# GrammaMT👵: Improving Machine Translation with Grammar-Informed In-Context Learning

Anonymous ACL submission

## Abstract

We introduce GrammaMT, a grammatically-aware prompting approach for machine translation that uses Interlinear Glossed Text (IGT), a common form of linguistic annotation describing lexical and functional morphemes of source sentences. GrammaMT proposes two prompting strategies: gloss-shot and chain-gloss. Both are training-free, require only a few examples, and involve minimal effort to collect, making them well-suited for low-resource setups. Experiments and ablation studies on open-source instruction-tuned LLMs, across three different benchmarks, demonstrate the benefits of leveraging interlinear gloss resources for machine translation. GrammaMT improves the translation performance for various low-resource to high-resource languages in the largest existing corpus of IGT data, on the challenging 2023 SIGMORPHON Shared Task data across rarely-seen, endangered languages, and even in an out-of-domain setting within FLORES.

## 1 Introduction

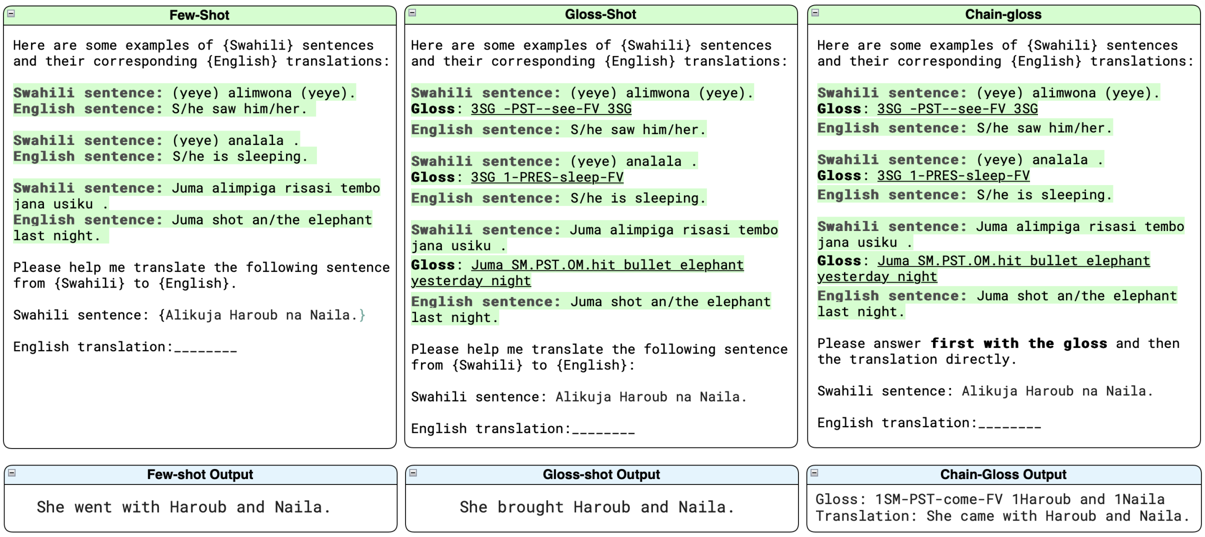
Large Language Models (LLMs) have taken over the NLP leaderboards (e.g., Zellers et al., 2019; Hendrycks et al., 2020; Li et al., 2023b). Training LLMs requires access to a plethora of datasets, a luxury accessible to only a few of the world's most high-resource languages. Consequently, for only a sliver of the world's languages, LLMs have been trained that have seen sufficient, let alone any, quantity of data to result in these impressive performance gains (Achiam et al., 2023; Üstün et al., 2024). To leverage the power 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 preferably (iii) **ease of 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 outlined 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 prevalent linguistic resource that represents input sentences as a sequence of functional and lemma morphemes, as illustrated in Figure 1.

Specifically, GrammaMT introduces two prompting strategies that augment few-shot machine translation using annotated glosses: gloss-shot and chain-gloss. In the gloss-shot strategy, an LLM is prompted with a small number of sentences that are paired not only with their translations but also with their glosses. In the chain-gloss strategy, the LLM first generates a gloss for the input sentence, breaking it down into its functional and lemma morphemes, before proceeding to produce the final translation. Importantly, GrammaMT adheres to all three design requirements:

**Training-free.** GrammaMT is training-free, working by simply prompting an instruction-tuned LLM with a grammatically annotated demonstration. This is especially important in low-resource 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 other-



Few-Shot

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

Swahili sentence: (yeye) alimwona (yeye).
English sentence: S/he saw him/her.

Swahili sentence: (yeye) analala .
English sentence: S/he is sleeping.

Swahili sentence: Juma alimpiga risasi tembo
jana usiku .
English sentence: Juma shot an/the elephant
last night.

Please help me translate the following sentence
from {Swahili} to {English}.

Swahili sentence: {Alikuja Haroub na Naila.}

English translation:________
```

Gloss-Shot

```
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: (yeye) 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.OM.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}:

Swahili sentence: Alikuja Haroub na Naila.

English translation:________
```

Chain-gloss

```
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: (yeye) 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.OM.hit bullet elephant
yesterday night
English sentence: Juma shot an/the elephant
last night.

Please answer first with the gloss and then
the translation directly.

Swahili sentence: Alikuja Haroub na Naila.

English translation:________
```

Few-shot Output

```
She went with Haroub and Naila.
```

Gloss-shot Output

```
She brought Haroub and Naila.
```

Chain-Gloss Output

```
Gloss: 1SM-PST-come-FV 1Haroub and 1Naila
Translation: She came with Haroub and Naila.
```

Figure 1: GRAMMAMT augments few-shot learning with Interlinear Gloss Text. In its gloss-shot strategy, the LLM is conditioned on a set of machine translation pairs associated with the glosses of the source sentences. In the chain-gloss strategy, the model is also prompted to first generate the gloss before producing the translation.

wise be insufficient for fine-tuning an LLM.

**Small number of examples.** GRAMMAMT does not require extensive linguistic annotation, such as grammar chapters or dictionaries that could be hard to fit into the prompt. Instead, our approach requires a minimal set of support examples with grammatical annotations in the form of interlinear glosses (e.g., only 21 examples). This differs from other few-shot methods, like retrieval-augmentation, which depend on acquiring large data stores to gather relevant samples.

**Ease of collection.** Our approach relies on a basic set of gloss annotations, ensuring ease of collection in low-resource settings. These sentences are not limited to linguistic annotators. They can also be sourced from grammar books or automatically generated using a specialised gloss generator model (e.g., Ginn et al., 2024). Our method is is also simpler than chain-of-thought prompting (Wei et al., 2022a). Breaking down a machine translation task into smaller steps demands costly human engineering, as it is less straightforward and more subjective (Wang et al., 2023a), compared to the structured rules of gloss notation.

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, namely LLaMA-3 (Meta, 2024) and Mixtral (Mistral, 2024). We find that GRAMMAMT outperforms or competes with established zero-shot and few-shot baselines 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 intensively explored in various studies (Zhang et al., 2023; Garcia et al., 2023; Peng et al., 2023; Pourkamali and Sharifi, 2024; Jiao et al., 2023; Kadaoui et al., 2023; Wang et al., 2023b). Notably, Zhu et al. (2023) assessed popular LLMs on machine translation for 102 languages and their results show that LLMs perform well for high-resource languages but underperform for low-resource languages. This finding is corroborated in Hendy et al. (2023) and Robinson et al. (2023). The majority of the above studies use zero-shot and few-shot prompting, and state the importance of carefully curating these prompts to get the most out of LLMs. However, these studies do not explore alternative ways of prompting LLMs for translation that go beyond exemplars for unseen and low-resource languages.

**Creatively prompting LLMs for Machine Translation** is crucial for unseen and/or low-resource languages. Gao et al. (2023) find that properly designed prompts improve the performance of LLMs on MT and use parts of speech tags in their prompt. However, they observe partial translation with this technique and leave it to future studies to explore its potential. Moslem et al. (2023) utilise fuzzy matches in the prompts and observe promising results using $N$-shots where $N \geq 7$. Nguyen et al. (2023) proposed having translation examples from different (linguistically-diverse) high-resource languages in prompts to help translation performance on low-resource language whereas Huang et al. (2023) introduced cross-lingual thought (XLT) prompting which, via a generic template prompt, requiring the system to translate the task to English before completing it. Our work follows the same vein of prompting LLMs with an emphasis on using glossary information for high-resource and low-resource languages.

**Using grammatical information with LLM** Modern techniques downplay grammatical information with different studies showing that this information is learnt implicitly by the model (Rogers et al., 2020; Jawahar et al., 2019). However, a few projects have shown that grammatical information introduced both during training and/or inference can improve the performance on models (Strubell et al., 2018; Cui et al., 2022; Stahlberg et al., 2016). Similar to our work, Zhou et al. (2020) use glosses while training to translate to low-resource languages. However, our work focuses on using glosses in a training free approach. Tanzer et al. (2024) incorporate grammatical information in the form of a grammar book and show that an unseen language can be translated by an LLM given the right context. In like manner, Zhang et al. (2024) use a grammar book along with morphologically-augmented input and a dictionary to translate low-resource languages. Our research also leverages morphological information; however, unlike Zhang et al. (2024)'s approach which utilises a morphological analyzer to generate a gloss, we exclusively use gold glosses. Another distinction is that we do not depend on grammar books and dictionaries as these resources are often unavailable in many low-resource languages. Instead, we focus on generating a few gloss examples, which is a more feasible solution for underrepresented languages.

## 3 GRAMMAMT

We propose GRAMMAMT, a simple grammatical-based prompt approach for machine translation, wherein examples of Interlinear Gloss Texts 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)).

**Interlinear Gloss Text Annotation.** GRAMMAMT augments an instruction-tuned LLM with in-context learning examples from simple morphologically annotated glosses, called Interlinear Gloss Text (IGT). IGT labels sentences by indicating their lexical (word-level) and functional (grammatical) morphemes. The corresponding interlinear gloss is placed between the source sentence and its translation, as demonstrated by the following Swahili example, for the sentence *Juma alimpiga risasi tembo jana usiku.* These glosses are the norm in linguistic descriptions of language, and hence very common to find and easy to create.

1. Source: Juma alimpiga risasi tembo jana usiku.
2. Gloss: Juma SM.PST.OM.hit bullet elephant yesterday night
3. Translation: Juma shot an/the elephant last night.

In this example, the morphological annotation SM.PST.OM denotes 'series marker' (SM), 'past tense' (PST), and 'object marker' (OM) and consists of functional morphemes that convey grammatical information. All such functional morphemes are labeled with uppercase letters. In contrast, lexical morphemes that convey semantic meaning are labeled in lowercase with their translation (e.g., *isasi* as *bullet*). 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 of language, and hence very common to find and easy to create.

**Prompting strategies.** We use interlinear glosses in our approach via the following two prompting strategies, illustrated in Figure 1.

In the first prompting strategy, **gloss-shot**, the model 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 model 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 strategies can offer some insights into how the LLM arrived at a specific translation.

We illustrate the format of the prompt in Figure 1 and can be seen in more detail in Appendix D.

## 4 Experimental setup

### 4.1 LLMs

We assess our GrammaMT approach on `Meta-Llama-3-70B-Instruct` (Meta, 2024), the recent instruction-tuned Lamma with 70B parameters. Our machine translation approach does not involve any training, the translations are generated at inference time using a single A100 SXM 80GB GPU. We also experiment with the smaller `Meta-Llama-3-8B-Instruct` variant, and `Mixtral-8x22B-Instruct-v0.1` (Mistral, 2024), a sparse Mixture-of-Experts model [1]. All the pretrained models were loaded with the HuggingFace hub client library (Wolf et al., 2020) using 4-bit quantisation. During inference, the models generate a translation using greedy decoding with the default temperature of 1. We release our code and the translations generated by each model at `anonymous-submission`.

### 4.2 Prompting Variants

We compare the GrammaMT approach against established MT baselines, specifically zero-shot and few-shot prompting strategies. Our approach augments few-shot prompting with grammatical information, where we explore our two novel variants, *gloss-shot* and *chain-gloss* prompting, described above.

1. **zero-shot:** The model is prompted to translate directly from the source language to the desired target language without any prior examples.

2. **few-shot:** The model translates the input sentence using a limited number of pairs of source-target examples.

3. **gloss-shot:** The model predicts based on examples that pair the source sentences not just with their translation but also with their gloss.

|  |  | Sigmorphon dataset |  |  |
|---|---|---|---|---|
| Language | Abbrev. | Script | Test size | Speakers[2] |
| Gitksan | Git | Latin | 37 | 1,110 |
| Lezgi | Lez | Cyrillic | 87 | 800,000 |
| Natugu | Ntu | Latin | 99 | 5,900 |
| Tsez | Ddo | Cyrillic | 445 | 18,000 |
|  |  | GlossLM dataset |  |  |
| Language | Abbrev. | Script | Test size | Speakers[2] |
| Swahili | Swa | Latin | 439 | 200 million |
| Yoruba | Yor | Latin with diac. | 135 | 47 million |
| Icelandic | Ice | Latin | 27 | 330,000 |
| Marathi | Mar | Devanagari | 43 | 83 million |
| Kannada | Kan | Kannada | 388 | 59 million |
| Urdu | Urd | Perso-Arabic | 259 | 232 million |
| Thai | Tha | Thai | 352 | 61 million |
| Greek | Gre | Greek | 59 | 13.5 million |
| Portuguese | Por | Latin | 309 | 264 million |
| Japanese | Jap | Japanese scripts[3] | 4,748 | 123 million |
| Russian | Rus | Cyrillic | 2,444 | 255 million |
| Arabic | Ara | Arabic | 136 | 274 million |

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

4. **chain-shot:** As in gloss-shot, the model receives triples of source, gloss, and target sentence examples to translate the input sentence. However, the model is additionally prompted to generate the gloss for the input sentence itself before its translation.

For each language in our study, we use the same 21 translation examples for each of the prompting strategies outlined above. We use 21 examples since we found to be the optimal value found in our ablation studies in Section 6). See Appendix D for detailed information on the selected prompting templates.

### 4.3 Datasets and Languages

We evaluate translation quality of GrammaMT on three different datasets from English to 16 different languages. These languages fall under endangered languages that are hardly seen during pre-training, as well low-resource languages and mid-to-high-resource languages. Table 1 provides an overview of the languages, scripts and test set sizes.

**Sigmorphon:** We use the dataset from the 2023 SIGMORPHON Shared Task for evaluating on unseen languages Ginn et al. (2023), selecting 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.

[1] Experiments reported in Appendix A.

[2] The number of speakers are referenced from Wikipedia.

[3] Kanji,Hiragana and Katakana.

| Method | BLEU | | | | | chrF++ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Git | Lez | Ntu | Ddo | Avg. | Git | Lez | Ntu | Ddo | Avg. |
| zero-shot | 1.41 | 1.46 | 0.26 | 0.39 | 0.88 | 23.88 | 17.71 | 13.77 | 16.84 | 18.05 |
| few-shot | 4.71 | 6.36 | **3.34** | 1.35 | 3.94 | 25.18 | 22.89 | 19.41 | 19.90 | 21.85 |
| gloss-shot | 4.96 | 5.81 | 1.32 | **1.55** | 3.41 | **25.87** | **23.08** | **20.24** | **20.79** | **22.50** |
| chain-gloss | **5.84** | **7.30** | 2.35 | 1.49 | **4.25** | 23.65 | 22.62 | 19.19 | 17.89 | 20.84 |

Table 2: BLEU and chrF++ performance on **unseen/endangered languages** using the 2023 SIGMORPHON Shared Task test split. Best results shown in bold and second-best underlined.

| Method | BLEU | | | | | | chrF++ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Swa | Yor | Ice | Mar | Kan | Avg. | Swa | Yor | Ice | Mar | Kan | Avg. |
| zero-shot | 16.96 | 4.48 | 4.92 | 0.70 | 5.84 | 6.58 | 40.33 | 18.87 | 27.97 | 13.28 | 25.65 | 25.22 |
| few-shot | 22.35 | 11.98 | **6.43** | **19.19** | 23.50 | 16.69 | 45.73 | 29.92 | **28.87** | 36.11 | 44.16 | 36.96 |
| gloss-shot | 22.20 | **16.32** | 3.50 | 17.53 | 22.40 | 16.39 | **46.49** | 33.24 | 25.79 | **36.18** | 42.69 | 36.88 |
| chain-gloss | **23.54** | 14.10 | 5.11 | 17.32 | **25.26** | **17.06** | 45.43 | **33.54** | 24.91 | 35.37 | **46.27** | **37.10** |

Table 3: BLEU and chrF++ performance on **low-resource languages**. We use the GlossLM data, the largest corpus of IGT data. Best results shown in bold and second-best are underlined.

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 indicates that $N \geq 21$ is optimal.)

**GlossLM:** 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 languages included in the SIGMORPHON Shared Task for unseen languages, as this data is the most consistent. For other languages ranging from low to high-resource, the dataset offers only a training split. To address this, we created evaluation splits by designating most of the training set for evaluation, reserving the first 21 examples for in-context learning. We have detailed the number of validation samples for each language in Table 1.

**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 common 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 annotator scores (Popović, 2017).

## 5 Results

**GrammaMT outperforms few-shot in unseen/endangered languages.** In Table 2, we show how GRAMMAMT performs on four unseen languages: Gitksan, Lezgi, Natugu and Tsez. The results show that the gloss-shot strategy outperforms the baselines on all unseen languages tested on using the chrF++ metric. We see an improvement of 0.69 for Gitksan, 0.19 for Lezgi, 0.84 for Natugu, and 0.89 chrF++ points for Tsez over few-shot. Additionally, BLEU scores improved for both Gitksan and Tsez. For chain-gloss, while few-shot outperforms with the chrF++ metric, we see BLEU score increases of 1.13 points for Gitksan, 0.94 for Lezgi, and 0.14 for Tsez. On average GRAMMAMT is competitive or 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 the low-resource languages Swahili, Yoruba, Icelandic, Marathi and Kannnada (see

| Method | BLEU | | | | | | | | chrF++ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Urd | Tha | Gre | Por | Jap | Rus | Ara | Avg. | Urd | Tha | Gre | Por | Jap | Rus | Ara | Avg. |
| zero-shot | 4.00 | 1.35 | 6.16 | 37.75 | 7.16 | 25.20 | 3.46 | 12.15 | 20.53 | 12.56 | 23.16 | 59.21 | 27.61 | 47.30 | 19.95 | 30.05 |
| few-shot | 26.19 | 7.68 | 10.62 | 44.14 | 13.72 | 24.95 | 5.35 | 18.95 | 43.36 | 19.76 | **27.55** | **63.89** | 35.93 | 48.58 | **21.28** | 37.19 |
| 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 | 35.70 | 48.13 | 21.19 | 36.95 |
| chain-gloss | **28.71** | **8.37** | **10.74** | 42.88 | **14.78** | **27.52** | 5.26 | **19.75** | **45.86** | **19.85** | 27.11 | 62.33 | **36.41** | **49.30** | 19.51 | **37.20** |

Table 4: BLEU and chrF++ performance on **mid-high-resource languages** also reported on the GlossLM data. Best results are shown in bold and second-best are underlined.

Table 3). Chain-gloss improves the performance on the majority of the low-resource languages we examine as seen in the average BLEU and chrF++ score. This improvement is similarly observed with gloss-shot, particularly in the chrF++ performance for Swahili and Marathi. Notably, Yoruba showed significant improvement 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.

**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. Chain-gloss outperforms few-shot and gloss-shot in 5 of the 7 languages, while gloss-shot achieves the best results for Portuguese and Arabic. 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 again for the Portuguese and Arabic, as well as Greek, for which few-shot outperforms both gloss-shot and chain-gloss. For these languages, gloss-shot also outperforms chain-gloss.

Overall, these 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 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 5, we observe that seg-shot improves gloss-shot on Natugu, Greek and Arabic. This improvement is also seen at average performance level, which provides motivation for GRAMMAMT to be explored with other grammatical augmentations.

We then combined glosses and segmentation in our prompts (*gloss w/ seg*) and found performance improvment 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.

### 5.2 OOD - Flores

The above experiments all use data from the same domain. In these experiments, we explore the performance of GRAMMAMT in an out-of-distribution setup with the FLORES dataset. We use the same examples from the GlossLM in the prompts and ask the model to translate the FLORES test set. Table 6 shows the results with gloss-shot being the best performer on average. We observe an improvement of over 1 BLEU point for Swahili, Icelandic, Greek, Portuguese, Japanese and Russian. This shows that the gloss-shot strategy can be applied even in the absence of annotated glosses for the current domain, and highlights its potential for out-of-domain settings. However, this seems not to be the case for chain-gloss. Without in-domain gloss examples, the model seems to struggle to generate good glosses, resulting in poorer translations.

| Method | 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 5: The effect of augmenting GRAMMAMT with other grammatical information than glosses. We find that morphological segmentation can be a viable alternative to annotated glosses.

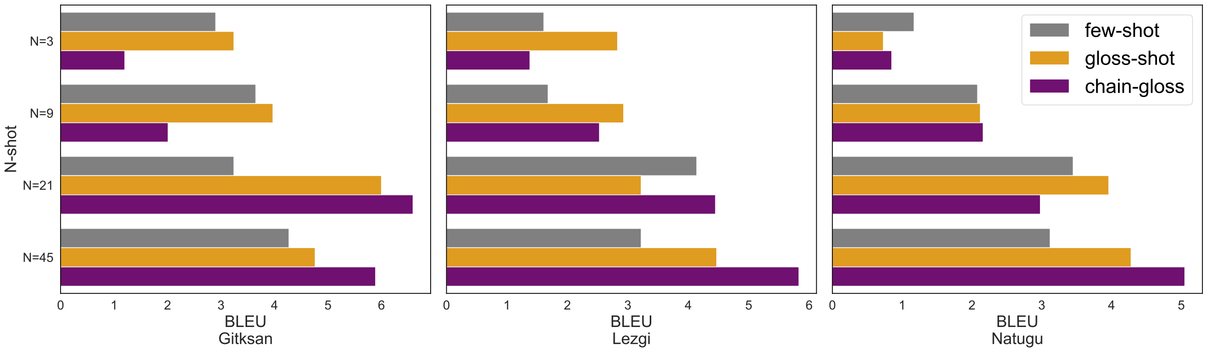


Figure 2: Varying the number of $N$-shot examples, ranging $N$ from 3 to 45. These ablation studies were performed on the validation split of the 2023 SIGMORPHON Shared Task data.

## 6 Discussion

We conduct ablations studies, discussing the impact of varying the number of N-shot examples in our approach. This analysis was performed on the validation split from the SIGMORPHON data, where we have access to a larger range of examples (greater than 21) from the training split. Additionally, we analyse the impact of the gloss component under an oracle setup, by providing a gold gloss for the source sentence to be translated.

**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 basic few-shot approach. An increase of $N$ leads to improvements in all strategies, see Figure 2. Interestingly, gloss-shot beats few-shot with the increase in $N$ with exception of Lezgi at $N = 21$. We also see significant gains on chain-gloss with the largest number of N suggesting also that chain-of-thought would benefit also from adding more examples to illustrate the task in question. These results provide a strong argument for adding more examples to both gloss-shot and chain-gloss, in order to maximise the benefit of GRAMMAMT.

**Oracle Performance.** In neither GRAMMAMT prompting strategy, is the model provided with an annotated gloss for the source sentence to be translated. In gloss-shot, glosses are provided for only the $N$-shot examples. In chain-gloss, the model is required to generate the gloss for the source sentence before translating, even though these LLMs were not specifically trained on that task. We examined the impact on performance, of providing the model access to an accurate gloss for the source sentence, using gold glosses available on the SIGMORPHON and GlossLM data. Results are presented in Table 7.

On average, scores jump by a massive 17.46 (± 6.6 st. dev) BLEU points across all languages. The highest improvement is 30.33 BLEU for Kannada, and the least gains seen for Portuguese and Lezgi, still improving by 6.2 and 7.64 BLEU respectively. In fact, large improvements are generally seen for each class of languages: unseen and low-, mid-high-resource languages. This points to the potential impact of exploring and training automatic gloss models like GlossLM (Ginn et al., 2024) to support tasks like machine translation.

We also conducted this oracle experiment in the zero-shot setting, where the model is no longer provided with any $N$-shot examples of machine-translation pairs or glosses, but only the source sentence to translate together with its gold gloss.

Once again, across all languages, the model outperforms the zero-shot setting by a massive margin: an average gain of 16.02 (± 8.89 st. dev.) BLEU

| Method | 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** | 20.7 |
| **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 |

Table 6: 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 generalize in an out-of-domain setting.

| Method | Git | Lez | Nat | Ddo | Swa | Yor | Ice | Mar | Kan | Urd | Tha | Gre | Por | Jap | Rus | Ara |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Few-shot** | 3.24 | 4.14 | 3.45 | 1.53 | 20.23 | 11.9 | 4.54 | 19.19 | 22.29 | 26.09 | 7.42 | 8.4 | 44.37 | 12.3 | 22.63 | 5.35 |
| **Gloss-oracle** | 19.71 | 11.78 | 16.52 | 13.95 | 40.0 | 38.32 | 22.56 | 38.69 | 52.62 | 38.29 | 28.31 | 34.68 | 50.57 | 28.46 | 37.57 | 24.45 |
| **ΔDelta** | 16.47 | 7.64 | 13.07 | 12.42 | 19.77 | 26.42 | 18.02 | 19.5 | 30.33 | 12.2 | 20.89 | 26.28 | 6.2 | 16.16 | 14.94 | 19.1 |

Table 7: Simulation of an oracle experiment with GRAMMAMT receiving a reference gloss. Previous experiments lacked glosses for input sentences (only available for the $N$-shot, $N = 21$, examples). Our aim was to investigate whether the model's performance could further improve with access to a correct gloss.

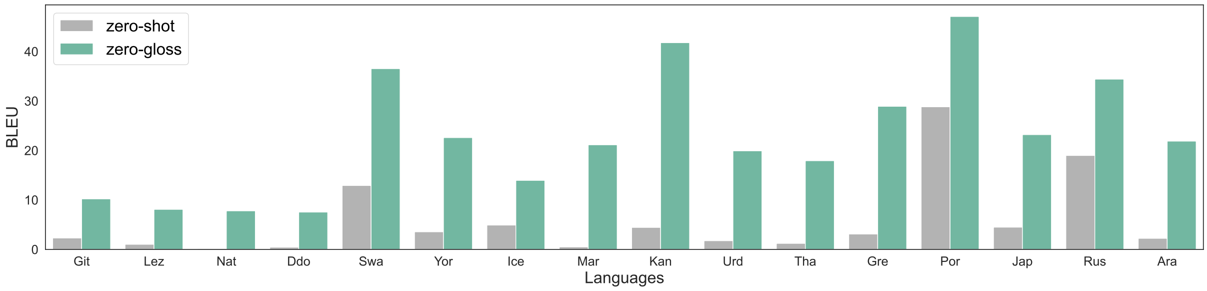


Figure 3: Oracle experiment in the zero-shot setup. We study if the performance could continue to improve with the gloss of the source sentence, without having access to any support examples. The experiments show significant BLEU gains across all languages, attesting for the potential of automatically generating glossed text.

points. The highest improvement is 37.34 points for Kannada again. Compared to the previous in-context learning experiments reported in Section 5, there are greater gains when the input sentence is accompanied by its gloss, compared to using few examples from machine translation pairs (few-shot) or from interlinear glosses (gloss-shot). This again underscores the potential of having specialised external gloss models automatically annotate the input sentences.

**Qualitative Examples** We present examples of translations produced by GRAMMAMT using the gloss-shot and chain-gloss strategies, and compare against the few-shot approach, as shown in Table 10 and Table 11 in Appendix E. Specifically, we showcase examples for gloss-shot in Yoruba and for chain-gloss in Swahili, where we observed respective gains of 4.25 and 1.19 BLEU points, respectively. In line with these quantitative results, we can observe that both gloss-shot and chain-gloss strategies generate more satisfactory translations compared to the few-shot approach.

## 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, likewise, 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, akin to a chain-of-thought prompting, provides more interpretability by generating step-by-step glosses specifically for the input sentence.

While our work covers a wide variety of languages, it only evaluates machine transliteration into English (→ en) and not from English (en →). This limitation is due to the available Interlinear Gloss Text datasets, which primarily contain glosses and translations in English for various source languages, but not the other way around. However, since most data on the internet is in English, it would be valuable for future work to evaluate our approach for translating from English as well.

Another limitation is the scope of metrics that our study covers. In the Results section, we occasionally observe inconsistencies between improvements in BLEU and chrF+. Therefore, expanding the range of metrics, such as incorporating COMET (Rei et al., 2020) and other model-based evaluations, would provide a more comprehensive evaluation of our approach.

## 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*, pages 3316–3333, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Yuan Gao, Ruili Wang, and Feng Hou. 2023. How to design translation prompts for chatgpt: An empirical study. *Preprint*, arXiv:2304.02182.

Xavier Garcia, Yamini Bansal, Colin Cherry, George Foster, Maxim Krikun, Fangxiaoyu Feng, Melvin Johnson, and Orhan Firat. 2023. The unreasonable effectiveness of few-shot learning for machine translation. *Preprint*, arXiv:2302.01398.

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. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

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.

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.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Wenxiang Jiao, Wenxuan Wang, Jen tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *Preprint*, arXiv:2301.08745.

Karima Kadaoui, Samar Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. TARJAMAT: Evaluation of bard and ChatGPT on machine translation of ten Arabic varieties. In *Proceedings of ArabicNLP 2023*, pages

52–75, Singapore (Hybrid). 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.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 227–237, Tampere, Finland. European Association for Machine Translation.

Xuan-Phi Nguyen, Sharifah Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2023. Democratizing llms for low-resource languages by leveraging their english dominant abilities with linguistically-diverse prompts. *Preprint*, arXiv:2306.11372.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5622–5633, Singapore. Association for Computational Linguistics.

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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Unbabel's participation in the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 911–920, Online. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Felix Stahlberg, Eva Hasler, Aurelien Waite, and Bill Byrne. 2016. Syntactically guided neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 299–305, Berlin, Germany. Association for Computational Linguistics.

Emma Strubell, Patrick Verga, Daniel Andor, David Weiss, and Andrew McCallum. 2018. Linguistically-informed self-attention for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5027–5038, Brussels, Belgium. Association for Computational Linguistics.

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*.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023b. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

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*.

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*.

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. 2023. Machine translation with large language models: Prompting, few-shot learning, and fine-tuning with QLoRA. In *Proceedings of the Eighth Conference on Machine Translation*, pages 468–481, Singapore. Association for Computational Linguistics.

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 Model Size

To assess the generalisability of GRAMMAMT, we carried out experiment with Llama-3-8B and Mixtral-8x22B. Majority of the languages exhibit better performance using GRAMMAMT with a few exceptions: 2 languages for LLaMA 70b and Mixtral, 4 languages for LLaMA 8b where few-shot was the best performer. Closer examination of Appendix A shows that chain-gloss in larger models, especially Mixtral, provides considerable gains compared to few-shot. We also see similar gains with chain-gloss with LLaMA 8b for most of the low-resource language. This is particularly attractive as with most low-resource languages often face double-bind (Ahia et al., 2021) of compute and data. The success of smaller model doing well with chain-gloss and gloss-shot means a lower barrier to achieving good translation for these languages. Additionally, gloss-shot exhibits the best performance in 3 high-resource languages for the smaller LLaMA model. Overall, these results provide evidence that GRAMMAMT is versatile and can produce good performance with both small and large models.

## B FLORES chrF++

Here we report chrF++ results on the FLORES test set. Consistently with BLEU scores, we also observe improvements of chrF++ for Swahili, Icelandic, Greek, Portuguese, Japanese and Russian.

## C 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.

**Low-resource languages.** Swahili, Yoruba, Icelandic, Marathi, and Kannada languages exhibit diverse morphological structure and word order. Swahili, Marathi, and Kannada are agglutinative, Yoruba is analytic, and Icelandic is fusional. In terms of word order, Swahili, Yoruba and Icelandic are characterised by a Subject-Verb-Object order

| Method | Model | 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 8: BLEU performance on reported on the GlossLM acrossed the different models (LLaMA-3 70b, LLaMA-3 8b, Mixtral-8x22B).

| Method | 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 |

Table 9: chrF++ performance on the Flores test set.

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.

## D Prompt-Template

Our prompt follows the *Hire a Linguist* 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:

### D.1 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: ###
```

### D.2 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: ###

### D.3 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: ###

### D.4 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>
Gloss:

### D.5 Segmentation-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>
**Segmentation:** <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: ###

### D.6 Segmentation-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>
**Segmentation:** <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 segmentation and then** with the translation directly and enclose your translation in ###.

Swahili sentence: <input sentence 1>
Segmentation:

## E Qualitative Examples

We provide examples of translations generated with GrammaMT in Table 10 and Table 11.

| Reference | Few-shot | Gloss-shot |
|---|---|---|
| Yoruba | | |
| I cut the bread with a knife. | I gave a knife to a butcher. | I used a knife to cut the bread. |
| S/he is bringing them. | S/he took them (a collection of small things) from home. | S/he gathered them and came. |
| Ade cannot sing. | Ade is not singing. | Ade doesn’t know how to sing. |
| Ojo bought a book for mother | The king gave a house to his mother. | The day broke and gave the mother. |
| The child accepted the present with joy. | The child threw away the entire gift. | The child used a joyful mood to receive a gift. |

Table 10: Qualitative examples of translations generated with the gloss-shot strategy for Yoruba compared to the few-shot strategy.

| Reference | Few-shot | Chain-gloss |
|---|---|---|
| Swahili | | |
| One large basket fell. | A large basket fell. | **Gloss:** 7.basket 7.big 7.one 7.PST-fall<br>**Translation:** A big basket fell . |
| I was still working. | I had a job that I was doing. | **Gloss:** 1SG-PST-be 1SG-PST-want 1SG-PST-do 9work<br>**Translation:**I wanted to do the work. |
| Juma had cooked food. | Juma had cooked food. | **Gloss:** Juma SM.PST-be 1SG-REL-cook food.<br>**Translation:**Juma had cooked food. |
| The leg of the table and the chair are broken. | The leg of the table and chair is broken. | **Gloss:** 13-leg 3-table and 7-chair 3S-PST-break-APPL-FV<br>**Translation:**The leg of the table and the chair broke. |
| I read for each author his book. | I read every author’s book. | **Gloss:** Gloss: 1sgS-PST-read-FV every 1-author 7-book 7-POSS<br>**Translation:**I read every author’s book. |

Table 11: Qualitative examples for the Swahili language of the translations generated with the chain-gloss strategy compared to the few-shot strategy.