Interlinear Glosses as a Multilingual Pivot for Machine Translation: An Updated Study on Turkish with Restricted Resources

Volkan Özer[1], Shu Okabe[1,2] and Alexander Fraser[1,2,3]
[1]School of Computation, Information and Technology, Technische Universität München (TUM)
[2]Munich Center for Machine Learning
[3]Munich Data Science Institute
Corresponding authors: {volkan.oezer, shu.okabe}@tum.de

Abstract

Translating very low-resource languages is a challenge that has been approached using available linguistic cues. Among them, interlinear glosses are linguistic annotations that can essentially bridge the gap between two languages thanks to both grammatical and lexical information. We perform a case study on a simulated low-resource condition for Turkish, a morphologically rich language, with a pipeline approach, following (Zhou et al., 2020). A source sentence is passed through a morphological analyzer and a bilingual dictionary to obtain a gloss-like representation. We then evaluate the current capacity of Neural Machine Translation systems and Large Language Models in performing the translation task from interlinear glosses into fluent English translations. We notably evaluate how performance scales with multilingual glossed data and how translation is affected by pseudo-glosses. Pivoting with glosses remains a better approach than a direct translation for languages with limited parallel data for training. Although glosses remain helpful resources, translations are sensitive to their quality, especially for lexical information.

Keywords: Low-Resource Machine Translation, Interlinear Glosses, Linguistic Representation

1. Introduction

In reaction to the ongoing context of language endangerment, language documentation is a field of linguistics that aims to collect, annotate, and archive resources. Linguists use Interlinear Glossed Text (IGT), which is a format of annotations to explain a language in a higher-resource language in grammar books or corpora. As shown in Figure 1, they label each morpheme (the smallest meaningful unit) in a sentence. They can be classified into two categories: grammatical glosses (e.g., 'GEN' for genitive) expressing the role of the morpheme and lexical glosses (e.g., 'go') indicating its meaning in the other language. As they are usually present in linguistic documents to support explanations, they can help understand the way the studied language functions.

This is why glosses have been shown to be a viable complementary resource for NLP as well, as it was pioneered by (Zhou et al., 2020) for a Neural Machine Translation (NMT) model. Successive attempts also focused on the use of glosses as additional context for Large Language Models (LLMs) for machine translation (Ramos et al., 2025), specifically for very low-resource languages. Most studies in this direction relied on fairly high-quality glosses from state-of-the-art glossing models.

In this article, we assess the extent of benefit from using glosses in very low-resource pipeline MT. We follow the three-step system from (Zhou et al., 2020) to translate a sentence in a studied language into a higher-resourced language. It first analyses the sentence to decompose each word into its lemma and a morphological analysis (tier 2 in Figure 1). Once the features are converted to follow the Leipzig Glossing Rules (Bickel et al., 2008), the most widely used convention to gloss a language, we replace all source lemmata with their translation in the target language using a bilingual dictionary (tier 3). Finally, we rely on an NMT implementation of the last step to convert a glossed sentence into a fluent translation. We update the last step by also evaluating task performance with three open-weight LLMs of restricted size (7-8B) to assess whether the bottleneck originates from the model. We analyse the impact of the input format and of the training datasets on translation.

We choose to perform a case study on Turkish, where we restrict the amount of resources we use, to simulate a low-resource MT scenario, following (Zhou et al., 2020). We namely rely on less than 5k Turkish-English parallel sentences for MT training explicitly and realistically available resources (morphological analyzer and bilingual dictionary). Although not intrinsically low-resource, Turkish reflects several challenges, especially a rich morphology and a different language family than English, which enable us to approximate the task difficulty for actual under-represented and low-resource languages.

We hence focus on two research questions: to what extent can machine translation be performed using glosses as a pivot with multilingual NMT models and LLMs? How does the pipeline approach compare with a direct translation or a translation

Proceedings of SIGUL 2026 Joint Workshop with ELE, EURALI, and DCLRL @ LREC 2026, pages 183–197
11–12 May 2026.

from the reference glosses?

Contributions are as follows: (i) we compare the performance of NMT models and LLMs in translating Turkish into a fluent English translation with the help of gloss-like annotations in between, and (ii) we underline the contribution of the gloss representation (mainly, lemma and grammatical tag number) in the final translation. We release our evaluation data and code material publicly.[1]

2. Related Works

Low-Resource Machine Translation Machine translation for low-resource languages remains challenging, particularly as parallel data is scarce in general. Prior work has explored multilingual models with parameter sharing (Firat et al., 2016; Zoph et al., 2016) and transfer learning from related languages (Nguyen and Chiang, 2017; *inter alia*), but substantial quality gaps remain (Koehn and Knowles, 2017). More recently, LLMs enable zero- and few-shot translation (Zhang et al., 2023; Garcia et al., 2023), yet performance degrades for low-resource and unseen languages (Hendy et al., 2023; Robinson et al., 2023).

Multilingual Neural Machine Translation Multilingual NMT models translate multiple language pairs using shared parameters and attention mechanisms (Johnson et al., 2017; Dong et al., 2015), which enables cross-lingual transfer and zero-shot translation (Firat et al., 2016; Al-Rfou' et al., 2013). Effectiveness, however, depends on parallel data and linguistic similarity between languages. This framework underlies gloss-based pivoting, where representations can be shared across typologically diverse languages.

Morphology-Aware Machine Translation Morphologically rich languages increase data sparsity in low-resource MT (Habash and Sadat, 2006; Lee, 2004). For such languages, common approaches include character- (Ling et al., 2015), byte- (Gillick et al., 2016), and subword-level modeling (Sennrich et al., 2016). Morpheme-level and factored models capture lemma, POS, and morphological information, reducing sparsity while preserving meaning (García-Martínez et al., 2016; Cotterell and Schütze, 2015; Koehn and Hoang, 2007). Morphological supervision has been shown to improve both efficiency and translation quality (Cui et al., 2022; Stahlberg et al., 2016), motivating the use of morphological analyzers in gloss-based pipelines.

Linguistically-Informed Translation with Glosses IGT provides linguistically informed source representations that can improve low-resource translation. Zhou et al. (2020) train multilingual gloss-to-target models across thousands of languages. Rapacz and Smywiński-Pohl (2025) explores morphological embeddings in interlinear translation, showing substantial gains even with limited training data. Recent extensions leverage external resources, including grammar books and morphological analyzers (Tanzer et al., 2024; Zhang et al., 2024), or use grammar-aware prompting without model training (Ramos et al., 2025). However, the sensitivity of translation systems to gloss reliability, automatic glossing errors, and annotation granularity remains underexplored. Our work systematically evaluates NMT and LLM approaches across varying gloss qualities, training sizes, and morphological representations. In terms of glossed data, we mainly rely on GlossLM (Ginn et al., 2024b), which provides the largest multilingual IGT corpus to date.

3. Data

Our experiments rely on two complementary resources: a large-scale Turkish–English parallel corpus, downsampled to replicate a low-resource scenario, and a multilingual IGT corpus.

3.1. Base Datasets

SETIMES We use the Turkish–English portion of the SETIMES parallel corpus (Tiedemann, 2012). This collection of news articles covers domains such as politics, economics, and social issues. The corpus contains 207k sentence pairs with complex syntactic structures and relatively long sentences.

GlossLM The GlossLM corpus (Ginn et al., 2024) is a large multilingual collection of IGT, comprising 450k instances covering around 1,800 languages. It comprises data from multiple sources, including ODIN (Lewis and Xia, 2010), IMTVault (Nordhoff and Krämer, 2022), and the SIGMORPHON Shared Task on interlinear glossing (Ginn et al., 2023) datasets. Each instance contains a transcription aligned with morphological glosses and a free translation, mainly in English.

In our work, SETIMES provides parallel Turkish–English data, while GlossLM provides multilingual IGT data for MT training.

3.2. Simplified SETIMES-S Corpus

The original SETIMES contains long, formally written news sentences that differ from the shorter and structurally simpler style typical of IGT data, creating a domain and style mismatch. To mitigate this, we construct a simplified variant, `SETIMES-S`, by systematically simplifying Turkish sentences and

[1] https://github.com/vlknozr/interlinear-gloss-mt-turkish

1. Source sentence (Turkish)	Dün	gittiğini	Fatma'nın	Ayşe	biliyor.
2. Source pseudo gloss (lemma)	Dün	git-GER-3SG-ACC	Fatma-GEN	Ayşe	bil-PROG-3S
3. Interlinear gloss (target-lemma)	Yesterday	go-GER-3SG-ACC	Fatma-GEN	Ayşe	know-PROG-3S
4. Target sentence (English)	It's Ayşe who knows that she, Fatma, left yesterday.				

Figure 1: Illustration of the MT pipeline using pseudo-glosses as a pivot. Example sentence from the Turkish subset of the GlossLM corpus.

regenerating aligned English translations under controlled constraints using GPT-4 (OpenAI et al., 2024).

Simplification follows explicit rules, such as a CEFR (Council of Europe, 2001) A1–A2 vocabulary and grammar. We limited sentence length to roughly 7–8 words (by splitting longer sentences). We also ensured semantic preservation, avoided complex constructions, and maintained parallel alignment. The resulting dataset contains 2,000 training pairs and 600 evaluation pairs.

3.3. GlossLM Turkish Subset

The Turkish portion of GlossLM originates primarily from ODIN and exhibits recurring quality issues, largely attributable to the digitization based on Optical Character Recognition (OCR) (Nordhoff and Krämer, 2022). We therefore sample 200 Turkish transcriptions to construct our Turkish evaluation set and manually inspect and correct them, leveraging native-speaker judgment to restore orthographic validity.

These transcription defects lead to wrongly out-of-vocabulary items and failures in downstream morphological analysis (e.g., no parse or spurious analyses). Corrections target three recurring error classes: (i) OCR artifacts such as incorrect diacritics and faulty segmentation, (ii) lexical inconsistencies including duplicated or nonsensical tokens and missing fields, and (iii) grammatical inconsistencies affecting agreement or gloss alignment. This curation improves reliability for downstream morphological processing.

3.4. Training Data

Multilingual IGT Training Sets We construct multilingual training sets by sampling unsegmented IGT examples from GlossLM while prioritizing Turkic languages. All available examples from 23 Turkic languages (1,962 examples, including 1,700 for Turkish) are included, and the remaining examples are uniformly sampled from other languages.

After preprocessing (removing entries with missing fields and duplicates), the number of usable Turkish examples is reduced to 1,500. From these, we sample 200 examples to form our Turkish evaluation set (`TR-GLM`) and use the remaining 1,300 Turkish examples in training. We then construct three progressively larger training splits from GlossLM–with 60k, 100k, and 150k instances–denoted as `Subset-60k`, `Subset-100k`, and `Subset-150k`.

Dataset	Langs	Train	Test
Turkish–English			
SETIMES (original)	1	207,678	600
`SETIMES-S`	1	2,000	600
Multilingual			
GlossLM (original)	1,800	451,000	—
`Subset-60k`	1,508	59,716	—
`Subset-100k`	1,515	97,877	—
`Subset-150k`	1,517	150,588	—
Evaluation			
`TR-GLM` (Gold)	1	—	200
`TR-Morph-GLM` (Silver)	1	—	200
`TR-SET600`	1	—	600

Table 1: Dataset statistics. The multilingual section lists the cumulative `GlossLM` training splits (each including the same 1.3k Turkish training examples).

Training Configurations Using `SETIMES-S` Using the `Subset-100k` training set as our base, we create three configurations by adding varying amounts of `SETIMES-S` Turkish–English data (600, 1.2k, and 1.8k sentence pairs).

3.5. Evaluation Data

We use two complementary sets. `TR-GLM` comprises 200 Turkish sentences from GlossLM with their original human-annotated glosses, serving as a gold-standard benchmark. `TR-SET600` is a 600-sentence test set drawn from `SETIMES-S`, designed to match the simplified style typical of IGT data.

Table 1 details the datasets used in our experiments for both training and testing.

4. Machine Translation Pipeline

We adopt the translation pipeline of (Zhou et al., 2020), illustrated in Figure 1, and update its components. The pipeline consists of three sequential steps, each described in detail below.

4.1. Source Sentence to Pseudo-gloss (1 → 2)

The first step converts each source word into a lemma–morphology representation following the UniMorph schema (Sylak-Glassman et al., 2015), which is subsequently transformed into a gloss format compliant with the Leipzig Glossing Rules (Bickel et al., 2008), the most widely used convention for interlinear glossing.

4.1.1. Morphological Analyzers

We consider three Turkish morphological analyzers: TRmorph (Çöltekin, 2010, 2014), TRMOR (Kayabaş et al., 2019), and turkish-morphology (Öztürel et al., 2019). All three are based on finite-state transducers, making them representative of tools realistically available for low-resource languages (Zhang et al., 2024).

Since analyzers may return multiple analyses for an inflected word, we select the most probable analysis given by the model subsequently.

4.1.2. Analyzer Comparison and Selection

We process Turkish sentences from both resources (SETIMES and GlossLM) with a morphological analyzer to derive gloss representations used in our experiments. To assess robustness and suitability for downstream MT, we evaluate all three Turkish morphological analyzers on 4,000 sentences (77,146 tokens) from the original (unsimplified) SETIMES corpus using three criteria: coverage (whether a token receives any analysis), ambiguity (the average number of analyses per token), and processing time.

Table 2 shows a trade-off between robustness and efficiency. TRmorph achieves the highest coverage, exceeding both turkish-morphology and TRMOR. Although slower, its superior coverage is decisive in our setting, since unanalyzed tokens would result in empty outputs and prevent morphological decomposition.

We therefore select TRmorph for the remainder of the pipeline. Common sources of analysis failure across tools include foreign proper nouns (Radiç, Rijeka), loanwords (European, Southeast), and abbreviations (CNN, BBC). Unanalyzed words are left unchanged in the analysis.

4.1.3. Format Conversion

TRmorph outputs a lemma and a sequence of morphological tags in its native format. We map these tags to UniMorph feature labels and combine them with the lemma into a hyphen-separated pseudo-gloss token. Figure 2 illustrates this conversion process.

Metric	TRmorph	turkish-mor.	TRMOR
Coverage			
Type Coverage	96.22%	90.92%	61.12%
Words Analyzed	72,449	70,145	47,151
Ambiguity			
Avg. Analyses	14.82	22.31	4.03
Max. Analyses	1,400	550	88
Total Analyses	1,143,339	1,565,250	310,947
Efficiency			
Processing Time	165 min	40 min	55 min
Relative Speed	4.1×	1.0×	1.4×

Table 2: Turkish morphological analyzer comparison for 4,000 sentences.

We discard predicted labels which correspond to a part-of-speech (PoS) tag (e.g., Adj, N), since PoS labels are typically not included in interlinear glosses. The remaining grammatical features are concatenated using hyphens to follow the Leipzig Glossing Rules.

Source word	*yeterliliklerine*
TRmorph output	`yeterli<Adj><lik><N><pl><p3s><dat>`
Converted output	yeterli–LIK–PL–P3S–DAT

Figure 2: Example of format conversion for the Turkish word *yeterliliklerine* ('to their competencies').

4.2. Source to Target Pseudo-gloss (2 → 3)

The second stage replaces source lemmata with their target-language equivalents using a bilingual dictionary. Such resources exist for approximately 70% of the world's 7,000 languages (Wang et al., 2022), making this step realistic even in low-resource settings.

4.2.1. Dictionary Construction

Following (Zhou et al., 2020), we simulate a bilingual dictionary by extracting frequent word associations from a parallel corpus. We use 40,000 Turkish–English sentence pairs from the SETIMES corpus[2] and apply the awesome-align multilingual aligner (Dou and Neubig, 2021) to find word pairs.

Since Turkish is supported by mBERT (Devlin et al., 2019), we fine-tune the aligner on the corpus to improve alignment quality. While such resources may not be available for most low-resource languages, our objective here is simply to construct a reliable Turkish–English dictionary.

4.2.2. Preprocessing and Filtering

To approximate actual dictionary entries, we first lemmatize Turkish tokens using TRmorph, as dic-

[2]They do *not* overlap with our evaluation dataset.

tionaries typically store canonical forms (e.g., 'go') rather than inflected forms (e.g., 'went'). We additionally remove English function words that lack direct Turkish equivalents due to typological differences in morphological complexity.

As noted in (Zhang et al., 2024), dictionary mapping is non-trivial and can be noisy due to incorrect word alignments. To improve reliability, we discard alignment links that occur only once in the corpus, as these are more likely to correspond to erroneous or idiosyncratic matches. Finally, when a Turkish lemma aligns to multiple English candidates, we retain only the most frequent counterpart as its translation.

Processing source pseudo glosses Words not found in the dictionary are left unchanged in the source sentence. After applying these steps, we obtain a dictionary containing 20,953 Turkish–English word pairs, achieving 99.7% token-level coverage on the SETIMES corpus used in dictionary creation. Coverage is 93.4% on the Turkish evaluation set (`TR-GLM`), and 98.3% on the `SETIMES-S` evaluation set (`TR-SET600`). Despite the domain difference, the dictionary achieves reasonable coverage.

4.2.3. Input variants

To assess which components of the gloss representation contribute most to performance, we evaluate controlled variants for `TR-GLM` and `TR-SET600`.

We use the same 200 Turkish transcriptions as `TR-GLM` and process them through the pipeline. Here, the source pseudo-gloss is generated from the morphological analysis and the dictionary-based lemma substitution, representing a silver-grade, realistic input. We denote this dataset `TR-Morph-GLM`.

Moreover, we use the label `GoldLemma` to denote configurations where pipeline target lemmata are replaced with gold (native) lemmata from `TR-GLM`. This guarantees a correct lemma in the glossed sentence and removes the influence of the dictionary quality.

4.3. Target Pseudo-gloss to Translation (3 → 4)

The final stage converts the target-language pseudo-gloss sequences into a fluent sentence. We implement this translation step using the OpenNMT toolkit (Klein et al., 2018), following the NMT architecture and hyperparameters of (Zhou et al., 2020). For this last step, we update the original pipeline by considering the larger and newer GlossLM corpus for training and evaluate on a controlled test dataset.

4.3.1. Model Architecture and Training

We use an RNN encoder–decoder model with global attention. Both encoder and decoder are two-layer LSTMs with 512 hidden units, 300-dimensional embeddings, and a dropout rate of 0.3. Training is performed with SGD using an initial learning rate of 0.8 and label smoothing of 0.1. We apply sentence length-based batching with a batch size of 32, gradient clipping with a maximum norm of 1.0, and train for up to 100,000 steps, validating every 250 steps. During inference, we use beam search with a beam size of 5.

4.3.2. Multilingual Setting

To leverage cross-lingual transfer and shared glossing conventions, we train the gloss-to-target model (with English as the target language) in a multilingual setting rather than fitting separate models per language. Each GlossLM instance is treated as a training example, where the source is a glossed sequence, and the target is the corresponding English free translation. To help the model handle cross-lingual variation in gloss representations and morphology, we prepend an ISO 639-3 language tag to each source sequence (e.g., tur for Turkish).

We train three NMT models on the multilingual IGT training splits described in Section 3.4, denoted `NMT60`, `NMT100`, and `NMT150`.

5. Experimental Results

5.1. Experimental setup

Baseline We compare our pipeline to a direct Turkish-English translation model (1 → 4) to assess the contribution of gloss information. These models are trained on `SETIMES-S` parallel data at matching sizes of 600, 1.2k, and 1.8k sentence pairs. We denote this model as the `baseline`.

Evaluation Protocol We evaluate IGT-to-English generation on the evaluation sets defined in Section 3.5, covering Turkish-specific test sets. We compare models under two input conditions: gloss-based settings where the input is IGT, and direct baselines where the input is the Turkish transcription (i.e., for `baseline`).

Evaluation Metrics We evaluate translation quality using BLEU (Papineni et al., 2002) computed with NLTK (Bird and Loper, 2004) and chrF++ (Popović, 2015) computed with sacreBLEU (Post, 2018). As our target language is English, we additionally report xCOMET-XL (Guerreiro et al., 2024), one of the top-performing MT metrics in the WMT24 Metrics Shared Task (Freitag et al., 2024).

5.2. Impact of Training Size on Gloss-to-Target Performance

Test set	NMT60	NMT100	NMT150
TR-GLM	19.23	**29.11**	25.46
TR-Morph-GLM	10.77	**14.01**	10.23

Table 3: BLEU scores across multilingual training dataset sizes (NMT60/100/150).

We evaluate three multilingual training settings with increasing training set sizes—NMT60, NMT100, and NMT150—on TR-GLM, and our pipeline test set, TR-Morph-GLM. The results are reported in Table 3.

We observe non-uniform scaling effects for the two test sets. Turkish-specific performance peaks at the intermediate setting: TR-GLM increases from 19 BLEU for NMT60 to 29 for NMT100, then drops to 25 for NMT150, and TR-Morph-GLM follows a similar pattern. Both are possibly due to the high degree of multilingualism in the training dataset, which seems to act as noise at 150k sentences.

Overall, NMT100 is optimal for Turkish-specific evaluation, which we will mainly consider below.

5.3. Comparing a Direct and Pipeline Translation

Data Size	baseline	NMT100		
	BLEU	BLEU	chrF	xC
0	—	3.19	25.89	46.30
600	**0.28**	11.12	33.17	62.58
1,200	0.11	17.98	39.50	70.60
1,800	0.18	**25.20**	**45.40**	**76.94**

Table 4: Scaling with Turkish parallel data on TR-SET600 (baseline vs. NMT100 with pipeline processing). xC stands for xCOMET.

We assess whether our pipeline improves Turkish–English translation relative to direct translation under limited Turkish-only supervision. Specifically, we compare the direct $1 \rightarrow 4$ translation baseline with the pipeline setting (NMT100) while varying the amount of additional Turkish–English training data from SETIMES-S, using 600, 1.2k, and 1.8k sentence pairs. We recall that the baseline directly uses the source Turkish sentence as input, while NMT100 processes the gloss-like output from the morphological analyser and dictionary substitution.

Table 4 shows that direct Turkish-only baselines fail in this low-resource setting, scoring near-zero BLEU on TR-SET600 across all settings due to the dataset sizes. In contrast, our pipeline produces substantial and consistent gains and scales reliably as more Turkish data is added. Starting from the same base model (NMT100), performance improves consistently as more SETIMES data is added, with steady gains on all three metrics.

Overall, these results indicate that the pipeline translation setting is effective for low-resource Turkish–English translation: it substantially outperforms direct translation and exhibits stable improvements as additional Turkish parallel data is introduced from the same domain.

Test Set	NMT100-w/oIMT	NMT100
TR-GLM	22.58	**29.11**
TR-Morph-GLM	9.76	**14.01**

Table 5: Effect of adding IMTVault to ODIN for Turkish IGT training (BLEU).

5.4. Impact of Dataset Diversity

We examine whether increasing within-language training diversity improves Turkish translation quality by experimenting with alternative Turkish IGT sources in the training corpus. Specifically, we compare NMT100, which used the two sources of Turkish IGT data, ODIN and IMTVault, where the latter accounts for 270 Turkish sentences (and ODIN, 1,280), with NMT100-w/oIMT, which is trained on Turkish data from ODIN only. This ablation isolates the effect of adding a second Turkish dataset—introducing broader coverage of genres, annotation styles, and lexical and morphological patterns—while keeping the target language and overall training procedure fixed.

As shown in Table 5, incorporating IMTVault leads to substantial improvements across all Turkish evaluation settings. On TR-GLM, BLEU increases from 23 to 29. The same trend is observed for pipeline evaluation: TR-Morph-GLM improves from 10 to 14.

Overall, these results indicate that adding a relatively small but diverse Turkish IGT dataset substantially benefits Turkish translation performance. This suggests that even within a single language, dataset diversity—rather than sheer volume alone—plays a critical role in improving robustness to variation in gloss format and morphological representation.

5.5. Study on Gloss Representations

5.5.1. Impact of Lemmatisation

To isolate the effect of lemma choice, we evaluate two hybrid configurations: GoldLemma, which replaces pipeline-generated *lemmata* with the correct

Test Set	BLEU	chrF++	xCOMET
TR-Morph-GLM	11.94	31.12	64.13
TR-Morph-GLM-GoldLemma	**23.76**	**43.77**	**77.38**
TR-Morph-GLM-TRLemma	1.99	13.36	45.82

Table 6: Effect of lemma source on pipeline translation quality using the NMT100+SET1.2k model.

original GlossLM TR-GLM lemmata (e.g., *sing*-AOR-3S), and TRLemma, which retains Turkish lemmata (e.g., *söyler*-AOR-3S, i.e., skipping step 2 → 3). In both cases, grammatical glosses are kept fixed, and results are compared against the standard pipeline (TR-Morph-GLM; e.g., *say*-AOR-3S). Results are reported in Table 6.

Using gold lemmata (TR-Morph-GLM-GoldLemma) substantially improves performance over the full pipeline, improving all three metrics by more than 10 points, confirming that dictionary-based lemma substitution is a major bottleneck. In contrast, retaining Turkish lemmata (TR-Morph-GLM-TRLemma) severely degrades performance, since source-language lemmata hinder generation while target-language lemmata simplify the task, as they match the training data format. Overall, producing accurate target-side lemmata remains a key challenge for the pipeline translation.

5.5.2. Impact of Gloss Representation

Test Set	BLEU	95% CI
Full (3+ glosses)	14.01	9.40–18.37
3Gloss (max 3)	**14.32**	9.79–18.76
2Gloss (max 2)	13.82	8.83–18.17

Table 7: Ablation result with gloss truncation on TR-Morph-GLM for NMT100 (BLEU).

To assess the effect of reduced morphological granularity, we evaluate two gloss truncation variants that limit the number of gloss tags per token: 2Gloss (max. 2 tags) and 3Gloss (max. 3 tags), compared to the full feature set. When truncation is applied, we retain glosses according to a priority hierarchy: inflectional features, case markers, derivational features, and other tags, so that tokens with more tags keep only the most crucial ones.

As shown in Table 7, limiting to three glosses slightly improves performance (around 0.3 point), while reducing to two yields a small drop (by around 0.2 point). These results indicate that moderate morphological detail is sufficient, whereas excessive tags provide little benefit and may introduce noise. Careful truncation, therefore, simplifies the representation without harming translation quality.

6. LLM for Gloss-to-Target

We update the last step of our pipeline with an LLM to reflect the recent change in state-of-the-art MT. This aligns with (Zhang et al., 2024; Ramos et al., 2025) in an attempt to extend MT to very low-resource languages using LLMs.

6.1. Base LLMs

We evaluate three open-weight multilingual LLMs with approximately 7–8 billion parameters from two model families: Llama-3.1-8B-Instruct (Llama) (Grattafiori et al., 2024), Qwen2.5-7B-Instruct (Qwen2.5) (Qwen et al., 2025), and Qwen3-8B (Qwen3) (Yang et al., 2025) in the non-thinking mode. All models are loaded with 4-bit quantization. During inference, translations are generated with a default temperature of 0.7. We report results from a single run.

6.2. Prompting Strategies

We evaluate three prompting configurations: zero-shot (ZS), few-shot (FS), and an advanced few-shot variant with additional guidance (FS-ADV).

In the FS prompt, rather than using fixed or random examples, we dynamically retrieve 3 examples based on character-level similarity between the input and training sentences, measured using chrF++ (Popović, 2015), following (Ginn et al., 2024a). This approach avoids language-specific assumptions and scales naturally to multilingual settings. The retrieved examples are included in the prompt as gloss–translation pairs to guide the model.

In FS-ADV, we further augment the FS prompt with brief explanations of the morphological gloss tags produced by the TRmorph analyzer (e.g., case, tense), helping the model interpret the glosses more effectively. We also consider an All-Tags variant, where the prompt lists all 95 grammatical tags available in TRmorph.

All prompts instruct the model to translate interlinear glossed text into fluent target-language output and to return only the translation. In summary, the ZS prompt includes only task instructions and the input, FS adds dynamically retrieved examples, and FS-ADV additionally provides tag descriptions. Full prompt templates are available in our repository.

6.3. Fine-tuning

We additionally fine-tune Qwen2.5 on a Turkish-specific IGT dataset consisting of 1,300 gloss–translation pairs from GlossLM-corpus. The resulting model is denoted as FT-Qwen2.5-TR. We fine-tune the models using QLoRA (Dettmers et al., 2023) with rank $r = 16$ and scaling factor $\alpha = 32$. Optimization is performed

Test set	Setting	Llama			Qwen2.5			Qwen3			FT-Qwen2.5-TR		
		BLEU	chrF++	xC	BLEU	chrF++	xC	BLEU	chrF++	xC	BLEU	chrF++	xC
TR-SET600	ZS	10.3	41.5	70.69	**14.9**	44.3	**74.78**	11.5	41.5	72.74	14.8	**44.6**	73.81
	FS	31.3	56.1	78.71	33.4	58.7	**81.03**	**33.9**	**58.9**	80.89	32.3	58.6	80.82
	FS-ADV	27.5	54.0	78.71	32.2	58.2	80.77	**34.4**	**58.5**	**81.02**	31.4	58.0	80.51
	FS-ADV(All-Tags)	25.3	55.9	68.41	31.3	58.3	**70.41**	**33.3**	**58.4**	70.09	28.8	57.6	70.13
TR-SET600-GoldLemma	ZS	10.4	40.2	70.33	14.5	43.2	**73.87**	9.1	38.8	71.18	**15.5**	**44.4**	73.50
	FS	30.6	54.8	78.29	32.7	58.4	80.12	**38.3**	**61.6**	**80.19**	32.6	58.1	79.98
	FS-ADV	26.5	54.7	78.64	30.5	58.2	80.12	**38.0**	**61.5**	**80.23**	29.0	57.7	79.68
TR-SET600-OrgSent	ZS	7.11	38.9	29.08	**8.81**	**39.2**	**31.23**	8.1	37.9	30.85	8.5	38.6	30.73
	FS	11.8	41.0	35.60	13.2	43.1	36.06	13.1	**43.5**	**36.09**	**13.4**	42.6	35.33
	FS-ADV	10.7	40.7	33.54	12.5	43.1	35.44	**13.1**	**43.8**	**36.09**	12.2	42.6	35.23

Table 8: LLM inference results on **TR-SET600** variants. xC is xCOMET. For reference, NMT100 scores 3.19 / 25.89 / 46.30, and NMT100+SET1.8k 25.20 / 45.40 / 76.94 on TR-SET600 (Table 4).

Test set	Setting	Llama			Qwen2.5			Qwen3			FT-Qwen2.5-TR		
		BLEU	chrF++	xC	BLEU	chrF++	xC	BLEU	chrF++	xC	BLEU	chrF++	xC
TR-GLM	ZS	14.8	41.3	73.13	**16.0**	**42.0**	**75.11**	8.7	39.4	70.43	15.2	41.8	74.03
	FS	27.2	50.0	75.90	27.0	51.5	**78.12**	**30.2**	**54.2**	77.33	27.8	51.9	77.82
	FS-ADV	26.0	48.9	75.20	**28.7**	51.7	78.51	28.0	**51.9**	76.52	27.2	51.3	**79.11**
TR-Morph-GLM	ZS	**6.8**	28.3	56.90	5.8	27.3	58.39	6.2	**29.0**	58.42	5.3	27.1	**59.10**
	FS	13.8	34.5	59.41	9.7	31.9	60.74	**15.4**	**35.8**	**62.30**	11.3	34.0	61.83
	FS-ADV	13.6	33.4	59.93	10.3	32.8	61.18	**14.3**	**36.1**	61.41	10.0	33.4	**62.23**
TR-Morph-GLM-GoldLemma	ZS	8.0	29.0	57.61	7.7	28.0	**59.13**	**8.1**	**29.1**	57.44	6.2	28.1	58.63
	FS	**64.3**	**73.8**	72.52	51.9	67.3	70.25	61.5	73.6	**73.81**	53.8	69.3	70.90
	FS-ADV(All-Tags)	40.0	63.6	68.66	49.0	66.3	70.99	**61.9**	**73.7**	**74.02**	52.1	68.5	71.42

Table 9: LLM inference results on **TR-GLM** variants. xC is xCOMET. For reference, NMT100 scores 29.11 (BLEU) on TR-GLM and 14.01 on TR-Morph-GLM (Table 3).

with an initial learning rate of 2×10^{-6}, a batch size of 6, and gradient accumulation over 3 steps. We train for 3 epochs, with early stopping based on a patience of 5. We used standard QLoRA hyperparameter settings with early stopping, given the small training set and computational constraints.

6.4. LLM Results

We evaluate several controlled input variants to understand which aspects of the gloss representation most affect LLM performance. OrgSent uses original SETIMES sentences instead of simplified SETIMES-S sentences to test language complexity effects. Results are shown in Tables 8 for TR-SET600 and 9 for TR-GLM.

Prompting Strategy As expected, FS consistently outperforms the zero-shot setting across all models and test sets. On TR-SET600, Qwen3 improves from 73 to 81 xCOMET score and Qwen2.5 from 75 to 81; similar gains hold on TR-GLM (e.g., Qwen3: 9 to 30). FS-ADV performs comparably to FS, indicating that brief tag descriptions provide little additional benefit. In contrast, the All-Tags variant, listing all gloss tags, consistently degrades performance (e.g., xCOMET drops from 80-81 to 68–70), suggesting that exhaustive tag inventories introduce noise.

Input representation Lemma substitution yields consistent improvements when combined with FS. If on TR-SET600-GoldLemma, performance is comparable with the standard representation, gains are larger on TR-Morph-GLM-GoldLemma, where FS significantly improves results: (e.g., Llama 64.3 BLEU; Qwen3 61.5), compared to 14–15 BLEU without lemma substitution. This confirms that lemma quality is a key bottleneck: reliable target-language lemmata allow LLMs to better exploit their language modelling capacity. In contrast, ZS gains little from Lemma, highlighting the importance of few-shot guidance.

Complexity Effects Using original SETIMES sentences (OrgSent) causes large performance drops across models and prompting strategies. On TR-SET600-OrgSent, xCOMET falls to 25–36 compared to 68–81 on simplified inputs, while chrF++ drops from 54–59 to 37–43. This confirms that longer and more complex sentences remain challenging when translating with glosses as a pivot.

Fine-tuning The fine-tuned FT-Qwen2.5-TR model performs comparably to the base model across settings. On TR-SET600 (FS), it achieves 58.6 chrF++ and 81 xCOMET, compared to 58.7 and 81 for base Qwen2.5; on TR-GLM (FS-ADV), it scores 27.2 compared to 28.7 BLEU. These results suggest that 1.3k Turkish gloss–translation pairs provide limited benefit beyond multilingual pretraining.

NMT vs. LLM With a few-shot approach (FS) on TR-SET600, LLMs (55–59 chrF++, 78–81 xCOMET) outperform the best NMT model, NMT100+SET1.8k (cf. Table 4, 45 chrF++, 77

xCOMET). On TR-GLM with gold glosses, LLMs are comparable to NMT (e.g., 30.2 vs. 29.1 BLEU with Qwen3). On TR-Morph-GLM with silver glosses (i.e., from our pipeline), FS LLM scores range from 9.7 to 15.4 BLEU, with NMT100 (14.0 BLEU) within this range, indicating comparable performance. However, with lemma substitution (with gold lemmata), LLMs surpass NMT substantially (up to 64.3 BLEU). Overall, NMT appears more robust to noisy representations, while LLMs (even fine-tuned) excel with cleaner inputs and benefit more from improved lemma quality.

7. Conclusion

We performed a case study of MT from Turkish into English using interlinear glosses as pivots, thereby updating the study from (Zhou et al., 2020). After processing the sentence with a morphological analyzer and a synthetic bilingual dictionary, we convert the source sentence into pseudo-glosses with lexical terms in English. We then pre-trained NMT models on subsets of GlossLM, a multilingual IGT corpus, to obtain fluent English translations. We show that pivoting through glosses remains better than a direct translation when parallel sentences are scarce, thanks to a multilingual pre-training on glosses. We also performed complementary analyses to observe the effect of pseudo-gloss quality on translation performance. Finally, we relied on LLMs as an alternative to NMT models for the gloss-to-target translation step.

Future work will extend the study to actual low-resource languages to go beyond the restricted resource scenario. We will mainly target languages with little available data, such as the SIGMORPHON Shared Task languages. We will also explore integrating multilingual automatic glossing models, such as GlossLM, to generate input glosses for translation.

8. Limitations

Our work studied Turkish, which is not a low-resource language, and higher MT performance can be achieved in absolute terms. We focused here on an artificially restricted resource scenario to simulate data availability for actual low-resource languages, allowing us to control the types of data we use. In this context, Turkish is an ideal candidate, as it is notably morphologically rich (an agglutinative language) and belongs to a different language family than English, our target language.

Our restricted-resource scenario also relies on two existing resources: a morphological analyzer and a bilingual dictionary (in our case, made from a larger parallel corpus through alignment). If both require linguistic knowledge and initial data to exist, they are realistically available. A lexicon is one of the first resources to be created during documentation and is widely available for languages. The analyzers we used rely on finite-state transducers rather than large-scale data training.

Moreover, our approach is designed with relatively short and simple sentences in mind, which are typical in many language documentation corpora. While this may be a limitation for real-world Turkish MT, where sentences are often more complex, it is less problematic for our extension to languages under documentation. Fieldwork corpora typically feature shorter and simpler sentences compared to domains such as news (e.g., SETIMES), which also motivated our use of a simplified version of the original SETIMES dataset. Although our simplified SETIMES dataset (SETIMES-S) is synthetic and generated by GPT-4, it is based on actual sentences, with a simplification task rather than generation from scratch.

9. Acknowledgments

We thank the anonymous reviewers for their insightful comments. This work was funded by the European Research Council (ERC) under grant agreement No. 101141712 - EPICAL. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

10. Bibliographical References

Rami Al-Rfou', Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Balthazar Bickel, Bernard. Comrie, and Martin Haspelmath. 2008. The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. Leipzig: Max Planck Institute for Evolutionary Anthropology, Department of Linguistics. https://www.eva.mpg.de/lingua/resources/glossing-rules.php.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Çağrı Çöltekin. 2010. A freely available morphological analyzer for Turkish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Çağrı Çöltekin. 2014. A set of open source tools for Turkish natural language processing. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1079–1086, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ryan Cotterell and Hinrich Schütze. 2015. Morphological word-embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1287–1292, Denver, Colorado. Association for Computational Linguistics.

Council of Europe. 2001. Common european framework of reference for languages: Learning, teaching, assessment. https://rm.coe.int/168045b15e.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128, Online. Association for Computational Linguistics.

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

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 *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Mercedes García-Martínez, Loïc Barrault, and Fethi Bougares. 2016. Factored neural machine translation architectures. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.

Dan Gillick, Cliff Brunk, Oriol Vinyals, and Amarnag Subramanya. 2016. Multilingual language processing from bytes. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1296–1306, San Diego, California. Association for Computational Linguistics.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024a. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

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. 2024b. GlossLM: A massively multilingual corpus and pretrained model for interlinear glossed text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12267–12286, Miami, Florida, USA. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine,

Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

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.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 49–52, New York City, USA. Association for Computational Linguistics.

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

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Ayla Kayabaş, Helmut Schmid, Ali E. Topcu, and Özlem Kılıç. 2019. Trmor: a finite-state-based

morphological analyzer for turkish. *Turkish Journal of Electrical Engineering & Computer Sciences*.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Philipp Koehn and Hieu Hoang. 2007. Factored translation models. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 868–876, Prague, Czech Republic. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proceedings of HLT-NAACL 2004: Short Papers*, pages 57–60, Boston, Massachusetts, USA. Association for Computational Linguistics.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Sebastian Nordhoff and Thomas Krämer. 2022. IMTVault: Extracting and enriching low-resource language interlinear glossed text from grammatical descriptions and typological survey articles. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 17–25, Marseille, France. European Language Resources Association.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly

Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Adnan Öztürel, Tolga Kayadelen, and Işın Demirşahin. 2019. A syntactically expressive morphological analyzer for turkish. In *Proceedings of the 14th International Conference on Finite-State Methods and Natural Language Processing*, pages 65–75, Dresden, Germany. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Rita Ramos, Everlyn Asiko Chimoto, Maartje Ter Hoeve, and Natalie Schluter. 2025. GrammaMT: Improving machine translation with grammar-informed in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29920–29940, Vienna, Austria. Association for Computational Linguistics.

Maciej Rapacz and Aleksander Smywiński-Pohl. 2025. Low-resource interlinear translation: Morphology-enhanced neural models for Ancient Greek. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 145–165, Abu Dhabi, United Arab Emirates. 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.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

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.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. A language-independent feature schema for inflectional morphology. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short*

Papers), pages 674–680, Beijing, China. 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.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

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

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

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

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

11. Language Resource References

Ginn, Michael and Tjuatja, Lindia and He, Taiqi and Rice, Enora and Neubig, Graham and Palmer, Alexis and Levin, Lori. 2024. *GlossLM: A Massively Multilingual Corpus and Pretrained Model for Interlinear Glossed Text*. Association for Computational Linguistics. PID https://huggingface.co/datasets/lecslab/glosslm-corpus.

Lewis, William D. and Xia, Fei. 2010. *Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages*. PID http://depts.washington.edu/uwcl/odin/.

Nordhoff, Sebastian and Krämer, Thomas. 2022. *IMTVault: Extracting and Enriching Low-resource Language Interlinear Glossed Text from Grammatical Descriptions and Typological Survey Articles*. European Language Resources Association. PID https://imtvault.org/?languageiso6393filter=tur.

Tiedemann, Jörg. 2012. *Parallel Data, Tools and Interfaces in OPUS*. European Language Resources Association (ELRA). PID https://opus.nlpl.eu/datasets/SETIMES?pair=en&tr.