Why do Large Language Models Fail in Low-resource Translation? Unraveling the Token Dynamics of Large Language Models for Machine Translation

Shenbin Qian and **Yves Scherrer**
Language Technology Group, Department of Informatics
University of Oslo, Norway
{shenbinq, yves.scherrer}@ifi.uio.no

Abstract

Large Language Models (LLMs) have recently demonstrated strong performance in machine translation (MT). However, most prior work focuses on improving or benchmarking translation quality, offering limited insight into when and why LLM-based translation fails. In this work, we systematically analyze failure modes of LLMs in MT by evaluating 15 models, including four reasoning LLMs, across 22 language pairs (LPs) with varying resource levels. We find that non-English-centric LPs consistently yield lower COMET scores than English-centric pairs. To investigate the underlying causes, we introduce **Token Activation Rate (TAR)**, a metric that captures how effectively a model utilizes language-specific tokens in its vocabulary during generation. We validate TAR as a proxy for language representation using models with known language distributions in the training data, and show that lower TAR is strongly associated with poorer translation performance. Furthermore, reasoning LLMs tend to generate more tokens when translating into low-TAR languages, suggesting a compensatory mechanism, although its impact on translation quality varies across models. Overall, our findings emphasize the importance of token-level dynamics in understanding MT performance of LLMs.

1 Introduction

Large Language Models (LLMs) have achieved significant advancements across various subfields of Natural Language Processing (NLP), including sentiment analysis, text summarization, and machine translation (MT) (Zhang et al., 2024; Pu et al., 2023; Zhang et al., 2023). More recently, LLMs trained via Reinforcement Learning with Verifiable Rewards (RLVR) (Lambert et al., 2024) have demonstrated reasoning capabilities that extend beyond language tasks to include coding and mathematical problem-solving (OpenAI et al., 2024; Guo et al., 2025; Ahn et al., 2024; Jiang et al., 2025).

Alongside these developments, numerous benchmarks have emerged to evaluate the state-of-the-art capabilities of LLMs on specific tasks (Wang et al., 2019; Hendrycks et al., 2021; OpenAI, 2024; Phan et al., 2025; Yue et al., 2025; Romanou et al., 2025; Huang et al., 2025). However, most benchmarks aim to assess how well LLMs perform on tasks with definitive correct answers, typically through multiple-choice formats or comparison with human-prepared references, but not on open-ended multilingual generation tasks like translation. Although MT evaluation datasets such as FLORES (Guzmán et al., 2019) or test sets from the Conference on Machine Translation (WMT[1]) can be leveraged for evaluating LLMs' translation abilities, relatively little work has investigated why LLMs fail on certain translation tasks, particularly in low-resource and non-English-centric settings.

To address this gap, we perform a large-scale empirical analysis of LLM-based translation, focusing on how performance varies across language pairs (LPs) with different resource availabilities. We observe that non-English-centric and lower-resource LPs consistently yield lower COMET

[1] https://www2.statmt.org/

(Rei et al., 2020; Stewart et al., 2020; Rei et al., 2022a; Rei et al., 2022b) and BLEU (Papineni et al., 2002) scores. We hypothesize that low token activation for these languages contributes to these failures, and that reasoning models may partially compensate by generating more tokens at inference time. Our contributions are as follows:

- We evaluate 15 models across 22 LPs and show that non-English-centric LPs exhibit significantly lower COMET scores compared to English-centric pairs.
- We propose **Token Activation Rate (TAR)**[2] as a metric for quantifying language representation in model vocabularies, and demonstrate its effectiveness as a proxy for language coverage. We further show that TAR and **typological distance** are strongly associated with COMET and BLEU scores.
- We investigate the relationships among TAR, **reasoning tokens**, and COMET and BLEU scores. Our findings suggest that low TAR of the target language is significantly correlated with the number of generated reasoning tokens, which for some LLMs is correlated with COMET or BLEU improvements.

2 Related Work

LLM Translation The emergence of LLMs has spurred extensive research on their application to machine translation (Zhang et al., 2023; Vilar et al., 2023; Castaldo and Monti, 2024; He, 2024). Early work (Zhang et al., 2023) explored prompting strategies and showed that well-designed prompts can yield performance comparable to traditional MT systems. Subsequent studies (Kocmi et al., 2024; Song et al., 2025) highlight that LLMs consistently underperform in low-resource settings, motivating approaches such as retrieval-augmented and context-aware translation (Court and Elsner, 2024). More recently, reasoning LLMs have been applied to translation tasks. Liu et al. (2025) argue that these models improve contextual coherence, cultural intentionality, and self-reflection, while Ye et al. (2025) show that they outperform instruction-tuned models in semantically complex domains, particularly for long-text and high-difficulty translation scenarios. Despite these advances, prior work largely focuses on improving translation quality rather than explaining the root causes of failure, particularly in low-resource settings.

Tokenization and Vocabulary Effects in MT A growing body of work attributes translation failures to tokenization and vocabulary design (Rust et al., 2021; Sindhujan et al., 2025; Lundin et al., 2025). Multilingual models often underperform on languages that are under-represented in the shared vocabulary, while dedicated or language-specific tokenizers can mitigate this gap (Rust et al., 2021). Tokenization inefficiency, commonly measured by high sub-word fertility, has also been shown to correlate with lower performance, especially for morphologically rich and low-resource languages (Lundin et al., 2025). Several methods have been proposed to address these issues including stochastic segmentation techniques, such as BPE-dropout, vocabulary refinement approaches to remove low-utility tokens, and targeted vocabulary expansion *etc* (Provilkov et al., 2020; Chizhov et al., 2024; Singh et al., 2025). Overall, prior work consistently links tokenization properties such as vocabulary coverage, or token efficiency, to downstream translation performance. However, these studies primarily focus on model design and optimization, leaving open the question of how token-level dynamics within LLMs contribute to systematic failures in translation, especially low-resource settings.

3 Experimental Setup

We describe our datasets in Section 3.1. Models and inference details are in Sections 3.2 and 3.3.

3.1 Data

To assess the translation capabilities of LLMs, we compiled multiple datasets covering different LPs and translation directions across resource-varying settings. Our test data comprises 10 **non-English-centric** LPs[3] and 12 **English-centric** LPs, with the latter consisting of 6 **en-XX** pairs and 6 **XX-en** pairs. These span high-, medium-, and low-resource languages, including **Arabic-Chinese (ar-zh)**, **Arabic-Hebrew (ar-he)**, **Chinese-French (zh-fr)**, **Chinese-Russian (zh-ru)**, **French-Italian (fr-it)**, **German-French (de-fr)**, **German-Italian (de-it)**, **Korean-Chinese (ko-zh)**, **Korean-French (ko-fr)**, and **Russian-French (ru-fr)**

[2] https://github.com/shenbinqian/llm4mt

[3] These datasets do not involve English during the process of their construction, unlike FLORES.

Model Name	Architecture	Instruction-tuned or Reasoning	Open Weights	Parameter Size
Qwen3-30B-A3B-Instruct-2507	decoder-only-moe	instruction-tuned	yes	30B in total, 3B active
Qwen3-30B-A3B-Thinking-2507	decoder-only-moe	reasoning	yes	30B in total, 3B active
Qwen3-4B-Instruct-2507	decoder-only-dense	instruction-tuned	yes	4B
Qwen3-4B-Thinking-2507	decoder-only-dense	reasoning	yes	4B
Llama-3.2-3B-Instruct	decoder-only-dense	instruction-tuned	yes	3B
gemma-3-27b-it	decoder-only-dense	instruction-tuned	yes	27B
Qwen2.5-32B-Instruct	decoder-only-dense	instruction-tuned	yes	32B
DeepSeek-R1-Distill-Qwen-32B	decoder-only-dense	reasoning	yes	32B
aya-expanse-32b	decoder-only-dense	instruction-tuned	yes	32B
Tower-Plus-72B	decoder-only-dense	instruction-tuned	yes	72B
t5gemma-xl-xl-prefixlm-it	encoder-decoder-dense	instruction-tuned	yes	4B
Deepseek-V3.2-Exp	decoder-only-moe	mixed	yes	671B
nllb-200-3.3B	encoder-decoder-dense	neither, translation only	yes	3.3B
nllb-moe-54b	encoder-decoder-moe	neither, translation only	yes	54B
Google Translate	unknown	neither, translation only	no	unknown

Table 1: Model details including names, architectures, size and either instruction-tuned or reasoning and open-weights or proprietary models.

from the TED Multilingual Parallel Corpora (Kulkarni, 2015), the multilingual corpus from the Swiss Federal Administration (SwissAdmin) (Scherrer et al., 2014), and the Chinese-Korean parallel corpus (Park and Zhao, 2019); as well as **English-Chinese (en-zh)**, **English-Czech (en-cs)**, **English-German (en-de)**, **English-Polish (en-pl)**, **English-Russian (en-ru)**, **English-Tamil (en-ta)**, **Chinese-English (zh-en)**, **Czech-English (cs-en)**, **German-English (de-en)**, **Khmer-English (km-en)**, **Russian-English (ru-en)**, and **Tamil-English (ta-en)** from the Quality Estimation Shared Task of the Fifth Conference on Machine Translation (WMT20) (Barrault et al., 2020). We randomly sampled 3,000 examples per LP from these corpora to form our test set, yielding 66,000 instances in total[4] (see Appendix A). We did not select these resources with the intention of benchmarking the latest LLMs, as they are publicly available online and may have been included in LLM training data. Rather, we use this data to investigate when and why models fail, even on potentially seen examples.

3.2 Methodology

Prompt Selection We initially adopted the prompt template from Zhang et al. (2023) to instruct LLMs to perform translation via in-context learning in both zero-shot and few-shot settings. However, preliminary experiments revealed that some models failed to adhere to the instruction, producing verbose and noisy outputs with explanatory text rather than translations in the target language (see Appendix B). Such behavior interferes with reliable automatic evaluation. To deal with this issue, we designed two additional prompt templates aimed at eliciting translation-only outputs. We denote the original prompt from Zhang et al. (2023) as Prompt 0, and our proposed templates as Prompt 1 and Prompt 2 (see Appendix C). These prompts are not intended to optimize translation performance, but to ensure output consistency for evaluation, which is critical for maintaining the validity of metric-based comparisons such as COMET and BLEU. We conducted experiments with all 3 prompts and assessed output noise using a rule-based detector followed by manual inspection (see Appendix D). We selected outputs from Prompt 2, which consistently produced the cleanest translations, for all subsequent analyses.

Model Selection We selected 15 models spanning a wide range of sizes, architectures, post-training methods, and levels of multilingual data coverage as shown in Table 1. These include decoder-only instruction-tuned (IT) models from the Qwen series, such as **Qwen3-30B-A3B-Instruct-2507** and **Qwen3-4B-Instruct-2507**, along with their corresponding reasoning variants post-trained using RLVR: **Qwen3-30B-A3B-Thinking-2507** and **Qwen3-4B-Thinking-2507** (Qwen Team, 2025). To compare instruction-tuned and reasoning models, we also include **Qwen2.5-32B-Instruct** (Qwen Team, 2024) versus **DeepSeek-R1-Distill-Qwen-32B**, which share the same base model but differ in post-training—the latter was trained via knowledge distillation (Hinton et al., 2015) using DeepSeek-R1 (Guo et al., 2025) as a teacher model trained with RLVR. Additionally, we compare the chat mode and reasoning mode of **DeepSeek-V3.2-Exp** (DeepSeek-V3.2-Exp-671B-chat and

[4]We treat language pairs with different translation directions as distinct, as we used separate data instances for each direction rather than swapping source and target.

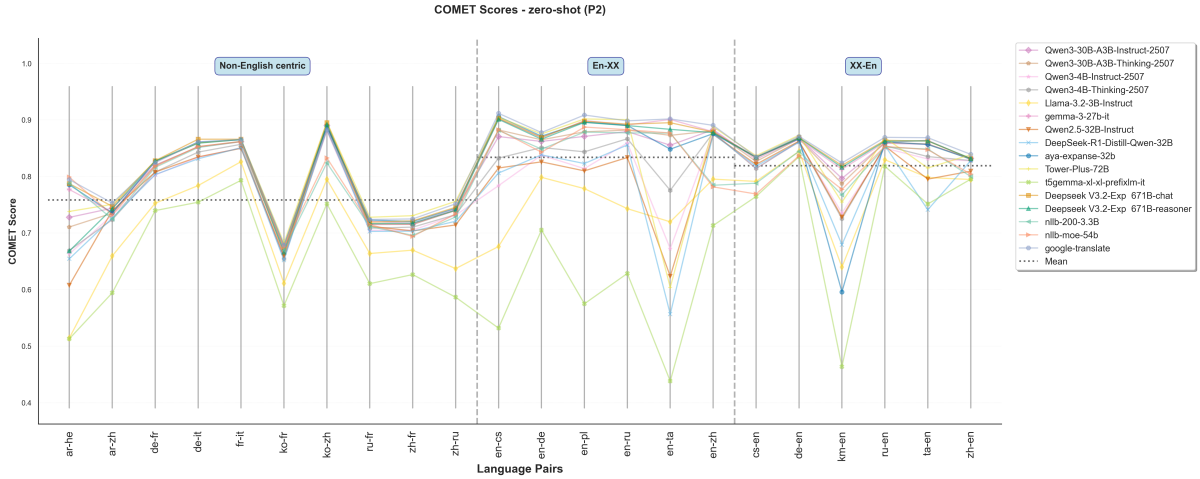

Figure 1: COMET scores of translations for 22 language pairs using Prompt 2 under zero-shot setting.

DeepSeek-V3.2-Exp-671B-reasoner, respectively) (DeepSeek-AI, 2025). **Llama-3.2-3B-Instruct** (Meta AI, 2024) and **gemma-3-27b-it** (Gemma Team et al., 2025) were selected as decoder-only dense IT models, while **t5gemma-xl-xl-prefixlm-it** (Zhang et al., 2025) serves as a representative of recent encoder-decoder IT models. Since most of these LLMs are predominantly English- and/or Chinese-centric, we included **aya-expanse-32b** (Dang et al., 2024), which was pre-trained on extensive multilingual data, and **Tower-Plus-72B** (Rei et al., 2025), a translation-specific LLM fine-tuned on Qwen-2.5-72B. For baseline comparison, we selected two neural machine translation models, **nllb-200-3.3B** and **nllb-moe-54b** (NLLB Team et al., 2022), along with a widely used proprietary system, **Google Translate**[5].

Evaluation Metrics Considering their popularity, we used COMET-22 (Rei et al., 2022a) and SacreBLEU (Post, 2018) as the main evaluation metrics for our LLM translation outputs. chrF++ scores (Popović, 2017) were included in Appendix E as references for morphologically-rich target languages.

3.3 Inference Details

We used vLLM (Kwon et al., 2023) for inference with most models, with the exception of DeepSeek-V3.2-Exp, t5gemma-xl-xl-prefixlm-it, and the baseline systems. For these models, we obtained inference results using their respective APIs or the HuggingFace Transformers library (Wolf et al., 2020). We initially conducted experiments using Prompt 0 with the temperature and top_p both set to 1. We further evaluated the effect of varying the temperature by increasing it to 1.5 and decreasing it to 0. Increasing the temperature to 1.5 resulted in a clear performance degradation across all language pairs, as measured by both COMET and BLEU scores. Conversely, setting the temperature to 0 led to slight performance improvements for nearly all language pairs. Consequently, all reported experiments were conducted with a temperature of 0. In the few-shot setting, we randomly selected 5 examples for each language pair from the rest of the corpora as demonstrations inserted in the prompt templates.

With the exception of DeepSeek-V3.2-Exp and Google Translate, all models were run without quantization on 4 NVIDIA GH200 GPUs. On average, an IT model requires approximately 10 minutes to process one LP (3,000 instances), whereas a reasoning model requires about 18 minutes.

4 Evaluation Results

This section presents the results of our evaluation. Figure 1 displays COMET scores for all 22 LPs under the zero-shot setting.

The parallel coordinates plot in Figure 1 reveals interesting patterns in COMET scores across language pairs of varying resource availability and across LLMs trained for general versus translation-specific purposes. Detailed tables of COMET and BLEU scores for both zero-shot and few-shot settings exhibit consistent patterns and are therefore provided in Appendix E.

[5]Available at `https://translate.google.com/`. We consider Google Translate as a translation LLM since Google claims it is supported by LLMs (Caswell, 2024).

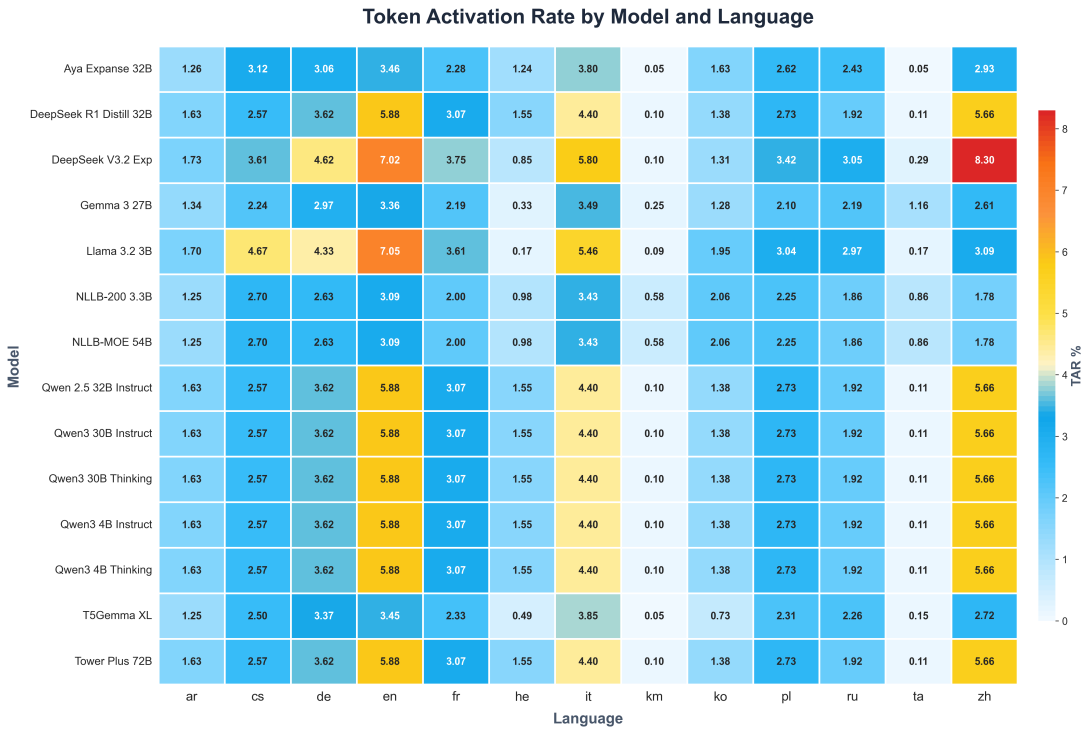

Figure 2: TAR for 13 different languages and 14 models (excluding Google Translate).

First, we observe that non-English-centric LPs have substantially lower average COMET scores than English-centric pairs, with greater performance variability across these LPs. This reflects the current state of the art in MT, namely the English-centricity of language resources. The figure also shows clear performance degradation for most LLMs on LPs involving lower-resource languages, such as Arabic-Hebrew, English-Tamil, and Khmer-English, suggesting that resource availability plays a key role in translation performance. However, we also observe that certain LPs, such as Chinese-French, yield notably lower COMET scores than French-Italian, despite both involving high-resource languages. We hypothesize that typological distance also influences COMET scores. In Section 5, we further investigate whether language resource availability, using TAR as a proxy, and typological distance are significant factors of LLM performance in translation.

Regarding model-wise performance, translation-specific LLMs such as Tower-Plus-72B and Google Translate achieve the highest COMET scores for most LPs, generally outperforming general-purpose LLMs. Among general-purpose models, those that are large in scale and trained on multilingual data such as aya-expanse-32b, gemma-3-27b-it, and DeepSeek-V3.2-Exp-671B-chat, achieve results comparable to translation-specific LLMs. This further suggests that greater exposure to diverse language data during training may positively impact translation performance, a hypothesis we explore in the following section.

5 Analysis and Findings

This section investigates factors associated with LLM failure in translation, especially for low-resource languages. The previous section suggests that factors such as language resource availability and typological distance between languages may be important predictors of LLM translation performance. We explore these factors in Sections 5.1 and 5.2. Assuming that language data representation in the training data is an important factor for LLM performance, we further investigate whether generating more tokens (*i.e.*, the number of reasoning tokens) at test time can compensate for limited TAR during pre-training in Section 5.3.

5.1 Token Activation Rate

Since we do not know the actual distribution of each language in the training data, we leveraged our test data as samples to calculate the Token Activation Rate (TAR) of the model vocabulary as an approximation, to understand language re-

Model	TAR	GENETIC	GEOGRAPHIC	SYNTACTIC	PHONOLOGICAL	INVENTORY	FEATURAL	MEAN
Qwen3-30B-A3B-Instruct-2507	**0.5352**	-0.1294	-0.2395	-0.2605	0.1940	0.0982	-0.4134	-0.1736
Qwen3-30B-A3B-Thinking-2507	**0.5339**	-0.1032	-0.2402	-0.2453	0.2225	0.1010	**-0.4275**	-0.1599
Qwen3-4B-Instruct-2507	**0.6575**	-0.1302	-0.0915	**-0.4974**	0.1583	0.0615	**-0.4723**	-0.1470
Qwen3-4B-Thinking-2507	**0.6490**	-0.1196	-0.1687	-0.4127	0.2140	0.0866	**-0.4963**	-0.1594
Llama-3.2-3B-Instruct	**0.7206**	-0.0682	-0.1286	-0.4216	0.2668	-0.0539	**-0.5666**	-0.1486
gemma-3-27b-it	**0.5164**	-0.1706	-0.3282	-0.1792	0.1478	0.1586	-0.3691	-0.2157
Qwen2.5-32B-Instruct	**0.6693**	-0.0761	-0.0799	**-0.4937**	0.1830	-0.0148	**-0.4720**	-0.1353
DeepSeek-R1-Distill-Qwen-32B	**0.6685**	-0.1932	-0.1026	**-0.5949**	0.0548	0.0666	**-0.4635**	-0.2038
aya-expanse-32b	**0.5545**	-0.2598	-0.3132	-0.2977	0.0355	0.1484	-0.3759	-0.2746
Tower-Plus-72B	**0.5954**	-0.2347	-0.2158	**-0.4974**	0.0117	0.1164	**-0.4281**	-0.2593
t5gemma-xl-xl-prefixlm-it	**0.5905**	-0.1857	-0.0649	**-0.5403**	0.1237	-0.0076	**-0.4533**	-0.1707
DeepSeek-V3.2-Exp-671B-chat	0.3166	-0.1842	-0.3243	-0.1863	0.1240	0.1495	-0.3528	-0.2234
DeepSeek-V3.2-Exp-671B-reasoner	**0.4700**	-0.0191	-0.2189	-0.2002	0.2706	0.0397	**-0.4292**	-0.1224
nllb-200-3.3B	**0.5643**	-0.2850	**-0.5233**	-0.2289	-0.0040	0.0968	**-0.4307**	-0.4101
nllb-moe-54b	**0.5080**	-0.2621	**-0.5037**	-0.2168	-0.0328	0.1037	-0.4011	-0.3950

Table 2: Pearson's r correlation between COMET scores and TAR, genetic, geographic, syntactic, phonological, inventory, featural and the mean of the latter six typological distances. **Bold values** are statistically significant.

source availability during training. TAR measures the proportion of a model's tokenizer vocabulary that is activated when processing text in a given language. Formally, given a model M with vocabulary V_M, a tokenizer function Tokenize_M, and text data D_l in language l, TAR is defined as:

$$\text{TAR}(l, M) = \frac{|\{t \in V_M : t \in \text{Tokenize}_M(D_l)\}|}{|V_M|} \quad (1)$$

We used the 3,000 instances per language pair from either the source or the target in the test set, and tokenized them into input IDs using the corresponding model tokenizers. We retained only unique input IDs for each language (13 in total) and divided this count by the vocabulary size of the model. For example, we used the source text of the 3,000 instances in Arabic-Hebrew, tokenizing them with the Qwen3-4B-Instruct-2507 tokenizer to obtain 2,469 unique input IDs. This count was then divided by the model vocabulary size of 151,669, resulting in a TAR of 1.63% for Arabic.

Figure 2 presents a heatmap of TAR across the 13 languages and 14 models. It reveals that Khmer, Tamil, and Hebrew exhibit notably low TAR across nearly all models, which corresponds precisely to the COMET score drops observed for Arabic-Hebrew, English-Tamil, and Khmer-English in Figure 1. Regarding model-wise coverage, neural MT models such as NLLB maintain better balance across languages compared to English- and Chinese-dominant LLMs, resulting in smaller performance disparities among LPs.

5.2 Typological Distance

We observe that although Chinese, French, and Italian exhibit high TAR, the average COMET scores for Chinese-French are lower than those for French-Italian. We hypothesize that other factors, such as typological distance, also affect LLM performance. To quantify these distances across LPs, we rely on URIEL (Littell et al., 2017), a database and toolkit that provides multiple distance measures between languages, including genetic, geographic, syntactic, phonological, inventory, and featural distances. These measures capture, respectively, genealogical relatedness within a language family, physical distance between speaker populations, divergence in grammatical structure, differences in sound systems, variation in phoneme inventories, and an overall typological distance derived from the full set of URIEL features. Details of the design and computation of the distances can be found in Littell et al (2017).

Table 2 displays Pearson's r correlation scores between COMET scores, TAR[6], the six typological distances and their mean. With the exception of DeepSeek-V3.2-Exp-671B-chat, TAR is highly correlated with COMET scores across all models. Syntactic and featural distances also exhibit moderate negative correlations with model performance for many models. That means, greater distance between two languages corresponds to lower COMET scores. The correlation patterns for BLEU and chrF++ scores are consistent with these observations, as shown in Tables F.1 and F.2 in Appendix F. These results align with prior findings reported by Khiu et al. (2024), Ploeger et al. (2025), and Hirak et al. (2026).

5.3 Reasoning Tokens

Given that low TAR in a model's vocabulary at the pre-training stage is highly correlated with translation performance, we analyze whether reasoning LLMs would generate more reasoning tokens for languages with lower TAR as a compen-

[6]TAR for a language pair is computed by summing the TAR values of the source and target languages.

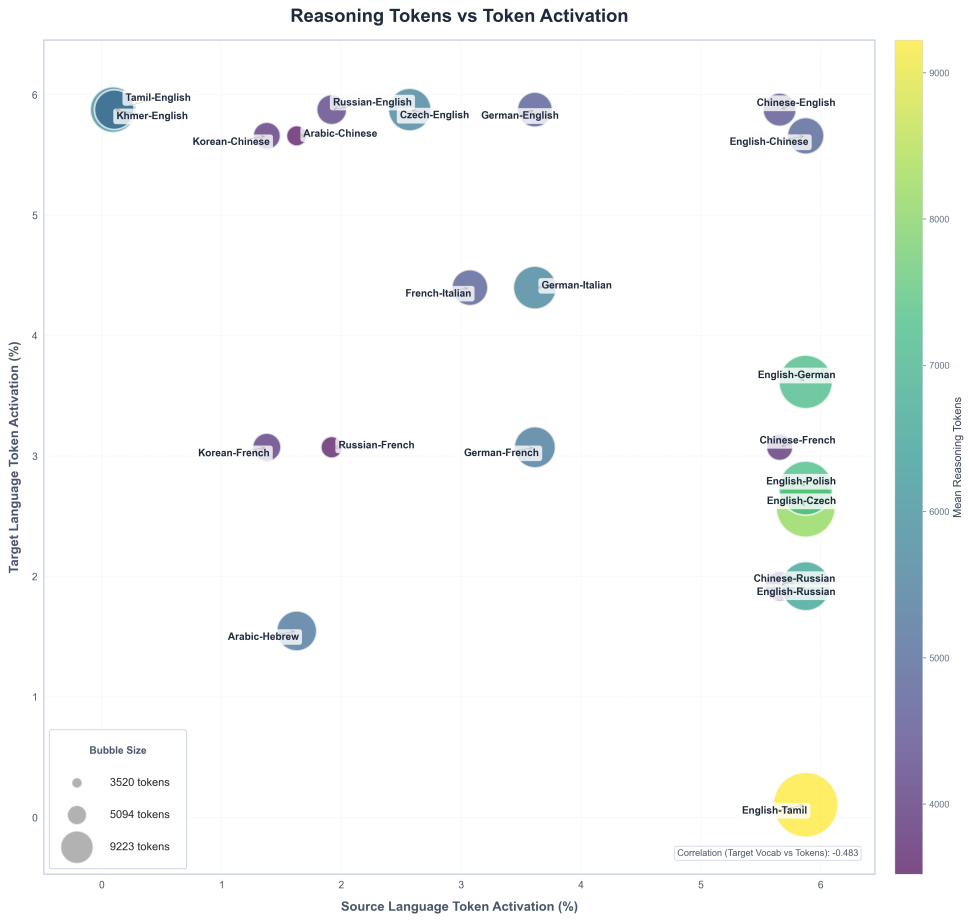

Figure 3: TAR of the vocabulary of Qwen3-4B-Thinking-2507 per language pair in the source (X axis) and target (Y axis) language against the average number of reasoning tokens.

satory mechanism. Furthermore, we also explore whether generating more reasoning tokens at test time would improve translation quality.

Reasoning Tokens *vs* TAR Figure 3 illustrates the relationship between TAR for each LP and the average number of reasoning tokens generated by Qwen3-4B-Thinking-2507, with source language TAR on the X-axis and target language TAR on the Y-axis. The figure clearly shows that Qwen3-4B-Thinking-2507 generates substantially fewer reasoning tokens for LPs with high TAR on the target side, such as Korean-Chinese and Russian-English. For LPs with high source-side TAR but medium or low target-side TAR, at the mid-right region of the figure, the model generates considerably more reasoning tokens. We further calculated correlations between the number of reasoning tokens and TAR on both the source and target sides for the 4 reasoning models. We find that TAR in the target language is indeed negatively correlated with the number of reasoning tokens (r=-0.2572, ρ=-0.3177, τ=-0.2306; all statistically significant). This indicates that lower TAR in the target language tends to elicit more reasoning tokens at test time as compensation.

Reasoning Tokens *vs* Metric Improvements We continued our investigation on whether more reasoning tokens generated at test time would benefit the performance of LLM translation, by examining the difference of COMET and BLEU scores (ΔCOMET and ΔBLEU) between reasoning models and their instruction-tuned counterparts. This analysis examines whether increases or decreases in COMET and BLEU scores correlate with the number of generated reasoning tokens.

Model Name	ΔCOMET	ΔBLEU
Qwen3-30B-A3B-Thinking-2507	**0.5734**	0.3273
Qwen3-4B-Thinking-2507	**0.7925**	**0.5900**
DeepSeek-R1-Distill-Qwen-32B	-0.1043	0.0177
DeepSeek-V3.2-Exp-671B-chat	**-0.9825**	**-0.9660**

Table 3: Pearson's r correlation between ΔCOMET and ΔBLEU and the average number of reasoning tokens for each LP. **Bold values** are statistically significant.

Table 3 presents Pearson's r correlation coeffi-

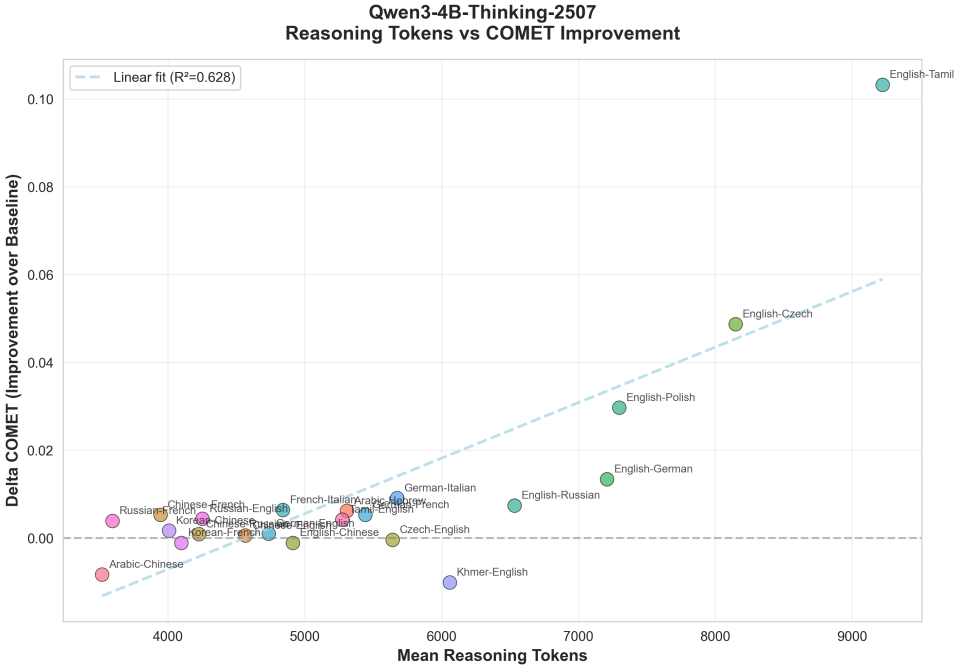

Figure 4: The average number of reasoning tokens from Qwen3-4B-Thinking-2507 *vs* the increase of COMET scores (ΔCOMET) compared to its IT model Qwen3-4B-Instruct-2507.

cients between the average number of reasoning tokens and ΔCOMET and ΔBLEU. The table reveals that their correlations are model-dependent. For Qwen models, more reasoning tokens exhibit a strong positive correlation with COMET score improvements, indicating that additional reasoning tokens contribute positively to translation quality. Figure 4 plots the relationship between ΔCOMET and the average number of reasoning tokens for Qwen3-4B-Thinking-2507, showing that a simple linear model could explain 62.8% of the variability in the response variable. For language pairs with low TAR at the target side like English-Tamil, the model generates a considerable amount of reasoning tokens, which correlates positively with the increase of COMET scores. However, DeepSeek models, in contrast, exhibit negative correlations. To further explore this model-specific difference, we continued our investigations in Section 7 on other reasoning models.

6 Validation on Token Activation Rate

The analyses in Section 5.1 rely on the assumption that TAR reflects how well a language is represented in the model's pre-training data. To validate this assumption, we sought open-source LLMs that disclose language-level data distributions. To our best effort, we identified Bloomz (BigScience Workshop et al., 2022) and EuroLLM (Martins et al., 2024), both of which report this information. Other open-source LLMs including Olmo (Groeneveld et al., 2024) and Apertus (Apertus Project et al., 2025) do not explicitly provide detailed language distributions in their training data.

Language	Actual	TAR
Arabic	4.65%	2.58%
English	30.11%	3.63%
French	12.94%	2.56%
Chinese	16.21%	4.17%
Tamil	0.50%	1.73%
Gujarati	0.07%	2.30%
Hindi	1.53%	2.88%
Malayalam	0.23%	2.46%
Portuguese	4.92%	4.78%
Telugu	0.19%	2.34%

Table 4: TAR and the actual language-level training data distribution (Actual) in bloomz-7b1.

As shown in Table 4 for bloomz-7b1, we computed TAR for Arabic, English, French, Chinese, and Tamil using the method and data described in Sections 5.1 and 3.1 respectively. To increase the number of languages for validation, we incorporated additional language data including Gujarati, Hindi, Malayalam, Portuguese and Telugu from the monolingual training data of WMT24 (Kocmi et al., 2024), as these are mostly from similar sources and of comparable length to our data. For EuroLLM-22B-Instruct-2512, the train-

Language	Actual	TAR
German	6.00%	6.06%
French	6.00%	4.23%
Italian	6.00%	7.28%
Chinese	3.50%	3.88%
Russian	2.50%	4.32%
Polish	2.50%	5.34%
Arabic	1.50%	1.87%
Korean	1.50%	2.27%
Czech	1.50%	4.99%
English	82.50%	6.52%

Table 5: TAR and the actual language-level training data distribution (Actual) in EuroLLM-22B-Instruct-2512.

ing data distributions for German, French, Italian, Chinese, Russian, Polish, Arabic, Korean, Czech and English are openly released. We computed their TAR using our data and present the results in Table 5.

We then applied a leave-one-language-out methodology: for each language, we remove it from the set and recompute the correlation between TAR and actual training data proportions. This tests whether the observed correlation is robust or driven by individual outlier languages.

left-out	r	ρ	τ
None	0.4980	**0.7697**	**0.5556**
Arabic	0.4925	**0.7500**	**0.5556**
English	0.5215	**0.7500**	**0.5556**
French	0.5444	**0.8167**	**0.6111**
Chinese	0.4166	**0.7500**	**0.5556**
Tamil	0.4514	**0.7833**	**0.6111**
Gujarati	0.4661	**0.7333**	0.5000
Hindi	0.5036	**0.7500**	**0.5556**
Malayalam	0.4761	**0.7333**	0.5000
Portuguese	**0.7544**	**0.8167**	**0.6111**
Telugu	0.4688	**0.7333**	0.5000

Table 6: Pearson's r, Spearman's ρ and Kendall's τ correlation coefficients between the actual language-level training data distribution and TAR of bloomz-7b1. Leave-one-language-out was applied to ensure the score stability. **Bold values** are statistically significant.

Tables 6 and 7 display the Pearson's r, Spearman's ρ and Kendall's τ correlation coefficients between the actual training data distribution and TAR, for bloomz-7b1 and EuroLLM-22B-Instruct-2512. The Spearman and Kendall rank correlations are consistently strong and statistically significant across most leave-one-language-out conditions for both models, indicating that the relationship is robust and not driven by individual outlier languages. The Pearson correlations are generally weaker, which is expected given the non-linear relationship between TAR and actual data proportions (e.g., English has a disproportionately high data share but its TAR is bounded). These results support using TAR as a reliable proxy for language representation in the training data, though we note the limitation that our validation is restricted to only two models with 10 languages each.

left-out	r	ρ	τ
None	0.4177	**0.6669**	**0.5320**
German	0.4581	0.5899	0.4490
French	0.4138	**0.7866**	**0.6286**
Italian	0.5389	0.6156	0.5089
Chinese	0.4077	**0.7105**	**0.5880**
Russian	0.4130	**0.7246**	**0.6086**
Polish	0.4417	**0.6901**	0.5477
Arabic	0.4159	0.5814	0.4490
Korean	0.4050	0.5814	0.4490
Czech	0.4314	**0.7695**	**0.6286**
English	0.6581	0.5719	0.4642

Table 7: Pearson's r, Spearman's ρ and Kendall's τ correlation coefficients between the actual language-level training data distribution and TAR of EuroLLM-22B-Instruct-2512. Leave-one-language-out was applied to ensure the score stability. **Bold values** are statistically significant.

7 Validation on Reasoning Tokens

To validate the generality of our findings on Qwen and DeepSeek models regarding the relationship between TAR, the number of reasoning tokens, and ΔCOMET and ΔBLEU, we replicated our analysis on two additional reasoning LLMs, Olmo-3-7B-Think and K2-Think-V2, along with their instruction-tuned counterparts, Olmo-3-7B-Instruct and K2-V2-Instruct (Olmo Team et al., 2025; K2 Team et al., 2026).

Reasoning Tokens *vs* TAR We observe consistent negative correlations between the TAR of the target language and the average number of reasoning tokens ($r = -0.3045$, $\rho = -0.4917$, $\tau = -0.3414$), all statistically significant. These results corroborate our earlier findings: reasoning LLMs tend to generate more tokens when translating into languages with lower token activation rates. This suggests that increased reasoning token usage may act as a compensatory mechanism for limited token availability on the target side.

Reasoning Tokens *vs* Metric Improvements Table 8 reports the Pearson, Spearman, and Kendall correlations between ΔCOMET, ΔBLEU, and the average number of reasoning tokens for K2-Think-V2 and Olmo-3-7B-Think. Consistent with our observations on Qwen and

Model	Metric	r	ρ	τ
K2-Think-V2	ΔCOMET	0.0698	0.3755	0.2814
	ΔBLEU	**-0.4367**	**-0.4241**	-0.2814
Olmo-3-7B-Think	ΔCOMET	-0.0376	-0.0271	-0.0087
	ΔBLEU	-0.0100	-0.0717	-0.0736

Table 8: Pearson's r, Spearman's ρ and Kendall's τ correlation scores between ΔCOMET, ΔBLEU and the average number of reasoning tokens for K2-Think-V2 and Olmo-3-7B-Think. **Bold values** are statistically significant.

DeepSeek models, the relationship between the number of reasoning tokens and translation quality measured by COMET and BLEU is highly model-dependent. For some models (e.g., Qwen3-4B-Thinking-2507), increased reasoning tokens are associated with improvements in COMET and BLEU scores, whereas for others (e.g., DeepSeek-V3.2-Exp-671B-reasoner and K2-Think-V2), the correlations are weak or negative.

This variability is expected, as translation performance of LLMs depends on multiple factors, including training data, model architecture, and alignment strategies *etc*. Furthermore, automatic metrics such as COMET and BLEU are sensitive to output noise. As observed in models like gemma-3-27b-it and K2-V2-Instruct, the inclusion of explanatory text alongside translations (see Appendix B) can distort metric scores and obscure the true relationship between reasoning and translation quality. These findings highlight the importance of careful model selection and output cleaning to ensure valid evaluation and reliable conclusions. Overall, our results suggest that while increased reasoning token usage consistently compensates for low TAR, its impact on translation quality is not universal, underscoring the need to jointly consider token dynamics and model-specific factors when evaluating reasoning LLMs for MT.

8 Conclusion

In this paper, we systematically evaluated the performance of LLMs on MT, with a focus on understanding their failures in low-resource and non-English-centric settings. To better characterize language representation within model vocabularies, we introduced TAR and validated it as a proxy using models with known training language distributions. Our analyses show that TAR and typological distance are both strongly associated with translation quality: lower TAR and greater typological distance consistently correlate with reduced COMET and BLEU scores. We further examined the relationship between TAR, the number of reasoning tokens, and translation quality. Our results indicate that increased reasoning token generation is closely associated with low TAR in the target language, suggesting a compensatory mechanism. However, the extent to which additional reasoning tokens improve COMET and BLEU scores is highly model-dependent, highlighting the influence of other factors such as training data, alignment, and output noise. Overall, our findings emphasize the importance of token-level dynamics in understanding multilingual performance in LLMs. For future work, we plan to develop robust methods for controlling output noise and to investigate additional factors affecting multilingual capabilities, particularly from an interpretability perspective.

Limitations

Despite our findings, several limitations should be noted. First, output noise remains a significant challenge. LLM-generated translations often include extraneous text, and the extent of such noise varies across models and prompting strategies. Although we design prompts and apply rule-based filtering to encourage translation-only outputs, we cannot guarantee complete removal of noise. As a result, automatic evaluation metrics such as COMET and BLEU may be affected, potentially introducing bias into our results. Second, while we show that TAR correlates with known language distributions and translation performance, it does not fully capture all aspects of multilingual competence. Therefore, TAR should be interpreted as a complementary signal rather than a complete explanation of model behavior. Third, metrics such as COMET and BLEU, while widely used, are sensitive to surface variation and may not fully capture semantic adequacy, especially in multilingual and low-resource settings. This limitation is further exacerbated by the presence of output noise and multiple valid translations.

Finally, our study focuses on correlation rather

than causation. While we identify strong relationships between TAR, reasoning token usage, and translation performance, we do not establish causal mechanisms. Future work is needed to develop controlled experiments and model interventions to better understand the causal role of token dynamics in multilingual generation.

Sustainability Statement

Following the principles of "Green AI" (Schwartz et al., 2020), we aim to minimize the environmental impact of our experiments by improving inference efficiency. Specifically, we leverage vLLM to accelerate inference and reduce computational overhead. In total, our experiments require approximately 200 GPU hours, corresponding to an energy consumption of 397.64 kWh and an estimated 3.03 kg of CO_2 emissions, calculated using the methodology of Lannelongue et al (2021).

Acknowledgments

This work has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101126636.

The computations were performed on resources provided through Sigma2 – the national research infrastructure provider for high-performance computing and large-scale data storage in Norway. We acknowledge Norway and Sigma2 for awarding this project access to the Olivia supercomputer, through Project nn9851k.

References

Ahn, Janice, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In Falk, Neele, Sara Papi, and Mike Zhang, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237, St. Julian's, Malta, March. Association for Computational Linguistics.

Apertus Project, Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Inés Altemir Mariñas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao, Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. 2025. Apertus: Democratizing open and compliant LLMs for global language environments. *arXiv preprint*, December.

Barrault, Loïc, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In Barrault, Loïc, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Yvette Graham, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, and Matteo Negri, editors, *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi

Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J Mielke, Wilson Y Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-parameter open-access multilingual language model. *arXiv preprint*, November.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with

subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Castaldo, Antonio and Johanna Monti. 2024. Prompting large language models for idiomatic translation. In Vanroy, Bram, Marie-Aude Lefer, Lieve Macken, and Paola Ruffo, editors, *Proceedings of the 1st Workshop on Creative-text Translation and Technology*, pages 32–39, Sheffield, United Kingdom, June. European Association for Machine Translation.

Caswell, Isaac. 2024. 110 new languages are coming to Google Translate. Accessed on 10, Dec 2025.

Chizhov, Pavel, Catherine Arnett, Elizaveta Korotkova, and Ivan P. Yamshchikov. 2024. BPE gets picky: Efficient vocabulary refinement during tokenizer training. In Al-Onaizan, Yaser, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16587–16604, Miami, Florida, USA, November. Association for Computational Linguistics.

Court, Sara and Micha Elsner. 2024. Shortcomings of LLMs for low-resource translation: Retrieval and understanding are both the problem. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1332–1354, Miami, Florida, USA, November. Association for Computational Linguistics.

Dang, John, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint*, December.

DeepSeek-AI. 2025. Deepseek-v3.2-exp: Boosting long-context efficiency with deepseek sparse attention. Accessed on 08, Dec 2025.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A Choquette-Choo, C J Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-Yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *arXiv preprint*, 3.

Groeneveld, Dirk, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya

Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, William Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the science of language models. In Ku, Lun-Wei, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15789–15809, Bangkok, Thailand, August. Association for Computational Linguistics.

Guo, Daya, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z F Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J L Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R J Chen, R L Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S S Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W L Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X Q Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y K Li, Y Q Wang, Y X Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y X Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z Z Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, September.

Guzmán, Francisco, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In Inui, Kentaro, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China, November. Association for Computational Linguistics.

He, Sui. 2024. Prompting ChatGPT for translation: A comparative analysis of translation brief and persona prompts. In Scarton, Carolina, 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 316–326, Sheffield, UK, June. European Association for Machine Translation (EAMT).

Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint*, March.

Hirak, Vitalii, Jaap Jumelet, and Arianna Bisazza. 2026. Assessing the Impact of Typological Features on Multilingual Machine Translation in the Age of Large Language Models. In Demberg, Vera, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2416–2434, Rabat, Morocco, March. Association for Computational Linguistics.

Huang, Xu, Wenhao Zhu, Hanxu Hu, Conghui He, Lei Li, Shujian Huang, and Fei Yuan. 2025. BenchMAX: A comprehensive multilingual evaluation suite for large language models. In Christodoulopoulos, Christos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 16751–16774, Suzhou, China,

November. Association for Computational Linguistics.

Jiang, Juyong, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2025. A survey on large language models for code generation. *ACM Trans. Softw. Eng. Methodol.*, July. Just Accepted.

K2 Team, Zhengzhong Liu, Liping Tang, Linghao Jin, Haonan Li, Nikhil Ranjan, Desai Fan, Shaurya Rohatgi, Richard Fan, Omkar Pangarkar, Huijuan Wang, Zhoujun Cheng, Suqi Sun, Seungwook Han, Bowen Tan, Gurpreet Gosal, Xudong Han, Varad Pimpalkhute, Shibo Hao, Ming Shan Hee, Joel Hestness, Haolong Jia, Liqun Ma, Aaryamonvikram Singh, Daria Soboleva, Natalia Vassilieva, Renxi Wang, Yingquan Wu, Yuekai Sun, Taylor Killian, Alexander Moreno, John Maggs, Hector Ren, Guowei He, Hongyi Wang, Xuezhe Ma, Yuqi Wang, Mikhail Yurochkin, and Eric P Xing. 2026. K2-V2: A 360-open, Reasoning-Enhanced LLM. *arXiv preprint*, January.

Khiu, Eric, Hasti Toossi, David Anugraha, Jinyu Liu, Jiaxu Li, Juan Flores, Leandro Roman, A. Seza Doğruöz, and En-Shiun Lee. 2024. Predicting machine translation performance on low-resource languages: The role of domain similarity. In Graham, Yvette and Matthew Purver, editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1474–1486, St. Julian's, Malta, March. Association for Computational Linguistics.

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

Kulkarni, Ajinkya. 2015. TED Multilingual Parallel Corpus. GitHub, 12. Accessed on 08, Dec 2025.

Kwon, Woosuk, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, SOSP '23, page 611–626, New York, NY, USA. Association for Computing Machinery.

Lambert, Nathan, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2024. TULU 3: Pushing frontiers in open language model post-training. *arXiv preprint*, November.

Lannelongue, Loïc, Jason Grealey, and Michael Inouye. 2021. Green algorithms: Quantifying the carbon footprint of computation. *Advanced Science*, 8(12):2100707.

Littell, Patrick, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In Lapata, Mirella, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain, April. Association for Computational Linguistics.

Liu, Sinuo, Chenyang Lyu, Minghao Wu, Longyue Wang, Weihua Luo, Kaifu Zhang, and Zifu Shang. 2025. New trends for modern machine translation with large reasoning models. *arXiv preprint*, March.

Lundin, Jessica M, Ada Zhang, Nihal Karim, Hamza Louzan, Victor Wei, David Adelani, and Cody Carroll. 2025. The token tax: Systematic bias in multilingual tokenization. *arXiv preprint*, September.

Martins, Pedro Henrique, 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. 2024. EuroLLM: Multilingual language models for europe. *arXiv preprint*, September.

Meta AI. 2024. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Accessed on 08, Dec 2025.

NLLB Team, Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint*, July.

Olmo Team, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman,

Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A Smith, and Hannaneh Hajishirzi. 2025. Olmo 3. *arXiv preprint*, December.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. 2024. OpenAI o1 system card. *arXiv preprint*, December.

OpenAI. 2024. Introducing SWE-bench Verified. Accessed on 11, Dec 2025.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Isabelle, Pierre, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Park, Jeonghyeok and Hai Zhao. 2019. Korean-to-Chinese Machine Translation using Chinese Character as Pivot Clue. *arXiv preprint*, November.

Phan, Long, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Dmitry Dodonov, Tung Nguyen, Jaeho Lee, Daron Anderson, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Oleksandr Pokutnyi, Oleg Iskra, Jessica P

Wang, John-Clark Levin, Mstyslav Kazakov, Fiona Feng, Steven Y Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Serguei Popov, Robert Gerbicz, Geoff Galgon, Johannes Schmitt, Will Yeadon, Yongki Lee, Scott Sauers, Alvaro Sanchez, Fabian Giska, Marc Roth, Søren Riis, Saiteja Utpala, Noah Burns, Gashaw M Goshu, Mohinder Maheshbhai Naiya, Chidozie Agu, Zachary Giboney, Antrell Cheatom, Francesco Fournier-Facio, Sarah-Jane Crowson, Lennart Finke, Zerui Cheng, Jennifer Zampese, Ryan G Hoerr, Mark Nandor, Hyunwoo Park, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Alexis C Garretson, Edwin Taylor, Damien Sileo, Qiuyu Ren, Usman Qazi, Lianghui Li, Jungbae Nam, John B Wydallis, Pavel Arkhipov, Jack Wei Lun Shi, Aras Bacho, Chris G Willcocks, Hangrui Cao, Sumeet Motwani, Emily de Oliveira Santos, Johannes Veith, Edward Vendrow, Doru Cojoc, Kengo Zenitani, Joshua Robinson, Longke Tang, Yuqi Li, Joshua Vendrow, Natanael Wildner Fraga, Vladyslav Kuchkin, Andrey Pupasov Maksimov, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Aleksandar Mikov, Andrew Gritsevskiy, Julien Guillod, Gözdenur Demir, Dakotah Martinez, Ben Pageler, Kevin Zhou, Saeed Soori, Ori Press, Henry Tang, Paolo Rissone, Sean R Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, Joseph Marvin Imperial, Ameya Prabhu, Jinzhou Yang, Nick Crispino, Arun Rao, Dimitri Zvonkine, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Tad Hogg, Carlo Bosio, Brian P Coppola, Julian Salazar, Jaehyeok Jin, Rafael Sayous, Stefan Ivanov, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Kelsey Van den Houte, Lynn Van Der Sypt, Brecht Verbeken, David Noever, Alexei Kopylov, Benjamin Myklebust, Bikun Li, Lisa Schut, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Ariel Ghislain Kemogne Kamdoum, Alvin Jin, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Gongbo Sun, Daniil S Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Joseph M Cavanagh, Daofeng Li, Jiawei Shen, Donato Crisostomi, Wenjin Zhang, Ali Dehghan, Sergey Ivanov, David Perrella, Nurdin Kaparov, Allen Zang, Ilia Sucholutsky, Arina Kharlamova, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Shankar Sivarajan, Dan Bar Hava, Aleksey Kuchkin, David Holmes, Alexandra Rodriguez-Romero, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Zakayo Kazibwe, Don Clarke, Dae Hyun Kim, Felipe Meneguitti Dias, Sara Fish, Veit Elser, Tobias Kreiman, Victor Efren Guadarrama Vilchis, Immo Klose, Ujjwala Anantheswaran, Adam Zweiger, Kaivalya Rawal, Jeffery Li, Jeremy Nguyen, Nicolas Daans, Haline Heidinger, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Christian Stump, Niv Cohen, Rafał Poświata, Josef Tkadlec, Alan Goldfarb, Chenguang Wang, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Ryan Stendall, Jamie Tucker-Foltz, Jack Stade, T Ryan Rogers, Tom Goertzen, Declan Grabb, Abhishek Shukla, Alan Givré, John Arnold Ambay, Archan Sen, Muhammad Fayez Aziz, Mark H Inlow, Hao He, Ling Zhang, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Harrison K Wang, Kalyan Ramakrishnan, Elliott Thornley, Antonio Terpin, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Ethan D L Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Martin Stehberger, Peter Bradshaw, J P Heimonen, Kaustubh Sridhar, Ido Akov, Jennifer Sandlin, Yury Makarychev, Joanna Tam, Hieu Hoang, David M Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K Cohen, Orr Paradise, Jan Hendrik Kirchner, Maksym Ovchynnikov, Jason O Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Shreyas Verma, Prashant Joshi, Eli Meril, Ziqiao Ma, Jérémy Andréoletti, Raghav Singhal, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Marco Piccardo, Hamid Mostaghimi, Qijia Chen, Virendra Singh, Tran Quoc Khánh, Paul Rosu, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Aline Menezes, Jonathan Roberts, William Alley, Kunyang Sun, Arkil Patel, Max Lamparth, Anka Reuel, Linwei Xin, Hanmeng Xu, Jacob Loader, Freddie Martin, Zixuan Wang, Andrea Achilleos, Thomas Preu, Tomek Korbak, Ida Bosio, Fereshteh Kazemi, Ziye Chen, Biró Bálint, Eve J Y Lo, Jiaqi Wang, Maria Inês S Nunes, Jeremiah Milbauer, M Saiful Bari, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Hossam Elgnainy, Guillaume Douville, Daniel Tordera, George Balabanian, Hew Wolff, Lynna Kvistad, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Andrew Favre D O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Sherwin Abdoli, Tim Santens, Shaul Barkan, Allison Tee, Robin Zhang, Alessandro Tomasiello, G Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Jiayi Pan, Emma Rodman, Jacob Drori, Carl J Fossum, Niklas Muennighoff, Milind Jagota, Ronak Pradeep, Honglu Fan, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Mohammadreza Mofayezi, Alexander Piperski, David K Zhang, Kostiantyn Dobarskyi, Roman Leventov, Ignat Soroko, Joshua Duersch, Vage Taamazyan, Andrew Ho, Wenjie Ma, William Held, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Claudio Di Fratta, Edson Oliveira, Joseph W Jackson, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen,

Bita Golshani, David Stap, Egor Kretov, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Nick Winter, Miguel Orbegozo Rodriguez, Robert Lauff, Dustin Wehr, Colin Tang, Zaki Hossain, Shaun Phillips, Fortuna Samuele, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Rayner Hernandez Perez, Daniel Pyda, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Mukhwinder Singh, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Harsh Kumar, Chiara Ceconello, Chao Zhuang, Haon Park, Micah Carroll, Andrew R Tawfeek, Stefan Steinerberger, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Jainam Shah, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Paolo Giordano, Philipp Petersen, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Zheng-Xin Yong, Florencia de la Rosa, Nathan Cho, Xiuyu Li, Guillaume Malod, Orion Weller, Guglielmo Albani, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Rai, Filippo Bigi, M C Boscá, Oleg Shumar, Kaniuar Bacho, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Thomas C H Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Stefano Cavalleri, Olle Häggström, Emil Verkama, Joshua Newbould, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Ting Wang, Yosi Kratish, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Christian Schroeder de Witt, Pablo Hernández-Cámara, Emanuele Rodolà, Jules Robins, Dominic Williamson, Vincent Cheng, Brad Raynor, Hao Qi, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y Wang, Kaylie Hausknecht, Michael P Brenner, Mao Mao, Christoph Demian, Peyman Kassani, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D P Shinde, Yan Carlos Leyva Labrador, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C B Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Earth Anderson, Rodrigo De Oliveira Pena, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Ross Finocchio, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Isaac C McAlister, Alejandro José Moyano, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Yana Malysheva, Daphiny Pottmaier, Omid Taheri, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M R Minissi, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Ronald Clark, Josh Ducey, Matheus Piza, Maja Somrak, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Antoine Jallon, I M J McInnis, Evan Chen, Avi Semler, Luk Gloor, Tej Shah, Marc Carauleanu, Pascal Lauer, Tran Đuc Huy, Hossein Shahrtash, Emilien Duc, Lukas Lewark, Assaf Brown, Samuel Albanie, Brian Weber, Warren S Vaz, Pierre Clavier, Yiyang Fan, Gabriel Poesia Reis e Silva, Long, Lian, Marcus Abramovitch, Xi Jiang, Sandra Mendoza, Murat Islam, Juan Gonzalez, Vasilios Mavroudis, Justin Xu, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Thorben Jansen, Antonella Pinto, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Tong Jiang, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Gang Zhang, Zhehang Du, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D Hartman, Tim Tarver, Stephen Mensah, Gautier Abou Loume, Wiktor Morak, Farzad Habibi, Sarah Hoback, Will Cai, Javier Gimenez, Roselynn Grace Montecillo, Jakub Łucki, Russell Campbell, Asankhaya Sharma, Khalida Meer, Shreen Gul, Daniel Espinosa Gonzalez, Xavier Alapont, Alex Hoover, Gunjan Chhablani, Freddie Vargus, Arunim Agarwal, Yibo Jiang, Deepakkumar Patil, David Outevsky, Kevin Joseph Scaria, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Ashley Cartwright, Sergei Bogdanov, Niels Mündler, Sören Möller, Luca Arnaboldi, Kunvar Thaman, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Tony Fruhauff, Glen Sherman, Mátyás Vincze, Siranut Usawasutsakorn, Dylan Ler, Anil Radhakrishnan, Innocent Enyekwe, Sk Md Salauddin, Jiang Muzhen, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Claire Sparrow, Jasdeep Sidhu, Sam Ali, Song Bian, John Lai, Eric Singer, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Dario Bezzi, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Imad Ali Shah, Jun Jin, Scott Creighton, Denis Peskoff, Zienab EL-Wasif, Ragavendran P, V, Michael Richmond, Joseph McGowan, Tejal Patwardhan, Hao-Yu Sun, Ting Sun, Nikola Zubić, Samuele Sala, Stephen Ebert, Jean Kaddour, Manuel Schottdorf, Dianzhuo Wang, Gerol Petruzella, Alex Meiburg, Tilen Medved, Ali ElSheikh, S Ashwin Hebbar, Lorenzo Vaquero, Xianjun Yang, Jason Poulos, Vilém Zouhar, Sergey Bogdanik, Mingfang Zhang, Jorge Sanz-Ros, David Anugraha, Yinwei Dai, Anh N Nhu, Xue Wang, Ali Anil Demircali, Zhibai Jia, Yuyin Zhou, Juncheng Wu, Mike He, Nitin Chandok, Aarush Sinha, Gaoxiang Luo, Long Le, Mickaël Noyé, Michał Perełkiewicz, Ioannis Pantidis, Tianbo Qi, Soham Sachin Purohit, Letitia Parcalabescu, Thai-Hoa Nguyen, Genta Indra Winata, Edoardo M

Ponti, Hanchen Li, Kaustubh Dhole, Jongee Park, Dario Abbondanza, Yuanli Wang, Anupam Nayak, Diogo M Caetano, Antonio A W L Wong, Maria del Rio-Chanona, Dániel Kondor, Pieter Francois, Ed Chalstrey, Jakob Zsambok, Dan Hoyer, Jenny Reddish, Jakob Hauser, Francisco-Javier Rodrigo-Ginés, Suchandra Datta, Maxwell Shepherd, Thom Kamphuis, Qizheng Zhang, Hyunjun Kim, Ruiji Sun, Jianzhu Yao, Franck Dernoncourt, Satyapriya Krishna, Sina Rismanchian, Bonan Pu, Francesco Pinto, Yingheng Wang, Kumar Shridhar, Kalon J Overholt, Glib Briia, Hieu Nguyen, David, Soler Bartomeu, Tony C Y Pang, Adam Wecker, Yifan Xiong, Fanfei Li, Lukas S Huber, Joshua Jaeger, Romano De Maddalena, Xing Han Lù, Yuhui Zhang, Claas Beger, Patrick Tser Jern Kon, Sean Li, Vivek Sanker, Ming Yin, Yihao Liang, Xinlu Zhang, Ankit Agrawal, Li S Yifei, Zechen Zhang, Mu Cai, Yasin Sonmez, Costin Cozianu, Changhao Li, Alex Slen, Shoubin Yu, Hyun Kyu Park, Gabriele Sarti, Marcin Briański, Alessandro Stolfo, Truong An Nguyen, Mike Zhang, Yotam Perlitz, Jose Hernandez-Orallo, Runjia Li, Amin Shabani, Felix Juefei-Xu, Shikhar Dhingra, Orr Zohar, My Chiffon Nguyen, Alexander Pondaven, Abdurrahim Yilmaz, Xuandong Zhao, Chuanyang Jin, Muyan Jiang, Stefan Todoran, Xinyao Han, Jules Kreuer, Brian Rabern, Anna Plassart, Martino Maggetti, Luther Yap, Robert Geirhos, Jonathon Kean, Dingsu Wang, Sina Mollaei, Chenkai Sun, Yifan Yin, Shiqi Wang, Rui Li, Yaowen Chang, Anjiang Wei, Alice Bizeul, Xiaohan Wang, Alexandre Oliveira Arrais, Kushin Mukherjee, Jorge Chamorro-Padial, Jiachen Liu, Xingyu Qu, Junyi Guan, Adam Bouyamourn, Shuyu Wu, Martyna Plomecka, Junda Chen, Mengze Tang, Jiaqi Deng, Shreyas Subramanian, Haocheng Xi, Haoxuan Chen, Weizhi Zhang, Yinuo Ren, Haoqin Tu, Sejong Kim, Yushun Chen, Sara Vera Marjanović, Junwoo Ha, Grzegorz Luczyna, Jeff J Ma, Zewen Shen, Dawn Song, Cedegao E Zhang, Zhun Wang, Gaël Gendron, Yunze Xiao, Leo Smucker, Erica Weng, Kwok Hao Lee, Zhe Ye, Stefano Ermon, Ignacio D Lopez-Miguel, Theo Knights, Anthony Gitter, Namkyu Park, Boyi Wei, Hongzheng Chen, Kunal Pai, Ahmed Elkhanany, Han Lin, Philipp D Siedler, Jichao Fang, Ritwik Mishra, Károly Zsolnai-Fehér, Xilin Jiang, Shadab Khan, Jun Yuan, Rishab Kumar Jain, Xi Lin, Mike Peterson, Zhe Wang, Aditya Malusare, Maosen Tang, Isha Gupta, Ivan Fosin, Timothy Kang, Barbara Dworakowska, Kazuki Matsumoto, Guangyao Zheng, Gerben Sewuster, Jorge Pretel Villanueva, Ivan Rannev, Igor Chernyavsky, Jiale Chen, Deepayan Banik, Ben Racz, Wenchao Dong, Jianxin Wang, Laila Bashmal, Duarte V Gonçalves, Wei Hu, Kaushik Bar, Ondrej Bohdal, Atharv Singh Patlan, Shehzaad Dhuliawala, Caroline Geirhos, Julien Wist, Yuval Kansal, Bingsen Chen, Kutay Tire, Atak Talay Yücel, Brandon Christof, Veerupaksh Singla, Zijian Song, Sanxing Chen, Jiaxin Ge, Kaustubh Ponkshe, Isaac Park, Tianneng Shi, Martin Q Ma, Joshua Mak, Sherwin Lai, Antoine Moulin, Zhuo Cheng, Zhanda Zhu, Ziyi Zhang, Vaidehi Patil, Ketan Jha, Qiutong Men, Jiaxuan Wu, Tianchi Zhang, Bruno Hebling Vieira, Alham Fikri Aji, Jae-Won Chung, Mohammed Mahfoud, Ha Thi Hoang, Marc Sperzel, Wei Hao, Kristof Meding, Sihan Xu, Vassilis Kostakos, Davide Manini, Yueying Liu, Christopher Toukmaji, Jay Paek, Eunmi Yu, Arif Engin Demircali, Zhiyi Sun, Ivan Dewerpe, Hongsen Qin, Roman Pflugfelder, James Bailey, Johnathan Morris, Ville Heilala, Sybille Rosset, Zishun Yu, Peter E Chen, Woongyeong Yeo, Eeshaan Jain, Ryan Yang, Sreekar Chigurupati, Julia Chernyavsky, Sai Prajwal Reddy, Subhashini Venugopalan, Hunar Batra, Core Francisco Park, Hieu Tran, Guilherme Maximiano, Genghan Zhang, Yizhuo Liang, Hu Shiyu, Rongwu Xu, Rui Pan, Siddharth Suresh, Ziqi Liu, Samaksh Gulati, Songyang Zhang, Peter Turchin, Christopher W Bartlett, Christopher R Scotese, Phuong M Cao, Ben Wu, Jacek Karwowski, Davide Scaramuzza, Aakaash Nattanmai, Gordon McKellips, Anish Cheraku, Asim Suhail, Ethan Luo, Marvin Deng, Jason Luo, Ashley Zhang, Kavin Jindel, Jay Paek, Kasper Halevy, Allen Baranov, Michael Liu, Advaith Avadhanam, David Zhang, Vincent Cheng, Brad Ma, Evan Fu, Liam Do, Joshua Lass, Hubert Yang, Surya Sunkari, Vishruth Bharath, Violet Ai, James Leung, Rishit Agrawal, Alan Zhou, Kevin Chen, Tejas Kalpathi, Ziqi Xu, Gavin Wang, Tyler Xiao, Erik Maung, Sam Lee, Ryan Yang, Roy Yue, Ben Zhao, Julia Yoon, Sunny Sun, Aryan Singh, Ethan Luo, Clark Peng, Tyler Osbey, Taozhi Wang, Daryl Echeazu, Hubert Yang, Timothy Wu, Spandan Patel, Vidhi Kulkarni, Vijaykaarti Sundarapandiyan, Ashley Zhang, Andrew Le, Zafir Nasim, Srikar Yalam, Ritesh Kasamsetty, Soham Samal, Hubert Yang, David Sun, Nihar Shah, Abhijeet Saha, Alex Zhang, Leon Nguyen, Laasya Nagumalli, Kaixin Wang, Alan Zhou, Aidan Wu, Jason Luo, Anwith Telluri, Summer Yue, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's Last Exam. *arXiv preprint*, September.

Ploeger, Esther, Johannes Bjerva, Jörg Tiedemann, and Robert Oestling. 2025. A cross-lingual perspective on neural machine translation difficulty. In Haddow, Barry, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 340–354, Suzhou, China, November. Association for Computational Linguistics.

Popović, Maja. 2017. chrF++: words helping character n-grams. In Bojar, Ondřej, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer, editors, *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark, September. Association for Computational Linguistics.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes,

Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October. Association for Computational Linguistics.

Provilkov, Ivan, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In Jurafsky, Dan, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online, July. Association for Computational Linguistics.

Pu, Xiao, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint*, September.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models!, 9. Accessed on 08, Dec 2025.

Qwen Team. 2025. Qwen3: Think Deeper, Act Faster, 4. Accessed on 08, Dec 2025.

Rei, Ricardo, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In Webber, Bonnie, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online, November. Association for Computational Linguistics.

Rei, Ricardo, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022a. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In Koehn, Philipp, 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 578–585, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Rei, Ricardo, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In Koehn, Philipp, 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 634–645, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Rei, Ricardo, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F T Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual LLMs. *arXiv preprint*, June.

Romanou, Angelika, Negar Foroutan, Anna Sotnikova, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Zeming Chen, Mohamed A. Haggag, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Danylo Boiko, Michael Chang, Jenny Chim, Gal Cohen, Aditya Kumar Dalmia, Abraham Diress, Sharad Duwal, Daniil Dzenhaliou, Daniel Fernando Erazo Florez, Fabian Farestam, Joseph Marvin Imperial, Shayekh Bin Islam, Perttu Isotalo, Maral Jabbarishiviari, Börje F. Karlsson, Eldar Khalilov, Christopher Klamm, Fajri Koto, Dominik Krzemiński, Gabriel Adriano de Melo, Syrielle Montariol, Yiyang Nan, Joel Niklaus, Jekaterina Novikova, Johan Samir Obando Ceron, Debjit Paul, Esther Ploeger, Jebish Purbey, Swati Rajwal, Selvan Sunitha Ravi, Sara Rydell, Roshan Santhosh, Drishti Sharma, Marjana Prifti Skenduli, Arshia Soltani Moakhar, Bardia soltani moakhar, Ayush Kumar Tarun, Azmine Toushik Wasi, Thenuka Ovin Weerasinghe, Serhan Yilmaz, Mike Zhang, Imanol Schlag, Marzieh Fadaee, Sara Hooker, and Antoine Bosselut. 2025. INCLUDE: Evaluating Multilingual Language Understanding with Regional Knowledge. In *The Thirteenth International Conference on Learning Representations*.

Rust, Phillip, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In Zong, Chengqing, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online, August. Association for Computational Linguistics.

Scherrer, Yves, Luka Nerima, Lorenza Russo, Maria Ivanova, and Eric Wehrli. 2014. SwissAdmin: A multilingual tagged parallel corpus of press releases. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios

Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1832–1836, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Schwartz, Roy, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2020. Green ai. *Commun. ACM*, 63(12):54–63, November.

Sindhujan, Archchana, Diptesh Kanojia, Constantin Orasan, and Shenbin Qian. 2025. When LLMs struggle: Reference-less translation evaluation for low-resource languages. In Hettiarachchi, Hansi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage, editors, *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 437–459, Abu Dhabi, United Arab Emirates, January. Association for Computational Linguistics.

Singh, Telem Joyson, Ranbir Singh Sanasam, and Priyankoo Sarmah. 2025. An information-theoretic approach to reducing fertility in LLMs for Manipuri machine translation. In Inui, Kentaro, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh, editors, *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 2394–2404, Mumbai, India, December. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Song, Yewei, Lujun Li, Cedric Lothritz, Saad Ezzini, Lama Sleem, Niccolo Gentile, Radu State, Tegawendé F Bissyandé, and Jacques Klein. 2025. Is small language model the silver bullet to low-resource languages machine translation? *arXiv preprint*, August.

Stewart, Craig, Ricardo Rei, Catarina Farinha, and Alon Lavie. 2020. COMET - deploying a new state-of-the-art MT evaluation metric in production. In Campbell, Janice, Dmitriy Genzel, Ben Huyck, and Patricia O'Neill-Brown, editors, *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 78–109, Virtual, October. Association for Machine Translation in the Americas.

Vilar, David, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2023. Prompting PaLM for translation: Assessing strategies and performance. In Rogers, Anna, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15406–15427, Toronto, Canada, July. Association for Computational Linguistics.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Liu, Qun and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Ye, Yongshi, Biao Fu, Chongxuan Huang, Yidong Chen, and Xiaodong Shi. 2025. How well do large reasoning models translate? a comprehensive evaluation for multi-domain machine translation. *arXiv preprint*, May.

Yue, Xiang, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2025. MMMU-pro: A more robust multi-discipline multimodal understanding benchmark. In Che, Wanxiang, 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 15134–15186, Vienna, Austria, July. Association for Computational Linguistics.

Zhang, Biao, Barry Haddow, and Alexandra Birch. 2023. Prompting large language model for machine translation: a case study. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zhang, Wenxuan, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico, June. Association for Computational Linguistics.

Zhang, Biao, Fedor Moiseev, Joshua Ainslie, Paul Suganthan, Min Ma, Surya Bhupatiraju, Fede Lebron, Orhan Firat, Armand Joulin, and Zhe Dong. 2025. Encoder-decoder Gemma: Improving the quality-efficiency trade-off via adaptation. *arXiv preprint*, April.

A Appendix: Dataset Details

Lang_pairs	Test_size	Source
Arabic-Chinese (ar-zh)	3,000	TED Multilingual Parallel Corpora
Arabic-Hebrew (ar-he)	3,000	TED Multilingual Parallel Corpora
Chinese-French (zh-fr)	3,000	TED Multilingual Parallel Corpora
Chinese-Russian (zh-ru)	3,000	TED Multilingual Parallel Corpora
French-Italian (fr-it)	3,000	SwissAdmin
German-French (de-fr)	3,000	SwissAdmin
German-Italian (de-it)	3,000	SwissAdmin
Korean-Chinese (ko-zh)	3,000	Chinese-Korean Parallel Corpus
Korean-French (ko-fr)	3,000	TED Multilingual Parallel Corpora
Russian-French (ru-fr)	3,000	TED Multilingual Parallel Corpora
English-Chinese (en-zh)	3,000	WMT20 QE Shared Task
English-Czech (en-cs)	3,000	WMT20 QE Shared Task
English-German (en-de)	3,000	WMT20 QE Shared Task
English-Polish (en-pl)	3,000	WMT20 QE Shared Task
English-Russian (en-ru)	3,000	WMT20 QE Shared Task
English-Tamil (en-ta)	3,000	WMT20 QE Shared Task
Chinese-English (zh-en)	3,000	WMT20 QE Shared Task
Czech-English (cs-en)	3,000	WMT20 QE Shared Task
German-English (de-en)	3,000	WMT20 QE Shared Task
Khmer-English (km-en)	3,000	WMT20 QE Shared Task
Russian-English (ru-en)	3,000	WMT20 QE Shared Task
Tamil-English (ta-en)	3,000	WMT20 QE Shared Task

Table A.1: The size of our test set for each language pair and their corresponding sources.

B Appendix: LLM Output Noise

English Source: WTI crude futures were last down by 0.9% to $55.91 after falling as low as $54.79. Brent crude futures dropped 1.4% to $61.89.
Czech Translation: Závěry futures na ropu WTI klesly o 0,9 % na 55,91 dolarů poté, co dosáhly minima 54,79 dolarů. Futures na ropu Brent klesly o 1,4 % na 61,89 dolarů.\n\n**Explanation of choices:**\n\n* **Závěry futures:** F̈uturesẗranslates well to f̈uturesïn Czech, but to be clearer we specify it's the *conclusions* or *settlements* of the futures contracts.\n* **klesly o:** ẅere down byẗranslates naturally to k̈lesly o(̈fell by).\n* **poté, co dosáhly minima:** äfter falling as low asïs best expressed as äfter reaching a minimum.\̈n* **dolarů:** d̈ollarsïs d̈olarůïn Czech.

Figure B.1: Noise in LLM Output from gemma-3-27b-it using Prompt 0.

C Appendix: Prompt Templates

Prompt 0

{src_lang}: {src_txt}
{tgt_lang}:

Prompt 1

Translate the following {src_lang} into {tgt_lang}: {src_text}

Prompt 2

Translate the following {src_lang} into {tgt_lang} and only output the target text: {src_text}

D Appendix: LLM Output Noise Detection

We introduce a rule-based method to calculate the proportion of instances that contain only the translation without extra explanatory text or text in an incorrect language, to quantitatively detect noise in LLM outputs. We term this metric the clean translation rate (Clean%). MT outputs containing extra explanatory text were detected using regular expressions matching explanatory terms such as "explanation", "indicate", and "analysis". Outputs in the wrong target language were identified based on a language identification model from fastText (Bojanowski et al., 2017) with a confidence threshold of 60%. An instance is classified as a clean translation only when it contains neither extra explanatory text nor text in an incorrect target language with more than 60% confidence. The clean translation rate is formally defined in Equation 2:

$$\text{Clean}\% = \frac{N - |E \cup W|}{N} \tag{2}$$

where is N is the total number of instances, E is the set of instances containing explanatory text and W is the set containing text in the wrong language. $\text{Exp}\%$ and $\text{WrongL}\%$ are defined as $\frac{|E|}{N}$ and $\frac{|W|}{N}$.

	Clean% ↑			Expl% ↓			WrongL% ↓		
Model_name	Prompt 0	Prompt 1	Prompt 2	Prompt 0	Prompt 1	Prompt 2	Prompt 0	Prompt 1	Prompt 2
Qwen3-30B-A3B-Instruct-2507	95.98%	96.69%	**96.72%**	2.95%	2.69%	**2.62%**	1.18%	**0.64%**	0.68%
Qwen3-30B-A3B-Thinking-2507	96.70%	**96.74%**	96.72%	2.81%	**2.76%**	2.78%	0.65%	**0.51%**	0.52%
Qwen3-4B-Instruct-2507	95.35%	**96.23%**	96.21%	**3.03%**	3.09%	3.08%	1.77%	**0.70%**	0.77%
Qwen3-4B-Thinking-2507	**96.49%**	96.34%	96.38%	**2.84%**	2.94%	2.91%	0.99%	0.75%	**0.73%**
Llama-3.2-3B-Instruct	59.42%	30.46%	**92.91%**	29.47%	67.60%	**3.30%**	21.67%	15.87%	**4.03%**
gemma-3-27b-it	2.68%	0.27%	**96.76%**	97.32%	99.72%	**2.79%**	31.82%	31.82%	**0.46%**
Qwen2.5-32B-Instruct	44.38%	71.95%	**90.94%**	53.93%	26.71%	**7.21%**	15.22%	8.08%	**4.10%**
DeepSeek-R1-Distill-Qwen-32B	73.25%	84.10%	**95.81%**	22.22%	14.42%	**2.89%**	11.92%	5.18%	**1.36%**
aya-expanse-32b	83.37%	68.29%	**96.35%**	15.03%	27.46%	**3.10%**	4.93%	12.47%	**0.60%**
Tower-Plus-72B	94.85%	94.74%	**96.12%**	3.34%	4.58%	**3.33%**	2.05%	0.94%	**0.58%**
t5gemma-xl-xl-prefixlm-it	53.68%	72.36%	**82.65%**	33.88%	12.86%	**4.11%**	25.41%	19.35%	**14.23%**
DeepSeek-V3.2-Exp-671B-chat	39.28%	/	**96.81%**	59.17%	/	**2.86%**	11.65%	/	**0.37%**
DeepSeek-V3.2-Exp-671B-reasoner	/	/	95.41%	/	/	4.29%	/	/	1.90%

Table D.1: The clean translation rate (Clean%), the rate of generating extra explanatory texts (Expl%) and the rate of outputting wrong language (WrongL%) for different prompts and LLMs. We did not run all prompts on DeepSeek-V3.2-Exp as we see much better performance on other LLMs using Prompt 2.

Table D.1 shows Clean% across different prompts and models. The results demonstrate that DeepSeek-V3.2-Exp exhibits the strongest performance in translation instruction following, and Prompt 2 yields the cleanest translation output among the three prompt templates. This finding was confirmed by manual inspection, and Prompt 2 was therefore used for all subsequent experiments and analyses.

E Appendix: Additional Evaluation Results

	non-English centric											En-XX							XX-En						
model_name	ar-he	ar-zh	de-fr	de-it	fr-it	ko-fr	ko-zh	ru-fr	zh-fr	zh-ru	mean	en-cs	en-de	en-pl	en-ru	en-ta	en-zh	mean	cs-en	de-en	km-en	ru-en	ta-en	zh-en	mean
Qwen3-30B-A3B-Instruct-2507	0.7278	0.7440	0.8202	0.8524	0.8615	0.6684	0.8902	0.7152	0.7153	0.7401	0.7735	0.8707	0.8621	0.8710	0.8795	0.8554	0.8828	0.8703	0.8306	0.8680	0.7966	0.8589	0.8570	0.8323	0.8406
Qwen3-30B-A3B-Thinking-2507	0.7108	0.7352	0.8193	0.8528	0.8622	0.6636	0.8907	0.7163	0.7153	0.7401	0.7706	0.8824	0.8650	0.8791	0.8821	0.8729	0.8808	0.8771	0.8266	0.8668	0.7876	0.8602	0.8577	0.8309	0.8383
Qwen3-4B-Instruct-2507	0.6630	0.7321	0.8026	0.8341	0.8506	0.6555	0.8806	0.7048	0.7052	0.7316	0.7560	0.7839	0.8379	0.8140	0.8594	0.6724	0.8785	0.8077	0.8145	0.8611	0.7354	0.8502	0.8311	0.8278	0.8200
Qwen3-4B-Thinking-2507	0.6692	0.7238	0.8079	0.8432	0.8570	0.6544	0.8823	0.7087	0.7105	0.7325	0.7589	0.8326	0.8513	0.8437	0.8668	0.7756	0.8774	0.8412	0.8141	0.8621	0.7253	0.8546	0.8353	0.8284	0.8200
Llama-3.2-3B-Instruct	0.5145	0.6595	0.7528	0.7840	0.8263	0.6111	0.7953	0.6639	0.6698	0.6372	0.6914	0.6763	0.7986	0.7786	0.7432	0.7199	0.7954	0.7520	0.7914	0.8446	0.6399	0.8298	0.7982	0.7948	0.7831
gemma-3-27b-it	0.7771	0.7363	0.8259	0.8596	0.8655	0.6713	0.8894	0.7208	0.7175	0.7423	0.7806	0.9021	0.8686	0.8989	0.8912	0.9010	0.8793	0.8902	0.8348	0.8659	0.8129	0.8602	0.8641	0.8309	0.8448
Qwen2.5-32B-Instruct	0.6078	0.7399	0.8079	0.8347	0.8499	0.6597	0.8888	0.7121	0.7038	0.7144	0.7519	0.8151	0.8260	0.8100	0.8337	0.6235	0.8746	0.7972	0.8215	0.8626	0.7284	0.8539	0.7954	0.8095	0.8119
DeepSeek-R1-Distill-Qwen-32B	0.6546	0.7266	0.8035	0.8314	0.8511	0.6521	0.8856	0.7029	0.7033	0.7209	0.7532	0.8062	0.8384	0.8230	0.8558	0.5566	0.8747	0.7925	0.8184	0.8606	0.6794	0.8551	0.7412	0.8280	0.7971
aya-expanse-32b	0.7854	0.7355	0.8270	0.8608	0.8650	0.6786	0.8884	0.7231	0.7204	0.7438	0.7828	0.9060	0.8714	0.8963	0.8904	0.8484	0.8759	0.8814	0.8341	0.8668	0.5960	0.8613	0.8567	0.8308	0.8076
Tower-Plus-72B	0.7383	0.7508	0.8275	0.8594	0.8659	**0.6827**	**0.8963**	**0.7280**	**0.7308**	**0.7561**	0.7836	0.9047	0.8753	0.9018	**0.9007**	0.6038	0.8880	0.8457	**0.8367**	**0.8731**	0.7547	0.8686	0.8151	0.8387	0.8312
t5gemma-xl-xl-prefixlm-it	0.5131	0.5943	0.7398	0.7546	0.7933	0.5715	0.7519	0.6106	0.6266	0.5866	0.6542	0.5319	0.7054	0.575	0.6286	0.4382	0.7134	0.5988	0.7647	0.8352	0.4637	0.8182	0.7514	0.7953	0.7381
DeepSeek-V3.2-Exp-671B-chat	0.7874	0.7473	**0.8280**	**0.8660**	**0.8660**	0.6738	0.8953	0.7193	0.7216	0.7457	0.7850	0.9031	0.8693	0.8979	0.8927	0.8952	0.8787	0.8895	0.8338	0.8688	0.8196	0.8640	0.8628	0.8318	0.8468
DeepSeek-V3.2-Exp-671B-reasoner	0.6683	0.7432	0.8272	0.8590	0.8654	0.6664	0.8923	0.7166	0.7175	0.7405	0.7696	0.9012	0.8661	0.8958	0.8901	0.8835	0.8776	0.8857	0.8342	0.8683	0.8163	0.8624	0.8632	0.8320	0.8461
nllb-200-3.3B	0.7882	0.7248	0.8162	0.8513	0.8618	0.6696	0.8242	0.7116	0.6960	0.7264	0.7672	0.8817	0.8485	0.8786	0.8772	0.8768	0.7846	0.8579	0.7882	0.8446	0.7675	0.8518	0.8487	0.8001	0.8168
nllb-moe-54b	**0.7989**	0.7323	0.8179	0.8531	0.8618	0.6753	0.8330	0.7152	0.6940	0.7340	0.7716	0.8825	0.843	0.8875	0.8828	0.8770	0.7817	0.8591	0.7691	0.8367	0.7773	0.8534	0.8476	0.8044	0.8148
Google Translate	0.7946	**0.7530**	0.8246	0.8613	0.8655	0.6788	0.8824	0.7241	0.7250	0.7517	**0.7861**	**0.9118**	**0.8778**	**0.9088**	0.8985	**0.9019**	**0.8905**	**0.8982**	0.8347	0.8711	**0.8240**	**0.8693**	**0.8687**	**0.8393**	**0.8512**

Table E.1: COMET scores of translations for 22 language pairs using Prompt 2 under **zero-shot setting**.

F Appendix: Correlation Between BLEU, chrF++, TAR and Typological Distances

	non-English centric											En-XX								XX-En						
model_name	ar-he	ar-zh	de-fr	de-it	fr-it	ko-fr	ko-zh	ru-fr	zh-fr	zh-ru	mean	en-cs	en-de	en-pl	en-ru	en-ta	en-zh	mean	cs-en	de-en	km-en	ru-en	ta-en	zh-en	mean
Qwen3-30B-A3B-Instruct-2507	10.34	17.72	28.87	26.46	26.69	11.31	32.95	20.09	14.50	7.50	19.64	25.39	32.34	21.96	24.34	5.50	39.88	24.9	28.27	37.60	17.20	37.79	19.89	28.02	28.13
Qwen3-30B-A3B-Thinking-2507	9.30	18.28	29.18	26.02	27.18	11.55	33.12	19.40	14.87	7.19	19.61	26.19	33.94	21.58	24.91	6.77	39.02	25.40	28.67	38.96	17.41	38.65	21.40	28.96	29.01
Qwen3-4B-Instruct-2507	4.44	15.68	24.52	20.88	23.94	10.05	30.08	17.67	12.28	6.62	16.62	15.86	27.15	12.84	20.73	0.65	37.83	19.18	25.26	36.07	9.70	35.09	14.99	26.99	24.68
Qwen3-4B-Thinking-2507	6.44	16.93	25.89	22.88	25.33	10.92	31.19	18.42	13.75	6.30	17.81	19.10	29.58	17.22	20.90	2.86	37.64	21.22	26.31	36.60	10.88	36.57	17.41	27.31	25.85
Llama-3.2-3B-Instruct	0.50	3.62	18.12	12.90	22.35	3.41	9.96	13.60	8.15	2.45	9.51	12.03	24.00	13.77	12.04	2.12	23.09	14.51	25.05	35.47	3.16	33.36	11.73	21.71	21.75
gemma-3-27b-it	13.45	18.48	31.38	28.88	28.40	11.83	34.61	21.45	14.28	7.24	21.00	32.31	36.23	26.59	27.24	9.71	41.07	28.86	30.49	40.98	20.49	39.48	23.78	29.67	30.81
Qwen2.5-32B-Instruct	0.05	18.21	17.69	19.58	25.06	4.59	33.51	6.62	5.09	0.48	13.09	12.31	25.07	2.37	4.13	0.47	40.94	14.21	29.43	40.08	11.12	36.71	15.03	14.40	24.46
DeepSeek-R1-Distill-Qwen-32B	1.96	16.96	26.59	21.53	24.70	10.28	32.83	17.81	13.53	6.50	17.27	19.58	27.43	17.19	20.64	0.95	39.54	20.89	28.12	32.10	8.05	36.34	10.49	28.23	23.89
aya-expanse-32b	13.46	16.29	32.81	30.23	28.95	12.34	34.52	21.68	14.06	7.22	21.16	31.39	35.51	25.14	26.08	5.45	40.88	27.41	30.42	39.44	3.13	38.85	20.65	28.46	26.83
Tower-Plus-72B	9.82	**19.90**	32.25	29.24	28.40	13.36	**38.12**	21.80	15.35	**8.15**	**21.64**	34.37	**39.59**	26.62	**30.55**	0.50	**45.32**	29.49	**32.41**	**43.40**	14.23	**42.58**	17.98	32.37	30.50
t5gemma-xl-xl-prefixlm-it	0.65	5.43	19.01	14.51	19.83	3.61	14.03	9.71	7.49	2.69	9.70	6.85	18.97	5.90	8.92	0.11	19.58	10.05	22.00	36.09	0.62	32.87	10.82	23.58	21.00
DeepSeek-V3.2-Exp-671B-chat	15.00	17.42	**33.12**	**30.52**	**29.29**	12.39	34.02	20.83	15.39	8.07	21.61	29.96	35.09	26.54	25.98	7.80	36.70	27.01	32.34	41.35	22.52	39.95	22.95	28.79	31.32
DeepSeek-V3.2-Exp-671B-reasoner	8.48	17.96	32.63	30.28	29.08	12.16	33.97	19.76	**15.52**	8.00	20.78	29.75	34.28	26.05	25.57	7.64	36.63	26.65	31.95	41.15	21.92	39.89	23.03	28.90	31.14
nllb-200-3.3B	15.80	16.69	27.68	26.18	26.29	13.55	22.93	20.89	13.29	6.91	19.02	29.73	33.90	24.67	26.07	10.26	28.05	25.45	22.71	35.75	18.52	38.34	21.88	25.49	27.12
nllb-moe-54b	**17.79**	18.11	28.29	27.33	26.62	**14.63**	24.92	**22.00**	13.60	7.32	20.06	28.80	31.63	25.52	27.10	10.43	26.48	24.99	18.62	34.50	20.88	39.42	21.50	26.96	26.98
Google Translate	14.91	18.57	28.69	29.07	28.10	13.64	27.81	21.21	14.45	7.55	20.40	**37.46**	36.93	**31.98**	29.13	**12.51**	43.74	**31.96**	32.35	40.22	**25.96**	42.11	**25.73**	**32.63**	**33.17**

Table E.2: BLEU scores of translations for 22 language pairs using Prompt 2 under **zero-shot setting**.

	non-English centric											En-XX								XX-En						
model_name	ar-he	ar-zh	de-fr	de-it	fr-it	ko-fr	ko-zh	ru-fr	zh-fr	zh-ru	mean	en-cs	en-de	en-pl	en-ru	en-ta	en-zh	mean	cs-en	de-en	km-en	ru-en	ta-en	zh-en	mean
Qwen3-30B-A3B-Instruct-2507	30.32	12.85	52.50	52.44	52.21	30.94	22.11	38.98	35.88	27.98	35.62	50.24	57.58	48.62	49.27	39.27	26.28	45.21	55.08	62.63	43.52	62.27	50.71	55.62	54.97
Qwen3-30B-A3B-Thinking-2507	29.40	12.57	52.77	52.49	52.68	31.09	22.24	38.68	36.62	28.57	35.71	51.91	58.98	49.12	49.61	42.57	25.56	46.29	55.34	63.44	44.43	62.88	51.80	55.97	55.64
Qwen3-4B-Instruct-2507	23.05	11.55	49.14	47.85	50.03	29.02	20.66	37.14	33.70	26.61	32.88	41.29	53.44	40.09	45.74	19.09	24.93	37.43	52.75	61.41	34.97	59.72	44.39	54.41	51.28
Qwen3-4B-Thinking-2507	25.58	11.73	50.46	50.08	51.38	30.12	21.40	37.87	35.51	27.35	34.15	45.53	55.85	44.80	46.56	34.16	24.65	41.93	53.44	61.82	36.35	61.59	47.79	55.05	52.67
Llama-3.2-3B-Instruct	10.14	5.85	44.35	42.05	49.38	21.44	11.67	33.39	29.84	19.15	26.73	38.44	51.69	41.07	36.58	29.80	16.65	35.70	51.34	60.42	23.62	58.10	39.61	48.69	46.96
gemma-3-27b-it	36.37	13.06	54.22	54.19	53.33	31.95	23.31	40.21	36.63	28.94	37.22	56.30	60.80	52.89	51.86	47.42	26.98	49.37	56.95	65.00	47.07	63.88	53.63	56.78	57.22
Qwen2.5-32B-Instruct	0.78	13.16	45.44	47.99	50.97	24.09	22.76	27.34	26.30	7.26	26.61	41.89	53.42	20.82	26.23	18.92	26.98	31.38	56.05	64.87	37.04	63.29	43.81	46.50	51.93
DeepSeek-R1-Distill-Qwen-32B	19.20	12.56	50.68	48.70	50.87	29.95	21.96	37.69	35.34	26.98	33.39	46.27	54.51	44.85	46.44	25.62	26.13	40.63	55.10	61.37	33.77	62.77	38.61	55.66	51.21
aya-expanse-32b	37.04	12.90	55.23	55.08	53.75	32.20	23.08	40.39	36.42	28.85	37.49	56.40	60.34	52.27	51.28	39.00	27.01	47.72	56.56	63.57	23.96	62.83	50.82	55.79	52.26
Tower-Plus-72B	30.69	**13.71**	54.73	54.31	53.23	**32.71**	**25.83**	**40.77**	**37.77**	29.52	37.33	57.25	**62.83**	52.42	**54.09**	17.85	**32.26**	46.12	58.21	**66.42**	39.90	**66.16**	46.37	58.69	55.96
t5gemma-xl-xl-prefixlm-it	8.17	4.69	42.86	39.50	45.23	17.50	9.93	27.13	26.12	15.84	23.70	26.46	43.58	24.51	25.41	3.46	14.07	22.91	48.28	60.34	13.06	57.42	35.40	49.98	44.08
DeepSeek-V3.2-Exp-671B-chat	37.73	12.61	**55.57**	**55.42**	**53.97**	31.77	22.66	39.94	37.16	**29.59**	**37.64**	54.69	59.95	52.75	50.52	44.97	23.91	47.80	**58.33**	65.73	48.71	64.42	54.12	55.77	57.85
DeepSeek-V3.2-Exp-671B-reasoner	26.35	12.52	55.10	55.25	53.78	31.47	22.62	39.14	36.84	29.15	36.22	54.52	59.23	52.26	50.12	43.95	23.93	47.34	58.06	65.52	47.87	63.98	54.17	55.87	57.58
nllb-200-3.3B	38.62	11.70	51.91	52.32	52.36	31.87	20.35	39.30	35.10	27.46	36.10	53.88	58.12	51.67	50.80	47.27	20.04	46.96	48.67	59.74	42.69	63.07	49.18	52.10	52.58
nllb-moe-54b	**40.50**	12.49	52.20	52.62	52.19	32.53	21.79	40.01	35.17	28.03	36.75	51.53	54.87	52.21	51.81	46.56	19.93	46.15	43.58	57.92	44.32	63.43	48.67	52.54	51.74
Google Translate	38.32	13.38	52.07	53.90	52.69	32.31	19.19	39.99	36.58	29.17	36.76	**59.93**	61.09	**56.44**	53.13	**49.98**	30.83	**51.90**	57.98	64.31	**51.28**	65.21	**55.08**	**59.14**	**58.83**

Table E.3: chrF++ scores of translations for 22 language pairs using Prompt 2 under **zero-shot setting**.

	Non-English centric											En-XX								XX-En						
model_name	ar-he	ar-zh	de-fr	de-it	fr-it	ko-fr	ko-zh	ru-fr	zh-fr	zh-ru	mean	en-cs	en-de	en-pl	en-ru	en-ta	en-zh	mean	cs-en	de-en	km-en	ru-en	ta-en	zh-en	mean
Qwen3-30B-A3B-Instruct-2507	0.7193	0.7410	0.8207	0.8393	0.8559	0.6667	0.8900	0.7018	0.7105	0.7369	0.7682	0.8146	0.8434	0.8702	0.8749	0.8457	0.8499	0.8498	0.8109	0.8595	0.7921	0.8511	0.8563	0.8316	0.8336
Qwen3-30B-A3B-Thinking-2507	0.7180	0.7358	0.8203	0.8538	0.8629	0.6640	0.8899	0.7122	0.7154	0.7420	0.7714	0.8832	0.8635	0.8765	0.8836	0.8737	0.8700	0.8751	0.8270	0.8661	0.7902	0.8590	0.8596	0.8323	0.8390
Qwen3-4B-Instruct-2507	0.6107	0.7268	0.7475	0.5583	0.4897	0.6385	0.7495	0.6763	0.6939	0.6689	0.6560	0.4323	0.4562	0.5010	0.5648	0.5440	0.6079	0.5177	0.4308	0.4685	0.7195	0.5034	0.8301	0.6873	0.6066
Qwen3-4B-Thinking-2507	0.6654	0.7227	0.8100	0.8438	0.8575	0.6499	0.8816	0.7066	0.7117	0.7320	0.7581	0.8303	0.8490	0.8446	0.8673	0.7806	0.8761	0.8413	0.8147	0.8589	0.7296	0.8530	0.8335	0.8293	0.8198
Llama-3.2-3B-Instruct	0.4080	0.4220	0.5018	0.5355	0.5567	0.3650	0.5050	0.3653	0.3503	0.4087	0.4418	0.5952	0.5816	0.5785	0.5640	0.6469	0.5965	0.5938	0.5475	0.5232	0.5013	0.4640	0.5285	0.5779	0.5237
gemma-3-27b-it	0.7788	0.7383	**0.8266**	0.8602	**0.8665**	0.6717	0.8892	0.7197	0.7179	0.7422	0.7811	0.9028	0.8686	0.9010	0.8924	**0.9028**	0.8792	0.8911	0.8341	0.8666	0.8120	0.8598	0.8646	0.8328	0.8450
Qwen2.5-32B-Instruct	0.6085	0.7327	0.8114	0.8385	0.8533	0.6230	0.8895	0.6826	0.6808	0.6847	0.7405	0.8309	0.8395	0.8261	0.8514	0.6637	0.8747	0.8144	0.8252	0.8626	0.7299	0.8550	0.8021	0.8154	0.8150
DeepSeek-R1-Distill-Qwen-32B	0.6656	0.7315	0.8072	0.8342	0.8524	0.6540	0.8873	0.7041	0.7007	0.7123	0.7549	0.8079	0.8376	0.8270	0.8611	0.5818	0.8698	0.7975	0.8228	0.8604	0.6895	0.8552	0.7499	0.8279	0.8009
aya-expanse-32b	0.7801	0.7155	0.8262	0.8340	0.8639	0.6739	0.8837	0.6992	0.7181	0.7416	0.7736	0.8880	0.8648	0.8962	0.7644	0.6101	0.8541	0.8129	0.7407	0.7520	0.5812	0.8565	0.8551	0.8298	0.7692
Tower-Plus-72B	0.7311	0.7453	0.8254	0.8506	0.8657	0.6780	**0.8933**	0.7226	**0.7251**	0.7511	0.7788	0.8491	0.8705	0.8998	0.8978	0.5889	0.8828	0.8315	0.8198	0.8649	0.7459	0.8622	0.8112	0.8352	0.8232
t5gemma-xl-xl-prefixlm-it	0.5019	0.5606	0.6953	0.7313	0.7436	0.5454	0.7126	0.5904	0.5844	0.5112	0.6177	0.5374	0.6581	0.6191	0.5594	0.4614	0.6247	0.5767	0.7033	0.7942	0.4178	0.7759	0.7209	0.7420	0.6924
nllb-200-3.3B	0.7882	0.7248	0.8162	0.8513	0.8618	0.6696	0.8242	0.7116	0.6960	0.7284	0.7672	0.8817	0.8485	0.8766	0.8772	0.6768	0.7846	0.8579	0.7862	0.8446	0.7675	0.8518	0.8487	0.8001	0.8168
nllb-moe-54b	**0.7989**	0.7323	0.8179	0.8531	0.8618	0.6753	0.8330	0.7152	0.6940	0.7340	0.7716	0.8825	0.8430	0.8875	0.8828	0.8770	0.7817	0.8591	0.7691	0.8367	0.7773	0.8534	0.8476	0.8044	0.8148
Google Translate	0.7946	**0.7530**	0.8246	**0.8613**	0.8655	**0.6788**	0.8824	**0.7241**	0.7250	**0.7517**	**0.7861**	**0.9118**	**0.8778**	**0.9088**	**0.8985**	0.9019	**0.8905**	**0.8982**	**0.8347**	**0.8711**	**0.8240**	**0.8693**	**0.8687**	**0.8393**	**0.8512**

Table E.4: COMET scores of translations for 22 language pairs using Prompt 2 under **few-shot setting**. The three baseline models do not have few-shot settings and they are here for comparison purpose.

	non-English centric											En-XX								XX-En						
model_name	ar-he	ar-zh	de-fr	de-it	fr-it	ko-fr	ko-zh	ru-fr	zh-fr	zh-ru	mean	en-cs	en-de	en-pl	en-ru	en-ta	en-zh	mean	cs-en	de-en	km-en	ru-en	ta-en	zh-en	mean
Qwen3-30B-A3B-Instruct-2507	4.79	17.55	28.68	18.02	23.14	11.40	33.50	19.40	13.78	6.90	17.71	11.34	22.29	18.46	18.75	2.04	12.54	14.24	27.28	35.09	16.50	32.63	20.92	28.31	26.79
Qwen3-30B-A3B-Thinking-2507	8.89	18.11	29.22	26.27	27.45	11.72	33.53	19.36	14.89	7.48	19.69	26.33	33.63	19.66	25.18	8.25	33.39	24.41	29.68	39.33	17.76	39.12	22.74	29.65	29.71
Qwen3-4B-Instruct-2507	1.39	15.53	20.60	0.45	2.07	8.43	21.69	16.25	6.59	3.48	9.65	0.11	0.58	2.66	1.46	0.34	2.88	1.34	0.45	1.49	3.80	10.15	5.02	19.62	6.76
Qwen3-4B-Thinking-2507	6.45	16.89	25.92	22.67	25.34	10.77	31.56	18.20	14.19	6.63	17.86	19.52	29.27	18.18	21.18	4.28	37.66	21.68	27.03	36.67	11.23	36.28	18.25	27.71	26.20
Llama-3.2-3B-Instruct	0.01	0.33	2.31	0.64	2.87	0.13	0.72	0.04	0.18	0.08	0.73	1.09	1.95	0.95	0.95	0.28	1.77	1.17	3.38	1.32	0.42	4.60	2.13	4.64	2.75
gemma-3-27b-it	13.15	18.68	31.44	**29.20**	**28.60**	12.54	34.76	21.43	14.75	7.30	**21.19**	32.70	36.65	28.20	27.53	10.89	41.34	29.55	30.99	**42.03**	21.41	40.12	25.15	30.79	31.75
Qwen2.5-32B-Instruct	0.04	17.94	21.40	15.69	25.47	4.80	34.46	4.74	9.27	1.12	13.49	15.62	29.37	2.99	8.45	1.11	41.02	16.43	29.87	40.22	11.75	16.37	16.47	26.50	23.53
DeepSeek-R1-Distill-Qwen-32B	3.09	17.71	26.84	21.58	24.98	10.35	33.77	17.85	13.23	6.04	17.54	20.11	27.67	18.18	21.31	1.81	35.39	20.75	29.08	38.01	8.44	36.84	11.45	28.15	25.33
aya-expanse-32b	8.70	12.85	**32.56**	20.29	28.33	11.67	31.07	16.67	13.85	6.87	18.28	26.24	32.29	25.42	13.72	1.86	23.30	20.47	14.41	9.33	2.93	37.57	21.31	28.50	19.01
Tower-Plus-72B	1.83	**20.09**	31.66	28.84	28.47	14.00	**38.30**	**22.06**	**16.50**	**8.41**	21.02	27.68	**38.95**	26.59	**29.26**	0.02	**43.80**	27.72	27.12	41.89	14.68	**43.18**	19.39	**34.30**	30.09
t5gemma-xl-xl-prefixlm-it	0.50	3.71	14.78	10.96	14.66	4.20	11.73	6.14	4.52	1.62	7.28	3.65	13.72	5.46	7.86	0.16	9.40	6.71	12.33	28.22	0.23	21.91	8.01	15.43	14.35
nllb-200-3.3B	15.80	16.69	27.68	26.18	26.29	13.55	22.93	20.89	13.29	6.91	19.02	29.73	33.90	24.67	26.07	10.26	28.05	25.45	22.71	35.75	18.52	38.34	21.88	25.49	27.12
nllb-moe-54b	**17.79**	18.11	28.29	27.33	26.62	**14.63**	24.92	22.00	13.60	7.32	20.06	28.80	31.63	25.52	27.10	10.43	26.48	24.99	18.62	34.50	20.88	39.42	21.50	26.96	26.98
Google Translate	14.91	18.57	28.69	29.07	28.10	13.64	27.81	21.21	14.45	7.55	20.40	**37.46**	36.93	**31.98**	29.13	**12.51**	43.74	**31.96**	**32.35**	40.22	**25.96**	42.11	**25.73**	32.63	**33.17**

Table E.5: BLEU scores of translations for 22 language pairs using Prompt 2 under **few-shot setting**. The three baseline models do not have few-shot settings and they are here for comparison purpose.

	non-English centric											En-XX								XX-En						
model_name	ar-he	ar-zh	de-fr	de-it	fr-it	ko-fr	ko-zh	ru-fr	zh-fr	zh-ru	mean	en-cs	en-de	en-pl	en-ru	en-ta	en-zh	mean	cs-en	de-en	km-en	ru-en	ta-en	zh-en	mean
Qwen3-30B-A3B-Instruct-2507	25.19	12.65	52.26	48.32	50.91	30.82	22.35	38.03	35.40	27.55	34.35	37.77	52.88	45.08	46.73	29.25	18.06	38.29	52.97	61.59	42.61	60.36	51.28	55.65	54.08
Qwen3-30B-A3B-Thinking-2507	29.81	12.52	52.82	52.56	52.82	30.91	23.01	38.81	36.35	28.46	35.81	52.14	58.59	48.00	49.81	43.44	24.99	46.16	55.77	63.52	44.51	62.89	52.63	56.65	55.99
Qwen3-4B-Instruct-2507	13.46	11.39	43.76	18.05	18.64	27.77	15.42	34.99	30.22	21.32	23.50	6.37	14.50	18.94	17.90	16.07	5.83	13.27	12.65	22.27	28.83	23.96	39.78	43.89	28.56
Qwen3-4B-Thinking-2507	25.86	11.82	50.46	49.94	51.46	29.87	21.32	37.78	35.79	27.11	34.14	45.82	55.37	44.82	46.78	35.07	24.92	42.13	53.64	61.67	35.93	60.74	47.76	55.28	52.50
Llama-3.2-3B-Instruct	0.85	1.36	22.64	19.44	25.71	4.05	2.39	0.87	5.34	3.03	8.57	13.55	18.13	13.46	14.63	11.19	5.08	12.67	26.73	18.26	11.05	29.46	22.61	31.13	23.21
gemma-3-27b-it	36.46	13.08	54.26	**54.56**	**53.48**	31.97	23.14	40.29	36.71	28.96	**37.29**	56.62	61.04	53.48	52.08	47.94	29.54	50.12	57.19	65.62	47.45	64.28	54.81	57.72	57.85
Qwen2.5-32B-Instruct	0.73	12.70	48.78	46.24	51.13	23.44	23.01	23.93	32.29	14.46	27.67	38.77	55.37	22.35	36.53	24.16	27.21	34.07	56.30	64.80	36.77	50.70	44.55	55.33	51.41
DeepSeek-R1-Distill-Qwen-32B	21.60	12.92	50.83	48.66	50.99	30.00	22.49	37.66	35.00	26.64	33.68	46.59	54.46	45.30	46.87	25.94	25.53	40.78	55.40	62.59	33.68	62.65	39.66	55.44	51.57
aya-expanse-32b	35.22	11.55	**54.95**	50.68	53.36	31.71	22.33	39.02	36.22	28.63	36.37	52.94	58.67	52.07	35.32	21.68	24.34	40.84	43.86	40.95	22.84	62.37	51.06	55.83	46.15
Tower-Plus-72B	17.78	**13.77**	54.37	53.56	53.26	**32.50**	**25.84**	**40.61**	**37.64**	**29.65**	35.90	49.86	**61.96**	52.20	**53.10**	2.44	**31.89**	41.91	55.24	**65.96**	38.88	**66.16**	45.93	**59.91**	55.35
t5gemma-xl-xl-prefixlm-it	4.88	3.20	36.81	32.69	37.81	17.73	8.25	19.77	18.68	8.73	18.85	20.53	34.53	21.55	17.79	2.56	6.79	17.29	31.66	51.48	2.49	37.37	31.13	36.04	31.69
nllb-200-3.3B	38.62	11.70	**51.91**	**52.32**	**52.36**	31.87	20.35	39.30	35.10	**27.46**	**36.10**	53.88	58.12	51.67	50.80	47.27	20.04	46.96	**48.67**	59.74	42.69	63.07	49.18	52.10	52.58
nllb-moe-54b	**40.50**	12.49	52.20	52.62	52.19	**32.53**	21.79	40.01	35.17	28.03	36.75	51.53	54.87	52.21	51.81	46.56	19.93	46.15	43.58	57.92	44.32	63.43	48.67	52.54	51.74
Google Translate	38.32	13.38	52.07	53.90	52.69	32.31	19.19	39.99	36.58	29.17	36.76	**59.93**	61.09	**56.44**	**53.13**	**49.98**	30.83	**51.90**	**57.98**	64.31	**51.28**	65.21	**55.08**	59.14	**58.83**

Table E.6: chrF++ scores of translations for 22 language pairs using Prompt 2 under **few-shot setting**. The three baseline models do not have few-shot settings and they are here for comparison purpose.

Model	TAR	GENETIC	GEOGRAPHIC	SYNTACTIC	PHONOLOGICAL	INVENTORY	FEATURAL	MEAN
Qwen3-30B-A3B-Instruct-2507	**0.6330**	-0.2691	-0.1451	**-0.5860**	0.0095	-0.0311	**-0.4997**	-0.2780
Qwen3-30B-A3B-Thinking-2507	**0.6406**	-0.2767	-0.1485	**-0.5781**	0.0194	-0.0550	**-0.5101**	-0.2839
Qwen3-4B-Instruct-2507	**0.6398**	-0.2202	-0.0140	**-0.6034**	0.0374	-0.0408	**-0.4573**	-0.1883
Qwen3-4B-Thinking-2507	**0.6405**	-0.2452	-0.0671	**-0.5962**	0.0202	-0.0530	**-0.4781**	-0.2308
Llama-3.2-3B-Instruct	**0.7496**	-0.3750	-0.2964	**-0.6180**	0.0052	-0.0136	**-0.6268**	-0.4013
gemma-3-27b-it	**0.6505**	-0.3172	-0.2405	**-0.5495**	-0.0040	-0.0310	**-0.5216**	-0.3419
Qwen2.5-32B-Instruct	**0.5121**	-0.2194	0.0102	-0.3912	0.0967	-0.1340	-0.2798	-0.1301
DeepSeek-R1-Distill-Qwen-32B	**0.6780**	-0.1356	0.0072	**-0.5655**	0.0626	-0.0381	**-0.4296**	-0.1417
aya-expanse-32b	**0.7273**	-0.3371	-0.2486	**-0.5934**	-0.0297	0.0020	**-0.5259**	-0.3571
Tower-Plus-72B	**0.6571**	-0.2941	-0.1528	**-0.5965**	-0.0161	-0.0268	**-0.4945**	-0.2943
t5gemma-xl-xl-prefixlm-it	**0.6607**	-0.3454	-0.1811	**-0.6178**	0.0132	-0.0373	**-0.5507**	-0.3259
DeepSeek-V3.2-Exp-671B-chat	**0.4763**	-0.3527	-0.2998	**-0.5874**	-0.0212	-0.0219	**-0.5598**	-0.3937
DeepSeek-V3.2-Exp-671B-reasoner	**0.5390**	-0.2675	-0.2388	**-0.5624**	0.0479	-0.0670	**-0.5624**	-0.3292
nllb-200-3.3B	**0.7019**	**-0.4490**	**-0.4479**	**-0.6232**	-0.1641	-0.0904	**-0.6212**	**-0.5564**
nllb-moe-54b	**0.6178**	-0.4115	**-0.4240**	**-0.6382**	-0.2169	-0.0871	**-0.5966**	**-0.5461**

Table F.1: Pearson's r correlation between BLEU scores and TAR, genetic, geographic, syntactic, phonological, inventory, featural and the mean of the latter six typological distances. **Bold values** are statistically significant.

Model	TAR	GENETIC	GEOGRAPHIC	SYNTACTIC	PHONOLOGICAL	INVENTORY	FEATURAL	MEAN
Qwen3-30B-A3B-Instruct-2507	0.3074	-0.3853	**-0.6543**	**-0.4471**	0.0770	0.1005	**-0.7294**	**-0.5461**
Qwen3-30B-A3B-Thinking-2507	0.3033	-0.3740	**-0.6473**	**-0.4237**	0.0975	0.0965	**-0.7249**	**-0.5319**
Qwen3-4B-Instruct-2507	0.3934	-0.3851	**-0.5475**	**-0.5740**	0.0595	0.0799	**-0.7388**	**-0.5141**
Qwen3-4B-Thinking-2507	0.3533	-0.3787	**-0.6066**	**-0.4995**	0.0891	0.0909	**-0.7488**	**-0.5267**
Llama-3.2-3B-Instruct	**0.7884**	-0.3469	**-0.5676**	**-0.4961**	0.0972	0.0102	**-0.7484**	**-0.5110**
gemma-3-27b-it	**0.5798**	-0.4111	**-0.6963**	-0.3872	0.0595	0.1233	**-0.6990**	**-0.5638**
Qwen2.5-32B-Instruct	**0.4320**	-0.2736	-0.4198	**-0.4284**	0.1984	-0.1251	**-0.6222**	-0.3906
DeepSeek-R1-Distill-Qwen-32B	**0.4480**	-0.3407	**-0.5446**	**-0.5632**	0.0882	0.0663	**-0.7549**	**-0.4979**
aya-expanse-32b	**0.5986**	**-0.4607**	**-0.6943**	**-0.4771**	-0.0015	0.1246	**-0.7142**	**-0.6025**
Tower-Plus-72B	**0.4246**	**-0.4255**	**-0.5883**	**-0.5748**	0.0153	0.1180	**-0.7278**	**-0.5482**
t5gemma-xl-xl-prefixlm-it	**0.7047**	-0.3670	**-0.4366**	**-0.5764**	0.0503	-0.0046	**-0.6704**	**-0.4608**
DeepSeek-V3.2-Exp-671B-chat	0.0937	-0.4214	**-0.7108**	-0.3917	0.0426	0.1200	**-0.6946**	**-0.5788**
DeepSeek-V3.2-Exp-671B-reasoner	0.1837	-0.3331	**-0.6474**	-0.3893	0.1240	0.0697	**-0.7204**	**-0.5156**
nllb-200-3.3B	**0.6381**	**-0.4340**	**-0.7431**	-0.3980	-0.0117	0.1280	**-0.6960**	**-0.6121**
nllb-moe-54b	**0.5872**	-0.4136	**-0.7356**	-0.4119	-0.0403	0.1480	**-0.6929**	**-0.6080**

Table F.2: Pearson's r correlation between chrF++ scores and TAR, genetic, geographic, syntactic, phonological, inventory, featural and the mean of the latter six typological distances. **Bold values** are statistically significant.