GRAMMAMT: Improving Machine Translation with Grammar-Informed In-Context Learning

Rita Ramos[◇] **Everlyn Asiko Chimoto**[†*] **Maartje ter Hoeve**[◇] **Natalie Schluter**[◇]
[◇]Apple
[†]University of Cape Town, South Africa
rita_ramos@apple.com

Abstract

We introduce GRAMMAMT, a grammatically-aware prompting approach for machine translation that uses Interlinear Glossed Text (IGT), a common form of linguistic description providing morphological and lexical annotations for source sentences. GRAMMAMT proposes three prompting strategies: gloss-shot, chain-gloss and model-gloss. All are training-free, requiring only a few examples that involve minimal effort to collect, and making them well-suited for low-resource setups. Experiments show that GRAMMAMT enhances translation performance on open-source instruction-tuned LLMs for various low- to high-resource languages across three benchmarks: (1) the largest IGT corpus, (2) the challenging 2023 SIGMORPHON Shared Task data over endangered languages, and (3) even in an out-of-domain setting with FLORES. Moreover, ablation studies reveal that leveraging gloss resources could substantially boost MT performance (by over 17 BLEU points) if LLMs accurately generate or access input sentence glosses.

1 Introduction

Large Language Models (LLMs) have taken over the NLP leaderboards (e.g., Zellers et al., 2019; Hendrycks et al., 2020; Li et al., 2023b). Training LLMs requires access to a plethora of datasets, a luxury accessible to only a few of the world's most high-resource languages. Consequently, only a sliver of the world's languages have sufficient data for LLMs to achieve these impressive performance gains (Achiam et al., 2023; Üstün et al., 2024). To leverage the capabilities of these existing, high-resource LLMs in a low-resource context, one needs to design an approach that requires: (i) **little to no training** (to avoid overfitting and catastrophic forgetting), (ii) only a **small amount of data**, and/or (iii) **ease in data collection**.

Recent studies have shown the capability of LLMs to perform complex tasks, when provided with only a small amount of high quality language data. This data comes in the form of instruction-answer pairs for instruction fine-tuning (e.g, Li et al., 2023a; Yuan et al., 2024) or in the form of high quality prompts (e.g, Wei et al., 2022b). For example, for machine translation of languages unseen during training, performance gains have been achieved by only providing a dictionary and grammar book for the unseen languages as input to the LLM (Tanzer et al., 2024; Zhang et al., 2024).

Motivated by these results and the three requirements above, we propose GRAMMAMT, an in-context learning approach that leverages grammatical information from Interlinear Glossed Text (IGT) to improve machine translation in both low and high-resource settings. IGT is a triplet of source sentence, gloss, and target translation, commonly used by grammarians and linguists in linguistic description. The gloss represents the source sentence as a sequence of morphological and lexical annotations, as illustrated in Figure 1.

GRAMMAMT introduces three prompting strategies that augment few-shot machine translation using annotated glosses: (i) **gloss-shot**, (ii) **chain–gloss** and (iii) **model-gloss**. In gloss-shot, the LLM is prompted with examples pairing source sentences both with their translations and their glosses. In chain-gloss, the LLM first generates a gloss of the source sentence before translating. Model-gloss uses an external gloss model to generate the gloss, reducing the risk of incorrect glosses in chain-gloss, especially if a specialised gloss model is available for the target language. Importantly, GRAMMAMT adheres to all three of the above design requirements as follows.

Training-free. GRAMMAMT works by simply prompting an LLM with a grammatical demonstration. This is especially important in low-resource

[*]Work done while at Apple.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29920–29940

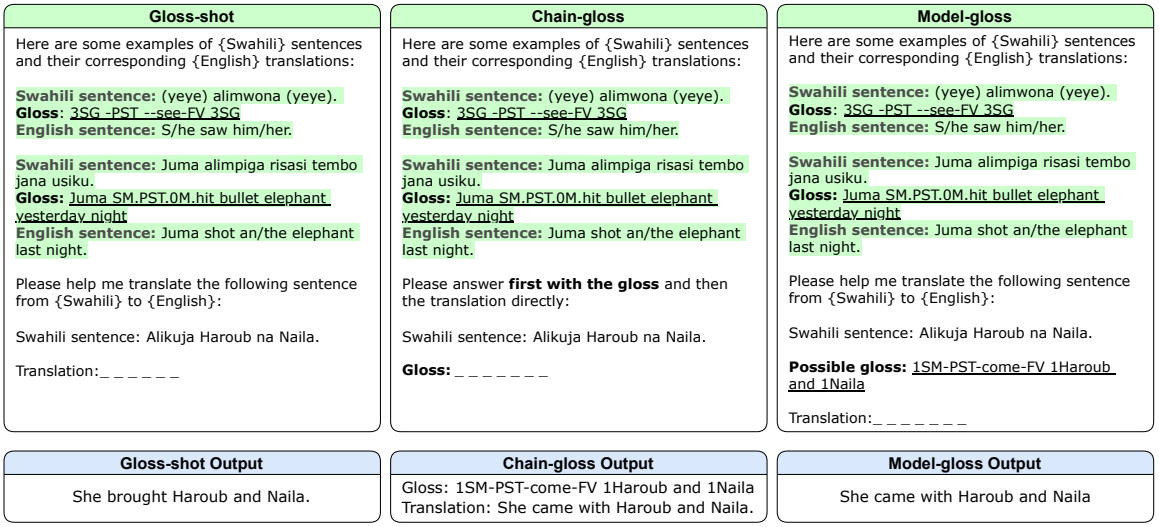

Figure 1: GRAMMAMT augments few-shot learning with Interlinear Gloss Text. In gloss-shot, the LLM is conditioned on translation pairs with source glosses. In chain-gloss, the LLM first generates the gloss before translating. Lastly, in model-gloss, the LLM receives an input gloss from an external gloss generation model.

settings, where sufficiently large training datasets are scarce, but minimal linguistic annotations exist or can be obtained. By incorporating linguistic knowledge directly into the prompt, we effectively leverage limited linguistic data that would otherwise be insufficient for fine-tuning an LLM.

Small number of examples. GRAMMAMT needs only a small number of grammatical annotations (e.g., 21 interlinear glosses examples). This differs from other few-shot methods, which depend on acquiring large data stores to gather relevant samples (e.g. retrieval-augmentation) or extensive resources like dictionaries or grammar chapters.

Ease of collection. Unlike chain-of-thought examples (Wei et al., 2022a), which require costly and subjective human engineering to break down machine translation into smaller steps, GRAMMAMT relies on basic gloss notation. These annotations are more straightforward–easier to either manually collect in low-resource settings, or can be sourced from grammar books or automatically generated (e.g., Ginn et al., 2024).

We benchmark our approach on three different datasets, including the 2023 SIGMORPHON Shared Task data (Ginn et al., 2023), the GlossLM dataset (Ginn et al., 2024) that has the most extensive corpus of IGT available, and also FLORES (Goyal et al., 2022); using state-of-the-art open-source instruction-tuned models, mainly Llama-3 (Meta, 2024) as well as Mixtral (Mistral, 2024). We find that GRAMMAMT can improve machine translation performance in low-resource setups, including endangered languages rarely encountered during pre-training. Even in high-resource languages, where the model has increased exposure and deeper understanding of the grammatical structure, we can observe substantial improvements from incorporating linguistic gloss resources into the prompt.

2 Related work

Machine translation with LLMs has been extensively explored (Zhang et al., 2023b; Garcia et al., 2023; Peng et al., 2023; Pourkamali and Sharifi, 2024). Although LLMs perform well for high-resource languages they underperform for low-resource languages (Hendy et al., 2023; Robinson et al., 2023; Zhu et al., 2023). While previous works study in-context learning for MT (Garcia et al., 2023; Puduppully et al., 2023; Zhang et al., 2023a; Sun et al., 2022), effective alternatives that leverage linguistic information for unseen and low-resource languages remain underexplored.

Using grammatical information with LLM Introducing grammatical information during training or inference can improve model performance (Strubell et al., 2018; Cui et al., 2022; Stahlberg et al., 2016). Similar to our work, Zhou et al. (2020) use glosses while training low-resource translation models. However, we use glosses in a training free approach and in the context of LLMs. Tanzer et al. (2024) and LingoLLM (Zhang et al., 2024) use grammar books along with other resources to translate unseen

and low-resource languages. Unlike these methods—which depend on grammar books, morphological analyzers, and dictionaries that are often unavailable—we use only a small number of gold or generated glosses, offering a more feasible solution for underrepresented languages.

3 GRAMMAMT

We propose GRAMMAMT, a simple grammar-informed prompting approach for machine translation, wherein examples of Interlinear Gloss Texts (IGT) are used as a prompt to instruction-tuned LLMs. In doing so, our approach is essentially **training-free**. The approach also requires a **small set of support examples** and **minimal annotation time** (a handful of glosses by a linguistic or automatically generated by a model (Ginn et al., 2024)). In this section, we provide an overview of IGT and describe the proposed prompting of GRAMMAMT.

Interlinear Gloss Text Annotation. IGT annotations are triplets of source text, glosses for the source text, and fluent target translations for the source text. The gloss consists of a sequence of target morphological annotations and (semantically full) lemmata for source words, indicating their grammatical morphemes and lexemes, shown by the following Swahili example.

1. `Source: (yeye) alimwona (yeye).`
2. `Gloss:` 3SG -PST –see-FV 3SG
3. `Translation:` *S/he saw him/her*.

In this example, the morphological annotation `3SG` stands for third-person singular and `PST` denotes the past tense of "see". Grammatical morphemes are labeled with uppercase letters. In contrast, lexemes (English lemma translations that convey semantic meaning) are labeled in lowercase (e.g., *see*). In this way, IGT captures the syntax and morphology of a sentence, aiding to grasp the structure of the source language and to understand the relationship between input sentence and the translation. These glosses are the norm in linguistic descriptions, and hence very common to find and easy to create.

Prompting strategies. GRAMMAMT augments an instruction-tuned LLM with in-context learning examples of interlinear glosses via three prompting strategies: **gloss-shot**, **chain-shot** and **model-gloss**, as illustrated in Figure 1.

In the first prompting strategy, **gloss-shot**, the LLM is prompted to generate the translation $\mathbf{y}$ for the input sentence $\mathbf{x}$ based on a set of N interlinear-glossed text exemplars $\mathbf{g}$ (i.e., triples of source sentence, gloss line, translation), essentially predicting $(\mathbf{g}_1, \cdots, \mathbf{g}_N, \mathbf{x}) \rightarrow \mathbf{y}$.

In the second prompting strategy, **chain-gloss**, the LLM is also conditioned on a set of N interlinear-glossed text exemplars $\mathbf{g}$ to generate the translation, but in this strategy, the model first produces the gloss $\mathbf{y}_g$ before formulating the translation $\mathbf{y}$, essentially $(\mathbf{g}_1, \cdots, \mathbf{g}_N, \mathbf{x}) \rightarrow (\mathbf{y}_g, \mathbf{y})$. This prompting strategy can offer some insights into how the LLM arrived at a specific translation.

In the **model-gloss** strategy, a specialised gloss generation model (e.g., GlossLM (Ginn et al., 2024)) provides the gloss for the source sentence, rather than relying on the LLM to generate it itself. As with the other strategies, this one also includes in-context examples of interlinear-glossed text, followed by the source sentence. However, here the source sentence is paired with a gloss predicted by the external model $\mathbf{y}_{ge}$, before the LLM produces the final translation: $(\mathbf{g}_1, \cdots, \mathbf{g}_N, \mathbf{x}, \mathbf{y}_{ge}) \rightarrow \mathbf{y}$

We illustrate the format of the prompt in Figure 1 and in more detail in Appendix L.

4 Experimental setup

4.1 LLMs

We assess our GRAMMAMT approach using `Meta-Llama-3-70B-Instruct` (Meta, 2024), the recent instruction-tuned Llama with 70B parameters. Our machine translation approach does not involve any training. The translations are generated at inference time using a single A100 80GB GPU. We also report experiments with the smaller `Meta-Llama-3-8B-Instruct`, and `Mixtral-8x22B-Instruct-v0.1` (Mistral, 2024), as well as the closed-source GPT-4o model (OpenAI, 2024) in Appendix E. The open-source LLMs were loaded via the HuggingFace Hub library (Wolf et al., 2020) using 4-bit quantization, while the GPT-4o model was accessed through the OpenAI API[1]. During inference, the models generate a translation using greedy decoding with a default temperature setting of 1.

4.2 Prompting strategies and baselines

Baselines. We first compare GRAMMAMT against other established in-context learning strategies, which use no explicit grammatical information:

- **zero-shot:** Translation from the source to the target language without examples.

[1] `https://platform.openai.com/`

- **zero-CoT:** The LLM is prompted to think step by step before translating, again without examples.
- **few-shot:** The LLM translates the input using a few source-target example pairs.

We select zero-CoT over Chain-of-Thought, because our data lacked the detailed steps needed for MT breakdown. We also compare GRAMMAMT to the training-free LingoLLM (Zhang et al., 2024), which uses more linguistic resources, including a grammar book, morphological analyzer, and a dictionary. For a thorough evaluation, we report performance of a state-of-the-art MT model, NLLB-200 (`nllb-200-distilled-600M`), while emphasising that it is not an LLM, as our focus is on improving LLMs for MT. Finally, we compare against a parallel dictionary baseline in Appendix C.

GRAMMAMT prompting. Our own approach augments few-shot prompting with grammatical information, where we explore three novel variants:

- ***gloss-shot*:** The LLM predicts based on examples that pair the source sentences not just with their translation but also with their gloss.
- ***chain-gloss*:** As in gloss-shot, but the LLM is additionally prompted to generate the gloss for the input sentence before translating.
- ***model-gloss*:** As in chain-gloss, but the gloss of the source sentence is obtained from an external gloss generation model and not from the LLM itself. For this, we use GlossLM (Ginn et al., 2024) that was trained to generate glosses.[2]

For all prompting strategies[3], we use the same 21 translation examples per language, identified as the optimal value in our ablation studies (see Section 6). Prompt templates are provided in Appendix L.

4.3 Datasets and Languages

We evaluate translation quality across three datasets, involving endangered, low-resource, and mid-to-high-resource languages, with English as the target language. Table 1 summarises the languages, scripts and test set sizes. For completeness, we also evaluate the reverse translation direction, with English as source language, in Section 6.

[2]See Appendix A for details.

[3]Except zero-shot and zero-CoT that have no examples.

Language	Abbr.	Script	Test	Speakers
Sigmorphon dataset				
Gitksan	Git	Latin	37	1,110
Lezgi	Lez	Cyrillic	87	800K
Natugu	Ntu	Latin	99	5,900
Tsez	Ddo	Cyrillic	445	18K
GlossLM dataset				
Swahili	Swa	Latin	439	200M
Yoruba	Yor	Latin w/ diac.	135	47M
Icelandic	Ice	Latin	27	330K
Marathi	Mar	Devanagari	43	83M
Kannada	Kan	Kannada	388	59M
Urdu	Urd	Perso-Arabic	259	232M
Thai	Tha	Thai	352	61M
Greek	Gre	Greek	59	13.5M
Portuguese	Por	Latin	309	264M
Japanese	Jap	Japanese [3]	4,748	123M
Russian	Rus	Cyrillic	2,444	255M
Arabic	Ara	Arabic	136	274M

Table 1: Overview of the languages and the test split sizes used in GRAMMAMT evaluation.

Sigmorphon: We use the dataset from the 2023 SIGMORPHON Shared Task for evaluating on unseen, endangered languages (Ginn et al., 2023), with Gitksan, Lezgi, Natugu, and Tsez. This dataset includes translation pairs from each source language to English, together with the interlinear glosses and morphological segmentation of the source sentences. We report performance on the test set, while the validation split is used for ablation studies. In both cases, support examples are drawn from the training split, specifically the first 21 sentences (Section 6 shows that $N = 21$ is optimal).

GlossLM corpus: For evaluating on low to high-resource languages, we use the GlossLM dataset (Ginn et al., 2024), a recent and extensive compilation of interlinear glossed text (IGT) from six different IGT corpora. This dataset includes 250k unique sentences across 1800 languages. We selected languages from different scripts, specifically considering Swahili, Yoruba, Icelandic, Marathi, and Kannada for low-resource languages. For mid-to-high-resource languages, we included Urdu, Thai, Greek, Portuguese, Japanese, Russian, and Arabic. However, the GlossLM dataset only provides evaluation splits (dev/test) for the endangered languages included in the SIGMORPHON Shared Task, as this data is the most consistent. For other languages ranging from low to high-resource, the dataset of-

Method	BLEU					chrF++					xCOMET
	Git	Lez	Ntu	Ddo	Avg.	Git	Lez	Ntu	Ddo	Avg.	Avg.
NLLB-200	0.9	0.8	0.4	0.1	0.55	23.65	18	12.3	10.10	13.80	12.82
LingoLLM w/ GPT-4	14.3	-	12.9	**15.1**	14.1	-	-	-	-	-	-
zero-shot	1.26	1.46	0.26	0.39	0.88	23.90	17.71	13.76	16.84	18.05	15.21
zeroCoT	2.84	1.74	0.37	0.32	1.32	21.21	15.27	13.95	15.68	16.53	14.50
few-shot	4.71	6.36	3.34	1.46	3.94	25.18	22.89	19.41	20.03	21.85	16.76
gloss-shot	4.96	5.80	1.32	1.72	3.41	25.87	23.08	20.24	20.95	22.50	18.21
chain-gloss	5.71	7.29	2.35	1.63	4.25	24.66	22.62	19.19	18.01	20.84	16.78
model-gloss	**18.7**	**13.94**	**16.96**	14.28	**15.97**	**47.89**	**39.65**	**41.56**	**42.30**	**41.45**	**40.83**

Table 2: GRAMMAMT's performance (using Llama-3 70B) for **unseen/endangered languages** on the 2023 SIGMORPHON test split, against in-context baselines and SOTA models like NLLB-200 and LingoLLM. Best results are in bold and second-best are underlined.

Method	BLEU						chrF++						xC
	Swa	Yor	Ice	Mar	Kan	Avg.	Swa	Yor	Ice	Mar	Kan	Avg.	Avg.
NLLB-200	6.9	0.5	3.5	0.3	0.8	2.4	24.2	10.8	21.1	10	10.7	15.36	20.21
zero-shot	16.99	4.48	4.92	0.70	5.84	6.58	40.35	18.87	27.97	13.28	25.65	25.22	27.10
zero-CoT	15.78	1.93	4.64	1.08	4.99	5.69	39.15	18.84	28.02	14.87	25.20	25.22	27.76
few-shot	22.41	11.98	**6.43**	**19.19**	23.50	16.69	45.75	29.92	**28.87**	36.11	44.16	36.96	34.52
gloss-shot	22.18	**16.32**	3.50	17.53	22.35	16.39	**46.50**	33.24	25.79	**36.18**	42.68	36.88	35.65
chain-gloss	**23.53**	14.10	5.05	17.32	**25.25**	**17.06**	45.44	**33.54**	24.90	35.37	**46.27**	**37.10**	**35.77**

Table 3: GRAMMAMT's performance (using Llama-3 70B) for **low-resource languages** on the GlossLM data, the largest corpus of IGT data. Best results are in bold; second-best underlined. xC is xCOMET.

fers only a training split. To address this, we created evaluation splits by designating most of the training set for testing, reserving the first 21 examples for in-context learning (Section 6 provides empirical evidence that $N = 21$ is optimal). We have detailed the number of test samples for each language in Table 1. To avoid unfair evaluation, results for the model-gloss strategy are not provided on our test split, since the GlossLM model (Ginn et al., 2024) used in this strategy was exposed to those training samples. But we report model-gloss results for these languages in the subsequent dataset.

FLORES-200: We also report results on the FLORES dataset (Goyal et al., 2022) (test split). We use the same languages we considered from the GlossLM dataset, and the same set of 21 examples since FLORES does not contain the annotated glosses, to assess our approach's ability to generalise in the absence of in-domain glosses.

4.4 Metrics

For evaluation, we report MT evaluation metrics, namely BLEU (Papineni et al., 2002) with SacreBLEU tokenisation (Post, 2018), and the chrF++ metric, which exhibits a stronger correlation with human scores (Popović, 2017). To further strengthen our evaluations, we include a model-based metric using xCOMET-XXL (Guerreiro et al., 2024), the latest version of the widely adopted COMET model (Rei et al., 2020). We report significance tests over these metrics in Appendix J.

5 Results

GRAMMAMT outperforms in unseen/endangered languages. In Table 2, we show how GRAMMAMT performs on four endangered languages: Gitksan, Lezgi, Natugu and Tsez (all unseen by the LLM during pre-training). The results demonstrate that the model-gloss strategy consistently outperforms the baselines across the three metrics. Focusing on BLEU, this strategy shows a large improvement of 15.09, 14.65, and 12.03 BLEU points against zero-shot, zero-CoT and the few-shot approach on average, respectively. Additionally, it surpasses the specialised NLLB translation model, which struggles with unseen languages. Furthermore, the model-gloss strategy outperforms LINGOLLM (Zhang et al., 2024), the state-of-the-art training-free method in this shared task, by over 4 BLEU points for Gitksan and Lezgi, while being only

Method	BLEU								chrF++								xC
	Urd	Tha	Gre	Por	Jap	Rus	Ara	Avg.	Urd	Tha	Gre	Por	Jap	Rus	Ara	Avg.	Avg.
NLLB-200	0.2	0.2	0.5	26.2	0.4	2.4	1.4	4.47	9.1	9.3	11.3	47	14.4	17	11.5	17.09	24.25
zero-shot	4.00	1.35	6.13	37.75	7.17	25.12	3.46	12.15	20.53	12.56	23.14	59.21	27.62	47.31	19.95	30.05	35.42
zero-CoT	4.71	1.80	8.22	37.20	7.26	23.42	3.81	12.35	22.60	12.91	25.56	56.50	27.30	45.07	19.18	29.87	35.10
few-shot	26.19	7.68	10.62	44.14	13.74	24.94	5.35	18.95	43.36	19.76	**27.55**	**63.88**	35.94	48.59	**21.28**	37.19	41.03
gloss-shot	26.86	6.26	9.56	**44.37**	13.65	23.99	**5.60**	18.61	43.49	19.27	27.17	63.72	35.71	48.13	21.19	36.95	41.05
chain-gloss	**28.71**	**8.34**	**10.74**	42.88	**15.41**	**27.92**	5.26	**19.75**	**45.86**	**19.81**	27.11	62.33	**37.29**	**50.22**	19.51	**37.20**	**41.46**

Table 4: GRAMMAMT's performance (using Llama-3 70B) for **mid-high-resource languages** on the GlossLM data. Best results are in bold; second-best underlined. xC is xCOMET.

slightly outperformed by 0.82 points for Tsez. This is despite LingoLLM's leveraging vastly more extensive linguistic resources, such as grammar books and dictionaries.

Within the GRAMMAMT strategies, model-gloss is more robust compared to relying on the LLM for gloss prediction (chain-gloss)[4] or using glosses only for examples (gloss-shot). This is most likely because it relies on a specialised gloss model tailored to these languages. However, both these methods still show promising results. We see that the gloss-shot strategy outperforms the prompting baselines across all unseen languages tested on using the chrF++ metric. Additionally, BLEU scores improve for both Gitksan and Tsez. For chain-gloss, while few-shot outperforms with the chrF++ metric, we observe BLEU score increases of 1 point for Gitksan, 0.93 for Lezgi, and 0.17 for Tsez. Overall, GRAMMAMT outperforms translation for unseen languages in our experiments, indicating the benefits in this challenging language setup.

Chain-gloss improves translation of low-resource languages. We also assess GRAMMAMT on low-resource languages, including Swahili, Yoruba, Icelandic, Marathi and Kannnada (see Table 3). Chain-gloss improves the performance on the majority of them as seen in the average BLEU, chrF++ and the xCOMET score. This improvement is similarly observed with gloss-shot, particularly in the chrF++ performance for Swahili and Marathi. Notably, we observed a large improvement for Yoruba from adding the gloss to the context, with an increase of more than 4 BLEU points and 3 chrF++ points compared to few-shot. Icelandic and Marathi, exhibited the best performance using few-shot based on BLEU. We exclude the model-gloss strategy, as it leverages glosses from GlossLM (Ginn et al., 2024). As GlossLM was pre-trained on this data, including the model-gloss strategy would lead to unfair evaluation, due to prior exposure to the test set.

Chain-gloss also improves mid-high-resource languages. In Table 4, we observe that GRAMMAMT improves the performance for all of the high-resource languages on BLEU, with the best performing method being either chain-gloss or gloss-shot. Notably, Urdu and Russian show substantial improvements, with chain-gloss surpassing few-shot by more than 2.5 BLEU points. Using chrF++, consistent with the BLEU results, we have chain-gloss outperforming the other methods except for Portuguese, Arabic and Greek, for which few-shot outperforms both gloss-shot and chain-gloss. For these languages, gloss-shot also outperforms chain-gloss. We again excluded results for the model-input strategy as the gloss model had prior exposure to the test set. Overall, results show that augmenting the context with grammatical information is not only beneficial in low-resource settings, but also for mid-to-high-resource languages.

5.1 Out of domain evaluation: Flores

We also evaluate GRAMMAMT on the FLORES test set, where in-domain glosses are unavailable, by reusing the same GlossLM examples in the translation prompts. Table 5 shows that gloss-shot achieves the highest average BLEU score, followed by model-gloss, with both achieving notable improvements of 2 points for Portuguese, Japanese, and Russian over few-shot. This suggests that both strategies can be effective even without annotated glosses for the current domain. In contrast, chain-gloss often struggles to predict accurate glosses and translations, likely due to a distributional shift from the short, simple GlossLM examples, to the more complex and lengthy input sentences in the FLORES dataset. The example in Figure 7 of Appendix F illustrates. The model-gloss strategy also performs poorly for low-resource languages. Thus, in out-of-domain settings, it is preferable to use glosses as examples (gloss-shot) rather than having

[4]See Section 6 for a comparison of gloss performance.

Method	BLEU												
	Swa	Yor	Ice	Mar	Kan	Urd	Tha	Gre	Por	Jap	Rus	Ara	Avg.
few-shot	20.99	3.94	18.23	**18.72**	3.63	**19.78**	**21.34**	28.07	41.24	16.62	27.27	**28.59**	20.69
gloss-shot	**22.37**	5.00	**19.40**	18.26	3.04	18.65	20.35	**30.08**	43.30	19.61	31.14	28.43	**21.64**
chain-gloss	20.26	**5.03**	18.05	17.20	**4.40**	18.13	19.32	27.82	41.62	18.23	30.26	26.74	20.59
model-gloss	18.30	3.67	16.92	17.73	3.09	18.75	20.21	29.15	**43.64**	**19.77**	**31.16**	**28.59**	20.92

Table 5: BLEU performance on the FLORES test set. We select the 21-shot examples from the GlossLM data, as FLORES lacks annotated glosses. Results show that GRAMMAMT can generalise in an out-of-domain setting.

Method	Model	BLEU					chrF++				
		Git	Lez	Ntu	Ddo	Avg.	Git	Lez	Ntu	Ddo	Avg.
few-shot	Llama-3 70B	4.71	6.36	3.34	1.46	3.94	25.18	22.89	19.41	20.03	21.85
gloss-shot	Llama-3 70B	4.96	5.80	1.32	1.72	3.41	25.87	23.08	20.24	20.95	22.50
chain-gloss	Llama-3 70B	5.71	7.29	2.35	1.63	4.25	24.66	22.62	19.19	18.01	20.84
model-gloss	Llama-3 70B	**18.7**	**13.94**	**16.96**	**14.28**	**15.97**	**47.89**	**39.65**	**41.56**	**42.30**	**41.45**
few-shot	Llama-3 8B	2.30	5.03	1.70	0.47	2.38	25.28	20.4	18.9	18.64	20.81
gloss-shot	Llama-3 8B	2.83	4.63	1.46	0.60	2.38	23.59	20.8	17.8	18.8	20.25
chain-gloss	Llama-3 8B	4.21	8.2	2.60	0.80	3.95	23.26	32.6	17.30	16.81	22.49
model-gloss	Llama-3 8B	**7.11**	**10.68**	**7.44**	**8.2**	**8.36**	**37.72**	**34.41**	**32.80**	**35.09**	**35.00**
few-shot	Mixtral-8x22B	3.32	7.05	3.80	2.46	4.16	25.04	23.45	21.59	20.76	22.71
gloss-shot	Mixtral-8x22B	4.58	6.56	4.63	3.08	4.71	24.80	22.45	22.14	21.52	22.73
chain-gloss	Mixtral-8x22B	3.12	1.50	4.56	1.06	2.56	18.55	10.88	21.94	15.28	16.66
model-gloss	Mixtral-8x22B	**16.27**	**13.74**	**19.24**	**18.89**	**17.03**	**48.45**	**40.45**	**44.71**	**44.87**	**44.62**
few-shot	GPT-4o	5.49	8.25	4.14	1.64	4.88	27.58	23.32	21.50	19.29	22.71
gloss-shot	GPT-4o	5.87	7.29	3.93	2.04	4.78	28.49	22.51	21.68	19.77	22.73
chain-gloss	GPT-4o	4.16	7.46	5.72	1.91	4.81	26.23	23.15	21.77	20.29	22.86
model-gloss	GPT-4o	**22.24**	**14.45**	**21.25**	**17.10**	**18.69**	**49.59**	**38.64**	**45.03**	**41.37**	**43.66**

Table 6: BLEU performance of GRAMMAMT on the 2023 SIGMORPHON test split across the different models (Llama-3 70B, Llama-3 8B, Mixtral-8x22B, GPT-4o).

the model generating the gloss without in-domain examples, to avoid misleading translations.

GRAMMAMT generalizes effectively across different LLM architectures and sizes. In addition to evaluating our approach using Llama-3 70B, we assess its ability to generalize to other models. Specifically, we report results for Llama-3 8B and Mixtral-8x22B, as well as the closed-source model GPT-4o. Table 6 shows the performance of these models for the unseen, endangered languages on the 2023 SIGMORPHON test split. Refer to Appendix E for the results across the remaining languages. As shown in Table 6, GRAMMAMT generalizes well to other LLMs, yielding a stronger performance with GPT-4o and Mixtral. The smaller Llama-3 8B model also benefits from incorporating grammatical information, with model-gloss and chain-gloss outperforming the few-shot baseline on average across both BLEU and chrF++. Overall, these results provide evidence that GRAMMAMT is a versatile approach that achieves good performance with both small and large models.

6 Further analysis and discussion

We conduct a series of ablation studies on the validation splits of the aforementioned datasets to better understand the impact of GRAMMAMT on improving LLM performance in machine translation.

Varying N. We consider the impact of the number of examples provided in prompts and vary the number of shots, N, both in our proposed GRAMMAMT strategy and in the few-shot baseline. We illustrate this for Lezgi in Figure 3. An increase of N leads to improvements in all strategies, with optimal value being $N = 21$. We see large gains on chain-gloss by increasing N, suggesting that chain-gloss needs a sufficient number of examples to demonstrate the process of generating glosses.

Gloss Accuracy. Here we study to what extent do the glosses generated by chain-gloss and model-gloss strategies influence the translation output. We compare the glosses generated by Llama (used in

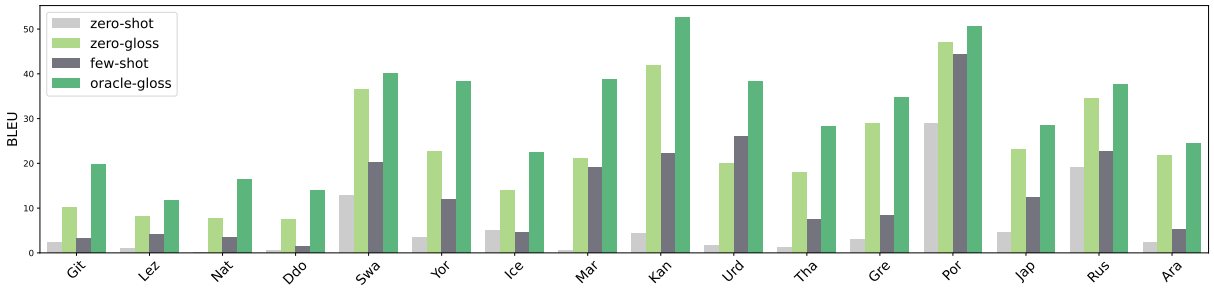

Figure 2: Simulation of an oracle experiment with GRAMMAMT using reference glosses (*oracle-gloss* with N-shot examples or *zero-gloss*) to assess if performance improves with accurate generation or access to correct glosses.

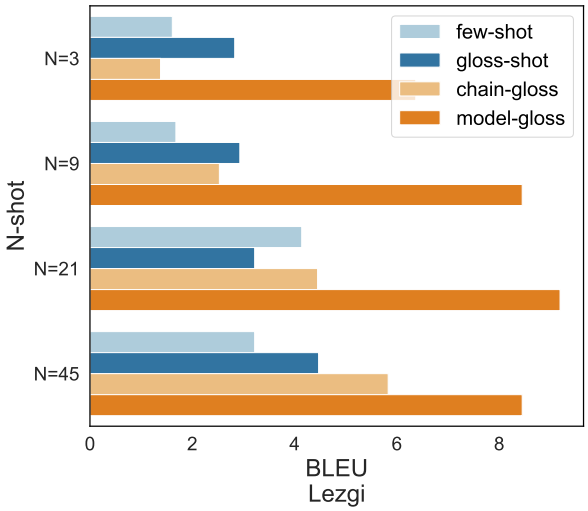

Figure 3: Varying N-shot examples from 3 to 45. These ablations were conducted on the validation split of the 2023 SIGMORPHON Shared Task data for Lezgi.

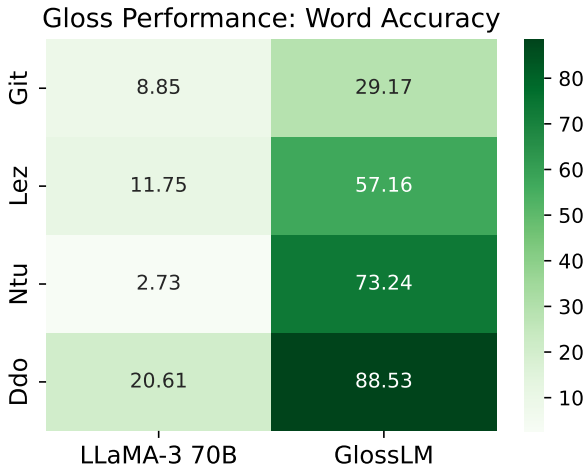

Figure 4: Evaluation of glosses generated by chain-gloss (Llama-3 70B) and model-gloss (GlossLM).

the chain-gloss strategy) with GlossLM (Ginn et al., 2024) (used in the model-gloss strategy). Figure 4 highlights that Llama struggles with gloss accuracy for rarely seen languages, achieving less than 21% accuracy for Tsez (Ddo). In contrast, GlossLM performs substantially better, achieving up to 88% for Tsez, directly contributing to the model-gloss strategy's superior MT performance in Table 2. We used word accuracy to assess gloss performance, consistent with the evaluation in GlossLM's work (Ginn et al., 2024), reporting further metrics in Appendix H.

Oracle Setup. Here we further study translation performance if the model could accurately generate or access the gloss of the source sentence. We conduct an oracle experiment where we replace the generated glosses in the chain-gloss or model-gloss strategies with gold-standard glosses (*oracle-gloss*). We also evaluate a zero-shot setup (*zero-gloss*), prompting the model to translate directly from the source with the gold gloss. Both are compared to their respective baselines (few-shot and zero-shot).

Oracle-gloss significantly improves by an average of 17.46 BLEU points ($\pm$ 6.6) over few-shot across all languages, and zero-gloss also outperforms zero-shot by a massive margin of 16.02 BLEU points ($\pm$ 8.89). Notably, zero-gloss even surpasses the few-shot setting that uses machine translation examples. Overall, these results highlight the potential of leveraging glosses for improving machine translation. A promising direction is the development of automatic gloss models, such as GlossLM (Ginn et al., 2024).

Role of grammatical annotations. We further analyze whether performance is solely due to the English lemmata or whether grammatical annotations actually matter. Figure 5 shows a performance drop when grammatical labels are removed, indicating their importance beyond mere word-by-word translation from lemmata. Moreover, we also present examples of translations produced by GRAMMAMT in Appendix K where we further observe that our strategies generate more satisfactory translations compared to the few-shot approach by

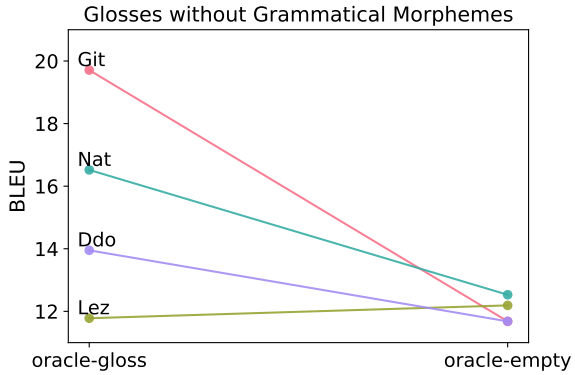

Figure 5: Performance drops when removing grammatical annotations (oracle-empty) compared to the original glosses (oracle-gloss). These ablations were also conducted on the validation split of the SIGMORPHON.

being grammatical-aware. In Appendix D, we also explore other grammatical augmentations.

MT from English (en →). Due to the limited availability of IGT datasets, we focus on translating into English (→ en). To translate from English (en →), we swap the source and target languages in our prompts, using the target language's gloss to guide the process.[5] Our prompting strategies continue to perform well in reverse translation, as shown in Figure 6. Future research should further explore our approach for translating from English.

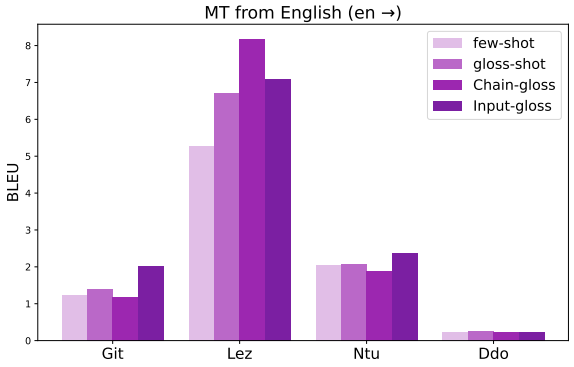

Figure 6: BLEU performance from English to target languages on the SIGMORPHON test set.

7 Conclusions

We propose GRAMMAMT, a machine translation prompting approach that augments instruction-tuned LLMs with grammatical information using interlinear gloss resources. This formulation of machine translation enables a range of desirable properties: it is training-free, efficient in terms of support examples, and requires minimal effort for data collection. Our results demonstrate improvements across low-resource contexts, including endangered languages that the model had minimal exposure to, as well as in high-resource languages where the model is already familiar with the grammatical structure.

Experiments further show the possibility of achieving large gains in BLEU across studied languages when an LLM has access to or can correctly generate a gloss for the input sentence. This attests for the potential impact of annotated glosses in machine translation, suggesting that exploring specialised models for automatic gloss generation could be an important avenue for future research.

8 Limitations

Our gloss-shot strategy builds upon few-shot prompting and, consequently, has limited interpretability. The glosses are derived from examples unrelated to the input image, making it unclear how these examples directly influence translation outcomes. In contrast, chain-gloss (and model-gloss), akin to chain-of-thought prompting, provides more interpretability by generating step-by-step glosses specifically for the input sentence. In Section 6, we conduct various ablation studies and qualitative analyses to provide insights into how GRAMMAMT helps LLMs generate better translations.

Although our work covers a wide range of languages, it focuses mainly on MT to English (→ en). This limitation is due to the availability of Interlinear Gloss Text datasets, which primarily contain glosses and translations in English. In Section 6 we also attempted translation from English (→ en) but this was not the focus of our research; future work should further evaluate our approach in this setup. Also, future research should explore our approach from a less English-centric perspective to assess its broader applicability.

[5] See an example in Appendix I.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Orevaoghene Ahia, Julia Kreutzer, and Sara Hooker. 2021. The low-resource double bind: An empirical study of pruning for low-resource machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word translation without parallel data. *Preprint*, arXiv:1710.04087.

Yiming Cui, Wanxiang Che, Shijin Wang, and Ting Liu. 2022. Lert: A linguistically-motivated pre-trained language model. *Preprint*, arXiv:2211.05344.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. In *International Conference on Machine Learning*, pages 10867–10878. PMLR.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer. 2023. Dictionary-based phrase-level prompting of large language models for machine translation. *Preprint*, arXiv:2302.07856.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. pages 186–201, Toronto, Canada.

Michael Ginn, Lindia Tjuatja, Taiqi He, Enora Rice, Graham Neubig, Alexis Palmer, and Lori Levin. 2024. Glosslm: Multilingual pretraining for low-resource interlinear glossing. *arXiv preprint arXiv:2403.06399*.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xcomet: Transparent Machine Translation Evaluation through Fine-grained Error Detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

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.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Xian Li, Ping Yu, Chunting Zhou, Timo Schick, Luke Zettlemoyer, Omer Levy, Jason Weston, and Mike Lewis. 2023a. Self-alignment with instruction backtranslation. *arXiv preprint arXiv:2308.06259*.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023b. Alpacaeval: An automatic evaluator of instruction-following models. `https://github.com/tatsu-lab/alpaca_eval`.

Meta. 2024. Llama 3 model card.

Mistral. 2024. Mixtral-8x22b.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Keqin Peng, Liang Ding, Qihuang Zhong, Li Shen, Xuebo Liu, Min Zhang, Yuanxin Ouyang, and Dacheng Tao. 2023. Towards making the most of ChatGPT for machine translation. Singapore.

Maja Popović. 2017. chrf++: words helping character n-grams. In *Proceedings of the second conference on machine translation*, pages 612–618.

Matt Post. 2018. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*.

Nooshin Pourkamali and Shler Ebrahim Sharifi. 2024. Machine translation with large language models: Prompt engineering for persian, english, and russian directions. *Preprint*, arXiv:2401.08429.

Ratish Puduppully, Anoop Kunchukuttan, Raj Dabre, Ai Ti Aw, and Nancy F Chen. 2023. Decomposed prompting for machine translation between related languages using large language models. *arXiv preprint arXiv:2305.13085*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. pages 911–920, Online.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. pages 392–418, Singapore.

Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne. 2016. Syntactically guided neural machine translation. pages 299–305, Berlin, Germany.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. pages 5027–5038, Brussels, Belgium.

Zewei Sun, Qingnan Jiang, Shujian Huang, Jun Cao, Shanbo Cheng, and Mingxuan Wang. 2022. Zero-shot domain adaptation for neural machine translation with retrieved phrase-level prompts. *arXiv preprint arXiv:2209.11409*.

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 *The Twelfth International Conference on Learning Representations*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, Quoc Le, and Denny Zhou. 2022a. Chain of thought prompting elicits reasoning in large language models. *CoRR*, abs/2201.11903.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Thomas Wolf, Julien Chaumond, Lysandre Debut, Victor Sanh, Clement Delangue, Anthony Moi, Pierric Cistac, Morgan Funtowicz, Joe Davison, Sam Shleifer, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2021. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Preprint*, arXiv:2105.13626.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Biao Zhang, Barry Haddow, and Alexandra Birch. 2023a. Prompting large language model for machine translation: A case study. In *International Conference on Machine Learning*, pages 41092–41110. PMLR.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *Preprint*, arXiv:2402.18025.

Xuan Zhang, Navid Rajabi, Kevin Duh, and Philipp Koehn. 2023b. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. Singapore.

Zhong Zhou, Lori Levin, David R. Mortensen, and Alex Waibel. 2020. Using interlinear glosses as pivot in low-resource multilingual machine translation. *Preprint*, arXiv:1911.02709.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

A GlossLM

GlossLM (Ginn et al., 2024) is a specialised gloss generation model trained on IGT corpora. To implement GlossLM, the authors used the ByT5 model (Xue et al., 2021). They continually pre-train the ByT5 model on their GlossLM data that consists of different IGT corpora. Their data includes 1.8k languages ranging from low- to high-resource. These languages are all included in their pre-training split; there are no separate development or test splits.

After this pre-training phase, the model is fine-tuned on endangered languages across the 2023 SIGMORPHON Shared Task dataset Ginn et al. (2023). This latter dataset has train, development, and test splits.

For our evaluation, in addition to the endangered languages, we are also interested in assessing low- to high-resource languages such as Swahili and Portuguese. To achieve this, we used most of the GlossLM training split as our test set (details in Section 4.3). As a result, we did not perform experiments with the model-gloss strategy for low- to high-resource languages, since this strategy leverages the GlossLM model and we are testing on the same corpus used for training GlossLM. Otherwise, GlossLM would just produce glosses over data it was trained on, biasing our results.

The authors directly provided the predictions over the test split of the SIGMORPHON Shared Task at `https://github.com/foltaProject/glosslm/tree/main/preds/glosslm-all-no_trans`. For Table 2, we used these predictions generated from the prompt below:

```
Provide the glosses for the transcription
in <lang>.

Transcription in <lang>: <transcription>
Transcription segmented: <yes/no/unknown>

Glosses:
```

We note GlossLM also offers a version that incorporates the translation in its prompt to generate glosses. Since we are interested in obtaining glosses specifically for translations, we chose to use the version of the model that excludes the translation (i.e., thus selecting "no-train" from `https://github.com/foltaProject/glosslm/tree/main/preds/glosslm-all-no_trans`).

Moreover, the authors also released the fine-tuned models without translations on huggingface through this link: `https://huggingface.co/lecslab`. We use their models to get the glosses for Flores (Section 5), as well as, to get the glosses for the ablation studies in Section 6 over the validation split of the SIGMORPHON Shared Task. Specifically, we used `lecslab/glosslm-gitx-all-no_trans`, `lecslab/glosslm-lezg-all-no_trans`, `lecslab/glosslm-natu-all-no_trans`, and `lecslab/glosslm-dido-all-no_trans` for Gitksan, Lezgi, Natugu and Tsez, respectively.

B xCOMET

Table 7 reports xCOMET-XXL (Guerreiro et al., 2024) scores for all languages using the `Unbabel/XCOMET-XXL` version available at the HuggingFace hub `https://huggingface.co/Unbabel/XCOMET-XXL`. Again, results for the model-gloss strategy are not provided for low- to high-resource languages, since the glosses are predicted by the GlossLM model, which was exposed to the GlossLM data during pre-training (i.e., to avoid unfair evaluation).

C Other baselines

We also considered the few-shot strategy of parallel dictionary, following Ghazvininejad et al. (2023) that prompts the LLM with the dictionary translations like so: “the word X means A; the word Y means B,C,D”. We report results for two high-resource languages using bilingual lexicons provided in Conneau et al. (2018), following Ghazvininejad et al. (2023) setup. We note that this baseline is also hard to fully compare against ours, as word-by-word mapping from Conneau et al. (2018) is unavailable for the unseen endangered languages and the low-resource languages used therefore we only show results for Portuguese and Russian. Results show that translation benefits more from glosses than dictionaries.

Similar to this baseline, in Section 6, we removed all grammatical labels such as "1SG", leaving only the (semantically full) lemmata, and observed a drop in performance (Table 8). This again suggests that there are gains from using more information than word-by-word translations, and that grammatical information plays a positive role.

Method	xCOMET-XXL																
	Git	Lez	Nat	Tse	Swa	Yor	Ice	Mar	Kan	Urd	Tha	Gre	Por	Jap	Rus	Ara	Avg.
NLLB-200	14.09	11.56	13.13	12.51	27.19	17.82	27.78	13.33	14.93	16.86	16.95	15.19	66.60	18.52	19.30	16.36	20.13
zero-shot	18.32	14.19	15.10	13.24	40.05	18.55	38.53	15.96	22.43	27.03	17.69	25.22	80.08	29.02	48.48	20.41	27.77
zero-CoT	17.34	13.77	14.35	12.55	39.88	22.03	36.13	16.39	24.38	28.46	18.00	29.33	75.08	29.57	45.92	19.36	27.66
few-shot	20.46	15.92	16.40	14.25	43.92	32.52	38.26	28.24	29.65	46.12	22.10	28.56	84.04	36.86	49.20	20.31	32.92
gloss-shot	22.10	18.47	17.23	15.04	45.79	33.48	39.62	28.94	30.42	46.60	21.75	28.33	84.11	37.04	49.73	19.79	33.65
chain-gloss	19.75	17.84	16.71	12.81	44.90	36.63	35.96	29.97	31.41	47.53	21.92	27.41	84.23	38.24	50.53	20.37	33.51
model-gloss	48.72	36.51	40.19	37.88	-	-	-	-	-	-	-	-	-	-	-	-	40.83

Table 7: xCOMET-XXL across all languages. Results for the model-gloss strategy are not provided for low- to high-resource languages, as the GlossLM model used in this approach was exposed to GlossLM data during pre-training.

Method	BLEU		chrF++	
	Por	Rus	Por	Rus
Dict	35.80	22.09	57.90	43.44
Gloss-shot	**44.37**	23.99	**63.72**	48.13
Chain-gloss	42.88	**27.52**	62.33	**49.30**

Table 8: GRAMMAMT compared to the parallel dictionary baseline on the GlossLM data.

D Segmentation

We further explore the use of morphological segmentation, which is also commonly adopted in IGT, where sentences may be accompanied both by the gloss as well as its segmentation. In this setup, we propose *seg-shot*, where instead of the gloss of the input sentence, we use morphological segmentation, as illustrated below:

1. `Source:` Juma alimpiga risasi tembo jana usiku .
2. `Segmentation:` Juma a-li-m-pig-a risasi tembo jana usiku
3. `Translation:` Juma shot an/the elephant last night.

In Table 9, we observe that seg-shot improves gloss-shot on Natugu, Greek and Arabic. We then combined glosses and segmentation in our prompts (*gloss w/ seg*) and found performance improvement on both gloss-shot and seg-shot for three languages (Gitksan, Marathi and Russian), suggesting that prompting strategies may be language specific. We also use segmentation in the chain-of-segmentation set-up (*chain-seg*), similarly to chain-gloss, and find that while on average chain-gloss outperforms chain-seg, chain-seg is competitive and outperforms the remaining methods. These improvements provide motivation for GRAMMAMT to be explored with other grammatical augmentations.

E Model Size

Previously, in Table 6, we reported the performance of models beyond Llama-3 70B, including Llama-3 8B, Mixtral-8x22B, and GPT-4o, on unseen languages. Here, we present results for the remaining languages in Table 10 on the GlossLM data, excluding GPT-4o to avoid additional costs. Across low- to high-resource languages, we again observe consistent improvements with the smaller models. Mixtral, in particular, shows substantial gains with the chain-gloss strategy. Similarly, Llama-3 8B benefits from chain-gloss over few-shot for most low-resource languages. This is particularly attractive since most low-resource languages often face double-bind (Ahia et al., 2021) of compute and data. The success of smaller models doing well with chain-gloss and gloss-shot means a lower barrier to achieving good translation for these languages.

F FLORES chrF++

Here we report chrF++ results over the FLORES test set. chrF++ performance is consist with BLEU scores; we also observe improvements of chrF++ for Swahili, Icelandic, Greek, Portuguese, Japanese and Russian (Table 11 and Figure 7).

G Languages

We discuss the various languages we consider below:

Unseen, Endangered languages. Gitksan, Lezgi, Natugu, and Tsez languages cover a diverse range of linguistics characteristics. Specifically, Gitksan language is polysynthetic with Verb-Subject-Object word order whereas Natugu languages is analytic with Subject-Verb-Object word order. Lezgi and Tsez are both agglutinative and use the Subject-Object-Verb word order.

Method	BLEU																
	Git	Lez	Nat	Tse	Swa	Yor	Ice	Mar	Kan	Urd	Tha	Gre	Por	Jap	Rus	Ara	Avg.
gloss-shot	4.96	5.81	1.32	**1.55**	22.20	**16.32**	3.50	17.53	22.40	26.86	6.26	9.56	**44.37**	13.65	23.99	5.60	14.12
seg-shot	2.23	5.96	**2.38**	1.32	22.15	13.27	3.56	18.59	24.58	28.04	7.00	**13.44**	43.60	13.30	25.85	**6.20**	14.47
gloss w/ seg	5.20	5.65	1.81	**1.55**	21.67	14.56	3.63	**18.71**	21.28	28.54	6.89	11.38	43.93	12.90	26.01	4.56	14.27
chain-gloss	**5.84**	**7.30**	2.35	1.49	**23.54**	14.10	5.11	17.32	**25.26**	28.71	**8.37**	10.74	42.88	14.78	27.52	4.51	**15.00**
chain-seg	5.17	6.68	2.00	1.05	23.34	13.08	**6.38**	16.35	23.59	**28.91**	7.91	11.18	43.04	**15.40**	**29.25**	3.30	14.79

Table 9: The effect of augmenting GRAMMAMT with other grammatical information than glosses. We find that morphological segmentation can be a viable alternative to annotated glosses.

Method	Model	BLEU											
		Swa	Yor	Ice	Mar	Kan	Urd	Tha	Gre	Por	Jap	Rus	Ara
few-shot	Llama-3 70B	22.35	11.98	**6.43**	**19.19**	23.50	26.19	7.68	10.62	44.14	13.72	24.95	5.35
gloss-shot	Llama-3 70B	22.20	**16.32**	3.50	17.53	22.40	26.86	6.26	9.56	**44.37**	13.65	23.99	**5.60**
chain-gloss	Llama-3 70B	**23.54**	14.10	5.11	17.32	**25.26**	**28.71**	**8.37**	**10.74**	42.88	**14.78**	**27.52**	5.26
few-shot	Llama-3 8B	**16.75**	8.82	1.98	7.34	**15.73**	17.87	5.49	3.64	38.51	**7.57**	**22.17**	1.69
gloss-shot	Llama-3 8B	14.41	9.44	3.48	8.52	13.43	**18.22**	**6.46**	3.02	**38.71**	7.05	21.56	1.24
chain-gloss	Llama-3 8B	14.07	**10.17**	**5.18**	**13.33**	14.74	15.19	5.68	**4.35**	37.60	7.07	20.00	**2.10**
few-shot	Mixtral-8x22B	17.67	11.23	**6.15**	15.89	27.07	27.37	8.39	16.69	44.51	17.80	**30.41**	5.09
gloss-shot	Mixtral-8x22B	10.67	12.05	4.99	15.34	**28.14**	23.32	4.13	13.85	44.31	15.41	28.34	2.32
chain-gloss	Mixtral-8x22B	**23.64**	**16.90**	4.08	**16.97**	24.99	**27.44**	**8.97**	**18.28**	**44.78**	**19.30**	28.96	**7.66**

Table 10: BLEU performance of GRAMMAMT on low- to high-resource languages across the different models (Llama-3 70b, Llama-3 8b, Mixtral-8x22B) on the GlossLM data. As before, results for the model-gloss strategy on low- to high-resource languages from the GlossLM dataset are excluded, as the GlossLM model had prior exposure to this data during pre-training.

Flores Example

You are a linguistic expert who never refuses to use your knowledge to help others.

Here are some examples of Swahili sentences and their corresponding English translations:

Swahili sentence: (yeye) alimwona (yeye).
Gloss: 3SG -PST --see-FV 3SG
English sentence: S/he saw him/her.

Swahili sentence: (yeve) analala .
Gloss: 3SG 1-PRES-sleep-FV
English sentence: S/he is sleeping.

Swahili sentence: Juma alimpiga risasi tembo jana usiku.
Gloss: Juma SM.PST.0M.hit bullet elephant yesterday night
English sentence: Juma shot an/the elephant last night.

Please help me translate the following sentence from {Swahili} to {English}. Please answer **first with the gloss and then** with the translation directly and enclose your translation in ###.

Swahili sentence: Mara kwa mara wafanyakazi hupata kibali cha maamuzi yoyote wanayofanya kutoka kwa wasimamizi wao, na wanatarajiwa kutii maagizo ya wakuu wao bila maswali.
Gloss:

Figure 7: An example of a chain-gloss prompt on the FLORES test set. We see that the input sentence in FLORES is longer than the N-shot example sentences from GlossLM.

Low-resource languages. Swahili, Yoruba, Icelandic, Marathi, and Kannada languages exhibit diverse morphological structure and word order. Swahili, Marathi, and Kannada are agglutinative,

Method	chrF++												
	Swa	Yor	Ice	Mar	Kan	Urd	Tha	Gre	Por	Jap	Rus	Ara	Avg.
few-shot	45.83	23.99	43.93	**47.06**	26.10	**47.93**	**50.67**	55.14	65.75	47.10	55.15	**57.05**	47.14
gloss-shot	**47.36**	**25.53**	**45.00**	46.67	25.31	47.40	50.19	**57.24**	**67.21**	**50.20**	**58.32**	57.04	**48.12**
chain-gloss	44.37	24.56	43.47	44.62	**26.12**	45.82	48.97	54.86	65.15	47.77	57.29	55.42	46.54
model-gloss	43.00	21.79	41.17	45.11	23.10	46.30	49.50	56.62	67.33	49.85	**58.32**	56.54	46.55

Table 11: chrF++ performance on the Flores test set.

Yoruba is analytic, and Icelandic is fusional. In terms of word order, Swahili, Yoruba and Icelandic are characterised by a Subject-Verb-Object order while Marathi and Kannada by Subject-Object-Verb.

Mid-to-high-resource languages. We also experiment on 7 mid-to-high-resource languages namely: Urdu, Thai, Greek, Portuguese, Japanese, Russian, and Arabic. Urdu, Greek, Portuguese, Russian have fusional mophological typology. Japanese is agglutinative while Thai is analytic. In terms of word order, all languages have a Subject-Verb-Object order, except Urdu and Arabic, which follow Subject-Object-Verb and Verb-Subject-Object orders respectively.

H Gloss Performance

Here we also report morpheme/lexeme level accuracy and chrF++ metrics for the glosses generated by chain-gloss and model-gloss (Table 12).

I Reverse translation (en →)

To address translations from English, we implemented a strategy where, given source-gloss-target triples (**x**, **g**, **y**), we swap the source and target languages in our prompts (**y**, **g**, **x**). This means that instead of using the gloss for the input sentence, we now use the gloss of the target language to guide the translation process. Here is an example:

> Swahili Sentence: [source sentence];
> Gloss: [gloss];
> A translation for this Swahili sentence in English is: [translation].

This changes to:

> English Sentence: [target sentence];
> **Swahili Gloss**: [gloss];
> A translation for this English sentence in Swahili is: [translation].

J Significance test

To show the significance of our results, we ran additional evaluation and report the statistical significance results with paired bootstrap resampling using sacreBLEU (Koehn, 2004). We compared few-shot and GRAMMAMT and find that in the unseen languages GRAMMAMT, particularly model-gloss, demonstrates a statistically significant performance improvement compared to few-shot. See Tables 13, 14 and 15. We do not report results for the model-gloss strategy on low- to high-resource languages from the GlossLM data, as the GlossLM model was exposed to this data during pre-training.

K Qualitative Examples

Table 16 shows qualitative examples from the Leiz language across the different methods. For larger N-shot settings (N=45), all our methods correctly used the past verb tenses ("the mother was" and "the father was"), whereas the few-shot method incorrectly used the present tense ("my mother is"). When N=3, it becomes evident that our strategies require a sufficient number of examples to perform well, which aligns with the overall qualitative results. For instance, at N=3, the gloss-shot method incorrectly generated "1," likely due to confusion with gloss annotations (e.g., 1SG), and chain-gloss failed now to produce a correct gloss (while successfully identified the verb as past tense (PST) at N=45). For smaller N, the model-gloss strategy proves more robust, as it consistently uses the correct past tense by leveraging a model that generates more reliable glosses.

In Table 17, similar to the qualitative results, we observe that, in larger N-shot settings (N=45), both gloss-shot and model-gloss, guided by glosses from the source sentence, tend to generate better translations than few-shot or gloss-shot. However, for N=3, few-shot, gloss-shot, and chain-gloss struggle to produce meaningful sentences in this endangered language due to insufficient exposure to the language by the LLM. This underscores the impor-

Method	Morpheme accuracy					chrF++				
	Git	**Lez**	**Ntu**	**Ddo**	Avg.	**Git**	**Lez**	**Ntu**	**Ddo**	Avg.
Llama-3 70b	6.95	11.46	10.40	3.86	8.17	22.88	23.81	28.67	22.83	16.53
GlossLM	**15.48**	**44.12**	**59.53**	**85.50**	**51.16**	**37.00**	**60.12**	**76.28**	**90.66**	**66.02**

Table 12: Morpheme/lexeme level accuracy and chrF++ scores for the glosses generated by Llama-3 70b (chain-gloss) compared to GlossLM (model-gloss).

Language	BLEU			
	Few-shot	Gloss-shot	Chain-gloss	Model-gloss
Gitksan	4.5 ± 3.0	**4.8 ± 3.0 (p = 0.1508)**	5.5 ± 3.2 (p = 0.0150)*	18.2 ± 6.4 (p = 0.0010)*
Lezgi	6.2 ± 5.5	**5.7 ± 4.8 (p = 0.1029)**	**7.2 ± 5.3 (p = 0.1219)**	13.9 ± 6.0 (p = 0.0010)*
Natugu	3.3 ± 1.4	1.3 ± 0.3 (p = 0.0030)*	**2.2 ± 1.2 (p = 0.0999)**	17.0 ± 3.2 (p = 0.0010)*
Tsez	1.3 ± 0.4	**1.5 ± 0.4 (p = 0.1349)**	**1.5 ± 0.5 (p = 0.2088)**	14.2 ± 1.2 (p = 0.0010)*
Swahili	22.4 ± 3.2	**22.2 ± 3.1 (p = 0.3586)**	**23.5 ± 3.0 (p = 0.1229)**	-
Yoruba	11.9 ± 4.3	16.1 ± 5.1 (p = 0.0060)*	**14.0 ± 4.7 (p = 0.1149)**	-
Icelandic	6.5 ± 6.0	**3.9 ± 4.4 (p = 0.0729)**	**5.0 ± 4.9 (p = 0.2068)**	-
Marathi	19.1 ± 7.7	**19.1 ± 7.7 (p = 0.1938)**	**17.0 ± 7.2 (p = 0.2028)**	-
Kannada	23.5 ± 4.2	**22.4 ± 4.4 (p = 0.1269)**	25.2 ± 4.1 (p = 0.0260)*	-
Urdu	26.3 ± 4.2	**26.9 ± 4.3 (p = 0.1568)**	28.8 ± 4.3 (p = 0.0160)*	-
Thai	7.6 ± 2.6	6.3 ± 2.4 (p = 0.0020)*	**8.2 ± 2.7 (p = 0.0529)**	-
Greek	10.6 ± 4.2	**9.5 ± 4.8 (p = 0.1828)**	**10.7 ± 4.9 (p = 0.4026)**	-
Portuguese	44.2 ± 4.4	**44.5 ± 4.3 (p = 0.3197)**	**42.9 ± 4.2 (p = 0.1548)**	-
Japanese	13.7 ± 0.6	**13.7 ± 0.7 (p = 0.2937)**	15.4 ± 0.7 (p = 0.0010)*	-
Russian	25.0 ± 1.3	24.0 ± 1.5 (p = 0.0300)*	27.8 ± 1.2 (p = 0.0010)*	-
Arabic	5.4 ± 2.7	**5.5 ± 3.0 (p = 0.3007)**	**5.2 ± 3.9 (p = 0.3906)**	-

Table 13: BLEU statistical significance test of all languages with the null hypothesis: mean score of few-shot is equal to the mean of GRAMMAMT. The values with the asterisks (p-value < 0.05) show that few-shot is significantly different from GRAMMAMT, while the values with p-value > 0.05 (bolded values) indicate that GRAMMAMT is equivalent to few-shot. Results for the model-gloss strategy on low- to high-resource languages from the GlossLM data are omitted, as the GlossLM model had prior exposure to this data during pre-training.

tance of model-gloss, which leverages an external gloss generation model to guide the LLM more effectively, resulting in improved translation quality. Additional examples in Table 18 for N=3 further reveal that, apart from model-gloss, few-shot and other strategies perform poorly in generating translations, underscoring the importance of having a sufficient number of examples.

L Prompt-Template

Our prompt follows the LingoLLM (Zhang et al., 2024) template, starting with a system message that sets the LLM into a linguistic mode: "You are a linguistic expert who never refuses to use your knowledge to help others.". We also request in the prompt that the model encloses its translation. For the baselines and our proposed prompting strategies, we ensure that the prompt is as similar as possible by including the same prefix and suffix: "Here are some examples of {language} sentences and their corresponding English translations:" and "A translation for this {language} sentence in English is:}". We just make minimal changes depending on the specific prompting strategy. For example, the zero-shot strategy does not include examples. In gloss-shot, we provide the gloss, while in chain-gloss, we ask the model to generate the gloss first. We show below the Swahili prompt for the different strategies. For other languages, it can be tailored by naming the corresponding language. See the prompt templates we used in Figures 8 and 9.

Language	**chrF++**			
	Few-shot	Gloss-shot	Chain-gloss	Model-gloss
Gitksan	25.1 ± 3.0	**25.8 ± 3.8 (p = 0.1638)**	**24.6 ± 4.1 (p = 0.1948)**	47.7 ± 3.9 (p = 0.0010)*
Lezgi	23.0 ± 4.3	**23.2 ± 4.2 (p = 0.2897)**	**22.7 ± 4.3 (p = 0.2567)**	39.6 ± 4.0 (p = 0.0010)*
Natugu	19.4 ± 1.3	**20.2 ± 1.3 (p = 0.0639)**	**19.2 ± 1.3 (p = 0.2468)**	41.5 ± 2.8 (p = 0.0010)*
Tsez	19.9 ± 0.5	20.8 ± 0.5 (p = 0.0010)*	17.9 ± 0.6 (p = 0.0010)*	42.3 ± 1.0 (p = 0.0010)*
Swahili	45.7 ± 2.9	**46.4 ± 2.9 (p = 0.1139)**	**45.4 ± 2.8 (p = 0.2607)**	-
Yoruba	29.8 ± 3.9	33.2 ± 4.2 (p = 0.0010)*	33.5 ± 4.1 (p = 0.0050)*	-
Icelandic	29.0 ± 8.2	26.1 ± 8.8 (p = 0.0090)*	**24.9 ± 8.1 (p = 0.0739)**	-
Marathi	36.0 ± 7.2	**36.0 ± 6.7 (p = 0.4066)**	**35.2 ± 7.0 (p = 0.2957)**	-
Kannada	44.2 ± 3.6	42.7 ± 3.7 (p = 0.0230)*	46.3 ± 3.5 (p = 0.0030)*	-
Urdu	43.5 ± 3.8	**43.6 ± 3.9 (p = 0.3417)**	45.9 ± 3.8 (p = 0.0060)*	-
Thai	19.8 ± 2.5	**19.3 ± 2.4 (p = 0.1159)**	**19.8 ± 2.6 (p = 0.3427)**	-
Greek	27.7 ± 4.8	**27.3 ± 5.0 (p = 0.2597)**	**27.2 ± 5.1 (p = 0.3177)**	-
Portuguese	63.9 ± 3.2	**63.8 ± 3.1 (p = 0.2747)**	62.4 ± 3.1 (p = 0.0260)*	-
Japanese	35.9 ± 0.6	**35.7 ± 0.6 (p = 0.0769)**	37.2 ± 0.6 (p = 0.0010)*	-
Russian	48.6 ± 1.0	**48.1 ± 1.1 (p = 0.0569)**	50.2 ± 1.1 (p = 0.0010)*	-
Arabic	21.5 ± 3.7	**21.3 ± 3.5 (p = 0.3726)**	**19.7 ± 5.1 (p = 0.0619)**	-

Table 14: chrF++ statistical significance test of all languages with the null hypothesis: mean score of few-shot is equal to the mean of GRAMMAMT. The values with the asterisks (p-value < 0.05) show that few-shot is significantly different from GRAMMAMT, while the values with p-value > 0.05 (bolded values) indicate that GRAMMAMT is equivalent to few-shot. We exclude results for the model-gloss strategy on low- to high-resource languages from the GlossLM data, as the GlossLM model used in this approach had prior exposure to this data during pre-training.

Language	xCOMET			
	Few-shot	Gloss-shot	Chain-gloss	Model-gloss
Gitksan	20.41	**22.23 (p = 0.1216)**	**19.92 (p = 0.4456)**	48.82 (p = 0.0000)*
Lezgi	15.95	**18.36 (p = 0.0128)**	**17.74 (p = 0.0872)**	36.42 (p = 0.0000)*
Natugu	16.31	17.22 (p = 0.2637)*	**16.69 (p = 0.6610)**	40.10 (p = 0.0000)*
Tsez	14.33	15.13 (p = 0.0002)*	12.84 (p = 0.0000)**	37.90 (p = 0.0000)*
Swahili	43.80	45.76 (p = 0.0008)*	**44.86 (p = 0.0855)**	-
Yoruba	32.71	**33.60 (p = 0.4145)**	36.72 (p = 0.0069)*	-
Icelandic	38.09	**39.34 (p = 0.4734)**	**35.69 (p = 0.3714)**	-
Marathi	28.02	**28.68 (p = 0.4898)**	**29.70 (p = 0.2239)**	-
Kannada	29.69	**30.49 (p = 0.1194)**	31.43 (p = 0.0057)*	-
Urdu	46.29	**46.76 (p = 0.3772)**	47.71 (p = 0.0472)*	-
Thai	22.09	**21.75 (p = 0.4797)**	**21.93 (p = 0.6982)**	-
Greek	28.43	**28.28 (p = 0.8618)**	**27.56 (p = 0.5252)**	-
Portuguese	84.15	**84.19 (p = 0.9342)**	**84.31 (p = 0.7712)**	-
Japanese	36.89	**37.06 (p = 0.2375)**	38.27 (p = 0.0000)*	-
Russian	49.20	49.76 (p = 0.0152)*	50.53 (p = 0.0000)*	-
Arabic	20.25	**19.86 (p = 0.4800)**	**20.43 (p = 0.9308)**	-

Table 15: xCOMET statistical significance test of all languages with the null hypothesis: mean score of few-shot is equal to the mean of GRAMMAMT. The values with the asterisks (p-value < 0.05) show that few-shot is significantly different from GRAMMAMT, while the values with p-value > 0.05 (bolded values) indicate that GRAMMAMT is equivalent to few-shot. All values with asterisk indicate GRAMMAMT is better that few-shot. Values with double asterisks(**) show few-shot being better than GRAMMAMT. Results for the model-gloss strategy on low- to high-resource languages from the GlossLM data are not included, as the GlossLM model used in this approach had prior exposure to the data during pre-training.

Method	Output (N=45)	Output (N=3)
few-shot	My mother is a teacher, my father is an accountant.	My mother is a teacher, for my father is an accountant.
gloss-shot	My mother was a teacher, my father was an accountant.	My mother was a teacher, to my father was an accountant.
chain-gloss	Gloss: My mother teacher cop-PST, my father accountant cop-PST. Translation: My mother was a teacher, my father was an accountant.	Gloss: my mother teacher-ERG be-AOR, my father accountant-ERG be-AOR. Translation: My mother is a teacher, and my father is an accountant.
model-gloss	Gloss: my mother teacher was, my uncle-FOC province-PL was Translation: My mother was a teacher, my uncle was a provincial accountant.	Gloss: my mother teacher was, my uncle-FOC province-PL was Translation: My mother was a teacher, a relative of my uncle's was a provincial accountant.
Reference: "My mother was a teacher, my father was an accountant."		

Table 16: Comparison of methods for N=45 and N=3.

Method	Output (N=45)	Output (N=3)
few-shot	"I," he said, "will go to the city, and I'll ask for a meeting with the inspector."	"Here," said the mother-in-law, "a human being was killed in our yard, to the east of our house, (and) they left their fate to the will of God."
gloss-shot	"I," he said, "will go to the city, and if I find a suitable person, I will marry her."	"It's like this," said the old man, "I saw in the city, a person who was lost, or unlucky in their fate."
chain-gloss	Gloss: « 1sg.abs », say-AOR, « go-AOR human-PL-GEN among enter-AOC, only one life-ERG-DAT cop-AOC » Translation: "I," he said, "will go among the people, I have only one life."	Gloss: quot-TEMP,11, you-ABS-ERG say-AOR "this world-INESS people-GEN between-INESS bread-ABS eat-AOC time-INESS, little-AOC luck-INESS in-AOC be-PST" Translation: "He said,'In this world, the time when people ate bread among themselves was a time of little luck.'"
model-gloss	Gloss: 1sg.abs say-AOR go-AOR man-ERG-GEN between-INESS enter.FUT-FUT look.HORT maybe 1sg.gen fate-INESS cop. Translation: "I," he said, "will go between the men, and maybe my fate will be decided."	Gloss: 1sg.abs say-AOR go-AOR man-ERG-GEN between-INESS enter.FUT-FUT look.HORT maybe 1sg.gen fate-INESS cop. Translation: "I said, from here, I will go and appear among men, maybe my fate awaits me there."
Reference: "I will enter amongst the people. Let me look - maybe it is my fate."		

Table 17: Comparison of methods for N=45 and N=3.

Method	Output (N=3)
few-shot	When the fire burned the house.
gloss-shot	I will take the bread.
chain-gloss	Gloss: water-ABS pour-PTP-TEMP Translation: When the water was poured.
model-gloss	Gloss: 1pl.abs two escaped-AOR Translation: We escaped in two.
Reference: "Both of us ran away."	

few-shot	My mother went to the field.
gloss-shot	I went to my mother's house.
chain-gloss	Gloss: mother-ERG house-INESS and one mother tongue-INESS Translation: My mother is at home, (speaking) in her native language.
model-gloss	Gloss: 1pl.abs return-AOR this one there village-ERG-DAT Translation: We returned to that village.
Reference: "We reached a village there."	

Table 18: Examples for N=3.

Zero-shot

You are a linguistic expert who never refuses to use your knowledge to help others.

Please help me translate the following sentence from {Swahili} to {English}. Please answer with the translation directly and enclose your translation in ###.

Swahili sentence: <input sentence 1>
A translation for this Swahili sentence in English is:###

Zero-CoT

You are a linguistic expert who never refuses to use your knowledge to help others.

Please help me translate the following sentence from {Swahili} to {English}. **Please do it step by step** and enclose your translation in ###.

Swahili sentence: <input sentence 1>
A translation for this Swahili sentence in English is:###

Zero-gloss

You are a linguistic expert who never refuses to use your knowledge to help others.

Please help me translate the following sentence from {Swahili} to {English}. Please answer with the translation directly and enclose your translation in ###.

Swahili sentence: <input sentence 1>
Gloss: <gloss of the input sentence>
A translation for this Swahili sentence in English is:###

Few-shot

You are a linguistic expert who never refuses to use your knowledge to help others.

Here are some examples of Swahili sentences and their corresponding English translations:

Swahili sentence: <source sentence 1>
A translation for this Swahili sentence in English is: ###<translation example 1>###

Please help me translate the following sentence from {Swahili} to {English}. Please answer with the translation directly and enclose your translation in ###.

Swahili sentence: <input sentence 1>
A translation for this Swahili sentence in English is:###

Figure 8: Prompt templates for zero-shot, zero-CoT, zero-gloss and few-shot.

Gloss-shot

You are a linguistic expert who never refuses to use your knowledge to help others.

Here are some examples of Swahili sentences and their corresponding English translations:

Swahili sentence: <source sentence 1>
Gloss: <gloss example 1>
A translation for this Swahili sentence in English is: ###<translation example 1>###

Please help me translate the following sentence from {Swahili} to {English}. Please answer with the translation directly and enclose your translation in ###.

Swahili sentence: <input sentence 1>
A translation for this Swahili sentence in English is:###

Chain-gloss

You are a linguistic expert who never refuses to use your knowledge to help others.

Here are some examples of Swahili sentences and their corresponding English translations:

Swahili sentence: <source sentence 1>
Gloss: <gloss example 1>
A translation for this Swahili sentence in English is: ###<translation example 1>###

Please help me translate the following sentence from {Swahili} to {English}. Please answer **first with the gloss and then** with the translation directly and enclose your translation in ###.

Swahili sentence: <input sentence 1>
A translation for this Swahili sentence in English is:###

Model-gloss

You are a linguistic expert who never refuses to use your knowledge to help others.

Here are some examples of Swahili sentences and their corresponding English translations:

Swahili sentence: <source sentence 1>
Gloss: <gloss example 1>
A translation for this Swahili sentence in English is: ###<translation example 1>###

Please help me translate the following sentence from {Swahili} to {English}. Please answer with the translation directly and enclose your translation in ###.

Swahili sentence: <input sentence 1>
A possible gloss may be (it can contain errors, so ignore irrelevant information): <input gloss example>
A translation for this Swahili sentence in English is:###

Oracle-gloss

You are a linguistic expert who never refuses to use your knowledge to help others.

Here are some examples of Swahili sentences and their corresponding English translations:

Swahili sentence: <source sentence 1>
Gloss: <gloss example 1>
A translation for this Swahili sentence in English is: ###<translation example 1>###

Please help me translate the following sentence from {Swahili} to {English}. Please answer with the translation directly and enclose your translation in ###.

Swahili sentence: <input sentence 1>
Gloss: <gold gloss of the input sentence>
A translation for this Swahili sentence in English is:###

Figure 9: Prompt templates for gloss-shot, chain-gloss, model-gloss and oracle-gloss.