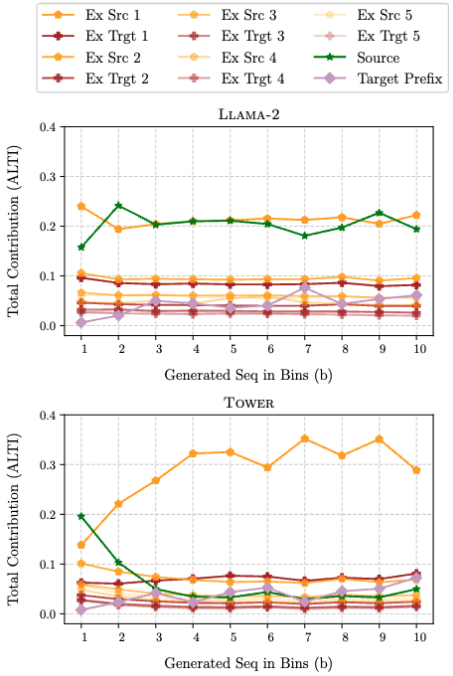

Figure 8: Illustration of an example exhibiting anomalous source contributions for TOWER — which hallucinates, followed by LLAMA-2's contributions, which performs normally.

The results reveal consistent patterns: example sources contribute more strongly than example targets, and earlier examples exert a stronger influence due to positional bias. Fine-tuning seems to be shifting reliance more heavily toward the correct source sentence. Moreover, abnormal contribution patterns align with translation errors, suggesting that attribution analysis can be used as a tool for error detection. This work advances interpretability in LLM-based MT and provides insights and guidance for more robust prompting and model tuning.

This work has been reported in Zaranis et al. (2024).

3.1.2 When natural language is not enough: The limits of in-context learning demonstrations in multilingual reasoning

To gain a deeper understanding of the role of reasoning methods for in-context demonstrations, we investigate how well Chain-of-Thought (CoT) and Program-Aided Language Models (PAL) perform across languages for arithmetic and symbolic reasoning tasks. Specifically, we examine when natural language is not enough for multilingual reasoning with in-context learning. Building on CoT and PAL, we conduct a systematic study of how the type (natural language and programme-like), quantity, and language (native, English, or cross-lingual) of demonstrations affect performance across tasks, models, and languages. Our analysis shows that structured, code-style demonstrations travel better across languages than natural-language rationales, and that asking models to reason in English while receiving prompts in other languages can be especially beneficial, particularly for low-resource settings.

We evaluate five benchmarks spanning mathematical reasoning (MGSM, MSVAMP), natural-language understanding (XNLI, PAWS-X), and commonsense causality (XCOPA), covering 26 languages in total (especially, low-resource Telugu, Bangla, Swahili, Thai, among others). We compute the accuracy by exact match after normalising answer formats to the target language. We operate with different models, including closed-source GPT-4 and GPT-3.5, as well as open-weight families, Llama-2 (7B/70B), Mistral/Mixtral, and StarCoder2/CodeLlama, to probe scale and proficiency. We use different prompting conditions, including Direct (no reasoning), native-language CoT/PAL, English-rationale variants (En-CoT/En-PAL), and cross-lingual settings in which the question is in the target language. Still, the step-wise rationale is elicited in English (Cross-CoT / Cross-PAL).

We find that reasoning methods reliably outperform Direct prompting beyond English, with gains observed in mathematical, understanding, and commonsense tasks across model families. PAL generally surpasses CoT for multilingual arithmetic, reflecting the transferability of structured, programme-like demonstrations. The language of reasoning matters. When we prompt models to produce the rationale in English while keeping the question in the target language, Cross-CoT/PAL deliver consistent improvements over their native counterparts (e.g., for GPT-4 on MGSM, Cross-CoT outperforms Native-CoT by +11.2 points). Analyses with OpenLID further show that CoT often drifts to English in low-resource settings and that the number of intermediate "hops" differs by language; by contrast, PAL yields more stable hop counts across languages. The quantity of demonstrations exhibits diminishing returns for larger models (GPT-series often plateau by 4-shot vs 6-shot), whereas smaller models (e.g., Llama-2-7B, StarCoder2) continue to benefit as shots increase. Beyond arithmetic, CoT improves XNLI and PAWS-X, and Cross-CoT is competitive on XCOPA, indicating that the cross-lingual rationale strategy extends to non-numeric reasoning. Collectively, these results recommend structured demonstrations (PAL) for multilingual arithmetic, English as the reasoning lingua franca when robustness matters, and task- and scale-aware shot counts to balance stability and efficiency.

This work is further reported in Ranaldi et al. (2025a).

3.1.3 XAMPLER: Learning to retrieve cross-lingual in-context examples

In this work, we focused on a central limitation of in-context learning: although LLMs can benefit from demonstrations provided in the prompt, their performance depends critically on the choice of examples. Naive strategies, such as random sampling or nearest-neighbor retrieval in embedding space, often result in suboptimal prompts that fail to capture the relevant linguistic or semantic cues. This problem is especially acute in the multilingual setting, where the availability of parallel data varies, and examples drawn from high-resource languages may not transfer effectively to low-resource ones.

To address this challenge, we introduced XAMPLER, a retrieval-based framework that learns to select cross-lingual in-context examples. Instead of relying on static heuristics, XAMPLER is trained end-to-end with a contrastive objective that directly rewards retrieval strategies leading to better translation quality. The system leverages both source- and target-side representations, aligning retrieval with task-specific signals rather than surface-level similarity. Importantly, XAMPLER is designed to operate in the multilingual space, enabling it to retrieve useful demonstrations even across language boundaries.

We evaluated our approach on multiple translation benchmarks spanning high- and low-resource

language pairs. Results show that XAMPLER consistently outperforms heuristic baselines, leading to more accurate and fluent translations. The improvements are particularly notable in low-resource scenarios, where careful retrieval of demonstrations provides the model with much-needed guidance. Beyond raw translation accuracy, we also observed gains in discourse-level features such as lexical consistency and pronoun resolution, demonstrating that targeted retrieval can enhance the depth of context exploitation.

Overall, XAMPLER shows that principled retrieval is essential for unlocking the full potential of in-context learning in multilingual settings. By learning to identify the most informative demonstrations, our approach mitigates positional biases, improves robustness across languages, and complements our broader investigations into how LLMs process context in machine translation.

This work is reported in more detail in Lin et al. (2025).

3.2 Contextualisation for document-level Machine Translation

While LLMs achieve strong results at the sentence level, document-level translation introduces additional challenges, such as maintaining discourse coherence, resolving pronouns, and preserving stylistic consistency. The ability to exploit context beyond the sentence is therefore a critical test of contextualisation. In this section, we examine whether LLMs genuinely benefit from extended context in translation, both through perturbation-based diagnostics and multilingual evaluations (§3.2.1 and §3.2.2). We then explore decoding strategies that explicitly surface latent discourse knowledge encoded in LLMs (§3.2.3). Collectively, these studies reveal that while LLMs have the capacity to handle discourse phenomena, naive inference methods underuse this potential, highlighting the need for context-aware evaluation and decoding.

3.2.1 Multilingual Contextualization of Large Language Models for Document-Level Machine Translation

Specifically for translation, we address the challenge of adapting large language models to document-level machine translation, where maintaining discourse coherence, cross-sentence consistency, and contextual appropriateness goes beyond the capabilities of standard sentence-level models. To this end, we curated DocBlocks, a multilingual dataset of high-quality document-level parallel corpora spanning domains such as news, parliamentary proceedings, and literary texts. We fine-tune strong baseline models, including Tower, EuroLLM, and Qwen2.5, using a multi-paradigm instruction framework that combines document-to-document translation, chunked translation with context, and sentence-level examples. This strategy allows models to learn from diverse contextual settings while preserving sentence-level accuracy. At inference time, we explore different paradigms—direct document translation, context-augmented chunking, and agent-based prompting—and systematically evaluate them across multiple language pairs.

Our results show that contextual fine-tuning on DocBlocks yields substantial improvements in discourse-level quality and consistency without sacrificing local accuracy, outperforming both prompting and chunk-based baselines in accuracy and efficiency. By releasing DocBlocks and demonstrating its effectiveness, we provide a practical pathway for leveraging multilingual LLMs in document-level translation tasks where long-range dependencies are critical.

This work is further reported in Ramos et al. (2025).

3.2.2 Context-Aware or Context-Insensitive? Assessing LLMs' Performance in Document-Level Translation

Large language models (LLMs) are increasingly strong contenders in machine translation. In this study, we focused on document-level translation, a setting where certain words cannot be translated correctly without information from beyond the current sentence. Specifically, we investigated the ability of prominent LLMs to exploit surrounding document context during translation. To do so, we carried out both a perturbation analysis, examining models' robustness to noisy or randomized contexts, and an attribution analysis, assessing the contribution of relevant antecedents to the translation of context-sensitive tokens. Our evaluation covered nine LLMs from diverse model families and training paradigms, alongside two encoder–decoder baselines.

We considered three model categories in order to disentangle the effects of large-scale training, multilingual pretraining, and translation-specific finetuning. For translation-finetuned LLMs, we evaluated the Tower family (Alves et al., 2024), which is based on Llama-2. TowerBase continues pretraining on multilingual data, while TowerInstruct adds further finetuning on translation-related tasks. We also analysed ALMA (Xu et al., 2024), which applies a two-step finetuning procedure on Llama-2 using multilingual and parallel data. As a baseline, we included Llama-2 itself (Touvron et al., 2023) to measure the effect of translation-specific finetuning. All of these models were tested in both 7B and 13B parameter versions where feasible. For multilingual LLMs, we evaluated EuroLLM-9B-Inst (Martins et al., 2024), trained on 35 languages and instruction-tuned to cover all EU official languages and others. This allowed us to compare multilingual pretraining and instruction tuning against the translation-focused adaptation of the Tower models. As encoder–decoder baselines, we used NLLB-3.3B (Costa-jussà et al., 2022), which is trained at the sentence level, and a context-aware transformer-small model trained on IWSLT2017 TED data (Cettolo et al., 2012a).

We used two complementary datasets to assess document-level translation and pronoun resolution.

General translation assessment. We employed the IWSLT2017 TED data Cettolo et al. (2012a) in EN→DE and EN→FR. For EN→DE, we combined `tst2016--2017` into a test set containing 2,271 sentences across 23 documents. For EN→FR, we used `tst2015`, with 1,210 sentences in 12 documents. In both settings, we provided a context window of five previous source–target pairs.

Pronoun resolution. We relied on ContraPro, an annotated subset of OpenSubtitles (Müller et al., 2018; Lopes et al., 2020), which contains examples with ambiguous pronouns, their gold translations, and automatic annotations of antecedents. For EN→DE, the dataset covers the translation of the English pronoun "it" into the German pronouns "er", "sie", or "es". For EN→FR, it concerns the English pronouns "it" and "they" and their French counterparts "il", "elle", "ils", and "elles". The dataset is balanced, with 12K instances for EN→DE and 14K for EN→FR. We restricted experiments to cases where antecedents occurred one to five sentences earlier, again providing five source–target pairs as context at inference time.

Our methodology compared translation performance under original, perturbed, and random contexts. Perturbed contexts were constructed by randomly sampling sentences from other documents, producing individually legible but incoherent discourse, while random contexts consisted of token-level resampling. We then measured how context alteration affected both standard translation metrics and pronoun translation accuracy. In parallel, attribution analysis using ALTI-Logit (Ferrando

et al., 2023a) and input erasure (Li et al., 2016) quantified the share of attribution assigned to known antecedents during pronoun translation.

Our findings indicate that LLMs' improved document-translation performance compared to encoder-decoder models does not correspond to a similar improvement in pronoun translation performance, and in fact, standard encoder-decoder translation models seem currently better from the perspective of pronoun translation (over long contexts). Our analysis highlights the need for context-aware finetuning of LLMs with a focus on relevant parts of the context to improve their reliability for document-level translation.

This work is further reported in Mohammed and Niculae (2025).

3.2.3 Unlocking Latent Discourse Translation in LLMs Through Quality-Aware Decoding

Extending previous analysis we also investigated a key limitation of large language models (LLMs) in machine translation: their difficulty in handling discourse phenomena such as pronoun resolution, lexical cohesion, formality, and verb consistency at the document level. While LLMs have achieved strong sentence-level results, they often fall short compared to specialised neural machine translation systems in capturing cross-sentence dependencies, which are essential for human-like document-level quality.

To address this, we designed a comprehensive evaluation setup using the discourse-rich DELA corpus, TED2020, and WMT24++ datasets, covering six target languages (Brazilian Portuguese, German, French, Korean, Arabic, and Russian). We evaluated both LLMs specialised for translation (TowerInstruct-13B, EuroLLM-9B-Inst) and strong encoder–decoder baselines (NLLB-3.3B), complemented by general-purpose models (Gemma3-12B, Qwen3-14B). Our main methodological contribution lies in adopting Quality-Aware Decoding (QAD), specifically minimum Bayes risk (MBR) decoding with translation and discourse-specific utility functions. This allows the model to select translations that maximise expected quality rather than relying on greedy decoding, which often fails to exploit latent discourse knowledge. We further performed ablation studies with alternative inference strategies such as sample fusion and automatic post-editing, and validated results with both automatic metrics (BLEU, COMET, docCOMET) and human assessments.

The findings show that under greedy decoding, encoder–decoder models outperform LLMs in discourse-sensitive tasks. However, when applying QAD, LLMs significantly improve in discourse performance, surpassing encoder–decoder baselines across multiple languages and datasets. QAD-enhanced LLMs maintained lexical cohesion, improved pronoun resolution, and produced more semantically consistent outputs that aligned better with human preferences. Importantly, our analysis demonstrated that discourse knowledge is already encoded in LLMs, but requires quality-aware inference to be unlocked. We release human annotations on discourse phenomena for TED2020 to support further research. Our work contributes both methodological insights and resources to improve the discourse capabilities of LLM-based translation systems, paving the way for more coherent document-level translation.

This work is currently under submission to ARR (October 2025 edition).

3.3 Long-Context Modelling

Recent advances in model architecture and training have extended the maximum input length of LLMs to tens of thousands of tokens. However, longer context windows do not automatically translate into robust reasoning: models trained on shorter sequences often fail to generalise to much longer inputs, dispersing attention across irrelevant tokens and struggling with long-term consistency. This section addresses the challenges of long-context modelling from multiple perspectives: architectural innovations that promote sparse attention (§3.3.1 and §3.3.2), adapting context windows of multilingual LLMs (§3.3.3, analyses of long-term memory (§3.3.4), and retrieval-based methods for knowledge-intensive tasks (§3.3.5). These contributions demonstrate that long-context modelling requires more than scaling input length—it demands new mechanisms for attention, memory, and retrieval to ensure reliable performance.

3.3.1 Long-Context Generalization with Sparse Attention

We address the limitations of transformers in extrapolating to long-context inputs. Standard softmax attention produces dense distributions where all tokens receive nonzero weight, causing irrelevant tokens to accumulate probability mass, and reducing focus as sequence length increases. This dispersion impairs the retrieval of fixed-size patterns and hinders long-context generalization. To overcome this, we investigate sparse alternatives, specifically replacing softmax with α-entmax (Peters et al., 2019), which allows irrelevant tokens to receive exactly zero attention. Building on this, we introduce Adaptive-Scalable Entmax (ASEntmax), a variant that learns head-specific sparsity regimes, flexibly interpolating between dense and sparse attention. We further analyze how different positional encoding strategies interact with attention under long-context conditions. Experiments on synthetic pattern-retrieval tasks and long-sequence language modeling show that ASEntmax substantially outperforms softmax and fixed entmax, maintaining attention on salient tokens and avoiding degradation as inputs grow. Our findings highlight the need for joint design of sparse attention and positional encodings, enabling more robust long-context generalization in large language models.

This work is further reported in Vasylenko et al. (2025).

3.3.2 AdaSplash: Adaptive Sparse Flash Attention

A persistent challenge for adaptable multimodal generation and translation is the efficient processing of extended contexts, which frequently occur in discourse-level translation and multimodal inputs. Conventional transformer attention mechanisms, even when accelerated through FlashAttention, remain dense and assign non-negligible probability mass to a large number of irrelevant tokens, leading to reduced selectivity and inefficiencies at scale.

To address this limitation, we introduced **AdaSplash**, an adaptive sparse variant of FlashAttention. AdaSplash incorporates a dynamic sparsity mechanism that enables each attention head to interpolate between dense and sparse regimes, depending on input characteristics. This design allows the model to concentrate computation on salient tokens while discarding irrelevant ones, thereby combining the computational efficiency of FlashAttention with the representational benefits of sparse attention.

Empirical evaluations on long-context language modelling and translation benchmarks demonstrate that AdaSplash improves efficiency and scalability while maintaining or surpassing the

accuracy of dense baselines. In particular, the method preserves discourse-relevant information across extended inputs and reduces inference latency, making it directly applicable to document-level and multimodal translation scenarios.

By coupling adaptive sparsity with efficient attention computation, AdaSplash advances the architectural foundations required for scalable, context-aware multimodal generation. It also complements the long-context generalisation strategies explored in Vasylenko et al. (2025), together establishing a coherent framework for robust translation under extreme context lengths.

This work has been recognized as a oral paper at ICML 2025 and is further detailed in Gonçalves et al. (2025).

3.3.3 Adaptation of EuroLLM to long-context

Large language models (LLMs) are typically trained with limited context windows (a few thousand tokens) to reduce computational cost. However, these models often do not generalize natively to sequences longer than their training context window, leading to severe performance degradation. For real-world applications, the ability to process long documents or sustain extended conversations without loss of quality is crucial.

The most common method for long context adaptation is exact attention with rotary positional embeddings (RoPE) (Su et al., 2024). This approach involves training the model on long sequences while rescaling the frequency of the positional encoding, as in Position Interpolation (PI), NTK scaling, or YARN (Peng et al., 2024). While effective, this method is computationally expensive. Despite these advances, relatively little is known about the impact of data mixtures in long-context training. In particular, it remains unclear whether multilingual data can improve performance on long-context benchmarks.

Our goal was to extend EuroLLM-1.7B (Martins et al., 2024) to a 64K context length while minimizing degradation on short-context tasks.

We conducted experiments using the EuroLLM-1.7B models. All models were trained for 10 billion tokens with a maximum sequence length of 64K, using the position interpolation method for context extension.

To investigate the impact of multilingual long-context data, we created two distinct data mixes. The first is an English-centric mix composed of data from the code and book domains. The second is a multilingual mix containing up-sampled sequences from non-English data. We filter the annealing data based on sequence length. The annealing data contains a small fraction of high-quality data that are used to further train in the annealing stage (cosine schedule). After filtering on long sequences, 98% of the tokens are from the Books domain. We also create a third data mix by adding high-quality short sequences ($\leq 4K$ tokens) to the English-centric mix to mitigate deterioration on short context benchmarks.

Impact of multilingual sequences in datamix. We compare the performance of models adapted on long context with long english-centric data only and multilingual data. Both models are trained with same hyperparameters on 10B tokens with maximal sequence length. Models are tested on Ruler (Hsieh et al., 2024), Needle in a Haystack and LongBenchV2 benches (Bai et al., 2025).

Impact of short mix. We measure how adding a high quality short mix data help preserve performances of models on short context tasks. We compare one model trained with long sequences only from code and book domains with models trained with an additional 30% of tokens coming from short annealing data. We evaluated the performance on MMLU and ARC for short tasks and LongbenchV2 for long context. We report performance and compare with the Instruct version of EuroLLM-2B (4K).

Our experiments revealed several key insights, which we summarise below.

- Short-sequence data is beneficial. Including a mix of short-sequence data mitigates performance degradation on short-context tasks like MMLU and ARC.
- Multilingual data showed no improvement on English tasks. The multilingual data mix did not improve performance on the English-based long-context benchmarks we tested.
- Longer sequences and more tokens improve performance. We confirmed that training on longer sequences and for more tokens consistently increases model performance, a finding supported by prior work.
- RoPE frequency scaling is effective. Further increasing the RoPE frequency beyond the initial scaling factor also boosts performance.
- Training strategy matters. A one-step continued pretraining strategy appears more efficient for context extension than a multi-step approach.

This work, which extends the contributions of Martins et al. (2024), is currently in progress and will be submitted for publication soon.

3.3.4 What Makes Memory Work? Evaluating Long-Term Memory for Large Language Models

Large language model (LLM) agents need to maintain coherence over long interactions, learn from experience, and operate effectively in long-context settings. While research has focused on the development of more complex memory systems, it remains unclear which memory architectures are most effective for long-context conversational tasks. In this paper (Terranova et al., 2025) we present a systematic evaluation of memory-augmented methods across open-weight LLMs using LoCoMo, a benchmark of long-context synthetic dialogues annotated for diverse question-answering tasks. We analyse different strategies: full-context prompting, retrieval-augmented generation (RAG), agentic memory (A-Mem), episodic memory through in-context learning, and procedural memory through prompt optimization. Our findings confirm that while full-context prompting performs strongly in more straightforward QA tasks, its performance in complex questions and its token efficiency lag behind memory-augmented approaches. Instruction-tuned models benefit most from episodic and agentic memory, while base models perform well with simpler retrieval. These results can offer insights into developing efficient, scalable, and reliable memory systems and highlight trade-offs between performance, interpretability, and resource use.

Our goal is to test different memory strategies on a task which models realistic long-context conversations, which is why we decided to use LoCoMo (Long-term Conversational Memory) (Maharana et al., 2024). We replicate the A-Mem architecture as introduced by Xu et al. (2025), where the

system maintains structured memory notes representing each utterance in the conversational data. As the memory agent is presented with a new utterance for the same conversation, it updates its memory with new entries and can make decisions about updating old ones and their connections.

We confirm prior findings and extend the literature by comparing simple and complex implementations of semantic memory, and minimal implementations of episodic and procedural memory. This allows us to identify the different purposes that memory can serve in knowledge-intensive and resource-constrained tasks, and identify what makes each method suited for specific use cases.

While appending the whole conversation context to the prompt serves as a strong baseline, this approach suffers from inefficiency, poor scalability, limited interpretability, and vulnerability to context-length issues that can degrade performance. Our results suggest that most foundation and instruction-tuned models can benefit from simple RAG approaches, but instruction-tuned models can make use of their improved instruction-following and reasoning capabilities and also show strong performance with more complex approaches to semantic memory like A-Mem and episodic memory integration.

This work is further detailed in Terranova et al. (2025).

3.3.5 Multilingual Retrieval-Augmented Generation for Knowledge-Intensive Task

Retrieval-augmented generation (RAG) has become a cornerstone of contemporary NLP, enhancing large language models (LLMs) by allowing them to access richer factual contexts through in-context retrieval. While effective in monolingual settings, especially in English, its use in multilingual tasks remains unexplored. We investigate multilingual RAG beyond the traditional English-centric setting by comparing four pipelines: monoRAG (monolingual retrieval in the query language), tRAG (question translation to English before retrieval), MultiRAG (retrieval over a multilingual index), and CrossRAG (MultiRAG followed by document-level translation to a single language—English—while still answering in the user's language). Our motivation is twofold: (i) naïve question translation can distort queries and misdirect retrieval; and (ii) heterogeneous, cross-lingual evidence retrieved by MultiRAG can confuse models at inference time (e.g., answer language drift), especially for low-resource languages. We therefore ask whether consolidating evidence into a single processing language can preserve the broader coverage of multilingual retrieval while improving the usability of evidence at generation time.

We evaluate on three knowledge-intensive QA benchmarks designed for multilingual open-domain settings: MKQA, MLQA, and XOR-TyDi QA. We operate using different retrieval engines adapted for multilingual embeddings over Wikimedia dumps, and we re-rank and pass the top-5 passages to the generator. Primary models are GPT-4o, Llama-3-8B-Instruct, and Command-R-35B, decoded greedily (temperature 0). Accuracy follows flexible exact match (with character 3-gram recall also reported), and answer language correctness is measured with OpenLID. Translation in tRAG/CrossRAG is performed using Google Translate (main results), with an ablation using GPT-4o as the translator. We further test Translation-Following fine-tuning to strengthen Llama-3-8B's multilingual translation ability. High- vs low-resource strata are defined using web and Wikipedia language distributions to contextualise retrieval coverage.

The main results demonstrate that:

- *RAG helps in every language, but breadth matters.* Relative to no-RAG, monoRAG already improves accuracy. Extending retrieval to MultiRAG yields further gains over monoRAG

(+5.4, +7.1, +5.7, respectively), reflecting the value of multilingual coverage; effects are larger for low-resource languages.

- *Consolidating evidence boosts robustness.* CrossRAG (document translation to English before generation) consistently outperforms MultiRAG, with bigger lifts in low-resource languages. Using GPT-4o as a translator further improves CrossRAG on most datasets, underscoring that the same retrieved facts are more usable once linguistically aligned.
- *Alternatives to translation at inference.* Lightweight Translation-Following fine-tuning of Llama-3-8B likewise narrows the gap to CrossRAG (e.g., +11.6 points over MultiRAG on MKQA; +7.5 on MLQA), but introduces training cost.
- *Better language control and stability.* CrossRAG raises the rate of answers produced in the correct query language and is robust to document order, while MultiRAG is more order-sensitive (notably for Llama-3-8B).
- *Retriever choice is not the bottleneck.* Changing the Retrieval system yields comparable averages, suggesting that the principal success comes from how multilingual evidence is presented to the LLM rather than from the retriever per se.

Overall, the results recommend a practical recipe for multilingual, knowledge-intensive QA: retrieve widely; then normalise the evidence language (CrossRAG) to stabilise reasoning and output, with robust returns in low-resource settings.

This work is further reported in Ranaldi et al. (2025b).

3.4 Conversation-aware, chat translation

Following from RP1, we also expanded our contributions towards the better use of context for dialogue translation in bilingual settings. This section focuses on methods for conversation-aware translation as well as the evaluation of chat MT. We present frameworks for incorporating conversational history into translation systems (§3.4.1), practical adaptations for bilingual customer support scenarios (§3.4.2), large-scale benchmarking through the WMT 2024 Shared Task on Chat Translation (§3.4.3), and critical analyses of evaluation practices (§3.4.4). Together, these works demonstrate how conversational MT can progress from system design, to applied use cases, to community-wide evaluation, offering a comprehensive view of context exploitation in dialogue.

3.4.1 A context-aware framework for translation-mediated conversations

Automatic translation systems offer a powerful solution to bridge language barriers in scenarios where participants do not share a common language. However, these systems can introduce errors leading to misunderstandings and conversation breakdown. A key issue is that current systems fail to incorporate the rich contextual information necessary to resolve ambiguities and omitted details, resulting in literal, inappropriate, or misaligned translations. In this work, we present a framework to improve large language model-based translation systems by incorporating contextual information in bilingual conversational settings during training and inference. We validate our proposed framework on two task-oriented domains: customer chat and user-assistant interaction. Across both settings, the system produced by our framework-TowerChat-consistently results in

better translations than state-of-the-art systems like GPT-4o and TowerInstruct, as measured by multiple automatic translation quality metrics on several language pairs. We also show that the resulting model leverages context in an intended and interpretable way, improving consistency between the conveyed message and the generated translations.

This work is further reported in Pombal et al. (2024a).

3.4.2 Improving context usage for translating bilingual customer support chat with large language models

This work describes Unbabel+IT's submission to the Chat Shared Task held at the Workshop of Machine Translation 2024. The task focuses on translating customer support chats between agents and customers communicating in different languages. We present two strategies for adapting state-of-the-art language models to better utilize contextual information when translating such conversations. Our training strategy involves finetuning the model on chat datasets with context-augmented instructions, resulting in a specialized model, TOWERCHAT. For inference, we propose a novel quality-aware decoding approach that leverages a context-aware metric, CONTEXTCOMET, to select the optimal translation from a pool of candidates. We evaluate our proposed approach on the official shared task datasets for ten language pairs, showing that our submission consistently outperforms baselines on all and competing systems on 8 out of 10 language pairs across multiple automated metrics. Remarkably, TOWERCHAT outperforms our contrastive submission based on the much larger TOWER-V2-70B model while being 10× smaller. According to human evaluation, our system outperforms all other systems and baselines across all language pairs. These results underscore the importance of context-aware training and inference in handling complex bilingual dialogues.

This work has been published in (Pombal et al., 2024b).

3.4.3 Findings of the WMT 2024 Shared Task on Chat Translation

We recognise that translating bilingual chat conversations presents unique challenges compared to standard machine translation. Customer–agent dialogues are characterised by short, fragmented turns, omitted information, pragmatic nuances, and frequent shifts in speaker roles. These properties demand models that go beyond sentence-level adequacy and account for discourse phenomena, consistency, and conversational flow. To address these challenges and to stimulate progress in this underexplored domain, we co-organised the third edition of the WMT Shared Task on Chat Translation.

For this edition, we provided datasets spanning five language pairs: English–German, English–French, and English–Brazilian Portuguese from previous rounds, and two newly introduced pairs, English–Dutch and English–Korean. The data consisted of authentic customer-support conversations, annotated with speaker roles and turn boundaries, thereby enabling evaluation at both the turn and conversation levels.

We invited participants to develop systems for both translation directions (customer→agent and agent→customer), and received 22 primary and 32 contrastive submissions from eight teams worldwide. Submitted models included fine-tuned neural MT systems and large language model–based approaches, some explicitly incorporating dialogue context.

Our evaluation combined automatic metrics with human judgments through direct assessment. This design allowed us to assess not only adequacy and fluency at the level of individual turns, but also conversation-level properties such as consistency of terminology, handling of anaphora, and maintenance of register.

Our findings reveal that while state-of-the-art systems achieve strong sentence-level performance, their ability to leverage dialogue context remains limited. Context-aware approaches showed improvements in pronoun resolution and coherence across turns, yet challenges persist in ensuring conversation-level consistency, especially in lower-resource language pairs such as English–Korean and English–Dutch. These results confirm the importance of chat translation as a benchmark for contextual MT and highlight the need for further innovation in modelling conversational structure.

This work is further reported in Mohammed et al. (2024).

3.4.4 Assessing the Role of Context in Chat Translation Evaluation: Is Context Helpful and Under What Conditions?

Despite the recent success of automatic metrics for assessing translation quality, their application in evaluating the quality of machine-translated chats has been limited. Unlike more structured texts like news, chat conversations are often unstructured, short, and heavily reliant on contextual information. This poses questions about the reliability of existing sentence-level metrics in this domain as well as the role of context in assessing the translation quality. Motivated by this, we conduct a meta-evaluation of existing automatic metrics, primarily designed for structured domains such as news, to assess the quality of machine-translated chats. We find that reference-free metrics lag behind reference-based ones, especially when evaluating translation quality in out-of-English settings. We then investigate how incorporating conversational contextual information in these metrics for sentence-level evaluation affects their performance. Our findings show that augmenting neural learned metrics with contextual information helps improve correlation with human judgments in the reference-free scenario and when evaluating translations in out-of-English settings. Finally, we propose a new evaluation metric, Context-MQM, that utilizes bilingual context with a large language model (LLM) and further validate that adding context helps even for LLM-based evaluation metrics.

This work is further reported in Agrawal et al. (2024b).

4 Task 4.3: Simultaneous translation (UEDIN*, NAV)

Summary of completed work

As mentioned in the first reporting period (D4.1), our contibutions to Task 4.3 started earlier than originally planned and focused on advancing methods for simultaneous speech translation, with an emphasis on retranslation-based systems and reducing flicker effects. Initial work demonstrated that naïve retranslation often leads to unstable outputs, where partial hypotheses are repeatedly overwritten. To address this, the team investigated self-training and other stabilisation techniques, showing that exposing models to their own partial outputs during training can substantially reduce flicker and improve user experience. Participation in the IWSLT 2023 Simultaneous Speech

Translation Shared Task provided a valuable testbed for benchmarking these methods under realistic conditions, yielding insights into latency–quality trade-offs and the challenges of adapting instruction-tuned LLMs for real-time scenarios.

These early efforts established an empirical foundation for evaluating simultaneous translation systems, clarifying the limitations of current approaches and informing the decision to adjust the scope of Task 4.3 in RP2. In this sense, the RP1 achievements in Task 4.3 were crucial for identifying both the potential and the bottlenecks of simultaneous translation, paving the way for the strategic redirection pursued in RP2 as described in Section 4.1.

Previously Associated Publications

- Sen et al. (2023). *Self-Training Reduces Flicker in Retranslation-Based Simultaneous Translation*. EACL 2023.
- Agarwal et al. (2023). *Findings of the IWSLT 2023 Evaluation Campaign*. IWSLT 2023.

4.1 Redirection of research efforts

UTTER aims to provide unified multilingual, multimodal pretrained models to support real-world applications, such as online and hybrid meetings. In such a scenario, simultaneous speech translation is an important component. However, existing work on simultaneous translation is based on small models and does not fully exploit the strong capability of LLMs. Meanwhile, making LLMs to understand speech inputs remained challenging and underexplored. In the second-half of the project, we therefore prioritized our research focus to the broader paradigm of integrating speech into LLMs.

For reporting consistency we are summarising in Table 3 the contributions (publications and repositories) reported throughout UTTER in this task. However, as reported in Section 4.2 since the publications relevant to the new research direction of this task are reported in detail in other reporting documents (D2.2, D3.2, and D6.2), as well as in Section 2.5, this table only summarizes the associated publications in RP1.

Period	Venue	Paper	ACK	Code	ACK	citations
RP1						
	WMT	Sen et al. (2023)	✔	✗	-	5
	IWSLT	Agarwal et al. (2023)	NA	✗	-	NA

Table 3: Research outputs (manuscripts and code) from T4.3. Note that for ACK{nowledgements} *R* signifies requested/under process acknowledgement and *NA* signifies 'not applicable'. Citations refer to publications and are obtained from Google Scholar as of September 30, 2025.

4.2 Contributions

Below is a list of publications related to our shifted focus of integrating speech into LLM. We also provide the related WPs where their details are reported:

- *From TOWER to SPIRE: Adding the Speech Modality to a Text-Only LLM*, in D16 D3.2 Final report on XR models Task 3.1.
- *Prepending or Cross-Attention for Speech-To-Text? An Empirical Comparison*, in D22 D6.2 Final report on efficient inference Task 6.1
- *Findings of the IWSLT 2024-2025 Evaluation Campaign*, in D14 D2.2 Final report on data and resources Task 2.1.
- *IWSLT 2025 Instruction Following - IT submission*, in Task 4.1.
- *IWSLT 2025 Instruction Following - NAV submission*, in Task 4.1.

4.3 Ongoing work on Full-Duplex Conversational Models

In addition to speech-LLM that generates text, we are also developing full-duplex (FD) (D'efossez et al., 2024) conversational models that can support simultaneous listening and speaking.

FD model is important for real conversations, where people interrupt each other and do backchanneling. Despite a plethora of recent work on FD modeling, these models are never trained to interrupt/backchannel the user (Arora et al., 2025; Hu et al., 2025). Furthermore, it is unclear how they should be properly evaluated. Existing work mainly used ASR-TTS pipelines and LLMs as a judge (Peng et al., 2025), but these do not actually test the things we are looking for, such as real dialogues and the ability to converse like humans.

Therefore, we are building a model that can talk to other full-duplex models and react to them. Such a model can test how well the existing FD models respond to interruptions and backchannelings, and how well they are able to keep conversation flows. SOTA models right now cannot be used for this, because they are trained to never interrupt/backchannel. Currently, we have already examined various codecs and LLMs, which serve as the backbone of our model. In addition, we are examining the pros and cons between single-stream and multi-stream modelling for FD systems.

5 Impact

WP4 has delivered substantial scientific and practical impact across three core dimensions: adaptation strategies, contextualisation and context-aware models, and (upon our research redirection) the incorporation of speech as a modality in LLMs and speech translation models. Together, these efforts expanded the scientific frontier of language technology and seeded community resources that will outlast the project. By advancing state-of-the-art architectures, adaptation strategies, and evaluation practices, WP4 has contributed both foundational research and widely used community resources.

This work package has produced over 50 research papers across Tasks 4.1 (Adaptation), 4.2 (Context and emotion aware models), and 4.3 (Simultaneous / Speech Translation). These include publications at top NLP and ML venues (ACL, NAACL, EACL, EMNLP, ICML, NeurIPS, COLM) and journals (TACL), as well as specialised conferences and workshops (IWSLT, WMT, EAMT, MT Summit). Outputs from RP1 and RP2 together have already gathered over 2,000 citations, demonstrating broad uptake in the community. Beyond publications, from a dissemination and community standpoint, WP4 partners have released public code and datasets (counting 30 associated Github repositories), and engaged actively with shared task communities (counting 12 WMT Shared Task contributions and 3 IWSLT Shared Task contributions). Among the code-related contributions, it is worth emphasising xCOMET (RP1-D5.1 §3.2.3), which has been recognised by European Commission's Innovation Radar while relevant COMET variants overall count more than 10,000 downloads overall (on Unbabel's HuggingFace repository [1]. Altogether, these contributions ensure broad visibility and uptake, enabling reuse and extension by both academic researchers and industrial stakeholders.

Overall, WP4's impact lies in creating adaptable, explainable, and contextualised language generation methods that advance the scientific state of the art, while at the same time seeding practical pathways for more robust, context-aware, and multimodal translation technologies.

6 Conclusion

WP4 aimed to investigate adaptation and contextualisation strategies for large language models in multilingual, multimodal, and interactive translation scenarios, a key direction towards achieving UTTER's objectives. The work was structured across three tasks: Task 4.1 on adaptable multimodal generation and translation, Task 4.2 on context- and discourse-aware translation and interaction, and Task 4.3 on simultaneous translation, which was later redirected toward speech translation and conversational models with speech capabilities.

Our contributions in Task 4.1, advanced the state of the art by exploring new architectural directions (State Space Models) and their potential in MT, improving methods for alignment with human feedback, and exploring further inference-time strategies such as reranking. We tackled challenges that focused on cross-lingual transfer, multilingual reasoning, and low-resource translation. This line of work not only delivered strong scientific contributions but also resulted in the release of models, datasets, and evaluation tools that are now available for the broader community.

In Task 4.2, we elaborated the role of context in multilingual translation and generation. By developing approaches for document-level and conversational translation, releasing benchmarks for chat

[1] https://huggingface.co/Unbabel/collections

translation, and designing frameworks for translation-mediated conversations, we provided both methodological insights and practical tools. These contributions enhance the reliability of multilingual interaction and directly support future applications in extended-reality and human–computer communication.

In Task 4.3, we redirected the original focus on simultaneous translation toward the modelling of full-duplex multilingual conversations. This shift ensured continued relevance to UTTER's objectives and produced frameworks and prototypes that address real-world challenges such as latency, turn-taking, and robustness in interactive multilingual scenarios. The outcomes demonstrate not only technical feasibility but also open pathways for future research and innovation in real-time multilingual communication.

Progress across WP4 has been balanced across the two reporting periods: RP1 concentrated on laying the foundations with advances in adaptation methods and contextual resources, while RP2 consolidated and expanded the work with strong results in alignment, inference-time adaptation, cross-lingual modelling, and discourse-aware translation. The adjustment of Task 4.3 illustrates WP4's agility in responding to emerging needs and opportunities, ensuring that results remain impactful and relevant.

The outputs of WP4 include a strong portfolio of peer-reviewed publications, open-source models and datasets, and the co-organisation of benchmarking tasks that are influencing current research directions in multilingual and multimodal NLP. Overall, WP4 has not only advanced the state of the art in adaptability, contextualisation, and multimodal translation but also delivered tangible resources and insights that will support the research community and industrial innovation well beyond the lifetime of UTTER.

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.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.1. URL https://aclanthology.org/2023.iwslt-1.1.

Sweta Agrawal, José G. C. De Souza, Ricardo Rei, António Farinhas, Gonçalo Faria, Patrick Fernandes, Nuno M. Guerreiro, and André F. T. Martins. Modeling user preferences with automatic metrics: Creating a high-quality preference dataset for machine translation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.803.

Sweta Agrawal, Amin Farajian, Patrick Fernandes, Ricardo Rei, and André F. T. Martins. Assessing the role of context in chat translation evaluation: Is context helpful and under what conditions? *Transactions of the Association for Computational Linguistics*, 12:1250–1267, 2024b. doi: 10.1162/tacl_a_00700. URL https://aclanthology.org/2024.tacl-1.69/.

Duarte Alves, Nuno Guerreiro, João Alves, José Pombal, Ricardo Rei, José de Souza, Pierre Colombo, and André FT Martins. Steering large language models for machine translation with

finetuning and in-context learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11127–11148, 2023. URL https://aclanthology.org/2023.findings-emnlp.744/.

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 G. C. de Souza, and Andre Martins. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=EHPns3hVkj.

Antonios Anastasopoulos, Núria Bel, Marta R. Costa-jussà, Siddharth Dalmia, Mikel L. Forcada, Qin Gao, Ulrich Germann, Silja Hartmann, Gholamreza Haffari, Eduard Hovy, Javier Iranzo-Sánchez, Prachi Khare, Rebecca Knowles, Philipp Koehn, Xutai Li, Zuchao Li, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, John E. Ortega, Vassilis Papavassiliou, Qinlan Shen, Clifford Sikasote, Mengzhou Sun, Xiang Tong, Chara Tsoukala, Tong Wang, Zongyuan Wang, and Rui Zhang. Tico-19: the translation initiative for covid-19. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 4983–4992, Marseille, France, May 2020. European Language Resources Association. URL https://aclanthology.org/2020.lrec-1.614.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung yi Lee, Karen Livescu, and Shinji Watanabe. On the landscape of spoken language models: A comprehensive survey. *ArXiv*, abs/2504.08528, 2025. URL https://api.semanticscholar.org/CorpusID:277741224.

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/.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc V. Le, and Charles Sutton. Program synthesis with large language models. In *Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2108.07732.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.183. URL https://aclanthology.org/2025.acl-long.183/.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, Ann Arbor, Michigan, USA, 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023.

Christos Baziotis, Biao Zhang, Alexandra Birch, and Barry Haddow. When does monolingual data help multilingual translation: The role of domain and model scale, 2023. URL https://arxiv.org/abs/2305.14124.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/e995f98d56967d946471af29d7bf99f5-Abstract.html.

Sumanta Bhattacharyya, Matteo Negri, and Marco Turchi. Energy-based reranking: Improving neural machine translation using energy scores. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, Online, 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.acl-long.92.

Frederic Blain, Chrysoula Zerva, Ricardo Rei, Nuno M. Guerreiro, Diptesh Kanojia, José G. C. de Souza, Beatriz Silva, Tânia Vaz, Yan Jingxuan, Fatemeh Azadi, Constantin Orasan, and André Martins. Findings of the WMT 2023 shared task on quality estimation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 629–653, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.52. URL https://aclanthology.org/2023.wmt-1.52.

Nikolay Bogoychev and Pinzhen Chen. Terminology-aware translation with constrained decoding and large language model prompting. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 890–896, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.80. URL https://aclanthology.org/2023.wmt-1.80.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, et al. Findings of the 2015 workshop on statistical machine translation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal, 2015. Association for Computational Linguistics. URL https://aclanthology.org/W15-3001.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, et al. Findings of the 2016 conference on machine translation. In *Proceedings of the First Conference on Machine Translation (WMT16)*, Berlin, Germany, 2016. Association for Computational Linguistics. URL https://aclanthology.org/W16-2201.

Mauro Cettolo, Christian Girardi, and Marcello Federico. WIT3: web inventory of transcribed and translated talks. In Mauro Cettolo, Marcello Federico, Lucia Specia, and Andy Way, editors, *Proceedings of the 16th Annual conference of the European Association for Machine Translation, EAMT 2012, Trento, Italy, May 28-30, 2012*, pages 261–268. European Association for Machine Translation, 2012a. URL https://aclanthology.org/2012.eamt-1.60/.

Mauro Cettolo, Christian Girardi, and Marcello Federico. Wit3: Web inventory of transcribed and translated talks. In *Proceedings of the 16th Conference of the European Association for*

Machine Translation, Trento, Italy, 2012b. European Association for Machine Translation. URL https://aclanthology.org/2012.eamt-1.60.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsu Sudoh, and Koichiro Yoshino. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Workshop on Spoken Language Translation (IWSLT 2017)*, Tokyo, Japan, 2017. IWSLT. URL https://aclanthology.org/2017.iwslt-1.1.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc 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. *CoRR*, abs/2207.04672, 2022. doi: 10.48550/ARXIV.2207.04672. URL https://doi.org/10.48550/arXiv.2207.04672.

Alexandre D'efossez, Laurent Mazar'e, Manu Orsini, Am'elie Royer, Patrick P'erez, Herv'e J'egou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. *ArXiv*, abs/2410.00037, 2024. URL https://api.semanticscholar.org/CorpusID:273022979.

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.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. Findings of the AmericasNLP 2024 Shared Task on Machine Translation into Indigenous Languages. 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 236–246, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.americasnlp-1.28. URL https://aclanthology.org/2024.americasnlp-1.28/.

Bryan Eikema and Wilker Aziz. Sampling-based minimum bayes risk decoding for neural machine translation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.emnlp-main.624.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. Beyond English-Centric Multilingual Machine Translation, October 2020. URL http://arxiv.org/abs/2010.11125. arXiv:2010.11125 [cs].

Ana C Farinha, M Amin Farajian, Marianna Buchicchio, Patrick Fernandes, José GC De Souza, Helena Moniz, and André FT Martins. Findings of the wmt 2022 shared task on chat translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 724–743, 2022. URL https://aclanthology.org/2022.wmt-1.70/.

António Farinhas, José de Souza, and Andre Martins. An empirical study of translation hypothesis ensembling with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.733. URL https://aclanthology.org/2023.emnlp-main.733.

António Farinhas, Haau-Sing Li, and André Martins. Reranking laws for language generation: A communication-theoretic perspective. *Advances in Neural Information Processing Systems*, 37: 111074–111105, 2024.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André Martins, Graham Neubig, Ankush Garg, Jonathan Clark, Markus Freitag, and Orhan Firat. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1066–1083, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.100. URL https://aclanthology.org/2023.wmt-1.100/.

Patrick Fernandes, Aman Madaan, Emmy Liu, António Farinhas, Pedro Henrique Martins, Amanda Bertsch, José GC de Souza, Shuyan Zhou, Tongshuang Wu, Graham Neubig, et al. Bridging the gap: A survey on integrating (human) feedback for natural language generation. *Transactions of the Association for Computational Linguistics*, 11:1643–1668, 2023b. URL https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00626/118795.

Patrick Fernandes, Kayo Yin, Emmy Liu, André FT Martins, and Graham Neubig. When does translation require context? a data-driven, multilingual exploration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 606–626, 2023c. URL https://aclanthology.org/2023.acl-long.36/.

Javier Ferrando, Nuno M. Guerreiro, Marcos Treviso, and André F. T. Martins. Towards interpreting sequence-to-sequence models: A visual analytics approach. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2022. URL https://aclanthology.org/2022.acl-long.123.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. Explaining how transformers use context to build predictions. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5486–5513. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.ACL-LONG.301. URL https://doi.org/10.18653/v1/2023.acl-long.301.

Javier Ferrando, Nuno M. Guerreiro, Marcos Treviso, and André F. T. Martins. Attribution analysis for sequence-to-sequence models via logit lens. In *Proceedings of the 61st Annual Meeting of the*

Association for Computational Linguistics (ACL). Association for Computational Linguistics, 2023b. URL https://aclanthology.org/2023.acl-long.456.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.51. URL https://aclanthology.org/2023.wmt-1.51.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frédéric Blain, Tom Kocmi, Jiayi Wang, David I. Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. Are llms breaking mt metrics? results of the wmt24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 47–81, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.wmt-1.2.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Goncalo Emanuel Cavaco Gomes, Chrysoula Zerva, and Bruno Martins. Evaluation of multilingual image captioning: How far can we get with CLIP models? In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5156–5175, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.287. URL https://aclanthology.org/2025.findings-naacl.287/.

Nuno Gonçalves, Marcos V Treviso, and Andre Martins. Adasplash: Adaptive sparse flash attention. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=OWIPDWhUcO.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The FLORES-101 Evaluation Benchmark for Low-Resource and Multilingual Machine Translation. *arXiv:2106.03193 [cs]*, June 2021. URL http://arxiv.org/abs/2106.03193. arXiv: 2106.03193.

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

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *arXiv preprint arXiv:2310.10482*, 2023.

Sumire Honda, Patrick Fernandes, and Chrysoula Zerva. Context-aware neural machine translation for english-japanese business scene dialogues. In *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 272–285, 2023. URL https://arxiv.org/abs/2311.11976.

Cheng-Yu Hsieh, Yao Li, Ting-Han Chen, Yizhong Zhu, Kai-Wei Chang, and William Yang Wang. Ruler: Evaluating long-context language models with progressive retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.acl-long.101.

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. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Zelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. Efficient and direct duplex modeling for speech-to-speech language model. *ArXiv*, abs/2505.15670, 2025. URL https://api.semanticscholar.org/CorpusID:278782646.

J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233, 2020.

Vivek Iyer, Edoardo Barba, Alexandra Birch, Jeff Pan, and Roberto Navigli. Code-switching with word senses for pretraining in neural machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12889–12901, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.859. URL https://aclanthology.org/2023.findings-emnlp.859.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. Towards effective disambiguation for machine translation with large language models. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 482–495, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.44. URL https://aclanthology.org/2023.wmt-1.44.

Vivek Iyer, Bhavitvya Malik, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. Quality or Quantity? On Data Scale and Diversity in Adapting Large Language Models for Low-Resource Translation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1393–1409, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.128. URL https://aclanthology.org/2024.wmt-1.128/.

Samy Jelassi, Tri Dao, Albert Gu, Sara Sabour, and Jascha Sohl-Dickstein. Vision transformers need registers. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*. PMLR, 2023. URL https://arxiv.org/abs/2309.16588.

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.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. Metricx-23: The google submission to the wmt 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, 2023.

Mateusz Klimaszewski, Piotr Andruszkiewicz, and Alexandra Birch. Is Modularity Transferable? a Case Study through the Lens of Knowledge Distillation, 2024. Accepted to the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation.

Philipp Koehn and Rebecca Knowles. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, Vancouver, Canada, 2017. Association for Computational Linguistics. URL https://aclanthology.org/W17-3204.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi kiu Lo, Vilém Zouhar, Frédéric Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David I. Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. Findings of the wmt25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Seventh Conference on Machine Translation (WMT 2025)*, Miami, Florida, USA, 2025. Association for Computational Linguistics. URL https://www2.statmt.org/wmt25/mteval-subtask.html.

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/.

Chia-Hsuan Lee, Szu-Lin Wu, Chi-Liang Liu, and Hung-yi Lee. Spoken squad: A study of mitigating the impact of speech recognition errors on listening comprehension. *Proc. Interspeech 2018*, pages 3459–3463, 2018.

Jiwei Li, Will Monroe, and Dan Jurafsky. Understanding neural networks through representation erasure. *CoRR*, abs/1612.08220, 2016. URL http://arxiv.org/abs/1612.08220.

Peiqin Lin, Chengzhi Hu, Zheyu Zhang, André FT Martins, and Hinrich Schütze. mplm-sim: Better cross-lingual similarity and transfer in multilingual pretrained language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 276–310, 2024.

Peiqin Lin, Andre Martins, and Hinrich Schuetze. XAMPLER: Learning to retrieve cross-lingual in-context examples. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3968–3977, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.221. URL https://aclanthology.org/2025.findings-naacl.221/.

António V. Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. Document-level neural MT: A systematic comparison. In Mikel L. Forcada, André Martins, Helena Moniz, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof Arenas, Mary Nurminen, Lena Marg, Sara Fumega, Bruno Martins, Fernando Batista, Luísa Coheur, Carla Parra Escartín, and Isabel Trancoso, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*, pages 225–234. European Association for Machine Translation, 2020. URL https://aclanthology.org/2020.eamt-1.24/.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*, 2024.

Pedro Henrique Martins, João Alves, Tânia Vaz, Madalena Gonçalves, Beatriz Silva, Marianna Buchicchio, José GC de Souza, and André FT Martins. Empirical assessment of knn-mt for real-world translation scenarios. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 115–124, 2023. URL https://aclanthology.org/2023.eamt-1.12.

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, 2024. URL https://arxiv.org/abs/2409.16235.

John Mendonça, Patrícia Pereira, Miguel Menezes, Vera Cabarrão, Ana C Farinha, Helena Moniz, João Paulo Carvalho, Alon Lavie, and Isabel Trancoso. Dialogue quality and emotion annotations for customer support conversations. *arXiv preprint arXiv:2311.13910*, 2023. URL https://arxiv.org/abs/2311.13910.

Miguel Menezes, Amin Farajian, Lisbon Unbabel, Portugal Helena Moniz, and Joao Graça. A context-aware annotation framework for customer support live chat machine translation. *MT Summit 2023*, page 286, 2023.

Joel Mire, Zubin Trivadi Aysola, Daniel Chechelnitsky, Nicholas Deas, Chrysoula Zerva, and Maarten Sap. Rejected dialects: Biases against African American language in reward models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7468–7487, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.417. URL https://aclanthology.org/2025.findings-naacl.417/.

Wafaa Mohammed and Vlad Niculae. On measuring context utilization in document-level MT systems. In Yvette Graham and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1633–1643, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-eacl.113.

Wafaa Mohammed and Vlad Niculae. Context-aware or context-insensitive? assessing LLMs' performance in document-level translation. In Pierrette Bouillon, Johanna Gerlach, Sabrina Girletti, Lise Volkart, Raphael Rubino, Rico Sennrich, Ana C. Farinha, Marco Gaido, Joke Daems, Dorothy Kenny, Helena Moniz, and Sara Szoc, editors, *Proceedings of Machine Translation Summit XX: Volume 1*, pages 126–137, Geneva, Switzerland, June 2025. European Association for Machine Translation. ISBN 978-2-9701897-0-1. URL https://aclanthology.org/2025.mtsummit-1.10/.

Wafaa Mohammed, Sweta Agrawal, Amin Farajian, Vera Cabarrão, Bryan Eikema, Ana C Farinha, and José G. C. De Souza. Findings of the WMT 2024 shared task on chat translation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 701–714, Miami, Florida, USA, November

2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.59. URL https://aclanthology.org/2024.wmt-1.59/.

Mathias Müller, Annette Rios, Elena Voita, and Rico Sennrich. A large-scale test set for the evaluation of context-aware pronoun translation in neural machine translation. In Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 61–72. Association for Computational Linguistics, 2018. doi: 10.18653/v1/w18-6307. URL https://doi.org/10.18653/v1/w18-6307.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook fair's wmt19 news translation task submission. In *Proceedings of the Fourth Conference on Machine Translation (WMT19)*, Florence, Italy, 2019. Association for Computational Linguistics. URL https://aclanthology.org/W19-5303.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. Analyzing uncertainty in neural machine translation. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018. PMLR. URL http://proceedings.mlr.press/v80/ott18a.html.

Proyag Pal and Kenneth Heafield. Cheating to identify hard problems for neural machine translation. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1620–1631, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.120. URL https://aclanthology.org/2023.findings-eacl.120.

Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. Findings of the WMT 2023 Shared Task on Low-Resource Indic Language Translation. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.56. URL https://aclanthology.org/2023.wmt-1.56/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. URL https://aclanthology.org/P02-1040.

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=wHBfxhZu1u.

Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Eng Siong Chng. FD-Bench: A Full-Duplex Benchmarking Pipeline Designed for Full Duplex Spoken Dialogue Systems. In *Interspeech 2025*, pages 176–180, 2025. doi: 10.21437/Interspeech.2025-739.

Hugo Pitorro and Marcos Vinicius Treviso. LaTIM: Measuring latent token-to-token interactions in mamba models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pages 24478–24493, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1194. URL https://aclanthology.org/2025.acl-long.1194/.

Hugo Pitorro, Pavlo Vasylenko, Marcos Treviso, and André FT Martins. How effective are state space models for machine translation? In *Proceedings of the Ninth Conference on Machine Translation*, pages 1107–1124, 2024.

José Pombal, Sweta Agrawal, Patrick Fernandes, Emmanouil Zaranis, and André FT Martins. A context-aware framework for translation-mediated conversations. *arXiv preprint arXiv:2412.04205*, 2024a.

Jose Pombal, Sweta Agrawal, and André Martins. Improving context usage for translating bilingual customer support chat with large language models. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 993–1003, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.100. URL https://aclanthology.org/2024.wmt-1.100/.

Maja Popović. chrf: character n-gram f-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT15)*, Lisbon, Portugal, 2015. Association for Computational Linguistics. URL https://aclanthology.org/W15-3049.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Miguel Moura Ramos, Patrick Fernandes, António Farinhas, and Andre Martins. Aligning neural machine translation models: Human feedback in training and inference. In Carolina Scarton, Charlotte Prescott, Chris Bayliss, Chris Oakley, Joanna Wright, Stuart Wrigley, Xingyi Song, Edward Gow-Smith, Rachel Bawden, Víctor M Sánchez-Cartagena, Patrick Cadwell, Ekaterina Lapshinova-Koltunski, Vera Cabarrão, Konstantinos Chatzitheodorou, Mary Nurminen, Diptesh Kanojia, and Helena Moniz, editors, *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 258–274, Sheffield, UK, June 2024. European Association for Machine Translation (EAMT). URL https://aclanthology.org/2024.eamt-1.22/.

Miguel Moura Ramos, Patrick Fernandes, Sweta Agrawal, and Andre Martins. Multilingual contextualization of large language models for document-level machine translation. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Ah0U1r5Ldq.

Leonardo Ranaldi, Giulia Pucci, Barry Haddow, and Alexandra Birch. Empowering multi-step reasoning across languages via program-aided language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12171–12187, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.678. URL https://aclanthology.org/2024.emnlp-main.678/.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. When natural language is not enough: The limits of in-context learning demonstrations in multilingual reasoning. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7369–7396, Albuquerque, New Mexico, April 2025a. Association for Computational Linguistics. ISBN 979-8-89176-195-7. doi: 10.18653/v1/2025.findings-naacl.412. URL https://aclanthology.org/2025.findings-naacl.412/.

Leonardo Ranaldi, Barry Haddow, and Alexandra Birch. Multilingual retrieval-augmented generation for knowledge-intensive task. *arXiv preprint arXiv:2504.03616*, 2025b.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of the 4th International Conference on Learning Representations (ICLR)*, 2016. URL https://arxiv.org/abs/1511.06732.

Ricardo Rei, Ana C. Farinha, Alon Lavie, José G. C. de Souza, Duarte Alves, and André F. T. Martins. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation (WMT20)*, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.wmt-1.115.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT22)*, Abu Dhabi, United Arab Emirates, 2022a. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.52.

Ricardo Rei, José GC De Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André FT Martins. Comet-22: Unbabel-ist 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, 2022b. URL https://aclanthology.org/2022.wmt-1.52/.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luísa Coheur, and André F. T. Martins. COMETKiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates, December 2022c. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.60.

Ricardo Rei, Nuno M. Guerreiro, JosÃ© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.73. URL https://aclanthology.org/2023.wmt-1.73.

Elizabeth Salesky, Kareem Darwish, Mohamed Al-Badrashiny, Mona Diab, and Jan Niehues. Evaluating multilingual speech translation under realistic conditions with resegmentation and terminology. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat, editors, *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 62–78, Toronto, Canada (in-person and online), July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.iwslt-1.2. URL https://aclanthology.org/2023.iwslt-1.2/.

Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos, editors. *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.0. URL https://aclanthology.org/2025.iwslt-1.0/.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.acl-main.704.

Sukanta Sen, Rico Sennrich, Biao Zhang, and Barry Haddow. Self-training Reduces Flicker in Retranslation-based Simultaneous Translation. In Andreas Vlachos and Isabelle Augenstein, editors, *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3734–3744, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.270. URL https://aclanthology.org/2023.eacl-main.270.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, Paul Christiano, and Jan Leike. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2009.01325.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomput.*, 568(C), February 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Alessandra Terranova, Björn Ross, and Alexandra Birch. What Makes Memory Work? Evaluating Long-Term Memory for Large Language Models, 2025.

Jörg Tiedemann. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf.

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, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom.

Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288, 2023. doi: 10.48550/ARXIV.2307.09288. URL https://doi.org/10.48550/arXiv.2307.09288.

Marcos V Treviso, Nuno M Guerreiro, Sweta Agrawal, Ricardo Rei, José Pombal, Tania Vaz, Helena Wu, Beatriz Silva, Daan Van Stigt, and Andre Martins. xTower: A multilingual LLM for explaining and correcting translation errors. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15222–15239, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.892. URL https://aclanthology.org/2024.findings-emnlp.892/.

Pavlo Vasylenko, Marcos Treviso, and André FT Martins. Long-context generalization with sparse attention. *arXiv preprint arXiv:2506.16640*, 2025.

Jonas Waldendorf, Barry Haddow, and Alexandra Birch. Contrastive Decoding Reduces Hallucinations in Large Multilingual Machine Translation 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 2526–2539, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.eacl-long.155.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2: A massively multilingual speech-to-text translation corpus, 2020.

Weixuan Wang, Barry Haddow, and Alexandra Birch. Retrieval-augmented multilingual knowledge editing, 2023. URL https://arxiv.org/abs/2312.13040.

Sam Wiseman and Alexander M. Rush. Sequence-to-sequence learning as beam-search optimization. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, Austin, Texas, USA, 2016. Association for Computational Linguistics. URL https://aclanthology.org/D16-1330.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A Paradigm Shift in Machine Translation: Boosting Translation Performance of Large Language Models, September 2023. URL http://arxiv.org/abs/2309.11674. arXiv:2309.11674 [cs].

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=farT6XXntP.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.

Emmanouil Zaranis, Nuno Guerreiro, and André FT Martins. Analyzing context contributions in llm-based machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14899–14924, 2024.

Emmanouil Zaranis, Giuseppe Attanasio, Sweta Agrawal, and Andre Martins. Watching the watchers: Exposing gender disparities in machine translation quality estimation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25261–25284, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1228. URL https://aclanthology.org/2025.acl-long.1228/.

Chrysoula Zerva, Frédéric Blain, Ricardo Rei, Piyawat Lertvittayakumjorn, José G. C. de Souza, Steffen Eger, Diptesh Kanojia, Duarte Alves, Constantin Orăsan, Marina Fomicheva, André F. T. Martins, and Lucia Specia. Findings of the WMT 2022 shared task on quality estimation. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 69–99, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.3.

Chrysoula Zerva, Frédéric Blain, José G. C. de Souza, Diptesh Kanojia, Sourabh Deoghare, Nuno M. Guerreiro, Giuseppe Attanasio, Ricardo Rei, Constantin Oraşan, Matteo Negri, Marco Turchi, Rajen Chatterjee, Pushpak Bhattacharyya, Markus Freitag, and André F. T. Martins. Findings of the quality estimation shared task at wmt 2024: Are llms closing the gap in qe? In *Proceedings of the Ninth Conference on Machine Translation (WMT)*, pages 82–109, Miami, Florida, USA, 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.wmt-1.3.

Biao Zhang, Barry Haddow, and Alexandra Birch. Prompting large language model for machine translation: A case study. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 41092–41110. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhang23m.html.

Giulio Zhou, Tsz Kin Lam, Alexandra Birch, and Barry Haddow. Prosody in cascade and direct speech-to-text translation: a case study on korean wh-phrases. In *Findings of EACL*, 2024. URL https://arxiv.org/abs/2402.00632.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. Question translation training for better multilingual reasoning. *arXiv preprint arXiv:2401.07817*, 2024.

ENDPAGE

UTTER

HORIZON-CL4-2021-HUMAN-01 101070631

D4.2 Final Report on adaptable and context-aware models