Incorporating Abstract Meaning Representation and Uniform Meaning Representation into In-Context Learning Strategies for Machine Translation

Thu Hoang **Ellie Yang** **Weixin Lin**
Amherst College
{thuhoang28,eyang27,wlin28}@amherst.edu

Shira Wein
University of South Florida
shirawein@usf.edu

Abstract

While in-context machine translation approaches lead to notable progress for low-resource languages, it is not yet clear what kinds of linguistic information in prompts can consistently and reliably lead to performance gains. We evaluate whether English semantic representations, specifically Abstract Meaning Representation (AMR) and its extension Uniform Meaning Representation (UMR), can improve LLM-based MT via prompting. Across 12 languages, 7 models from 5 families, and multiple prompting strategies, we find that conclusions about their effectiveness are highly sensitive to the setting. We further observe that prompts using UMR tend to outperform those using AMR, particularly in few-shot scenarios and for certain models. These results highlight the need for comprehensive evaluation for in-context learning work.

1 Introduction

Instruction-tuned multilingual large language models are emerging as competitive MT systems (Zhu et al., 2024; Kocmi et al., 2024; Ataman et al., 2025). Their ability to follow natural-language instructions and demonstrations offers a promising, training-free approach to reducing the need for domain-specific large parallel corpora. As a result, this has led to a growing body of work on prompt-based and retrieval-based methods that leverage linguistic resources (such as parallel examples, dictionary entries, and grammar books) to improve LLM-based MT (Agrawal et al., 2023; Tanzer et al., 2024; Zhang et al., 2024; Court and Elsner, 2024; Pei et al., 2025). In comparison, semantic representations (SRs), which abstract away from the surface form and reflect the core elements of meaning, have received little attention as a linguistic resource in prompting, despite their potential to improve translation performance (Wein, 2025; Zhang et al., 2025a).

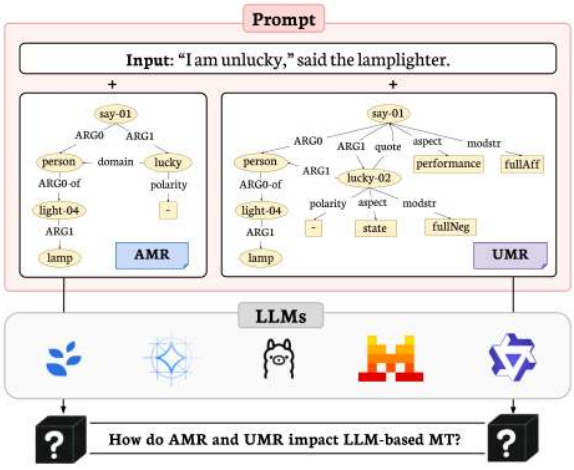


Figure 1: In addition to the input, AMR or UMR can be included in prompts as an in-context linguistic resource. For the same sentence, UMR is more complex due to its additional attribute roles.

Among SRs, Abstract Meaning Representation (AMR; Banarescu et al., 2013) is particularly well-resourced, with regard to parsing and generation technologies, and has been applied to a range of downstream tasks, including MT (Sadeddine et al., 2024; Wein and Opitz, 2024). Other tasks for which AMR has been shown to be useful include question answering (e.g., Wang et al., 2023; Pham et al., 2024) and NLG evaluation (e.g., Opitz and Frank, 2021; Ribeiro et al., 2022). However, results integrating AMR into MT are mixed. While some report gains, including for low-resource language pairs (Song et al., 2019; Nguyen et al., 2021), others find no such gain (Jin et al., 2024). Even so, because AMR was originally designed for English (Banarescu et al., 2013), its applicability to other languages is limited. Addressing this limitation, Uniform Meaning Representation (UMR; Van Gysel et al., 2021) is a flexible multilingual extension of AMR designed to be annotated in many languages (Wein and Bonn, 2023). Although UMR represents a sentence as a rooted,

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 2090–2124
November 6-10, 2026 ©2026 Association for Computational Linguistics

directed graph similar to AMR, it additionally annotates modality, aspect, quantification, and scope via paradigmatic lattices. Though available UMR resources are currently limited (Chun and Xue, 2024; Markle et al., 2025, 2026), the adaptable and comprehensive multilingual annotation schema of UMR, combined with recent work which finds that inserting UMR in the prompt helps GPT-4 translate three Indigenous languages to English (Wein, 2025), makes UMR a promising alternative to AMR for multilingual MT.

In this work, we investigate the following question (see Figure 1): **How does the inclusion of AMR and UMR affect prompt-based MT across languages, models, and evaluation metrics?** We perform a large-scale case study of AMR and UMR in prompt-based MT, evaluating English-to-X translation across 12 target languages, 7 models from 5 model families, a wide range of prompting protocols, and 3 evaluation metrics. We additionally provide a qualitative analysis of the model outputs, revealing some emergent noise and failure modes that can disrupt the model's performance on MT. Our results demonstrate that the inclusion of AMR and UMR vary substantially depending on the choice of language, model, and evaluation metric, highlighting the need for comprehensive evaluation when assessing SRs as prompting strategies in LLM-based MT. Our findings also motivate continued investment in UMR parsing infrastructure and downstream applications, given the relative performance.[1]

2 Background

In this section, we review relevant literature on machine translation via in-context learning (Section 2.1), as well as applications of AMR and UMR (Section 2.2).

2.1 In-Context Learning for MT

Large language models exhibit competence in machine translation across multiple languages, and in-context learning (ICL)—providing translation examples or other relevant information directly in the prompt—can improve LLM translation quality (Hendy et al., 2023). Zhu et al. (2024) show that LLMs can generate fair translations of languages scarcely represented in pre-training, with LLaMA (Grattafiori et al., 2024), Gemma (Team, 2025), Mistral (Jiang et al., 2023), Aya (Aryabumi et al., 2024), and Qwen (Team, 2024) specifically demonstrating competitive translation performance through prompting alone (Cui et al., 2025). Nevertheless, MT via in-context learning remains a challenge for low-resource languages due to the English-centric training bias in the aforementioned models (Zhu et al., 2024; Hendy et al., 2023). Novel approaches to prompt engineering for low-resource language MT incorporate: demonstrations retrieved based on syntactic coverage (Frontull and Ströhle, 2025), dictionary entries and parallel texts (Merx et al., 2024; Pei et al., 2025), and grammatical textbook content (Guo et al., 2024).

However, unlike encoder-decoder NMT systems, LLMs' in-context learning allows structured linguistic information to be supplied directly at inference time, without any need to modify the model architecture or provide additional training. Therefore, one question that arises is whether linguistic representations, provided directly at inference time, could help bridge the gap for low-resource languages.

2.2 AMR, UMR, and Their Applications

Abstract Meaning Representation (Banarescu et al., 2013) is a semantic representation that takes the form of a rooted, directed graph. It is designed to capture "who does what to whom," abstracting away from syntactic features. It has been incorporated into a range of downstream NLP tasks (Wein and Opitz, 2024), such as question answering (Kapanipathi et al., 2021; Wang et al., 2023), summarization (Liu et al., 2015), and information extraction (Zhang and Ji, 2021; Zhang et al., 2021).

Results of prior work integrating AMR into machine translation tasks are mixed. Jin et al. (2024) show that AMR hurts machine translation performance even when using the same dataset as Song et al. (2019), where AMR successfully improves model's performance. Moreover, it has yet to be determined whether the observed improvements in NMT (Song et al., 2019; Nguyen et al., 2021) can be generalized for in-context learning, on various domains and genres, or for low-resource languages.

Beyond these gaps in evaluation, there are also theoretical reasons to question the appropriateness of AMR specifically for multilingual machine translation. Wein and Schneider (2024) show that

[1]Our code is at https://github.com/ACNLPlab/umr-amr-mt-prompting.

source language quantifiably impacts AMR graph structure, demonstrating AMR's susceptibility to language-specific lexical differences, and suggests that an inherently multilingual semantic representation may be better suited for the cross-lingual task of translation.

UMR, on the other hand, is a multilingual extension of AMR, retaining AMR's graph-based predicate-argument structure while offering greater flexibility to represent the world's languages, including low-resource languages (Van Gysel et al., 2021).

Because annotated data, UMR parsing tools (Markle et al., 2025; Chun and Xue, 2024), and UMR-to-text generation models (Markle et al., 2026) are still nascent and have only recently become available, investigations into UMR's application to downstream NLP tasks remain limited. Most directly relevant to our work, Wein (2025) incorporates UMR into GPT-4 prompts for translating three low-resource languages (Navajo, Arápaho, and Kukama) into English, finding that this inclusion of UMR results in a statistically significant improvement. This finding provides an early indication that UMR can serve as an effective complementary representation for at least some models and languages, including low-resource ones.

3 Methodology

We present a large-scale experimental study to systematically evaluate and compare AMR and UMR in LLM-based MT.

3.1 Data

To support the comparison of AMR and UMR in prompts, we use annotated data of the *The Little Prince*.[2] We use this dataset, with English serving as the source language, because it is the only language where both AMR and UMR annotations are available for the entire corpus.

Due to the lack of multilingual UMR parsers and so that we may isolate the effect of AMR and UMR, we opt for gold annotations as opposed to large-scale automatically parsed data, reducing potential noise. One limitation this causes is that we are then only able to integrate the English graph and not the graph of the target language in demonstrations, which are found to be helpful in prior work (Wein, 2025). This experimental setup does, however, effectively reflect a potential real-world application, where only accurate English automatic parsing is available (Markle et al., 2026).

We evaluate 12 target languages in total, selected to ensure typological diversity and a breadth of resource levels, as shown in Table 1. Mandarin (Li et al., 2016) and Portuguese (Anchiêta and Pardo, 2022) come from existing AMR corpora, while the remaining languages are sourced from InterCorp,[3] a large parallel corpus containing *The Little Prince* in 26 languages (Cermák and Rosen, 2012). After aligning AMR and UMR annotations and removing non-matching cases, we retain only strictly one-to-one sentence pairs for fair comparison (see Table 1 for the number of sentences included in the test set of each language).

Language	Script	CC Size (%)	Dataset Size
High Resource			
German	Latin	5.9033	942
Japanese	Non-Latin	5.1572	899
Mandarin	Non-Latin	4.8564	1354
Portuguese	Latin	2.5116	1225
Dutch	Latin	1.8420	845
Czech	Latin	1.1493	973
Medium Resource			
Ukrainian	Non-Latin	0.8012	958
Hindi	Non-Latin	0.2259	622
Catalan	Latin	0.2010	960
Latvian	Latin	0.1172	942
Low Resource			
Belarusian	Non-Latin	0.0165	903
Upper Sorbian	Latin	N/A	934

Table 1: Target languages with their dataset sizes, ordered by how resourced they are. Following Lai et al. (2023), we categorize languages into high-resource ($\geq$1%), medium-resource ($\geq$0.1%), and low-resource (<0.1%) based on their proportion in CommonCrawl using the most recent snapshot, April 2026 via https://commoncrawl.github.io/cc-crawl-statistics/plots/languages. Given that we perform ICL experiments, the entire dataset for each of these languages serves as a test set (i.e., no training or validation is performed).

3.2 Prompts, Research Questions

We design prompting protocols organized around three research questions. Full prompts are provided in Appendix A.

[2]The AMR data was previously hosted at amr.isi.edu, which is no longer maintained. However, it can be retrieved by entering this URL into the Wayback Machine and selecting any snapshot captured in or before 2024. The corresponding UMR annotations are found in the UMR 2.0 Dataset.

[3]https://intercorp.korpus.cz

RQ1: How does incorporating a SR affect ICL MT performance? We establish two conditions without semantic representations. First, in the baseline, the model receives only the source sentence and a translation instruction. Second, our SENSE approach (An et al., 2025) does not provide SRs in the prompt but instead uses natural-language instructions in the form of *Please translate this English sentence into [target language] by utilizing its semantic parsing result which helps to understand grammar and semantics*.

We then evaluate how directly providing SRs to the model impacts performance. Specifically, we establish the SR-Direct approach, where the prompt includes the linearized AMR or UMR graph of the source sentence without additional instructions on how to use it. Both the base prompt and the linearization of SRs follow Wein (2025).

Next, we develop further approaches which include SRs at various shot levels.

RQ2: Does the usefulness of SRs increase with the number of demonstrations? More demonstrations including SRs may help the model leverage the semantic information. We therefore evaluate few-shot settings with $k \in \{1, 3, 5\}$ examples, each containing: the source sentence, its linearized AMR or UMR graph, and the target translation. Because gains may stem from the additional examples rather than the SRs themselves, we repeat the same few-shot settings for the baseline, where each example contains only the source sentence and target translation. Demonstrations are selected using chrF similarity between test sentences and candidate examples, following Wein (2025).

RQ3: Does instruction affect SR utility? Finally, we test whether different ways of instructing models to use SRs affect translation quality.

Extending SR-Direct, we develop two variants instructing the model how to make use of the SR:

- SR-Grounded: The model is instructed to rely on the information in SRs (e.g., predicate–argument structure) during translation.
- SR-Assisted: The model is instructed to prioritize the source sentence and use SRs only when the source sentence is unclear.

Additionally, we develop prompts which convert the SRs to natural language descriptions (NLDs). Following Zhang et al. (2025a), who find that converting SRs into NLDs improves translation quality over the baseline and SENSE, we provide a NLD of the linearized AMR or UMR graph in the prompt, instead of the graph itself. Appendix B details how SRs are converted into NLDs.

3.3 Models

We use multiple multilingual instruction-tuned LLMs: **Aya-23-8B** (Aryabumi et al., 2024); **Gemma-3-IT** variants (4B, 12B, and 27B) to examine potential effects of model size (Team, 2025); **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024); and **Qwen2.5-7B-Instruct** (Team, 2024). We also include **Mistral-7B-Instruct** (Jiang et al., 2023), which is primarily trained on English, effectively simulating a setting in which the model has limited exposure to the target languages.[4]

3.4 Evaluation

We evaluate translation quality using **BLEU** via SacreBLEU (Post, 2018), **COMET-22** (Rei et al., 2022), and **MetricX-24-Hybrid-XL** (Juraska et al., 2024). We use multiple metrics to enable a comprehensive evaluation of when SRs are beneficial, as the metrics are complementary and capture different aspects of translation quality, and to assess the sensitivity of the metrics in this setting. BLEU measures n-gram overlap with a reference, while COMET is an embedding-based metric that exploits both the source and reference. MetricX-24 is also embedding-based but trained on mixed input formats (source–hypothesis, hypothesis–reference, and source–hypothesis–reference), allowing it to determine how much to rely on the source and reference when evaluating the hypothesis, or even to disregard the reference when it is of low quality. This is particularly useful for our evaluation on *The Little Prince*, where the reference is not always strictly literal.

4 Results

Here we present our results in response to each of our three research questions.

4.1 Impact of Direct AMR/UMR Prompting

Overall, incorporating linearized AMR or UMR graphs directly into the prompt as additional information does not yield consistent improvements

[4]We do not include frontier models such as OpenAI's models because the InterCorp license agreement restricts sending the data to commercial services.

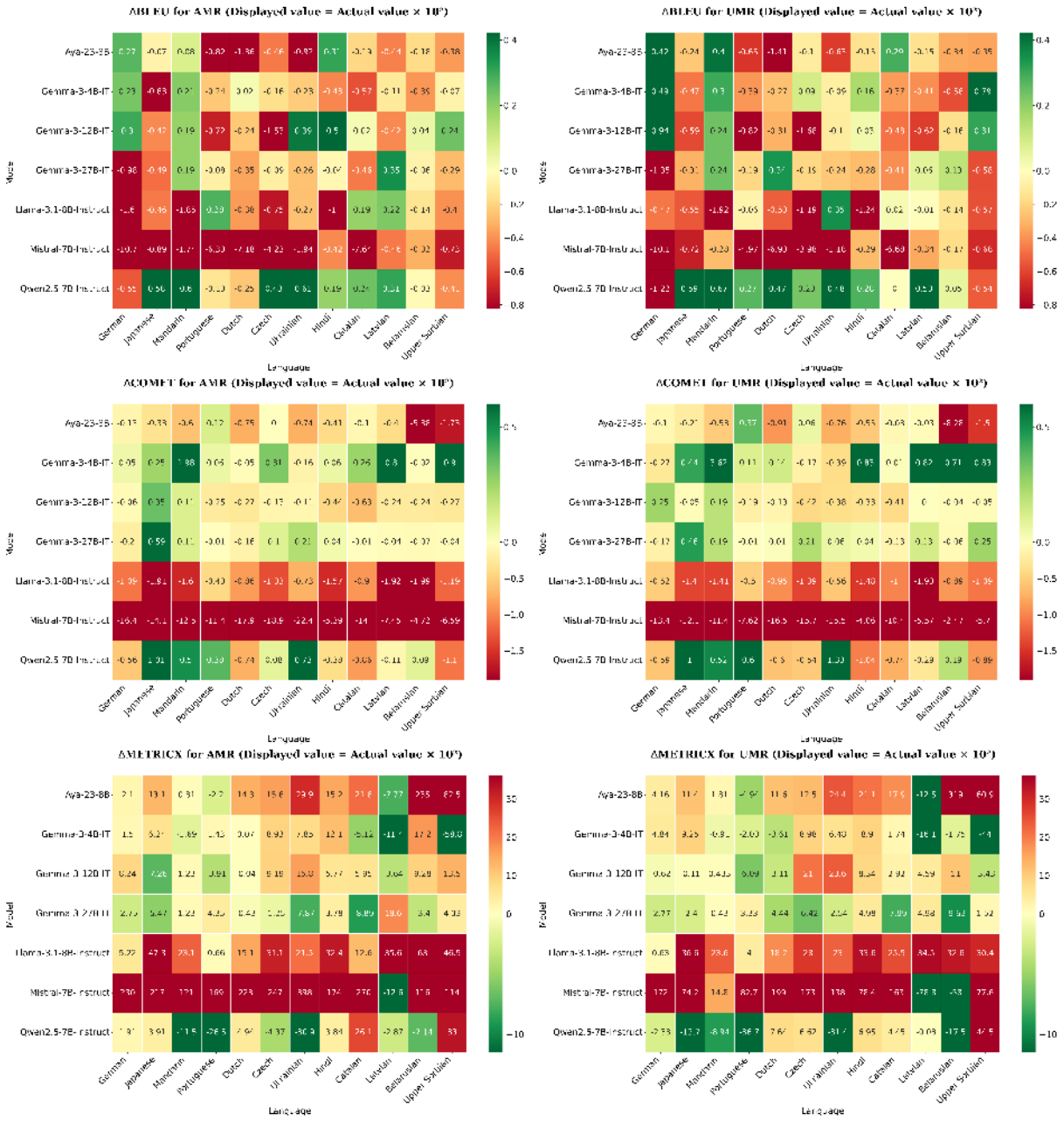


Figure 2: Heatmaps showing the impact of `SR-Direct` prompting (zero-shot), with AMR (left) and UMR (right), relative to the baseline, reported as ΔBLEU, ΔCOMET, and ΔMetricX (top to bottom). For each metric, the color scale is clipped at the 25th and 75th percentiles. Cell values are also scaled by 10^2 for readability. Note that for MetricX, lower values indicate better performance; therefore, decreases are positive, and its color scale is reversed relative to the other two metrics. Statistically significant results are reported in Appendix C.

over the baseline translation setup. As shown in Figure 2, the effects are mixed across languages and models, and are on the whole slightly negative-to-neutral across the metrics. We observe several more fine-grained patterns below:

Language-wise Across languages, there is no clear trend with respect to resource level or typology; effects are largely language-specific. For instance, five of seven models show improvements in BLEU when translating into Mandarin (high-resource, non-Latin) when either AMR or UMR is included, whereas for Japanese (also high-resource, non-Latin), only one model shows improvement in BLEU. In contrast, for both of our low-resource languages, AMR and UMR lead to BLEU decreases for most models.

Model-wise At the model level, **Qwen2.5-7B-Instruct** shows the most consistent BLEU im-

provements when AMR or UMR is included. **Gemma-3-4B-IT** shows moderate gains, particularly via COMET, but these are less consistent via BLEU and MetricX. Notably, this pattern does not hold for larger Gemma-3-IT variants, suggesting that gains observed when using the smaller model do not always translate to use of larger versions. In contrast, **Mistral-7B-Instruct** is particularly sensitive to the addition of AMR or UMR. This likely reflects its limited support for non-English languages.

Metric-wise We further observe that changes in one metric do not always align with changes in others. For example, several cases of BLEU improvements for **Qwen2.5-7B-Instruct** correspond to negative changes in COMET and MetricX.

Taken together, these results suggest that conclusions drawn from a limited set of languages, models, and evaluation metrics may not generalize to other settings.

4.2 Multi-shot Settings Increase SR Usefulness

We find that, although introducing SRs into our prompts regularly degrades MT performance, introducing examples through multi-shot prompting consistently reduces this degradation. This suggests that models gain more familiarity with the SR's structure and interpretation when provided multiple exemplars.

MetricX results across models, prompting strategies, and languages can be found in Table 2. These results highlight that, across prompting strategies and models, 5-shot prompts outperform 0-shot, often by over 1 point. In particular, we find patterns in the utility of multi-shot prompting in graph-based versus NLD-based prompts, and across language resource levels.

SRs as Graphs versus NLDs First, we find that multi-shot settings tend to improve performance more notably for graph-based prompts compared to those with the SRs converted to natural language descriptions. This is particularly apparent in Mistral's performance. In Table 2, under 0-shot Mistral is always the worst-performing model, oftentimes by a large margin. Mistral's MetricX score on graph-based prompts is also always higher, indicating weaker performance. However, under the 5-shot setting, Mistral is able to close its performance gap, with both graph and NLD approaches leveling to similar performances, showing that multi-shot prompting helps especially with prompts that included linearized graphs. Of all the models included in our experiments, Mistral is the only one whose only explicitly supported language is English. This show that the semantic information embedded in SRs introduce more unseen information. While introducing examples still did not result in SR-based prompts to outperform Mistral's baseline, the models' quick improvement on utilizing semantic information, especially graph-based information, shows the potential of SR graphs as new and compact data that can quickly enhance monolingual model's abilities to interpret multilingual semantic information.

Language-Resourcedness Second, we analyze the effects of multi-shot prompting across language resource levels. Table 3 shows the average MetricX score across differently resourced languages, in 0-shot and 5-shot settings. Compared to 0-shot performance, the improvement from moving to 5-shots is substantially higher in low-resource compared to high-resource languages, but does not result in overall better performance than that of the high-resource languages.

Overall, our experiments show that multi-shot settings increase LLM performance when integrating SRs for MT prompting. In particular, graph-based prompts and smaller, less multilingual models exhibit more improvement after providing examples. In terms of language resource level, MT from English into low-resource languages exhibits greater improvement from 0-shot to multi-shot prompting while high-resource languages improve more across 1/3/5-shot prompts.

4.3 Prompt Strategies

The results in Section 4.2 suggest that increasing exposure to AMR/UMR through demonstrations reduces the performance drop induced by including the SR, but not sufficiently to outperform the baseline. A natural follow-up question is whether this limitation can be addressed through alternate instructions.

Table 2 shows that among the three variants that incorporate SRs into the prompt, `SR-Direct` consistently achieves the best performance, followed by `SR-Assisted`. This suggests that detailed instructions on how to use SRs may interfere with the model's own strategy for leveraging them. While An et al. (2025) report that `SENSE` consistently enhances GPT-3.5 across four language

Model	Shot	No SR		SR-Direct		SR-Grounded		SR-Assisted	
		Baseline	SENSE	Graph	NLD	Graph	NLD	Graph	NLD
Aya-23-8B	0 Shot	**5.77**	6.73 ↓	6.20 ↓	6.28 ↓	6.48 ↓	6.18 ↓	6.28 ↓	6.16 ↓
	1 Shot	**5.72**	6.06 ↓	6.06 ↓	6.10 ↓	6.15 ↓	6.29 ↓	6.02 ↓	6.07 ↓
	3 Shot	**5.63**	5.85 ↓	5.79 ↓	5.79 ↓	5.84 ↓	5.90 ↓	5.81 ↓	5.76 ↓
	5 Shot	**5.59**	5.75 ↓	5.72 ↓	5.69 ↓	5.79 ↓	5.76 ↓	5.79 ↓	5.65 ↓
Gemma-3-4B-IT	0 Shot	**5.34**	5.48 ↓	**5.34** –	5.37 ↓	5.98 ↓	5.45 ↓	5.41 ↓	5.36 ↓
	1 Shot	4.91	**4.88** ↑	5.03 ↓	4.97 ↓	5.15 ↓	5.06 ↓	5.00 ↓	4.90 ↑
	3 Shot	4.65	**4.60** ↑	4.73 ↓	4.68 ↓	4.79 ↓	4.75 ↓	4.73 ↓	4.66 ↓
	5 Shot	4.55	**4.54** ↑	4.63 ↓	4.61 ↓	4.69 ↓	4.62 ↓	4.64 ↓	4.56 ↓
Gemma-3-12B-IT	0 Shot	4.09	4.14 ↓	4.10 ↓	**4.06** ↑	4.28 ↓	4.26 ↓	4.14 ↓	4.14 ↓
	1 Shot	3.84	**3.83** ↑	3.92 ↓	3.90 ↓	4.09 ↓	4.01 ↓	3.95 ↓	3.86 ↓
	3 Shot	3.71	**3.69** ↑	3.75 ↓	3.73 ↓	3.88 ↓	3.82 ↓	3.78 ↓	3.76 ↓
	5 Shot	3.67	**3.63** ↑	3.71 ↓	3.71 ↓	3.82 ↓	3.76 ↓	3.73 ↓	3.70 ↓
Gemma-3-27B-IT	0 Shot	3.56	3.62 ↓	3.53 ↑	**3.52** ↑	3.64 ↓	3.59 ↓	3.57 ↓	3.54 ↑
	1 Shot	**3.35**	3.37 ↓	3.40 ↓	3.39 ↓	3.46 ↓	3.41 ↓	3.41 ↓	3.39 ↓
	3 Shot	3.27	**3.25** ↑	3.28 ↓	3.26 ↑	3.33 ↓	3.30 ↓	3.31 ↓	3.28 ↓
	5 Shot	**3.21**	**3.21** –	3.24 ↓	3.22 ↓	3.29 ↓	3.24 ↓	3.26 ↓	3.22 ↓
Llama-3.1-8B-Instruct	0 Shot	**5.71**	5.96 ↓	5.98 ↓	6.00 ↓	7.25 ↓	6.64 ↓	6.06 ↓	6.06 ↓
	1 Shot	**5.19**	5.29 ↓	5.70 ↓	5.62 ↓	6.73 ↓	6.29 ↓	5.90 ↓	5.65 ↓
	3 Shot	**4.93**	5.12 ↓	5.20 ↓	5.15 ↓	6.13 ↓	5.46 ↓	5.47 ↓	5.18 ↓
	5 Shot	**4.86**	5.19 ↓	5.07 ↓	5.00 ↓	5.93 ↓	5.22 ↓	5.35 ↓	5.04 ↓
Mistral-7B-Instruct	0 Shot	**7.58**	7.75 ↓	8.61 ↓	8.29 ↓	14.48 ↓	10.73 ↓	9.72 ↓	8.06 ↓
	1 Shot	**6.77**	7.82 ↓	7.15 ↓	7.03 ↓	8.98 ↓	7.74 ↓	7.46 ↓	7.05 ↓
	3 Shot	**6.35**	6.92 ↓	6.52 ↓	6.46 ↓	7.50 ↓	7.07 ↓	6.64 ↓	6.57 ↓
	5 Shot	**6.21**	6.51 ↓	6.33 ↓	6.29 ↓	7.00 ↓	6.61 ↓	6.41 ↓	6.34 ↓
Qwen2.5-7B-Instruct	0 Shot	7.71	7.70 ↑	7.66 ↑	**7.65** ↑	7.89 ↓	7.73 ↓	7.74 ↓	7.67 ↑
	1 Shot	**6.80**	6.83 ↓	6.85 ↓	6.83 ↓	7.02 ↓	6.88 ↓	6.93 ↓	**6.80** –
	3 Shot	6.46	6.39 ↑	6.43 ↑	**6.38** ↑	6.55 ↓	6.42 ↑	6.54 ↓	6.39 ↑
	5 Shot	6.27	6.23 ↑	6.25 ↑	6.24 ↑	6.34 ↓	6.22 ↑	6.32 ↓	**6.20** ↑

Table 2: MetricX results across models and prompt strategies, averaged over 12 target languages and weighted by the number of sentences per language (including both AMR and UMR). Green upward arrows (↑) indicate improvements over the baseline, while red downward arrows (↓) indicate degradations. Bold values indicate the best-performing strategy in each row group. Note that lower scores indicate better performance.

Resource	0-shot	5-shot	Δ (5–0)	t
Low	11.1750	9.5227	-1.6523	13.2908
Medium	7.8027	6.3093	-1.4934	18.8776
High	3.7099	2.9869	-0.7230	12.4594

Table 3: Comparison of MetricX scores between 0-shot and 5-shot prompting across language resource levels. Statistical significance was evaluated using a two-sample t-test, all t-statistics have $p < 0.001$.

pairs, these gains do not appear in our experiments (see Table 2). We convert SRs into natural language descriptions (NLDs), but observe little improvement over the baseline in our settings. Overall, our results extend prior work through a broader evaluation across languages, models, and prompt templates, revealing that previously reported gains vary across settings and disappear when averaged across languages.

4.4 The Advantage of UMR

Out of all experimental conditions comparing AMR and UMR variants (across models, languages, number of shots, and prompts), totaling 2,016, UMR outperforms AMR in 1,156 cases (57.3%) under BLEU, 1,159 cases (57.5%) under COMET, and 1087 cases (53.9%) under MetricX. As the number of shots increase, win rates increase from 53.4% to 60.5% for BLEU, from 54.8% to 61.3% for COMET, and from 54.4% to 57.9% for MetricX. These relative improvements suggest that UMR inclusion benefits more from in-context examples than AMR, eventually enabling a clear majority of wins. We suspect that this is due to the substantial amount of linguistic information contained in UMR, making more demonstrations necessary. Regardless, these raw win rates only offer an initial signal that UMR may have an edge over AMR.

A natural follow-up question is whether this ad-

vantage is *statistically sound*, or whether it simply reflects noise from the results. To investigate this, we zoom in on the setting where linearized AMR or UMR graphs are incorporated directly into the prompt without any examples (zero-shot `SR-Direct`). For each model and language, we run a one-tailed paired t-test in each direction–one testing whether AMR scores significantly better than UMR, and the other testing the reverse–yielding the full set of p-values reported in Table 11.

Following Dror et al. (2017), we compute $\hat{k}_{Fisher}$, which is a statistically justified estimate of the number of languages for which SR outperforms the other, per model and metric (Table 4). For six of the seven models, $\hat{k}_{Fisher}$ is at or near zero in both directions, indicating little to no evidence that either representation is significantly better. Mistral is once again the clear exception: $\hat{k}_{Fisher}$ shows a clear advantage for UMR across nearly all 12 languages under every metric. This suggests that UMR's advantage, while not universal, can be substantial and consistent under specific conditions. We suspect that the model's limited support for non-English languages may be one factor at play for UMR's advantage here.

These findings reinforce our motivation for studying UMR as an alternative to AMR. While AMR likely benefits from greater exposure during pretraining, UMR shows larger gains with increasing in-context examples, and for Mistral, its advantage over AMR is significant. This suggests that UMR's benefits are not yet universal across models, but can emerge strongly under the right conditions. More broadly, these results are modest, but help justify investment in UMR infrastructure: more capable UMR parsers and thus larger-scale UMR resources would not only close this gap for more models, but also support broader downstream studies of UMR's utility.

Metric	AMR > UMR	UMR > AMR
Aya-23-8B		
BLEU	0	0
COMET	**1**	0
MetricX	**1**	0
Gemma-3-4B-IT		
BLEU	0	**1**
COMET	0	**1**
MetricX	0	0
Gemma-3-12B-IT		
BLEU	0	0
COMET	0	0
MetricX	0	0
Gemma-3-27B-IT		
BLEU	0	0
COMET	0	0
MetricX	0	0
Llama-3.1-8B-Instruct		
BLEU	0	0
COMET	0	**1**
MetricX	0	0
Mistral-7B-Instruct		
BLEU	0	**12**
COMET	0	**12**
MetricX	0	**11**
Qwen2.5-7B-Instruct		
BLEU	0	0
COMET	0	0
MetricX	0	0

Table 4: $\hat{k}_{Fisher}$ (out of 12 languages, $\alpha = 0.05$) in the case of zero-shot `SR-Direct` prompting, per model and metric. The *AMR > UMR* column reports the estimated number of languages for which AMR outperforms UMR; the *UMR > AMR* column reports the reverse. Non-zeros are bolded.

5 Error Analysis

To better understand how SRs affect translation quality in our experiments, we conduct a systematic diagnosis of the models' outputs, surfacing several error patterns that aggregate metric scores obscure.

After manually reviewing model outputs of all the configurations (model-method-language-graph type), we identify four recurring errors: (1) the model outputs an AMR/UMR graph in place of a natural-language translation, (2) the translation includes parenthetical romanization of source-script terms, (3) the translation includes parenthetical comments about the text, and (4) the translation uses slashes to list alternative word choices or masculine/feminine variants.

Mistral exhibits the first error mode most severely, reaching 63.7–97.9% graph-output errors with UMR and 52.6–89.3% with AMR. Critically, SR augmentation mostly amplifies rather than mitigates these errors in Mistral (Appendix E). Intuitively, little to no graph-output error is observed for configurations with NLDs. Error rates decrease as the number of shots increase across all prompting strategies and models, suggesting that demonstrations of how the SR is used help to re-

solve the ambiguity around the expected output type.

Unlike graph-output errors, the other three errors persist at similar rates regardless of shot counts. We argue that the models do not treat this additional content as erroneous, but rather as value-added service to the user. The romanization is not an arbitrary behavior but a targeted behavior at non-Latin-script languages (Appendix 11).

In order to understand the utlility of AMR and UMR in ICL, we now examine a case where UMR and AMR both improve model performance. Consider the following English-Japanese sentence pair, accompanied by the baseline translations, as well as the Llama 1-shot `SR-Direct` translations, with AMR/UMR:

> **Source (English):** For me, it was like a spring of fresh water in the desert.
> **Gold (Japanese):** 彼の笑いは僕にとって沙漠の泉のようなものだから。
> **Translated (baseline)** COMET=0.695, MetricX=10.780
> 砂漠の中の泉の方のように思えた。
> (It felt like an oasis in the desert.)
> **Translated (AMR)** COMET=0.796, MetricX=7.629
> 砂漠の中で、泉のような新鮮な水が感じられた。
> (In the midst of the desert, I sensed water as fresh as that of a spring.)
> **Translated (UMR)** COMET=0.846, MetricX=4.033
> 私にとっては、砂漠の中の清水の泉のように感じた。
> (To me, it felt like a spring of fresh water in the desert.)

While the baseline captures the metaphor of desert-spring, it omits the speaker entirely; both AMR and UMR improve the translation but to different degrees. Although there is no direct mention or translation of "fresh water" in the gold Japanese, both AMR and UMR recover the "fresh water" content because it is encoded in their graph structures (AMR: `(w / water :ARG1-of (f / fresh-04)`; UMR: `(s1414w / water :ARG1-of (s1414f / fresh-04)`). The difference in *sensing* via the AMR-prompted translation and *feeling* via the UMR-prompted translation is likely due to how this action is encoded in the corresponding English graphs. AMR labels the speaker as `:beneficiary`, which we argue is less suitable for the sentence given that the speaker does not benefit from the resemblance but rather feels it. In this regard, UMR labels the speaker as `:affectee`, a role that is more neutral.

6 Conclusion

In this work, we integrate English Abstract Meaning Representation and Uniform Meaning Representation gold annotations into prompts to assess their effectiveness as in-context learning strategies for machine translation. We present the first large-scale study of AMR and UMR in LLM-based MT, evaluating English-to-X translation across 12 target languages, 7 models from 5 model families, a wide range of prompting protocols, and 3 evaluation metrics. Our experiments show that conclusions about the effectiveness of AMR and UMR are highly sensitive to the choice of languages, models, and evaluation metrics, highlighting the importance of broad evaluation in this setting.

Our results further suggest that prompts using UMR tend to outperform those using AMR in more settings, especially for certain models and when provided demonstrations. This provides the first empirical evidence that UMR may hold relative advantage over AMR in a downstream task. This serves as motivation for investment in UMR resources, supporting their application to a broader range of tasks (as has been the case for AMR (Wein and Opitz, 2024)).

UMR's advantage may reflect greater information volume rather than representational superiority. Future work might explore the relative benefit of AMR and UMR graphs over randomized control conditions, such as a random graph, a shuffled graph, or unrelated graph, to reveal whether observed effects stem from the semantic content of the SR or simply from the presence of additional pertinent text. Future work could further analyze other tasks where SRs are have been shown to be most helpful, such as in low-resource settings (Guruaja et al., 2023; Hua et al., 2023; Wein, 2025) or for data augmentation/controlled generation of text (Ghazarian et al., 2022; Shou et al., 2022; Huang et al., 2022, 2023; Ghosh et al., 2024).

Limitations

Our experiments are restricted to English-to-target-language translation due to the availability of parallel gold AMR and UMR annotations. To our knowledge, *The Little Prince* is the only text

with both gold English AMR and UMR annotations, and published translations into many target languages, which is what makes a comparison at this scale possible. However, literary text differs from common machine translation benchmarks in style and complexity, and prior work has noted that automatic metrics can be less reliable in this setting (Thai et al., 2022; Zhang et al., 2025b). We recognize that automatic metrics can be unreliable especially when evaluating low-resource languages (Falcão et al., 2024); we include three metrics capturing lexical overlap, semantic adequacy, and error analyses in an effort to mitigate this issue. We hope our findings motivate future work extending to other domains as more parallel AMR/UMR data becomes available.

In order to perform a robust analysis across many models and languages, we constrict our scope to in-context learning for machine translation. We do not explore fine-tuning approaches, primarily due to the limited availability of high-quality UMR-annotated data and parsing tools. We evaluate existing methods along with a small set of variants introduced in this study. As a result, our findings are largely *analytical* in nature, highlighting the inherent challenges of effectively leveraging SRs as a linguistic resource within current LLM-based machine translation systems.

The use of AMR and UMR is motivated in part by their potential to improve interpretability in Transformer language models. We leave this to future work.

Ethical Considerations

Much of the data used in this work comes from the InterCorp parallel corpus, obtained by contacting the team maintaining it and signing a license agreement. These restrictions influence our experimental design: in particular, although the prior work we build upon evaluates commercial models such as GPT-4, GPT-4o-mini, and GPT-3.5 Turbo, we do not use these models in order to comply with the dataset license.

Given that our work contributes to LLM-based MT, we also acknowledge broader ethical concerns surrounding this topic. Expanding MT capabilities for low-resource languages may introduce risks of misuse, including surveillance or profiling of vulnerable language communities. Our work is exploratory and focuses on understanding the role of semantic representations in prompting-based MT rather than developing real-world models; nevertheless, future work in this area should carefully consider these societal implications.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2023. In-context examples selection for machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8857–8873, Toronto, Canada. Association for Computational Linguistics.

Kaikai An, Shuzheng Si, Helan Hu, Haozhe Zhao, Yuchi Wang, Qingyan Guo, and Baobao Chang. 2025. Rethinking semantic parsing for large language models: Enhancing LLM performance with semantic hints. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1020–1029, Vienna, Austria. Association for Computational Linguistics.

Rafael Torres Anchiêta and Thiago Alexandre Salgueiro Pardo. 2022. Abstract meaning representation parsing for the brazilian portuguese language. In *Computational Processing of the Portuguese Language*, pages 429–434, Cham. Springer International Publishing.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress. *Preprint*, arXiv:2405.15032.

Duygu Ataman, Alexandra Birch, Nizar Habash, Marcello Federico, Philipp Koehn, and Kyunghyun Cho. 2025. Machine translation in the era of large language models:a survey of historical and emerging problems. *Information*, 16(9).

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Frantisek Cermák and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics*, 13(3):411–427.

Jayeol Chun and Nianwen Xue. 2024. Uniform meaning representation parsing as a pipelined approach. In *Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing*, pages 40–52, Bangkok, Thailand. Association for Computational Linguistics.

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

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025. Multilingual machine translation with open large language models at practical scale: An empirical study. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5420–5443, Albuquerque, New Mexico. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. 2017. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486.

Júlia Falcão, Claudia Borg, Nora Aranberri, and Kurt Abela. 2024. COMET for low-resource machine translation evaluation: A case study of English-Maltese and Spanish-Basque. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3553–3565, Torino, Italia. ELRA and ICCL.

Samuel Frontull and Thomas Ströhle. 2025. Compensating for data with reasoning: Low-resource machine translation with llms. *Preprint*, arXiv:2505.22293.

Sarik Ghazarian, Nuan Wen, Aram Galstyan, and Nanyun Peng. 2022. DEAM: Dialogue coherence evaluation using AMR-based semantic manipulations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 771–785, Dublin, Ireland. Association for Computational Linguistics.

Sreyan Ghosh, Utkarsh Tyagi, Sonal Kumar, Chandra Kiran Evuru, Ramaneswaran S, S Sakshi, and Dinesh Manocha. 2024. ABEX: Data augmentation for low-resource NLU via expanding abstract descriptions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 726–748, Bangkok, Thailand. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and Heyan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15685–15697, Torino, Italia. ELRA and ICCL.

Sireesh Gururaja, Ritam Dutt, Tinglong Liao, and Carolyn Rosé. 2023. Linguistic representations for fewer-shot relation extraction across domains. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7502–7514, Toronto, Canada. Association for Computational Linguistics.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *Preprint*, arXiv:2302.09210.

Yilun Hua, Zhaoyuan Deng, and Kathleen McKeown. 2023. Improving long dialogue summarization with semantic graph representation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13851–13883, Toronto, Canada. Association for Computational Linguistics.

Kuan-Hao Huang, Varun Iyer, I-Hung Hsu, Anoop Kumar, Kai-Wei Chang, and Aram Galstyan. 2023. ParaAMR: A large-scale syntactically diverse paraphrase dataset by AMR back-translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8047–8061, Toronto, Canada. Association for Computational Linguistics.

Kuan-Hao Huang, Varun Iyer, Anoop Kumar, Sriram Venkatapathy, Kai-Wei Chang, and Aram Galstyan. 2022. Unsupervised syntactically controlled paraphrase generation with Abstract Meaning Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1547–1554, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhijing Jin, Yuen Chen, Fernando Gonzalez Adauto, Jiarui Liu, Jiayi Zhang, Julian Michael, Bernhard Schölkopf, and Mona Diab. 2024. Analyzing the role of semantic representations in the era of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages

3781–3798, Mexico City, Mexico. Association for Computational Linguistics.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Pavan Kapanipathi, Ibrahim Abdelaziz, Srinivas Ravishankar, Salim Roukos, Alexander Gray, Ramón Fernandez Astudillo, Maria Chang, Cristina Cornelio, Saswati Dana, Achille Fokoue, Dinesh Garg, Alfio Gliozzo, Sairam Gurajada, Hima Karanam, Naweed Khan, Dinesh Khandelwal, Young-Suk Lee, Yunyao Li, Francois Luus, and 11 others. 2021. Leveraging Abstract Meaning Representation for knowledge base question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3884–3894, Online. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. 2016. Annotating the little prince with Chinese AMRs. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 7–15, Berlin, Germany. Association for Computational Linguistics.

Fei Liu, Jeffrey Flanigan, Sam Thomson, Norman Sadeh, and Noah A. Smith. 2015. Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Emma Markle, Javier Gutierrez Bach, and Shira Wein. 2026. SETUP: Sentence-level English-to-uniform meaning representation parser. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 9710–9721, Palma de Mallorca, Spain. ELRA Language Resource Association.

Emma Markle, Reihaneh Iranmanesh, and Shira Wein. 2025. Generating text from uniform meaning representation. In *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 259–271, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Raphaël Merx, Aso Mahmudi, Katrina Langford, Leo Alberto de Araujo, and Ekaterina Vylomova. 2024. Low-resource machine translation through retrieval-augmented LLM prompting: A study on the Mambai language. In *Proceedings of the 2nd Workshop on Resources and Technologies for Indigenous, Endangered and Lesser-resourced Languages in Eurasia (EURALI) @ LREC-COLING 2024*, pages 1–11, Torino, Italia. ELRA and ICCL.

Long H. B. Nguyen, Viet H. Pham, and Dien Dinh. 2021. Improving neural machine translation with amr semantic graphs. *Mathematical Problems in Engineering*, 2021(1):9939389.

Juri Opitz and Anette Frank. 2021. Towards a decomposable metric for explainable evaluation of text generation from AMR. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1504–1518, Online. Association for Computational Linguistics.

Renhao Pei, Yihong Liu, Peiqin Lin, François Yvon, and Hinrich Schuetze. 2025. Understanding in-context machine translation for low-resource languages: A case study on Manchu. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8767–8788, Vienna, Austria. Association for Computational Linguistics.

Hai Pham, Isma Hadji, Xinnuo Xu, Ziedune Degutyte, Jay Rainey, Evangelos Kazakos, Afsaneh Fazly, Georgios Tzimiropoulos, and Brais Martinez. 2024. Graph guided question answer generation for procedural question-answering. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2501–2525, St. Julian's, Malta. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova,

Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. COMET-22: Unbabel-IST 2022 submission for the metrics shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. 2022. FactGraph: Evaluating factuality in summarization with semantic graph representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3238–3253, Seattle, United States. Association for Computational Linguistics.

Zacchary Sadeddine, Juri Opitz, and Fabian Suchanek. 2024. A survey of meaning representations – from theory to practical utility. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2877–2892, Mexico City, Mexico. Association for Computational Linguistics.

Ziyi Shou, Yuxin Jiang, and Fangzhen Lin. 2022. AMR-DA: Data augmentation by Abstract Meaning Representation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3082–3098, Dublin, Ireland. Association for Computational Linguistics.

Linfeng Song, Daniel Gildea, Yue Zhang, Zhiguo Wang, and Jinsong Su. 2019. Semantic neural machine translation using AMR. *Transactions of the Association for Computational Linguistics*, 7:19–31.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *International Conference on Learning Representations*, volume 2024, pages 18955–18985.

Gemma Team. 2025. Gemma 3.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jens E. L. Van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, Andrew Cowell, William Croft, Chu-Ren Huang, Jan Haji, James H. Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, Rosa Vallejos, and Nianwen Xue. 2021. Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz*, 35(3):343–360.

Cunxiang Wang, Zhikun Xu, Qipeng Guo, Xiangkun Hu, Xuefeng Bai, Zheng Zhang, and Yue Zhang. 2023. Exploiting Abstract Meaning Representation for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2083–2096, Toronto, Canada. Association for Computational Linguistics.

Shira Wein. 2025. Can uniform meaning representation help GPT-4 translate from indigenous languages? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 278–285, Vienna, Austria. Association for Computational Linguistics.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 23–33, Nancy, France. Association for Computational Linguistics.

Shira Wein and Juri Opitz. 2024. A survey of AMR applications. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6856–6875, Miami, Florida, USA. Association for Computational Linguistics.

Shira Wein and Nathan Schneider. 2024. Assessing the cross-linguistic utility of abstract meaning representation. *Computational Linguistics*, 50(2):419–473.

Jiahuan Zhang, Tianheng Wang, Hanqing Wu, Ziyi Huang, Yulong Wu, Dongbai Chen, Linfeng Song, Yue Zhang, Guozheng Rao, and Kaicheng Yu. 2025a. SR-LLM: Rethinking the structured representation in large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3443–3462, Vienna, Austria. Association for Computational Linguistics.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

Ran Zhang, Wei Zhao, and Steffen Eger. 2025b. How good are LLMs for literary translation, really? literary translation evaluation with humans and LLMs. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10961–10988, Albuquerque, New Mexico. Association for Computational Linguistics.

Zixuan Zhang and Heng Ji. 2021. Abstract Meaning Representation guided graph encoding and decoding for joint information extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 39–49, Online. Association for Computational Linguistics.

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. In *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 6261–6270, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Prompts

We present all zero-shot prompts in Figure 3 and all few-shot prompts in Figures 4 and 5, for the variant in which AMR or UMR is included in the prompt as a linearized graph. For the natural language description (NLD) variant, the overall prompt structure remains unchanged, with only minor wording edits. See the figure captions for details.

B Converting SRs to NLDs

Following Zhang et al. (2025a), we convert AMR or UMR graphs into natural language descriptions (NLDs) by transforming graph triples into a set of subsentences that collectively describe the original graph. The process consists of 4 steps:

1. We first extract all relation triples from the AMR or UMR graph.

2. Each triple is then converted into a subsentence using the relation mapping rules obtained from Zhang et al. (2025a). Because their mapping rules only cover AMR relations, we additionally define mappings for UMR-specific relations. For example, the relation `:modal-strength` between two entities is mapped to the template "{entity1}'s modal strength is {entity2}."

3. We then expand each entity using definitions derived from the corresponding PropBank entries, as well as the UMR guidelines[5]. For example, the entity `performance` is expanded to "performance(process that ends and reaches a result state)."

4. Finally, the generated subsentences are provided as input to GPT-4o Mini[6] to produce coherent natural language descriptions. Figure 6 shows the prompt we used.

Figure 7 illustrates a linearized UMR graph and its corresponding NLD.

C Statistical Significance Tests for Direct AMR/UMR Graph Prompting

We perform a two-tailed paired t-tests on the automatic metric scores to determine whether the effects of directly inserting linearized AMR or UMR graphs into the prompt in the zero-shot setting (`SR-Direct`) are statistically significant. These tests correspond to the heatmaps shown in Figure 2. We report the results in Table 5, covering all three metrics and both the AMR and UMR variants.

As shown in these tables, only a fraction of the results are statistically significant, indicating that the impact of direct AMR/UMR prompting tends to be noise. Among the statistically significant cases, most correspond to performance decreases, suggesting that AMR and UMR often significantly harm output quality. In particular, all results for Mistral-7B-Instruct across all metrics are statistically significant, indicating that inserting AMR or UMR into the prompt consistently degrades Mistral-7B-Instruct's performance. A similar trend can be observed for most COMET and MetricX results with Llama-3.1-8B-Instruct.

D Detailed Prompt Strategies Results

Here, we provide additional results extending Table 2 in the main paper.

D.1 BLEU and COMET Results

Tables 6 and 7 report BLEU and COMET results for all prompt strategies, respectively, paralleling Table 2 (which reports MetricX). We observe that BLEU is less aligned with COMET and MetricX:

[5] https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md

[6] The total cost was approximately USD 1.

many improvements in COMET and MetricX do not translate into BLEU improvements, and Table 6 therefore shows the prompt strategies as mostly detrimental compared to the baseline. This suggests that these strategies are less effective at increasing n-gram overlap with the reference, but more effective at improving semantic similarity.

In addition, while Table 2 shows that MetricX consistently favors the NLD version over their graph counterparts, BLEU and COMET show a more mixed pattern, with several cases where the graph versions outperform their NLD ones.

The BLEU and COMET results also make it clearer that the SR-Grounded prompt with linearized graphs in the zero-shot setting performs particularly poorly across both metrics, and that converting graphs into NLD is especially beneficial in these cases. Inspecting model outputs suggests that, in the graph version, models are more likely to translate the *graphs* themselves rather than the source sentence, especially when they are AMR graphs, leading to severely degraded outputs. This issue is substantially reduced when using UMR graphs or when graphs are converted into NLD, resulting in better performance. The fact that UMR is less harmful in this case may suggest that its additional linguistic information make it less likely to be misinterpreted as the text to be translated, thereby reducing the tendency to translate the *graph* rather than the source sentence. It may also be because these models have been exposed to cross-lingual adaptations of AMR, making them more likely to treat the given English AMR graphs as text to be reproduced in their outputs when predicting the next token.

D.2 Per-language MetricX Results

We report results per language for two prompt strategies: `SENSE` and `SR-Direct` where SRs are converted into NLD. We focus on SR-Direct in NLD form as it tends to yield the strongest overall performance among methods involving SRs. Figures 8 and 9 present heatmaps for both variants. Table 8 presents statistical significance tests comparing `SENSE` with the baseline. Tables 9 (SR-Direct with AMR converted to NLD) and 10 (SR-Direct with UMR converted to NLD) report statistical significance tests for SR-Direct when SRs are converted to NLD. Unlike prior work, we explicitly test significance to determine whether observed differences are meaningful.

For `SENSE`, we observe that improvements are mostly in the five-shot setting, while zero-shot setting tends to yield degradations. Meanwhile, An et al. (2025) report zero-shot gains for English–German and English–Chinese translation under BLEU and COMET22 using the WMT22 datasets.

For SR-Direct with NLD, we find that the zero-shot setting yields more improvements than the five-shot setting. In the zero-shot case, this strategy leads to BLEU improvements for Qwen2.5-7B-Instruct across most target languages, as well as a small number of additional model–language pairs. This partially aligns with Zhang et al. (2025a), who report BLEU gains for English–German translation using LLaMA3.1-8B-Instruct and GPT-3.5-turbo using the WMT16 dataset, in the sense that the method can be beneficial in specific model and language configurations. However, in our experiments on English–German translation, BLEU improvements are primarily observed for the Gemma-3-IT models, while LLaMA3.1-8B-Instruct does not show consistent gains.

Finally, we again observe that the automatic metrics do not always agree with each other, as improvements in one metric do not consistently correspond to improvements in another, and similarly for degradations. This suggests that evaluation conclusions can depend strongly on the choice of metric, highlighting the need for more comprehensive evaluation that uses multiple metrics.

E Error Heatmaps by Model-Method-Graph Configuration at 0-shot

This appendix provides the full heatmaps of the error rates of the four types across all model-method-graph configurations, split by Latin-script and non-Latin-script target languages, at zero shot. Cell values are mean error rates (%) averaged across the grouped languages.

F Graph-Output Error Heatmaps

Similarly, this appendix provides the full heatmaps of the graph-output error rates across language-method-graph configurations for each model.

G Error Heatmaps by Model-Method-Graph Configuration at 5-shot

Baseline (Zero-Shot)

Translate the following English sentence into {tgt_language}. Please output ONLY the translation.
English: {English sentence}
{tgt_language}:

SR-Direct (Zero-Shot)

Translate the following English sentence (accompanied by an X Meaning Representation parse) into {tgt_language}. Please output ONLY the translation.
English: {English sentence}
X Meaning Representation: {English XMR}
{tgt_language}:

SR-Grounded (Zero-Shot)

Task: Meaning-Preserving Translation Using Uniform Meaning Representation

Instructions:
- Use the Uniform Meaning Representation as the full semantic structure.
- Preserve predicate–argument relations.
- Preserve negation, modifiers, quantities, and named entities.
- Please output ONLY the translation.

Translate the following English sentence into {tgt_language}:
English: {English sentence}
X Meaning Representation: {English XMR}
{tgt_language}:

SR-Assisted (Zero-Shot)

Your task is to translate the English sentence into {tgt_language} using the Uniform Meaning Representation parse.

Instructions:
- Prioritize the English sentence.
- The Uniform Meaning Representation parse is provided only as a secondary aid.
- Use the Uniform Meaning Representation parse only if the meaning is unclear.
- Please output ONLY the translation.

Translate the following English sentence into {tgt_language}:
English: {English sentence}
X Meaning Representation: {English XMR}
{tgt_language}:

SENSE (Zero-Shot)

Please translate this English sentence into {tgt_language} by utilizing its semantic parsing result which helps to understand grammar and semantics (output ONLY the translation):
English: {English sentence}
{tgt_language}:

Figure 3: User prompts in the zero-shot setting for the graph variant. For the NLD variant, we prepend the following sentence to the user prompt: `You will be provided with an English sentence and its X Meaning Representation described in natural language`. We also replace later occurrences of `parse` with `description`, and replace `X Meaning Representation` with `Input X Meaning Representation description`. For all prompts, the system prompt is: `You are a helpful translation assistant.`

Baseline (Few-Shot)

Given the following examples of English sentences and their {tgt_language} translations:
English: {Sentence 1} {tgt_language}: {Translation 1}
English: {Sentence 2} {tgt_language}: {Translation 2}
...
English: {Sentence n} {tgt_language}: {Translation n}
Translate the following English sentence into {tgt_language}. Please output ONLY the translation.
English: {English sentence}
{tgt_language}:

SR-Direct (Few-Shot)

Given the following examples of English sentences (accompanied by X Meaning Representation parses) and their {tgt_language} translations:
English: {Sentence 1} X Meaning Representation: {XMR 1} {tgt_language}: {Translation 1}
English: {Sentence 2} X Meaning Representation: {XMR 2} {tgt_language}: {Translation 2}
...
English: {Sentence n} X Meaning Representation: {XMR n} {tgt_language}: {Translation n}
Translate the following English sentence (accompanied by an X Meaning Representation parse) into {tgt_language}. Please output ONLY the translation.
English: {English sentence}
X Meaning Representation: {English XMR}
{tgt_language}:

SR-Grounded (Few-Shot)

Task: Meaning-Preserving Translation Using Uniform Meaning Representation

Instructions:
- Use the Uniform Meaning Representation as the full semantic structure.
- Preserve predicate–argument relations.
- Preserve negation, modifiers, quantities, and named entities.
- Please output ONLY the translation.

Examples:
English: {Sentence 1} X Meaning Representation: {XMR 1} {tgt_language}: {Translation 1}
English: {Sentence 2} X Meaning Representation: {XMR 2} {tgt_language}: {Translation 2}
...
English: {Sentence n} X Meaning Representation: {XMR n} {tgt_language}: {Translation n}

Translate the following English sentence into {tgt_language}:
English: {English sentence}
X Meaning Representation: {English XMR}
{tgt_language}:

Figure 4: User prompts in the few-shot setting for the graph variant. For the NLD variant, we prepend the following sentence to the user prompt: `You will be provided with an English sentence and its X Meaning Representation described in natural language`. We also replace later occurrences of `parse` with `description`, and replace `X Meaning Representation` with `Input X Meaning Representation description`. For all prompts, the system prompt is: `You are a helpful translation assistant.`

SR-Assisted (Few-Shot)

Your task is to translate the English sentence into {tgt_language} using the Uniform Meaning Representation parse.

Instructions:
- Prioritize the English sentence.
- The Uniform Meaning Representation parse is provided only as a secondary aid.
- Use the Uniform Meaning Representation parse only if the meaning is unclear.
- Please output ONLY the translation.

Examples:
English: {Sentence 1} X Meaning Representation: {XMR 1} {tgt_language}: {Translation 1}
English: {Sentence 2} X Meaning Representation: {XMR 2} {tgt_language}: {Translation 2}
...
English: {Sentence n} X Meaning Representation: {XMR n} {tgt_language}: {Translation n}

Translate the following English sentence into {tgt_language}:
English: {English sentence}
X Meaning Representation: {English XMR}
{tgt_language}:

SENSE (Few-Shot)

Please translate the final English sentence into {tgt_language} by utilizing its semantic parsing result which helps to understand grammar and semantics (output ONLY the translation):
English: {Sentence 1} {tgt_language}: {Translation 1}
English: {Sentence 2} {tgt_language}: {Translation 2}
...
English: {Sentence n} {tgt_language}: {Translation n}
English: {English sentence}
{tgt_language}:

Figure 5: User prompts in the few-shot setting for the graph variant (cont.).

Produce Coherent NLD

[system_prompt]
Your task is to polish a collection of sub sentences and rewrite them based on the original sentence to make them read more like natural language. Please output your answer according to the examples.
###
Example 1
input_original_sentence: "He presented his research at the meeting yesterday."
input_sub_sentences: "

```
0 person is the doer of present-01(give).
1 3rd is the grammatical person of person.
2 singular is the grammatical number of person.
3 thing is the object of present-01(give).
4 thing is the object of research-01(study very hard).
5 person is the doer of research-01(study very hard).
6 present-01(give) takes place at meet-01(arrive at, achieve).
7 meet-01(arrive at, achieve) has aspect process.
8 present-01(give) is under the temporal circumstances of yesterday.
9 present-01(give) has aspect performance.
10 present-01(give)'s modal strength is full-affirmative.
```

" **output_sentences:** "

```
0 A person is the one who presented something.
1 The person is referred to in the third person.
2 The person is singular.
3 Something is being presented.
4 This thing is research.
5 The research is done by the person.
6 The presentation took place at a meeting.
7 The meeting is described as a process.
8 The presentation took place yesterday.
9 The presentation is completed and reaches a result.
10 There is full certainty that the presentation took place.
```

" ###
Example 2
input_original_sentence: "Are you still there?"
input_sub_sentences:"

```
0 you is the object of be-located-at-91(reification of :location).
1 there is the indirect object of be-located-at-91(reification of :location).
2 be-located-at-91(reification of :location) is described as still.
3 amr-unknown indicates a question about the polarity of be-located-at-91(reification of :location).
```

output_sentences:"

```
0 `you' is the one whose location is being described.
1 `there' is the location.
2 The state of being located is described as continuing (`still;)
3 This is a yes/no question about whether that state holds.
```

[user_prompt]
input_original_sentence: {original sentence}
input_sub_sentences: {generated subsentences}
output_sentences:

Figure 6: Prompt used to refine generated subsentences into coherent natural language descriptions (NLDs).

Including UMR as a linearized graph

Translate the following English sentence (accompanied by a Uniform Meaning Representation parse) into `{tgt_language}`. Please output ONLY the translation.
English: She had come in the form of a seed.
Uniform Meaning Representation: (s411c / come-01 :ARG1 (s411s / she) :manner (s411h / have-role-91 :ARG1 s411s :ARG3 (s411s2 / seed)) :aspect performance :modal-strength full-affirmative)
`{tgt_language}`:

Including UMR as a NLD

You will be provided with an English sentence and its Uniform Meaning Representation described in natural language.

Translate the following English sentence into `{tgt_language}`. Please output ONLY the translation.
English: She had come in the form of a seed.
Input Uniform Meaning Representation description:
0 "She" is described as the one who has arrived.
1 The arrival is characterized by a non-typical role.
2 "She" is portrayed as having this non-typical role.
3 This non-typical role is directed towards the form of a seed.
4 Her arrival is completed, indicating that a transformation has taken place.
5 There is complete certainty that she arrived in this form.
`{tgt_language}`:

Figure 7: Comparison between the linearized graph version and the NLD version of translating the sentence "`She had come in the form of a seed`", using the `SR-Direct` prompt.

Model	deu	jpn	zho	por	nld	ces	ukr	hin	cat	lav	bel	hsb
BLEU (AMR)												
Aya-23-8B	0.118	0.931	0.908	**0.001****	**0.023***	0.580	0.886	**0.041***	0.080	**0.041***	0.057	**0.002****
Gemma-3-4B-IT	0.384	0.216	0.216	0.711	0.799	0.726	0.143	0.382	0.258	0.993	0.197	0.488
Gemma-3-12B-IT	0.160	0.653	0.429	0.199	0.572	**0.004****	0.393	0.576	0.365	0.180	0.998	0.099
Gemma-3-27B-IT	0.149	0.625	0.429	0.598	0.387	0.875	0.524	0.312	0.178	0.184	0.630	0.566
Llama-3.1-8B-Instruct	**0.000*****	**0.044***	**0.000*****	0.966	**0.023***	0.464	0.919	**0.000*****	0.600	0.241	0.272	**0.000*****
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	**0.000*****	0.320	**0.028***	0.404	0.662	0.521	0.842	**0.049***	0.280	0.527	0.812	**0.000*****
BLEU (UMR)												
Aya-23-8B	0.149	0.406	0.165	**0.007****	**0.001****	0.783	0.722	0.137	0.850	0.482	0.071	**0.014***
Gemma-3-4B-IT	0.290	0.841	0.068	0.106	0.282	0.459	0.197	0.776	0.332	0.369	0.098	**0.000*****
Gemma-3-12B-IT	**0.000*****	0.427	0.751	**0.042***	0.213	**0.000*****	0.819	0.545	0.086	0.101	0.750	0.771
Gemma-3-27B-IT	0.086	0.954	0.751	0.706	0.326	0.769	0.833	0.209	0.646	0.295	0.164	0.321
Llama-3.1-8B-Instruct	**0.002****	**0.023***	**0.000*****	0.488	**0.002****	0.482	0.787	**0.000*****	0.638	**0.040***	0.298	**0.000*****
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	**0.000*****	0.148	**0.008****	0.677	0.565	0.864	0.737	0.069	0.153	0.367	0.937	**0.000*****
COMET (AMR)												
Aya-23-8B	0.476	0.065	**0.000*****	0.329	**0.001****	0.989	**0.007****	0.162	0.756	0.247	**0.000*****	**0.000*****
Gemma-3-4B-IT	0.836	0.330	**0.000*****	0.607	0.822	0.349	0.459	0.865	0.354	0.079	0.963	**0.002****
Gemma-3-12B-IT	0.707	0.080	0.397	**0.023***	0.098	0.633	0.545	0.082	**0.011***	0.424	0.445	0.308
Gemma-3-27B-IT	0.090	**0.000*****	0.397	0.880	0.207	0.553	0.144	0.861	0.970	0.868	0.752	0.892
Llama-3.1-8B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.006****	**0.000*****	**0.001*****	**0.024***	**0.000*****	**0.005****	**0.000*****	**0.000*****	**0.000*****
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	0.085	**0.043***	**0.017***	0.084	**0.033***	0.876	0.147	0.543	**0.043***	0.821	0.846	**0.000*****
COMET (UMR)												
Aya-23-8B	0.541	0.245	**0.001****	**0.002****	**0.000*****	0.798	**0.007****	0.080	0.813	0.926	**0.000*****	**0.000*****
Gemma-3-4B-IT	0.179	0.112	**0.000*****	0.392	0.527	0.608	0.092	**0.016***	0.970	0.061	0.097	**0.004****
Gemma-3-12B-IT	0.087	0.818	0.142	0.123	0.418	0.109	**0.048***	0.138	0.098	0.982	0.901	0.838
Gemma-3-27B-IT	0.163	**0.009****	0.142	0.927	0.937	0.212	0.666	0.826	0.528	0.543	0.763	0.395
Llama-3.1-8B-Instruct	**0.044***	**0.000*****	**0.000*****	**0.004****	**0.000*****	**0.000*****	0.082	**0.000*****	**0.001****	**0.041***	**0.051**	**0.000*****
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.028***	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	0.287	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	0.761	0.721	**0.003****	**0.000*****	0.534	0.785	**0.037***	0.803	**0.041***	0.852	0.671	**0.033***
MetricX (AMR)												
Aya-23-8B	0.537	**0.001*****	0.771	0.477	**0.006****	**0.048***	**0.000*****	**0.013***	**0.037***	0.414	**0.000*****	**0.000*****
Gemma-3-4B-IT	0.677	0.196	0.567	0.697	0.991	0.417	0.302	**0.034***	0.557	0.483	0.281	**0.000*****
Gemma-3-12B-IT	**0.016***	0.090	0.538	0.217	0.991	0.269	**0.007****	0.220	0.394	0.752	0.368	0.330
Gemma-3-27B-IT	0.177	0.094	0.538	0.087	0.863	0.830	0.056	0.316	0.087	**0.023***	0.627	0.727
Llama-3.1-8B-Instruct	0.207	**0.000*****	**0.000*****	0.894	**0.007****	**0.019***	0.052	**0.000*****	0.162	**0.032***	**0.000*****	**0.002****
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	0.287	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	0.761	0.721	**0.003****	**0.000*****	0.534	0.785	**0.037***	0.803	**0.041***	0.852	0.671	**0.033***
MetricX (UMR)												
Aya-23-8B	0.121	**0.002****	0.515	0.119	**0.014***	0.092	**0.000*****	**0.001****	0.084	0.202	**0.000*****	**0.000*****
Gemma-3-4B-IT	0.201	0.062	0.750	0.583	0.537	0.438	0.371	0.175	0.840	0.313	0.913	**0.006****
Gemma-3-12B-IT	0.825	0.980	0.820	**0.039***	0.387	**0.020***	**0.000*****	**0.047***	0.692	0.680	0.309	0.807
Gemma-3-27B-IT	0.122	0.490	0.820	0.162	0.062	0.240	0.548	0.169	0.115	0.548	0.159	0.897
Llama-3.1-8B-Instruct	0.885	**0.000*****	**0.000*****	0.458	**0.002****	**0.031***	**0.024***	**0.000*****	**0.006****	**0.041***	**0.041***	**0.039***
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.028***	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	0.287	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	0.714	0.217	**0.031***	**0.000*****	0.346	0.682	**0.037***	0.644	0.740	0.996	0.295	**0.004****

Table 5: Two-tailed paired t-test p-values for BLEU, COMET, and MetricX scores, computed over sentence-level outputs, comparing zero-shot `SR-Direct` prompting (with linearized AMR or UMR graphs) against the zero-shot baseline. Bolded values indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Model	Shot	No SR		SR-Direct		SR-Grounded		SR-Assisted	
		Baseline	SENSE	Graph	NLD	Graph	NLD	Graph	NLD
Aya-23-8B	0 Shot	**0.1283**	0.1239 ↓	0.1165 ↓	0.1076 ↓	0.1134 ↓	0.1159 ↓	0.1132 ↓	0.1193 ↓
	1 Shot	**0.1722**	0.1721 ↓	0.1615 ↓	0.1563 ↓	0.1607 ↓	0.1431 ↓	0.1627 ↓	0.1494 ↓
	3 Shot	**0.1872**	0.1854 ↓	0.1770 ↓	0.1775 ↓	0.1712 ↓	0.1625 ↓	0.1715 ↓	0.1745 ↓
	5 Shot	0.1902	**0.1910** ↑	0.1805 ↓	0.1823 ↓	0.1721 ↓	0.1709 ↓	0.1724 ↓	0.1795 ↓
Gemma-4B	0 Shot	**0.1379**	0.1259 ↓	0.1367 ↓	0.1367 ↓	0.1197 ↓	0.1321 ↓	0.1376 ↓	0.1365 ↓
	1 Shot	**0.1876**	0.1871 ↓	0.1810 ↓	0.1800 ↓	0.1791 ↓	0.1734 ↓	0.1787 ↓	0.1785 ↓
	3 Shot	0.2032	**0.2091** ↑	0.1989 ↓	0.1979 ↓	0.1977 ↓	0.1922 ↓	0.1979 ↓	0.1939 ↓
	5 Shot	0.2091	**0.2135** ↑	0.2030 ↓	0.2024 ↓	0.2019 ↓	0.1979 ↓	0.2015 ↓	0.1986 ↓
Gemma-12B	0 Shot	**0.1602**	0.1535 ↓	0.1580 ↓	0.1579 ↓	0.1491 ↓	0.1478 ↓	0.1557 ↓	0.1568 ↓
	1 Shot	**0.2118**	0.2041 ↓	0.1986 ↓	0.1938 ↓	0.1976 ↓	0.1878 ↓	0.1985 ↓	0.1956 ↓
	3 Shot	**0.2267**	0.2243 ↓	0.2158 ↓	0.2134 ↓	0.2142 ↓	0.2101 ↓	0.2153 ↓	0.2128 ↓
	5 Shot	0.2297	**0.2311** ↑	0.2206 ↓	0.2196 ↓	0.2175 ↓	0.2176 ↓	0.2194 ↓	0.2190 ↓
Gemma-27B	0 Shot	**0.1731**	0.1678 ↓	0.1716 ↓	0.1722 ↓	0.1652 ↓	0.1649 ↓	0.1710 ↓	0.1717 ↓
	1 Shot	**0.2300**	0.2246 ↓	0.2180 ↓	0.2115 ↓	0.2194 ↓	0.2061 ↓	0.2194 ↓	0.2091 ↓
	3 Shot	**0.2481**	0.2434 ↓	0.2364 ↓	0.2337 ↓	0.2356 ↓	0.2285 ↓	0.2362 ↓	0.2296 ↓
	5 Shot	**0.2531**	0.2505 ↓	0.2421 ↓	0.2393 ↓	0.2425 ↓	0.2373 ↓	0.2422 ↓	0.2375 ↓
Llama-3.1-8B-Instruct	0 Shot	**0.1215**	0.1151 ↓	0.1149 ↓	0.1138 ↓	0.0780 ↓	0.0923 ↓	0.1151 ↓	0.1142 ↓
	1 Shot	**0.1809**	0.1739 ↓	0.1632 ↓	0.1588 ↓	0.1460 ↓	0.1334 ↓	0.1607 ↓	0.1629 ↓
	3 Shot	**0.1971**	0.1897 ↓	0.1875 ↓	0.1864 ↓	0.1696 ↓	0.1727 ↓	0.1810 ↓	0.1883 ↓
	5 Shot	**0.1984**	0.1924 ↓	0.1944 ↓	0.1950 ↓	0.1759 ↓	0.1857 ↓	0.1860 ↓	0.1963 ↓
Mistral-7B-Instruct	0 Shot	**0.0674**	0.0549 ↓	0.0365 ↓	0.0389 ↓	0.0138 ↓	0.0178 ↓	0.0290 ↓	0.0394 ↓
	1 Shot	**0.1273**	0.0574 ↓	0.1016 ↓	0.1011 ↓	0.0569 ↓	0.0547 ↓	0.0926 ↓	0.0860 ↓
	3 Shot	**0.1628**	0.0914 ↓	0.1488 ↓	0.1475 ↓	0.0981 ↓	0.0833 ↓	0.1363 ↓	0.1166 ↓
	5 Shot	**0.1708**	0.1216 ↓	0.1658 ↓	0.1658 ↓	0.1185 ↓	0.1122 ↓	0.1514 ↓	0.1408 ↓
Qwen2.5-7B-Instruct	0 Shot	0.1020	0.0985 ↓	**0.1032** ↑	0.1029 ↑	0.1005 ↓	0.0914 ↓	0.1004 ↓	0.0982 ↓
	1 Shot	0.1560	0.1466 ↓	0.1539 ↓	0.1510 ↓	**0.1570** ↑	0.1472 ↓	0.1408 ↓	0.1489 ↓
	3 Shot	0.1743	0.1728 ↓	0.1749 ↑	0.1741 ↓	**0.1759** ↑	0.1701 ↓	0.1594 ↓	0.1706 ↓
	5 Shot	0.1809	0.1801 ↓	0.1800 ↓	0.1785 ↓	**0.1818** ↑	0.1757 ↓	0.1681 ↓	0.1765 ↓

Table 6: BLEU scores across models and prompt strategies, averaged over 12 target languages and the two SRs. Green arrows (↑) indicate improvements over the baseline, red arrows (↓) indicate degradations.

Model	Shot	No SR		SR-Direct		SR-Grounded		SR-Assisted	
		Baseline	SENSE	Graph	NLD	Graph	NLD	Graph	NLD
Aya-23-8B	0 Shot	**0.7517**	0.7247 ↓	0.7391 ↓	0.7358 ↓	0.7278 ↓	0.7382 ↓	0.7327 ↓	0.7408 ↓
	1 Shot	**0.7642**	0.7544 ↓	0.7532 ↓	0.7512 ↓	0.7491 ↓	0.7429 ↓	0.7534 ↓	0.7505 ↓
	3 Shot	**0.7683**	0.7631 ↓	0.7628 ↓	0.7629 ↓	0.7616 ↓	0.7577 ↓	0.7623 ↓	0.7639 ↓
	5 Shot	**0.7699**	0.7662 ↓	0.7654 ↓	0.7660 ↓	0.7644 ↓	0.7626 ↓	0.7640 ↓	0.7668 ↓
Gemma-4B	0 Shot	0.7597	0.7300 ↓	**0.7643** ↑	0.7636 ↑	0.7363 ↓	0.7638 ↑	0.7561 ↓	0.7611 ↑
	1 Shot	0.7854	**0.7876** ↑	0.7820 ↓	0.7837 ↓	0.7791 ↓	0.7827 ↓	0.7820 ↓	0.7859 ↑
	3 Shot	0.7932	**0.7963** ↑	0.7907 ↓	0.7918 ↓	0.7898 ↓	0.7921 ↓	0.7904 ↓	0.7938 ↑
	5 Shot	0.7972	**0.8001** ↑	0.7944 ↓	0.7949 ↓	0.7932 ↓	0.7965 ↓	0.7941 ↓	0.7970 ↓
Gemma-12B	0 Shot	**0.7970**	0.7938 ↓	0.7963 ↓	**0.7970** –	0.7892 ↓	0.7913 ↓	0.7946 ↓	0.7944 ↓
	1 Shot	0.8119	**0.8127** ↑	0.8081 ↓	0.8080 ↓	0.8035 ↓	0.8057 ↓	0.8066 ↓	0.8092 ↓
	3 Shot	0.8184	**0.8191** ↑	0.8159 ↓	0.8160 ↓	0.8113 ↓	0.8145 ↓	0.8143 ↓	0.8154 ↓
	5 Shot	0.8203	**0.8221** ↑	0.8178 ↓	0.8177 ↓	0.8143 ↓	0.8169 ↓	0.8160 ↓	0.8176 ↓
Gemma-27B	0 Shot	0.8109	0.8081 ↓	0.8120 ↑	**0.8125** ↑	0.8080 ↓	0.8092 ↓	0.8106 ↓	0.8115 ↑
	1 Shot	**0.8253**	0.8247 ↓	0.8224 ↓	0.8219 ↓	0.8214 ↓	0.8213 ↓	0.8223 ↓	0.8222 ↓
	3 Shot	0.8313	**0.8326** ↑	0.8291 ↓	0.8288 ↓	0.8284 ↓	0.8287 ↓	0.8282 ↓	0.8287 ↓
	5 Shot	0.8334	**0.8357** ↑	0.8319 ↓	0.8318 ↓	0.8312 ↓	0.8323 ↓	0.8305 ↓	0.8325 ↓
Llama-3.1-8B-Instruct	0 Shot	**0.7542**	0.7452 ↓	0.7430 ↓	0.7436 ↓	0.6752 ↓	0.7122 ↓	0.7394 ↓	0.7398 ↓
	1 Shot	**0.7750**	0.7700 ↓	0.7574 ↓	0.7564 ↓	0.7287 ↓	0.7337 ↓	0.7530 ↓	0.7583 ↓
	3 Shot	**0.7846**	0.7791 ↓	0.7748 ↓	0.7758 ↓	0.7488 ↓	0.7623 ↓	0.7670 ↓	0.7758 ↓
	5 Shot	**0.7878**	0.7790 ↓	0.7801 ↓	0.7816 ↓	0.7561 ↓	0.7748 ↓	0.7724 ↓	0.7811 ↓
Mistral-7B-Instruct	0 Shot	**0.6146**	0.5831 ↓	0.5245 ↓	0.5494 ↓	0.3275 ↓	0.4465 ↓	0.4498 ↓	0.5380 ↓
	1 Shot	**0.6966**	0.5979 ↓	0.6712 ↓	0.6749 ↓	0.5774 ↓	0.5791 ↓	0.6539 ↓	0.6523 ↓
	3 Shot	**0.7322**	0.6682 ↓	0.7249 ↓	0.7272 ↓	0.6720 ↓	0.6715 ↓	0.7157 ↓	0.7073 ↓
	5 Shot	**0.7401**	0.7065 ↓	0.7371 ↓	0.7392 ↓	0.7013 ↓	0.7071 ↓	0.7299 ↓	0.7296 ↓
Qwen2.5-7B-Instruct	0 Shot	0.6973	0.6915 ↓	0.6975 ↑	**0.6980** ↑	0.6903 ↓	0.6894 ↓	0.6930 ↓	0.6950 ↓
	1 Shot	0.7270	0.7245 ↓	0.7270 –	**0.7280** ↑	0.7211 ↓	0.7256 ↓	0.7192 ↓	0.7279 ↑
	3 Shot	0.7405	0.7427 ↑	0.7419 ↑	**0.7433** ↑	0.7383 ↓	0.7412 ↑	0.7343 ↓	0.7432 ↑
	5 Shot	0.7478	0.7490 ↑	0.7482 ↑	0.7485 ↑	0.7451 ↓	0.7484 ↑	0.7417 ↓	**0.7491** ↑

Table 7: COMET scores across models and prompting strategies, averaged over 12 target languages and the two SRs. Green arrows (↑) indicate improvements over the baseline, red arrows (↓) indicate degradations.

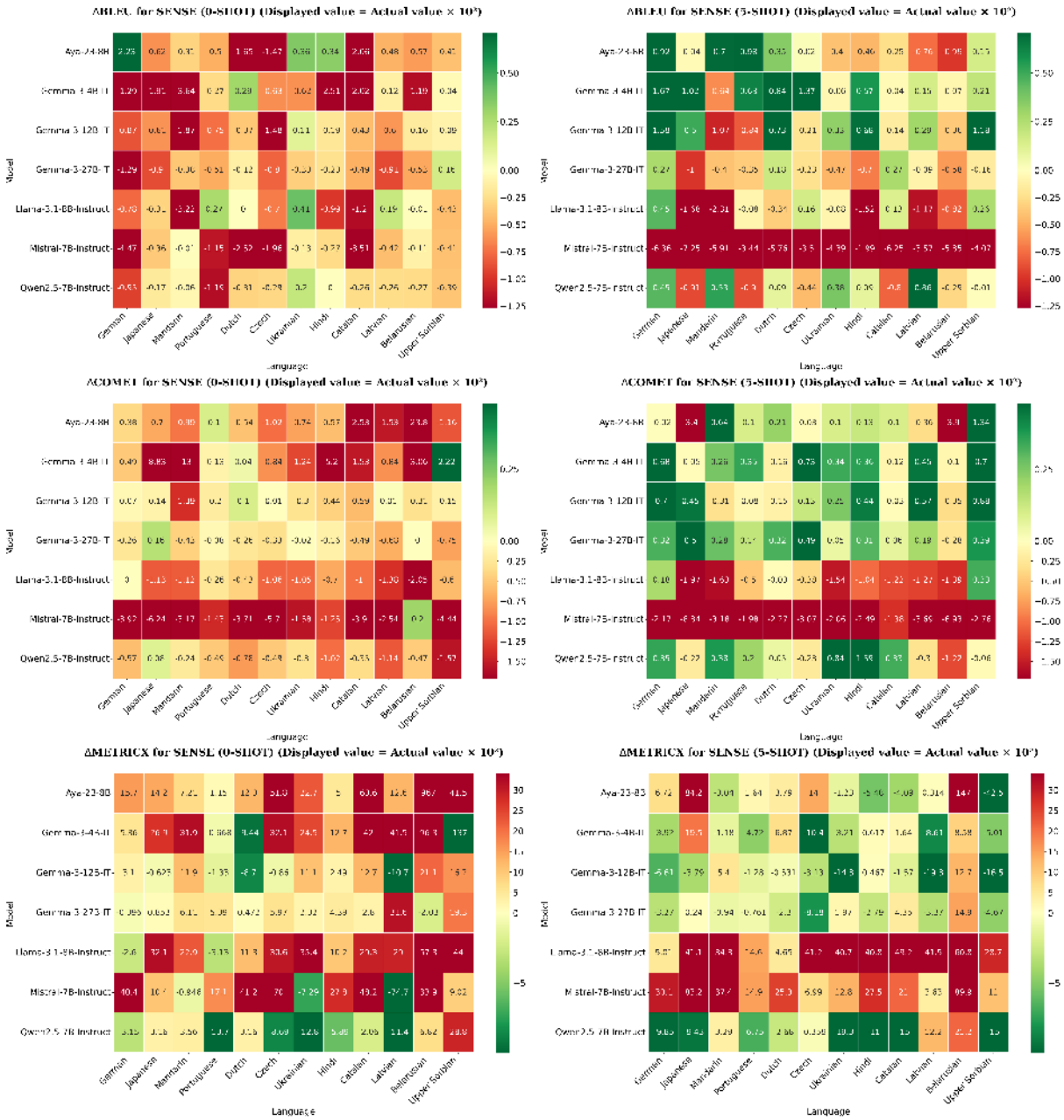


Figure 8: Heatmaps showing the impact of SENSE prompting, in the zero-shot (left) and five-shot (right) settings, relative to the baseline, reported as ΔBLEU, ΔCOMET, and ΔMetricX (top to bottom). For each metric, the color scale is clipped at the 25th and 75th percentiles. Cell values are also scaled by 10^2 for readability.

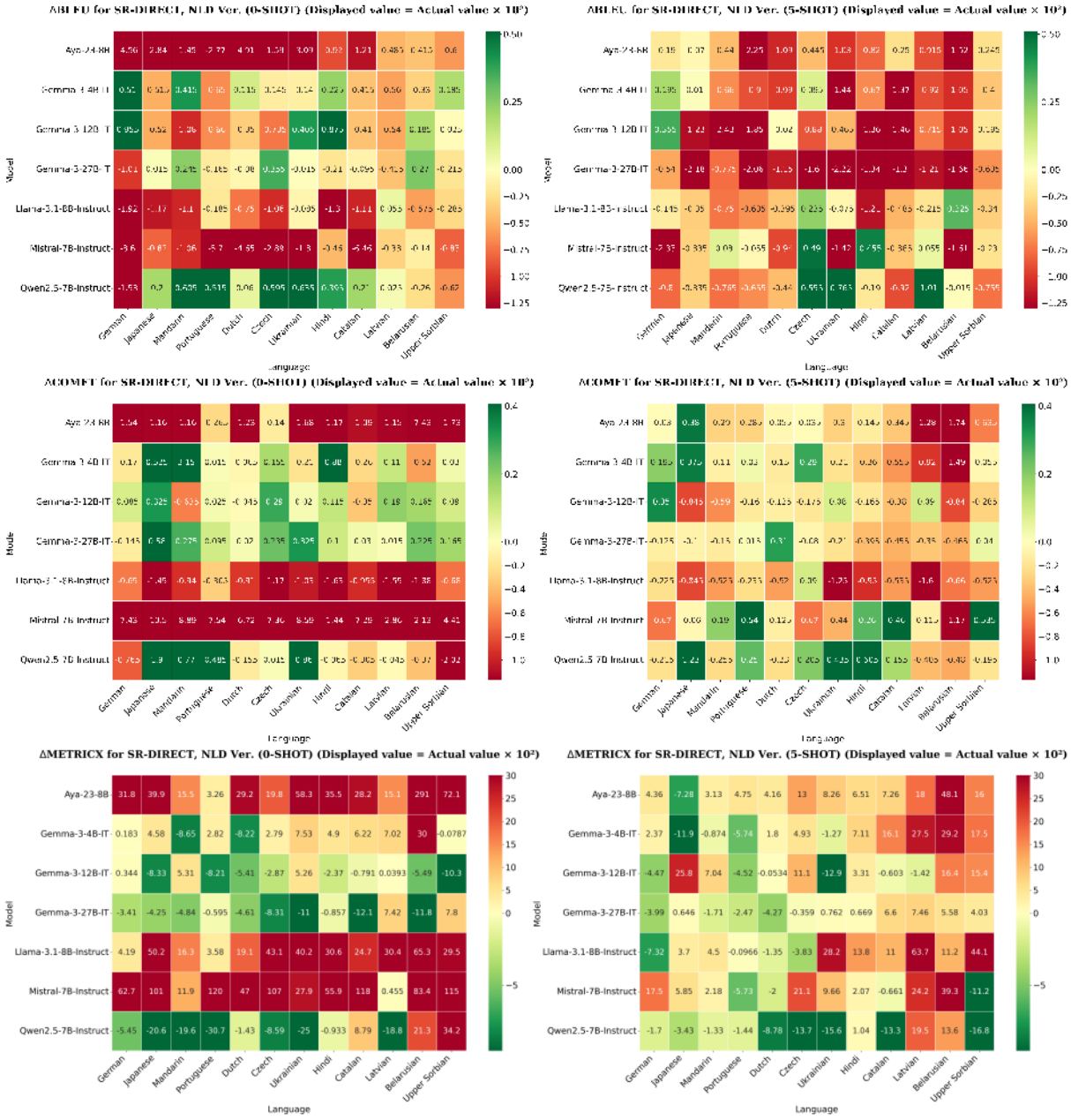


Figure 9: Heatmaps showing the impact of `SR-Direct` prompting when the SR is converted into NLD, in the zero-shot (left) and five-shot (right) settings, relative to the baseline, reported as ΔBLEU, ΔCOMET, and ΔMetricX (top to bottom). We compute these values by averaging the results of both AMR and UMR. For each metric, the color scale is clipped at the 25th and 75th percentiles. Cell values are also scaled by 10^2 for readability.

Model	deu	jpn	zho	por	nld	ces	ukr	hin	cat	lav	bel	hsb
BLEU (Zero-Shot)												
Aya-23-8B	0.118	0.931	0.908	**0.001****	**0.023***	0.580	0.886	**0.041***	0.080	**0.041***	0.057	**0.002****
Gemma-3-4B-IT	0.384	0.216	0.216	0.711	0.799	0.726	0.143	0.382	0.258	0.993	0.197	0.488
Gemma-3-12B-IT	0.160	0.653	0.429	0.199	0.572	**0.004****	0.393	0.576	0.365	0.180	0.998	0.099
Gemma-3-27B-IT	0.149	0.625	0.429	0.598	0.387	0.875	0.524	0.312	0.178	0.184	0.630	0.566
Llama-3.1-8B-Instruct	**0.000*****	**0.044***	**0.000*****	0.966	**0.023***	0.464	0.919	**0.000*****	0.600	0.241	0.272	**0.000*****
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	**0.000*****	0.320	**0.028***	0.404	0.662	0.521	0.842	**0.049***	0.280	0.527	0.812	**0.000*****
BLEU (Five-Shot)												
Aya-23-8B	**0.000*****	**0.000*****	**0.018***	0.273	**0.000*****	**0.000*****	0.756	0.558	**0.000*****	**0.008****	0.430	**0.000*****
Gemma-3-4B-IT	**0.011***	**0.000*****	**0.000*****	0.914	0.722	**0.026***	**0.004****	**0.000*****	**0.019***	0.107	**0.001*****	0.244
Gemma-3-12B-IT	0.388	0.084	**0.000*****	**0.004****	0.210	**0.000*****	0.997	0.226	0.332	**0.045***	0.869	0.075
Gemma-3-27B-IT	**0.001*****	0.073	**0.026***	0.059	0.547	0.214	0.151	0.294	0.337	0.079	0.253	0.139
Llama-3.1-8B-Instruct	0.100	0.249	**0.000*****	0.623	0.304	0.244	0.821	**0.001****	**0.030***	0.479	0.931	**0.012***
Mistral-7B-Instruct	**0.000*****	**0.000*****	0.171	**0.005****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	0.211	**0.000*****
Qwen2.5-7B-Instruct	**0.000*****	0.828	0.070	0.194	0.273	0.479	0.772	**0.032***	0.155	0.310	0.345	**0.000*****
COMET (Zero-Shot)												
Aya-23-8B	0.476	0.065	**0.000*****	0.329	**0.001****	0.989	**0.007****	0.162	0.756	0.247	**0.000*****	**0.000*****
Gemma-3-4B-IT	0.836	0.330	**0.000*****	0.607	0.822	0.349	0.459	0.865	0.354	0.079	0.963	**0.002****
Gemma-3-12B-IT	0.707	0.080	0.397	**0.023***	0.098	0.633	0.545	0.082	**0.011***	0.424	0.445	0.308
Gemma-3-27B-IT	0.090	**0.000*****	0.397	0.880	0.207	0.553	0.144	0.861	0.970	0.868	0.752	0.892
Llama-3.1-8B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.006****	**0.000*****	**0.001*****	**0.024***	**0.000*****	**0.005****	**0.000*****	**0.000*****	**0.000*****
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	0.085	**0.043***	**0.017***	0.084	**0.033***	0.876	0.147	0.543	**0.043***	0.821	0.846	**0.000*****
COMET (Five-Shot)												
Aya-23-8B	**0.010****	**0.000*****	**0.000*****	**0.000*****	0.187	0.940	0.538	0.941	0.174	0.926	**0.000*****	**0.002****
Gemma-3-4B-IT	0.248	**0.048***	0.628	0.209	0.185	0.296	0.638	0.914	0.850	0.556	0.537	0.707
Gemma-3-12B-IT	**0.003****	0.569	**0.007****	0.676	0.885	0.644	**0.002****	0.907	0.773	0.082	0.183	0.187
Gemma-3-27B-IT	**0.027***	0.946	0.545	0.790	0.380	**0.085**	0.555	0.396	0.264	0.658	**0.015***	0.638
Llama-3.1-8B-Instruct	0.313	**0.000*****	**0.000*****	**0.017***	0.395	**0.003****	**0.000*****	**0.000*****	**0.000*****	**0.012***	**0.000*****	**0.030***
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.043***	**0.002****	0.567	0.173	**0.035***	**0.029***	0.775	**0.000*****	0.402
Qwen2.5-7B-Instruct	**0.050***	0.322	0.255	0.185	0.716	0.980	0.118	0.394	0.208	0.407	0.188	0.243
MetricX (Zero-Shot)												
Aya-23-8B	**0.045***	**0.000*****	0.170	0.604	0.363	0.058	0.805	0.319	0.681	0.980	**0.000*****	**0.002****
Gemma-3-4B-IT	0.158	**0.000*****	**0.000*****	0.862	0.088	**0.007****	**0.001*****	**0.044***	**0.000*****	**0.012***	**0.000*****	**0.000*****
Gemma-3-12B-IT	**0.268**	0.868	**0.000*****	0.665	**0.004****	0.906	**0.016***	0.584	**0.047***	0.343	**0.044***	0.251
Gemma-3-27B-IT	**0.793**	0.790	**0.000*****	**0.021***	0.832	0.268	0.530	0.230	0.545	**0.000*****	0.764	0.104
Llama-3.1-8B-Instruct	0.533	**0.000*****	**0.000*****	0.488	**0.041***	**0.010****	**0.000*****	0.234	**0.002****	0.077	**0.001*****	**0.003****
Mistral-7B-Instruct	**0.000*****	0.209	0.856	**0.001****	**0.000*****	**0.000*****	0.461	**0.029***	**0.000*****	**0.000*****	**0.011***	0.505
Qwen2.5-7B-Instruct	0.602	0.762	0.393	**0.015***	0.687	0.575	0.393	0.690	0.869	0.437	0.677	0.073
MetricX (Five-Shot)												
Aya-23-8B	**0.010****	**0.000*****	**0.003****	0.717	**0.001****	**0.000*****	**0.000*****	0.408	**0.000*****	0.201	**0.000*****	**0.013***
Gemma-3-4B-IT	0.248	**0.048***	0.628	0.209	0.185	0.296	0.638	0.914	0.850	0.556	0.537	0.707
Gemma-3-12B-IT	**0.003****	0.569	**0.007****	0.676	0.885	0.644	**0.002****	0.907	0.773	0.082	0.183	0.187
Gemma-3-27B-IT	**0.027***	0.946	0.545	0.790	0.380	**0.085**	0.555	0.396	0.264	0.658	**0.015***	0.638
Llama-3.1-8B-Instruct	0.313	**0.000*****	**0.000*****	**0.017***	0.395	**0.003****	**0.000*****	**0.000*****	**0.000*****	**0.012***	**0.000*****	**0.030***
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.043***	**0.002****	0.567	0.173	**0.035***	**0.029***	0.775	**0.000*****	0.402
Qwen2.5-7B-Instruct	**0.050***	0.322	0.255	0.185	0.716	0.980	0.118	0.394	0.208	0.407	0.188	0.243

Table 8: Two-tailed paired t-test p-values for BLEU, COMET, and MetricX scores, computed over sentence-level outputs, comparing zero-shot SENSE prompting against the zero-shot baseline, and similarly for five-shot. Bolded values indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Model	deu	jpn	zho	por	nld	ces	ukr	hin	cat	lav	bel	hsb
BLEU (Zero-Shot)												
Aya-23-8B	0.053	0.041*	0.831	**0.000*** **	**0.000*** **	0.089	0.807	0.566	0.279	**0.000*** **	0.035*	**0.000*** **
Gemma-3-4B-IT	0.570	0.517	0.025*	0.026*	0.690	0.589	0.119	0.653	0.250	0.737	0.033*	0.615
Gemma-3-12B-IT	0.007**	0.946	**0.000*** **	0.224	0.409	0.017*	0.281	0.256	0.189	0.106	0.751	0.301
Gemma-3-27B-IT	0.043*	0.742	0.384	0.691	0.823	0.273	0.910	0.245	0.842	0.443	0.450	0.504
Llama-3.1-8B-Instruct	**0.000*** **	0.009**	0.002**	0.874	0.039*	0.111	0.733	**0.000*** **	0.023*	0.244	0.023*	0.001***
Mistral-7B-Instruct	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **
Qwen2.5-7B-Instruct	**0.000*** **	0.180	0.098	0.991	0.858	0.211	0.966	0.145	0.356	0.838	0.037*	**0.000*** **
BLEU (Five-Shot)												
Aya-23-8B	0.089	0.606	0.681	**0.000*** **	0.027*	0.521	0.162	0.063	0.860	0.253	0.005**	0.190
Gemma-3-4B-IT	0.981	0.782	0.007**	0.539	0.157	0.584	0.003**	0.124	0.098	0.015*	0.027*	0.565
Gemma-3-12B-IT	0.002**	0.002**	**0.000*** **	**0.000*** **	0.747	0.715	0.295	0.384	0.028*	0.206	0.127	0.985
Gemma-3-27B-IT	0.066	**0.000*** **	0.018*	**0.000*** **	0.108	**0.001*** **	**0.000*** **	0.147	**0.000*** **	**0.001** **	**0.001** **	0.347
Llama-3.1-8B-Instruct	0.053	0.002**	0.019*	0.119	0.310	0.178	0.010**	0.578	0.050	0.221	0.381	0.378
Mistral-7B-Instruct	0.968	0.850	**0.000*** **	0.613	0.545	0.002**	0.077	0.027*	0.975	0.293	0.163	0.526
Qwen2.5-7B-Instruct	0.285	0.108	0.056	0.816	0.111	0.987	0.241	0.532	0.776	0.598	0.397	0.248
COMET (Zero-Shot)												
Aya-23-8B	0.169	0.005**	**0.000*** **	**0.000*** **	**0.000*** **	0.294	**0.000*** **	0.307	0.006**	**0.000*** **	**0.000*** **	0.001***
Gemma-3-4B-IT	0.453	0.819	**0.000*** **	0.080	0.285	0.456	0.067	0.195	0.131	0.638	0.025*	0.057
Gemma-3-12B-IT	0.881	0.299	**0.000*** **	0.350	0.121	0.283	0.960	0.391	0.182	0.556	0.543	0.284
Gemma-3-27B-IT	0.713	0.232	0.443	0.730	0.199	0.087	0.608	0.689	0.143	0.743	0.380	0.825
Llama-3.1-8B-Instruct	**0.000*** **	**0.000*** **	**0.000*** **	0.123	**0.001** **	**0.000*** **	0.010*	**0.000*** **	0.021*	0.056	**0.000*** **	0.001***
Mistral-7B-Instruct	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **	**0.000*** **
Qwen2.5-7B-Instruct	**0.000*** **	0.011*	0.484	**0.000*** **	0.033*	0.471	0.303	0.708	0.001**	0.743	0.301	**0.000*** **
COMET (Five-Shot)												
Aya-23-8B	0.362	0.935	0.419	0.002**	0.043*	0.462	0.003**	0.066	0.362	0.001***	0.001***	0.910
Gemma-3-4B-IT	0.615	0.292	0.223	0.683	0.088	0.821	**0.000*** **	0.050*	0.274	**0.000*** **	**0.000*** **	0.664
Gemma-3-12B-IT	0.004**	**0.001*** **	**0.000*** **	**0.000*** **	0.076	0.024*	0.007**	0.204	0.005**	0.020*	**0.000*** **	0.069
Gemma-3-27B-IT	0.034*	0.006**	0.003**	**0.000*** **	0.323	0.032*	**0.000*** **	0.112	0.001**	**0.000*** **	**0.000*** **	0.053
Llama-3.1-8B-Instruct	0.002**	**0.000*** **	0.019*	0.001***	0.014*	0.481	0.007**	0.363	0.133	0.004**	0.814	0.628
Mistral-7B-Instruct	0.040*	0.466	0.076	0.005**	0.518	0.673	0.214	0.390	0.155	0.728	0.195	0.461
Qwen2.5-7B-Instruct	0.125	0.526	0.052	0.001**	0.885	0.556	0.332	0.220	0.753	0.164	0.735	0.001**
MetricX (Zero-Shot)												
Aya-23-8B	**0.001*** **	**0.000*** **	**0.003** **	0.830	**0.001*** **	0.029*	**0.000*** **	**0.001*** **	0.035*	0.310	**0.000*** **	**0.000*** **
Gemma-3-4B-IT	0.599	0.990	0.010**	0.159	0.311	0.768	0.443	0.291	0.437	0.396	0.198	0.752
Gemma-3-12B-IT	0.996	0.073	0.003**	0.005**	0.076	0.797	0.166	0.633	0.953	0.933	0.257	0.716
Gemma-3-27B-IT	0.033*	0.056	**0.000*** **	0.468	0.088	0.258	0.018*	0.562	0.005**	0.188	0.054	0.410
Llama-3.1-8B-Instruct	0.382	**0.000*** **	**0.000*** **	0.519	0.018*	0.028*	0.001**	0.006**	0.005**	0.233	0.007**	0.158
Mistral-7B-Instruct	**0.000*** **	**0.000*** **	0.234	**0.000*** **	**0.000*** **	**0.000*** **	0.918	**0.001*** **	**0.000*** **	0.814	**0.000*** **	**0.000*** **
Qwen2.5-7B-Instruct	0.284	0.020*	**0.000*** **	**0.000*** **	0.509	0.535	0.292	0.977	0.971	0.198	0.339	0.011*
MetricX (Five-Shot)												
Aya-23-8B	0.145	0.142	0.353	0.844	0.195	0.270	0.046*	0.311	0.993	0.210	0.073	0.181
Gemma-3-4B-IT	0.463	0.177	0.565	0.029*	0.761	0.762	0.985	0.070	0.052	0.051	0.109	0.491
Gemma-3-12B-IT	0.033*	**0.001*** **	**0.000*** **	0.175	0.653	0.288	0.011*	0.506	0.874	0.887	0.049*	0.768
Gemma-3-27B-IT	**0.003** **	0.565	0.396	0.448	0.269	0.246	0.590	0.826	0.199	0.121	0.117	0.710
Llama-3.1-8B-Instruct	0.456	0.293	0.507	0.816	0.935	0.933	0.005**	0.300	0.381	**0.000*** **	0.588	**0.002** **
Mistral-7B-Instruct	**0.002** **	0.264	0.110	0.471	0.546	0.116	0.417	0.943	0.699	**0.008** **	**0.017* **	0.347
Qwen2.5-7B-Instruct	0.911	0.765	0.824	0.843	0.515	0.532	0.300	0.849	0.242	0.187	0.460	0.260

Table 9: Two-tailed paired t-test p-values for BLEU, COMET, and MetricX scores, computed over sentence-level outputs, comparing zero-shot `SR-Direct` prompting (with AMR converted to NLD) against the zero-shot baseline, and similarly for five-shot. Bolded values indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Model	deu	jpn	zho	por	nld	ces	ukr	hin	cat	lav	bel	hsb
BLEU (Zero-Shot)												
Aya-23-8B	0.789	**0.000*****	0.069	**0.000*****	**0.000*****	**0.000*****	**0.001****	0.731	**0.003****	**0.000*****	**0.006****	**0.000*****
Gemma-3-4B-IT	0.161	0.901	0.077	0.062	0.937	**0.017***	0.437	0.100	0.206	0.432	0.126	0.386
Gemma-3-12B-IT	**0.001****	0.585	**0.013***	0.060	0.105	**0.034***	0.980	0.219	0.176	0.353	0.612	**0.005****
Gemma-3-27B-IT	0.162	0.313	0.816	0.640	0.667	0.609	0.966	0.623	0.248	0.334	0.344	0.568
Llama-3.1-8B-Instruct	**0.000*****	**0.005****	**0.000*****	0.938	**0.000*****	0.062	0.270	**0.001*****	**0.010***	0.669	**0.005****	**0.028***
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.003****	**0.000*****
Qwen2.5-7B-Instruct	**0.000*****	0.261	0.112	0.977	0.197	0.515	0.394	**0.049***	0.226	0.294	0.072	**0.000*****
BLEU (Five-Shot)												
Aya-23-8B	0.064	0.333	0.788	**0.000*****	**0.011***	0.511	0.108	0.991	0.787	0.144	0.149	0.287
Gemma-3-4B-IT	0.155	0.841	**0.015***	0.053	0.108	0.971	**0.012***	0.458	**0.001*****	**0.003****	**0.007****	0.652
Gemma-3-12B-IT	**0.008****	**0.003****	**0.000*****	**0.000*****	0.526	**0.026***	0.845	0.081	**0.017***	0.171	**0.038***	0.734
Gemma-3-27B-IT	0.565	**0.002****	**0.002****	**0.000*****	0.073	**0.000*****	**0.000*****	**0.007****	**0.001*****	0.107	**0.004****	0.873
Llama-3.1-8B-Instruct	0.578	**0.001****	**0.050***	**0.039***	0.466	0.289	0.502	0.155	**0.045***	0.853	0.883	0.804
Mistral-7B-Instruct	0.935	0.840	**0.000*****	0.613	0.422	**0.003****	0.585	**0.841**	0.625	0.257	0.151	0.526
Qwen2.5-7B-Instruct	0.106	0.059	**0.014***	0.816	0.063	0.657	**0.024***	0.316	0.529	0.347	0.397	0.248
COMET (Zero-Shot)												
Aya-23-8B	0.000***	0.000***	0.000***	0.000***	0.000***	**0.013***	0.000***	0.123	0.000***	0.000***	0.000***	**0.001****
Gemma-3-4B-IT	0.735	0.147	**0.000*****	0.056	0.382	0.887	0.710	**0.002****	0.105	0.096	**0.031***	0.249
Gemma-3-12B-IT	0.588	0.875	**0.014***	0.443	**0.016***	0.353	0.248	0.406	0.575	0.823	0.812	0.443
Gemma-3-27B-IT	0.503	0.310	0.508	0.411	0.957	**0.014***	0.804	0.364	0.363	0.427	0.329	0.241
Llama-3.1-8B-Instruct	**0.000*****	**0.000*****	**0.000*****	0.337	**0.010****	**0.000*****	**0.003****	**0.000*****	**0.005****	**0.021***	**0.000*****	**0.000*****
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	**0.000*****	**0.011***	0.130	**0.000*****	**0.033***	0.924	**0.001*****	0.428	0.386	0.576	0.459	**0.000*****
COMET (Five-Shot)												
Aya-23-8B	0.294	0.799	0.881	**0.015***	**0.035***	0.941	**0.050***	0.507	0.323	0.406	**0.000*****	0.760
Gemma-3-4B-IT	0.077	0.682	0.150	**0.043***	**0.015***	0.210	**0.001*****	0.431	0.058	**0.000*****	**0.000*****	0.178
Gemma-3-12B-IT	**0.015***	**0.001****	**0.000*****	**0.000*****	0.133	**0.001****	**0.011***	**0.020***	**0.005****	**0.005****	**0.000*****	**0.002****
Gemma-3-27B-IT	**0.182**	**0.001*****	**0.003****	**0.000*****	0.065	**0.011***	**0.000*****	**0.028***	**0.003****	**0.020***	**0.000*****	0.617
Llama-3.1-8B-Instruct	**0.101**	**0.000*****	**0.019***	**0.001****	**0.014***	0.481	**0.007****	0.363	0.133	**0.004****	0.814	0.628
Mistral-7B-Instruct	**0.040***	0.466	0.076	**0.005****	0.518	0.673	0.214	0.390	0.155	0.728	0.195	0.461
Qwen2.5-7B-Instruct	**0.125**	0.526	0.052	**0.001****	0.885	0.556	0.332	0.220	0.753	0.164	0.735	**0.001****
MetricX (Zero-Shot)												
Aya-23-8B	**0.001*****	**0.000*****	**0.003****	0.122	**0.001*****	**0.007****	**0.000*****	**0.000*****	**0.003****	0.061	**0.000*****	**0.000*****
Gemma-3-4B-IT	0.536	**0.049***	**0.000*****	0.949	0.069	0.839	0.254	0.585	0.539	0.986	**0.021***	0.753
Gemma-3-12B-IT	0.810	**0.022***	0.156	**0.004****	0.075	0.337	0.587	0.517	0.870	0.937	0.929	0.286
Gemma-3-27B-IT	0.081	0.566	0.088	0.211	**0.031***	**0.031***	**0.003****	0.899	**0.026***	0.599	0.159	0.590
Llama-3.1-8B-Instruct	0.383	**0.000*****	**0.000*****	0.497	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.011***	**0.020***	**0.000*****	**0.011***
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.031***	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	0.871	**0.000*****	**0.000*****
Qwen2.5-7B-Instruct	0.484	0.104	**0.000*****	**0.000*****	0.776	0.629	**0.022***	0.928	0.163	0.271	0.137	0.110
MetricX (Five-Shot)												
Aya-23-8B	0.186	0.433	0.150	**0.021***	0.565	**0.022***	0.348	0.224	0.203	0.184	**0.000*****	0.288
Gemma-3-4B-IT	**0.561**	0.166	0.883	0.462	0.660	0.546	0.714	0.800	0.107	0.091	**0.018***	0.054
Gemma-3-12B-IT	0.202	**0.001*****	**0.001*****	0.218	0.679	**0.050***	**0.002****	0.347	0.716	0.684	0.194	**0.045***
Gemma-3-27B-IT	**0.023***	0.377	0.215	0.346	0.124	0.298	0.892	0.541	0.118	0.767	0.864	0.676
Llama-3.1-8B-Instruct	**0.013***	0.886	**0.031***	0.846	0.691	0.600	**0.019***	**0.027***	0.062	**0.000*****	0.385	**0.002****
Mistral-7B-Instruct	**0.000*****	0.634	0.628	0.189	0.988	0.051	0.149	0.803	0.569	0.171	**0.004****	0.352
Qwen2.5-7B-Instruct	0.606	0.669	0.255	0.736	0.086	0.200	0.151	0.969	0.326	0.210	0.343	0.260

Table 10: Two-tailed paired t-test p-values for BLEU, COMET, and MetricX scores, computed over sentence-level outputs, comparing zero-shot `SR-Direct` prompting (with UMR converted to NLD) against the zero-shot baseline, and similarly for five-shot. Bold values indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

Model	deu	jpn	zho	por	nld	ces	ukr	hin	cat	lav	bel	hsb
BLEU (H_1: AMR significantly better than UMR)												
Aya-23-8B	0.440	0.115	0.973	0.755	0.071	0.667	0.247	0.295	0.966	0.891	0.587	0.833
Gemma-3-4B-IT	0.584	0.957	0.742	0.075	0.122	0.341	0.560	0.941	0.591	0.155	0.316	1.000
Gemma-3-12B-IT	0.992	0.315	0.280	0.242	0.214	0.124	0.122	0.115	0.251	0.291	0.367	0.936
Gemma-3-27B-IT	0.353	0.752	0.280	0.423	0.986	0.561	0.663	0.428	0.849	0.391	0.876	0.323
Llama-3.1-8B-Instruct	0.983	0.464	0.530	0.243	0.240	0.517	0.556	0.669	0.524	0.153	0.490	0.947
Mistral-7B-Instruct	0.993	1.000	1.000	1.000	0.970	1.000	1.000	0.999	1.000	1.000	1.000	0.996
Qwen2.5-7B-Instruct	**0.007****	0.646	0.652	0.958	0.912	0.130	0.232	0.589	0.295	0.646	0.626	0.564
BLEU (H_1: UMR significantly better than AMR)												
Aya-23-8B	0.560	0.885	**0.027***	0.245	0.929	0.333	0.753	0.705	**0.034***	0.109	0.413	0.167
Gemma-3-4B-IT	0.416	**0.043***	0.258	0.925	0.878	0.659	0.440	0.059	0.409	0.845	0.684	**0.000*****
Gemma-3-12B-IT	**0.008****	0.685	0.720	0.758	0.786	0.876	0.878	0.885	0.749	0.709	0.633	0.064
Gemma-3-27B-IT	0.647	0.248	0.720	0.577	**0.014***	0.439	0.337	0.572	0.151	0.609	0.124	0.677
Llama-3.1-8B-Instruct	**0.017***	0.536	0.470	0.757	0.760	0.483	0.444	0.331	0.476	0.847	0.510	0.053
Mistral-7B-Instruct	**0.007****	**0.000*****	**0.000*****	**0.000*****	**0.030***	**0.000*****	**0.000*****	**0.001*****	**0.000*****	**0.000*****	**0.000*****	**0.004****
Qwen2.5-7B-Instruct	0.993	0.354	0.348	**0.042***	0.088	0.870	0.768	0.411	0.705	0.354	0.374	0.436
COMET (H_1: AMR significantly better than UMR)												
Aya-23-8B	0.593	0.842	0.686	0.992	0.277	0.632	0.469	0.286	0.535	0.904	**0.000*****	0.783
Gemma-3-4B-IT	**0.048***	0.789	1.000	0.661	0.829	0.072	0.119	0.998	0.182	0.528	0.968	0.407
Gemma-3-12B-IT	0.982	**0.019***	0.791	0.702	0.852	0.115	**0.045***	0.692	0.847	0.790	0.771	0.806
Gemma-3-27B-IT	0.646	0.149	0.791	0.532	0.945	0.757	0.099	0.515	0.220	0.833	0.525	0.856
Llama-3.1-8B-Instruct	0.990	0.928	0.777	0.325	0.346	0.436	0.704	0.594	0.371	0.499	0.992	0.625
Mistral-7B-Instruct	1.000	1.000	0.999	1.000	0.991	1.000	1.000	0.999	1.000	1.000	1.000	0.997
Qwen2.5-7B-Instruct	0.465	0.482	0.566	0.864	0.679	0.076	0.897	0.122	0.620	0.347	0.583	0.759
COMET (H_1: UMR significantly better than AMR)												
Aya-23-8B	0.407	0.158	0.314	**0.008****	0.723	0.368	0.531	0.714	0.465	0.096	1.000	0.217
Gemma-3-4B-IT	0.952	0.211	**0.000*****	0.339	0.171	0.928	0.881	**0.002****	0.818	0.472	**0.032***	0.593
Gemma-3-12B-IT	**0.018***	0.981	0.209	0.298	0.148	0.885	0.955	0.308	0.153	0.210	0.229	0.194
Gemma-3-27B-IT	0.354	0.851	0.209	0.468	0.055	0.243	0.901	0.485	0.780	0.167	0.475	0.144
Llama-3.1-8B-Instruct	**0.010****	0.072	0.223	0.675	0.654	0.564	0.296	0.406	0.629	0.501	**0.008****	0.375
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.001****	**0.000*****	**0.009****	**0.000*****	**0.000*****	**0.001*****	**0.000*****	**0.000*****	**0.000*****	**0.003****
Qwen2.5-7B-Instruct	0.535	0.518	0.434	0.136	0.321	0.924	0.103	0.878	0.380	0.653	0.417	0.241
MetricX (H_1: AMR significantly better than UMR)												
Aya-23-8B	0.222	0.721	0.330	0.874	0.676	0.703	0.814	0.134	0.687	0.723	**0.000*****	0.936
Gemma-3-4B-IT	0.160	0.253	0.387	0.861	0.772	0.498	0.584	0.691	0.198	0.622	0.895	0.175
Gemma-3-12B-IT	0.997	**0.035***	0.701	0.778	0.809	0.070	0.061	0.270	0.681	0.226	0.429	0.881
Gemma-3-27B-IT	0.506	0.143	0.701	0.680	0.959	0.845	0.062	0.361	0.408	0.961	0.858	0.589
Llama-3.1-8B-Instruct	0.861	0.930	0.450	0.256	0.288	0.596	0.441	0.443	0.084	0.529	0.966	0.855
Mistral-7B-Instruct	1.000	1.000	1.000	1.000	0.949	1.000	1.000	1.000	1.000	1.000	1.000	0.996
Qwen2.5-7B-Instruct	0.771	0.962	0.207	0.979	0.347	0.222	0.514	0.415	0.962	0.424	0.738	0.231
MetricX (H_1: UMR significantly better than AMR)												
Aya-23-8B	0.778	0.279	0.670	0.126	0.324	0.297	0.186	0.866	0.313	0.277	1.000	0.064
Gemma-3-4B-IT	0.840	0.747	0.613	0.139	0.228	0.502	0.416	0.309	0.802	0.378	0.105	0.825
Gemma-3-12B-IT	**0.003****	0.965	0.299	0.222	0.191	0.930	0.939	0.730	0.319	0.774	0.571	0.119
Gemma-3-27B-IT	0.494	0.857	0.299	0.320	**0.041***	0.155	0.938	0.639	0.592	**0.039***	0.142	0.411
Llama-3.1-8B-Instruct	0.139	0.070	0.550	0.744	0.712	0.404	0.559	0.557	0.916	0.471	**0.034***	0.145
Mistral-7B-Instruct	**0.000*****	**0.000*****	**0.000*****	**0.000*****	0.051	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.000*****	**0.004****
Qwen2.5-7B-Instruct	0.229	**0.038***	0.793	**0.021***	0.653	0.778	0.486	0.585	**0.038***	0.576	0.262	0.769

Table 11: One-tailed paired t-test p-values for BLEU, COMET, and MetricX scores in the case of zero-shot `SR-Direct` prompting, comparing translations produced using linearized AMR graphs against translations produced using linearized UMR graphs (and vice versa), for the same model and language. For each metric, the top block tests whether AMR scores significantly betters than UMR; the bottom block tests the reverse. Bold values indicate statistical significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

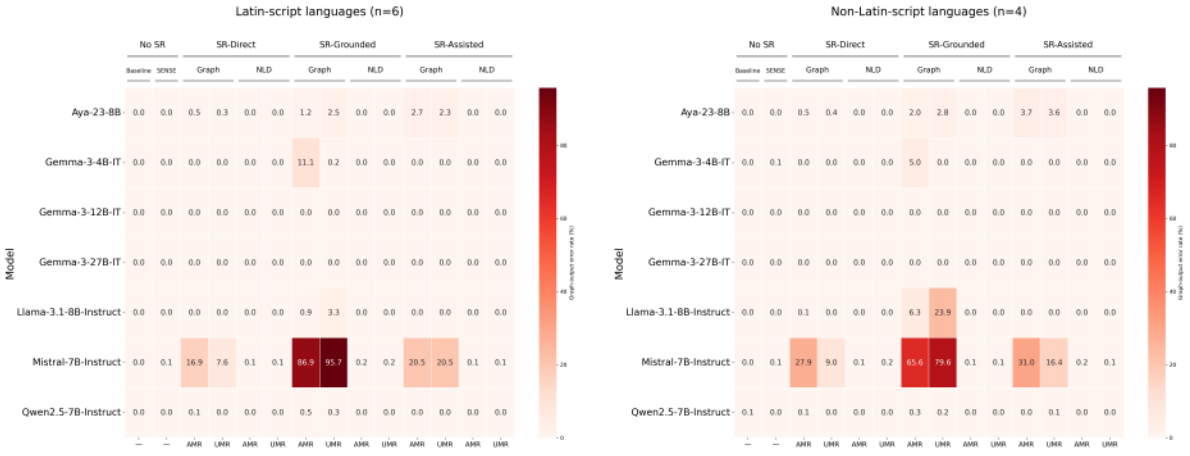


Figure 10: Graph-output error rate at 0-shot.

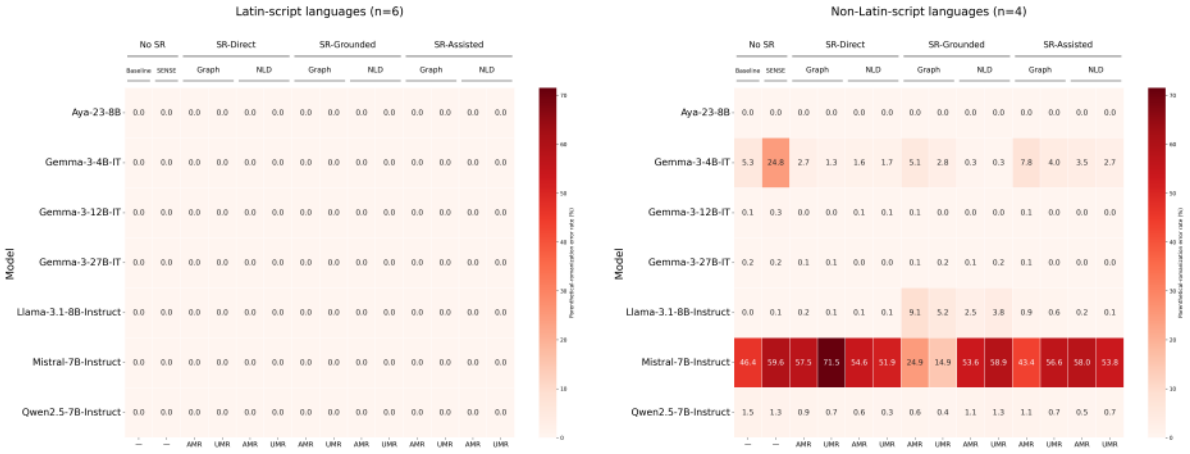


Figure 11: Parenthetical romanization error rate at 0-shot.

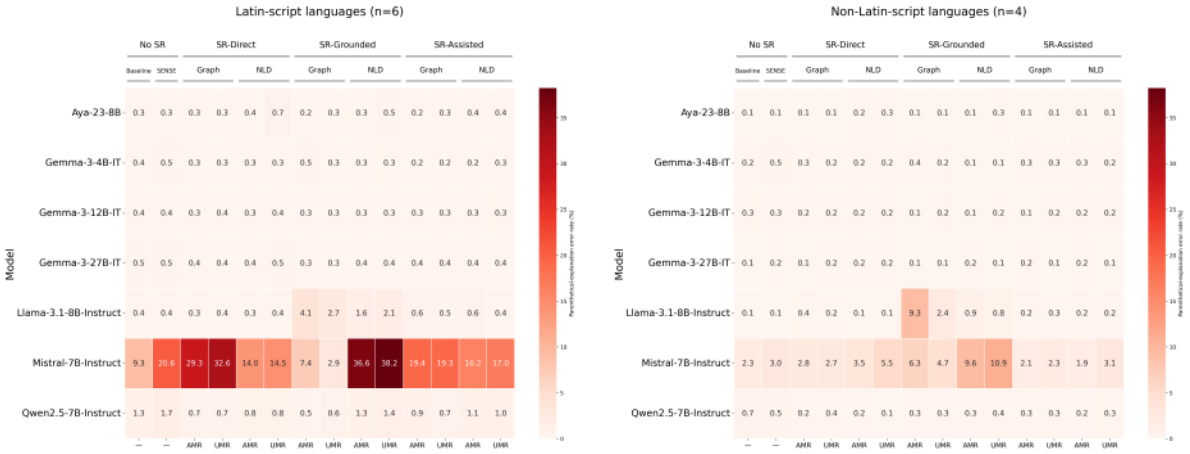


Figure 12: Parenthetical explanation error rate at 0-shot.

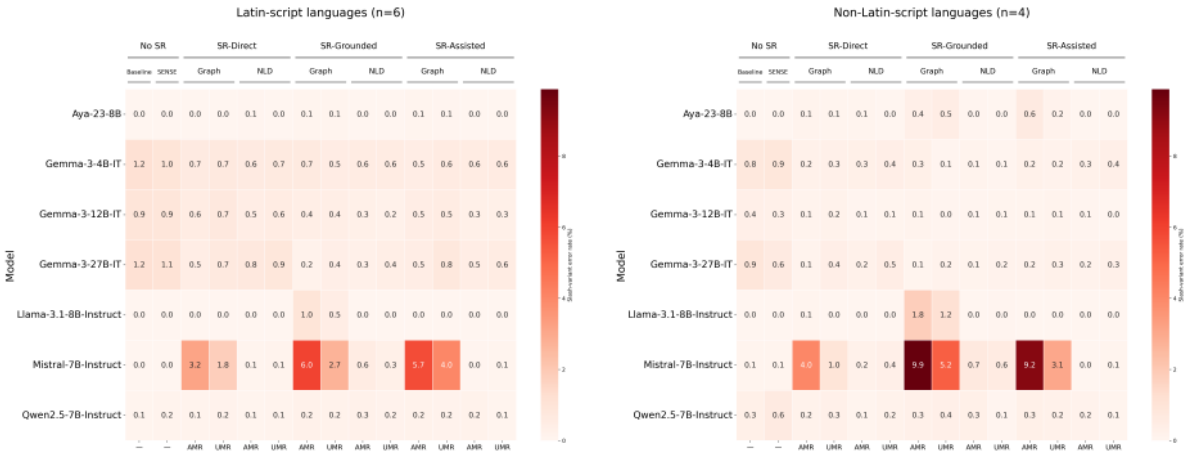


Figure 13: Slash-listed alternatives error rate at 0-shot.

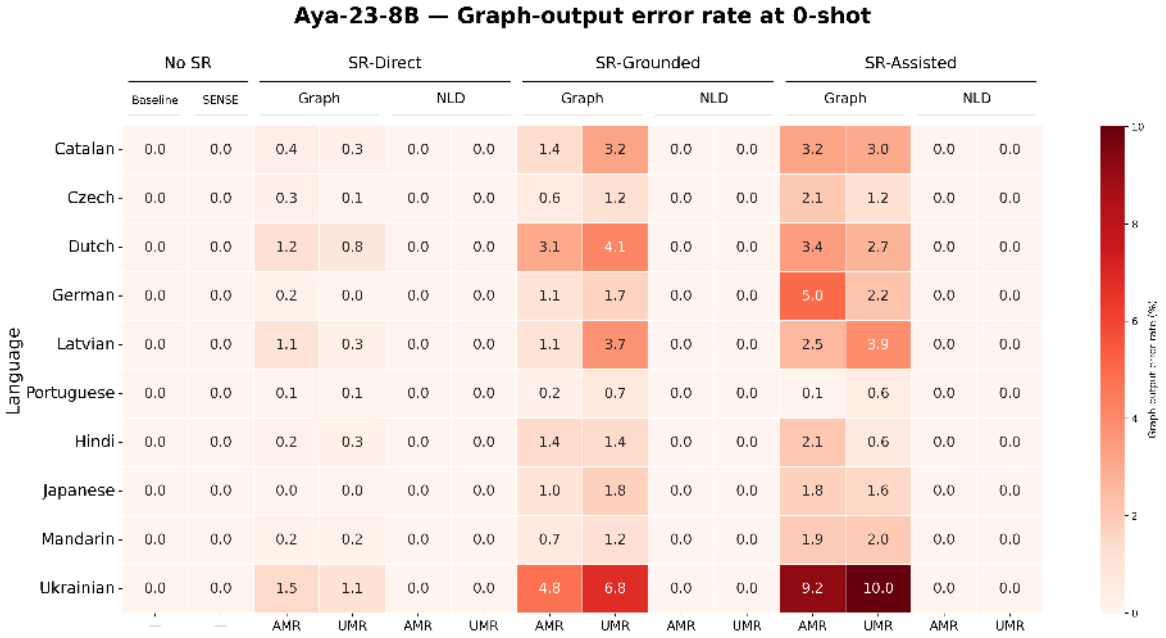


Figure 14: Aya-23-8B. Graph-output error rate at 0-shot.

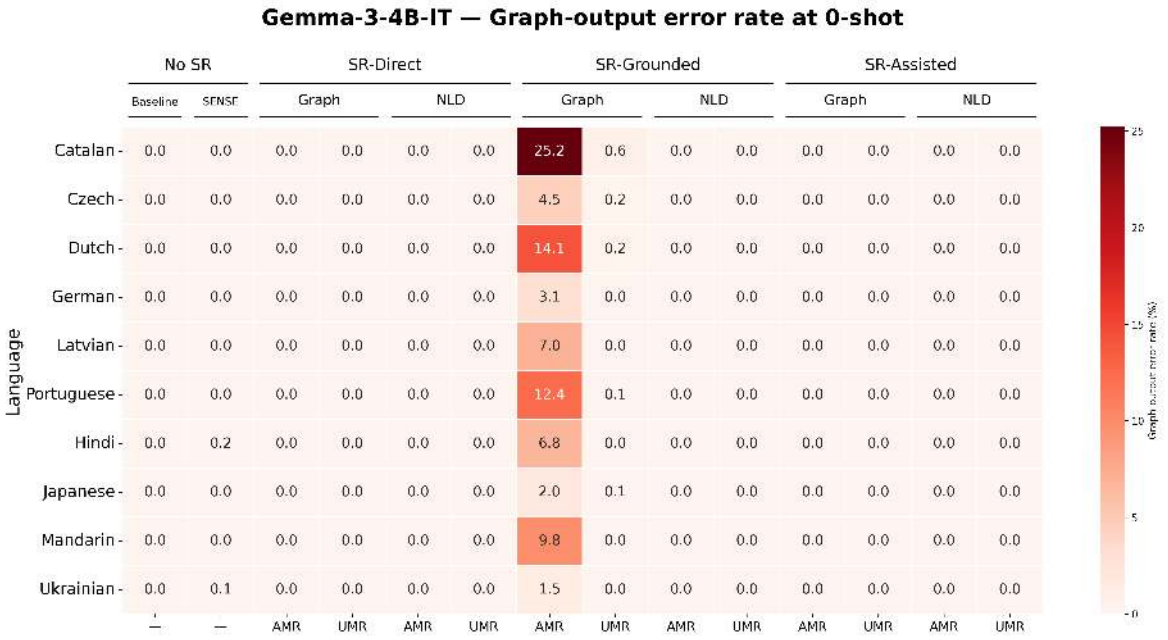


Figure 15: Gemma-3-4B-IT. Graph-output error rate at 0-shot.

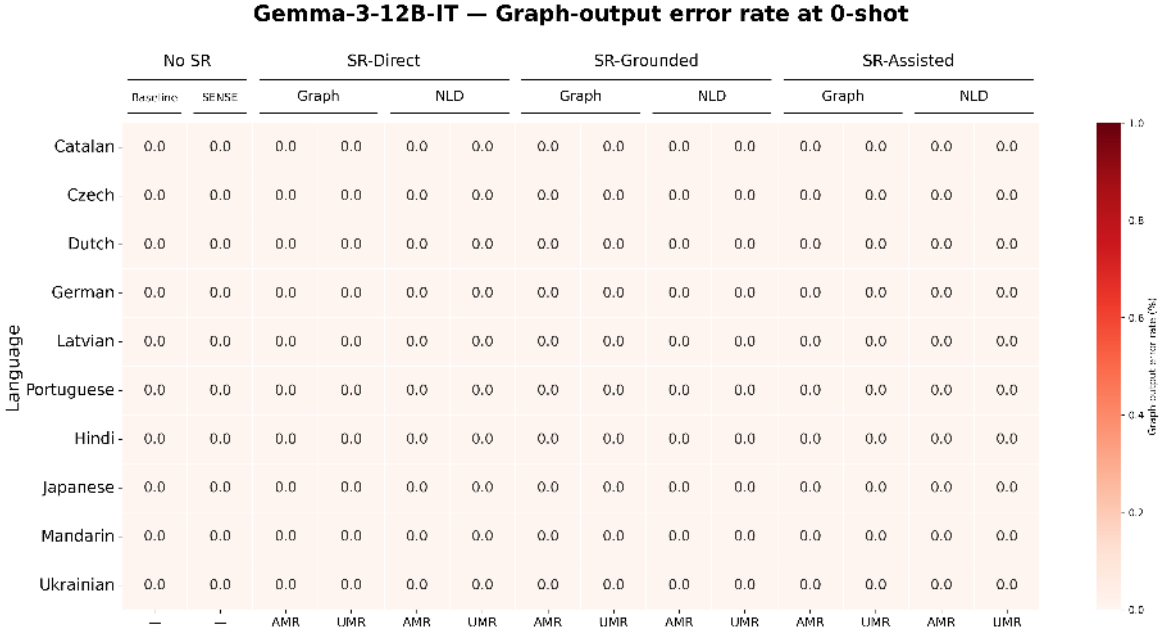


Figure 16: Gemma-3-12B-IT. Graph-output error rate at 0-shot.

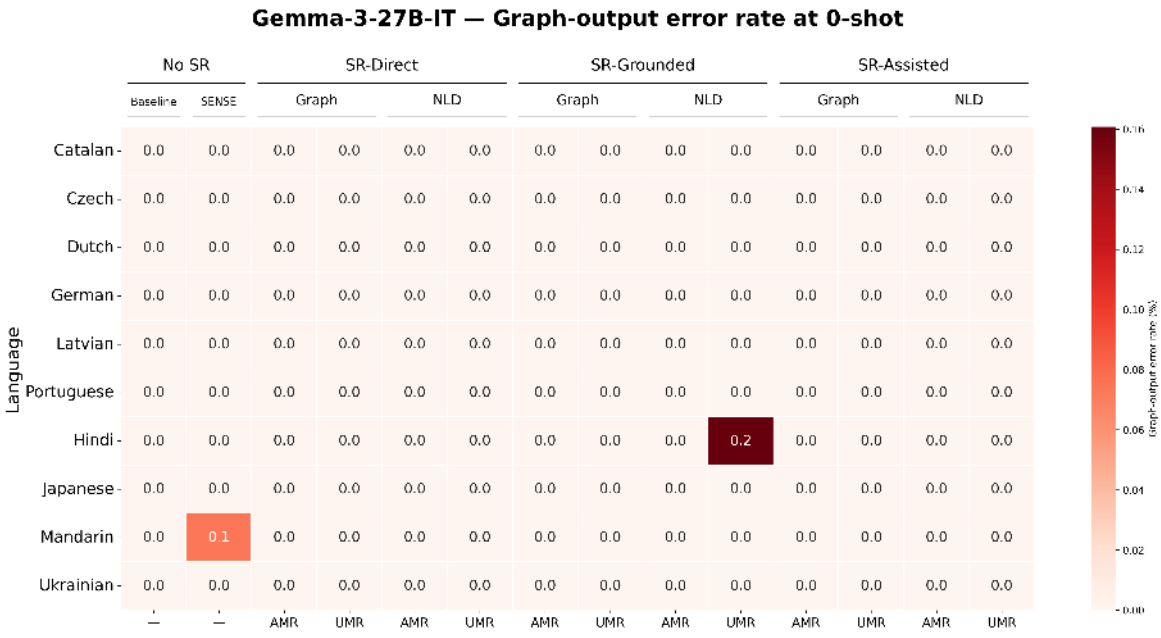


Figure 17: Gemma-3-27B-IT. Graph-output error rate at 0-shot.

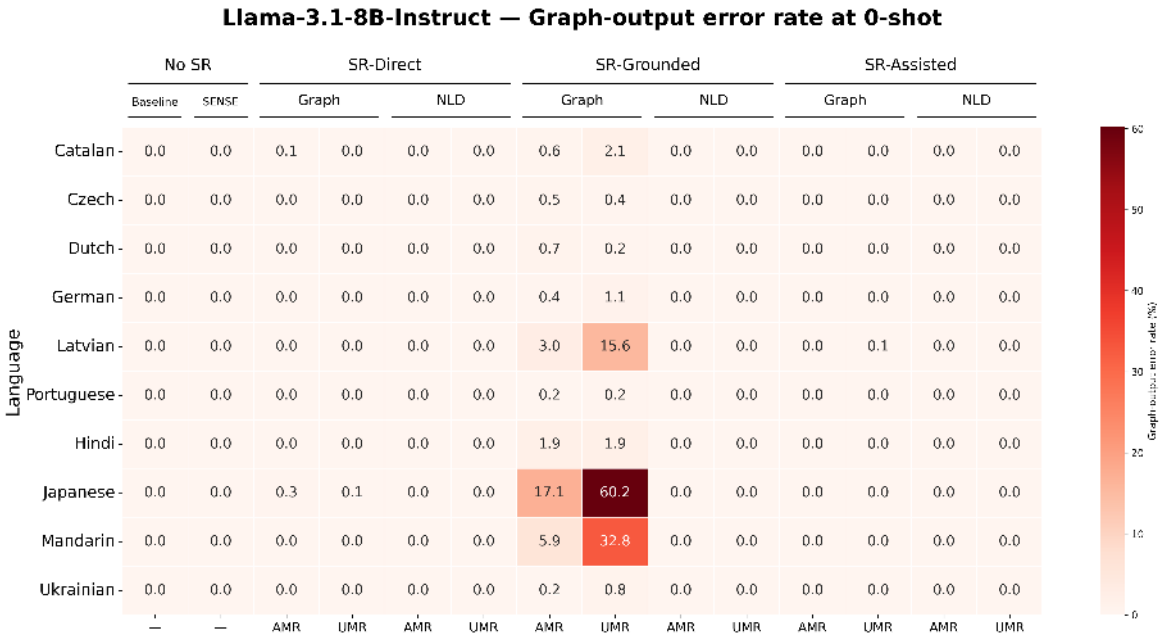


Figure 18: Llama-3.1-8B-Instruct. Graph-output error rate at 0-shot.

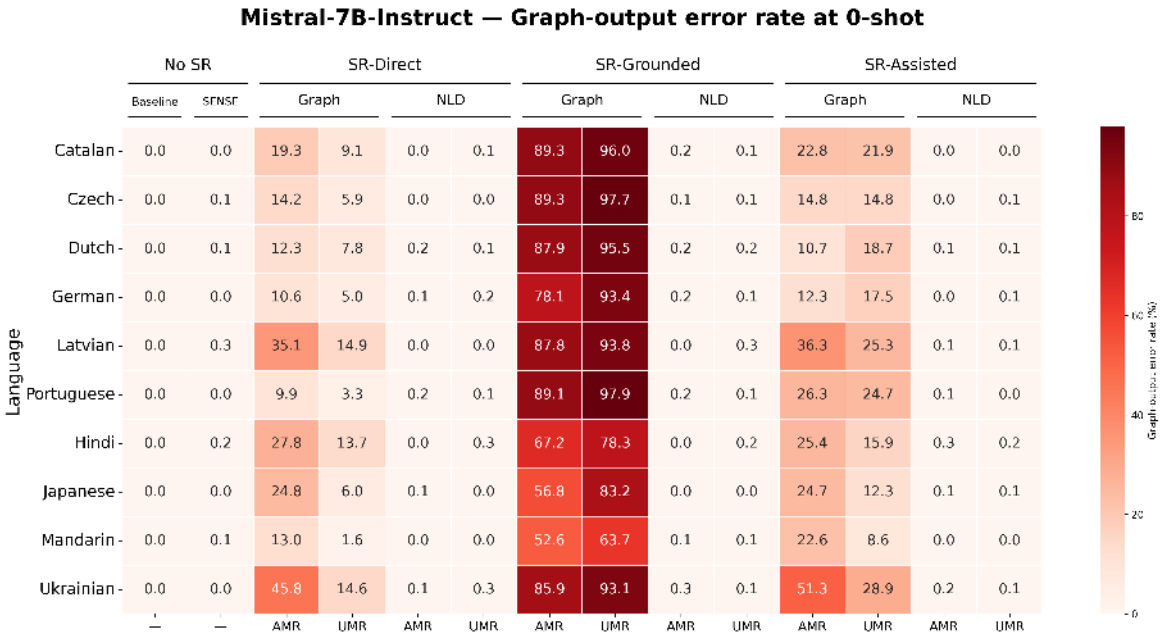


Figure 19: Mistral-7B-Instruct. Graph-output error rate at 0-shot.

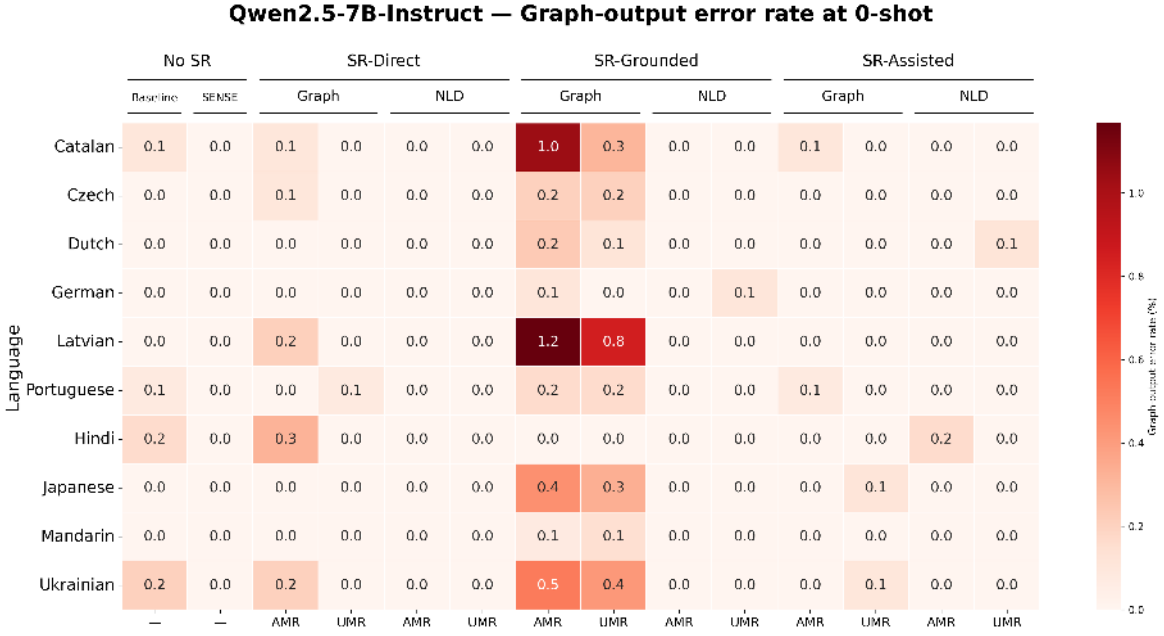


Figure 20: Qwen-2.5-7B-Instruct. Graph-output error rate at 0-shot.

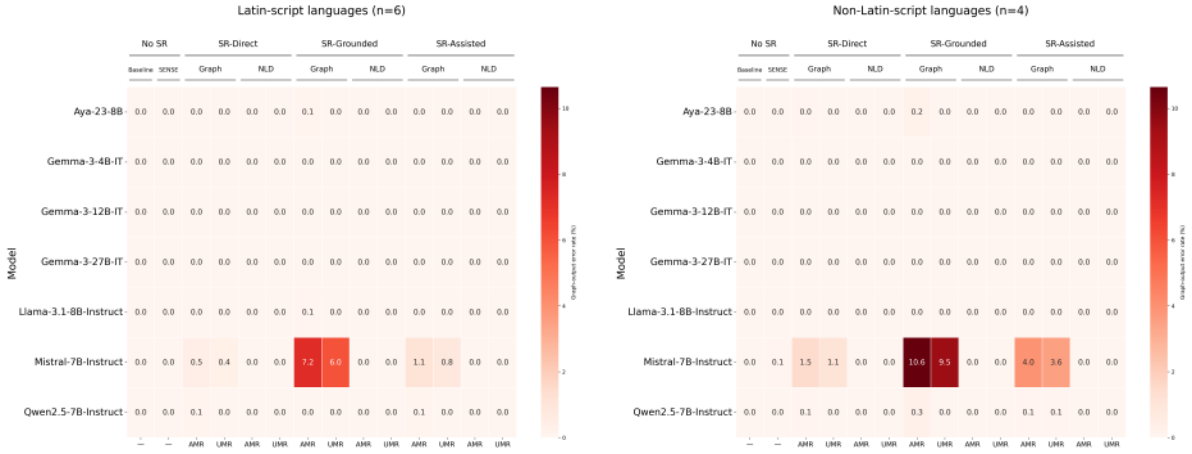


Figure 21: Graph-output error rate at 5-shot.

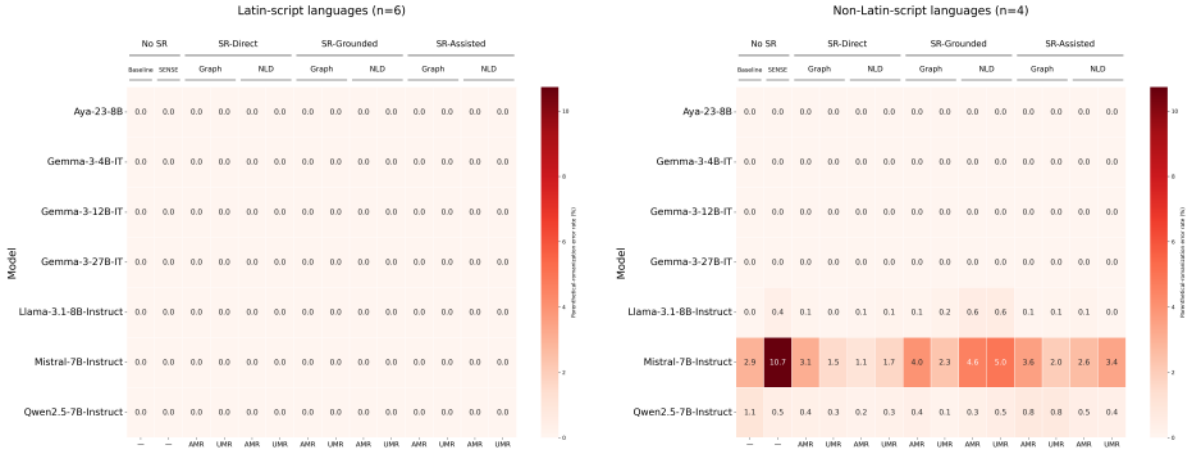


Figure 22: Parenthetical romanization error rate at 5-shot.

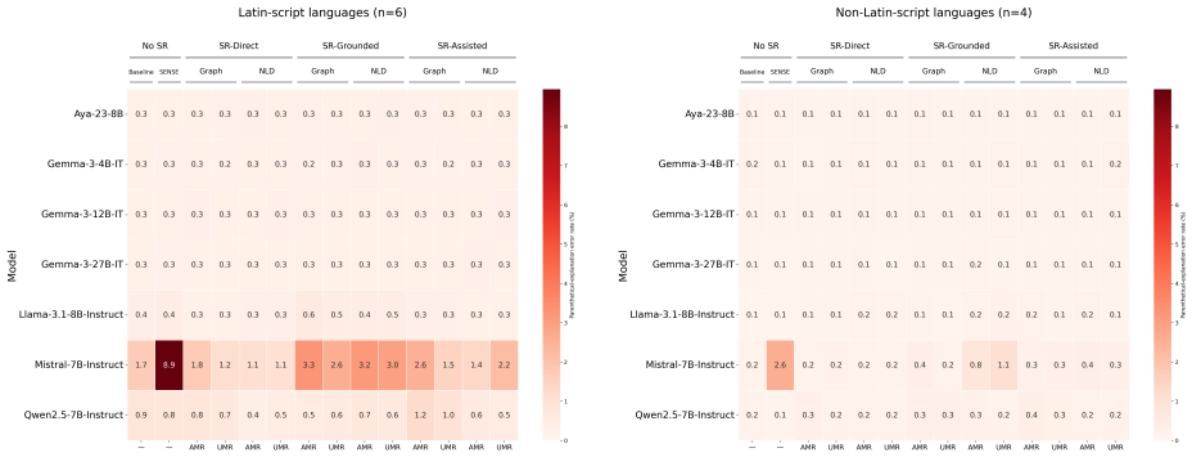


Figure 23: Parenthetical explanation error rate at 5-shot.

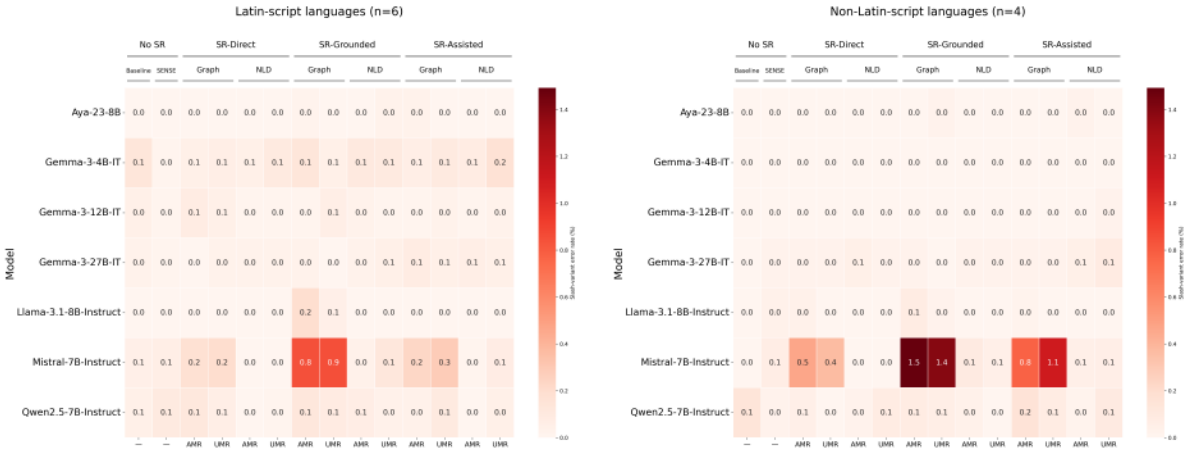


Figure 24: Slash-listed alternatives error rate at 5-shot.