GlossMATE: Multi-Agent Translator Explanations for Glosses

Changbing Yang[1], **Patrick Littell**[2], **Gabriel Bernier-Colborne**[2], **Yanfei Lu**[3], **Mengzhe Geng**[2]
[1]University of British Columbia,
[2]National Research Council Canada,
[3]University of Toronto
cyang33@mail.ubc.ca, patrick.littell@nrc-cnrc.gc.ca

Abstract

This paper introduces GlossMATE, a multi-agent critique-and-judge system that translates the gloss line in Interlinear Glossed Text (IGT) into fluent English using Large Language Models (LLMs). GlossMATE integrates linguist-provided resources (e.g., gloss-tag explanations, lexicon entries, curated IGT) with in-context learning and a multi-agent critique-and-judge procedure that iteratively evaluates and refines candidate translations. Our experiments show that leveraging analogous examples, explicit linguistic explanations, and collaborative agent interactions can enhance translation quality across several low-resource and polysynthetic languages. We also incorporate human linguists into the critique loop for selected languages. Case studies on three Indigenous languages further demonstrate the complementary strengths of human-in-the-loop feedback and multi-agent reasoning for language documentation tasks.

Keywords: Language documentation, Low-resource languages, Multi-agent systems

1. Introduction

Interlinear Glossed Text (IGT) is a standard format used by linguists for presenting linguistic data and morphological analysis. An IGT entry usually comprises four lines: (1) a phonological or orthographic transcription in the documented language, (2) a segmentation of words into morphemes, (3) corresponding grammatical glosses, and (4) a translation into English or another high-resource language. Here is an IGT example from Natügu:

(1) Orthography:	Mnctikr mrlcde wiki li
(2) Segmentation:	mnc-ti-kr mrlcde wiki li
(3) Gloss:	be-TR-1AUGI there week two
(4) Translation:	We stayed there two weeks.

Since the morpheme breakdowns and glosses are typically created by labor-intensive manual analysis by experts, most research on IGT has concentrated on generating lines (2) and (3) from (1) and (4) (Moeller and Hulden, 2018; Girrbach, 2023; Yang et al., 2024b; Zhao et al., 2020).

However, in this work, we address the complementary direction: turning glosses (3) into natural-language translations (4). This direction is motivated by a practical asymmetry in many documentation and revitalization workflows. Many languages already have tools that produce glosses at scale, including morphological analyzers that map surface forms to glosses (e.g. Harrigan et al., 2017; Micher, 2017; Wiemerslage et al., 2022) and conjugators that co-generate inflected forms and their gloss analyses (e.g. Kazantseva et al., 2018; Davis et al., 2021; Lu et al., 2024b). In such settings, glosses are useful as a sort of system-internal representation, but they can be difficult for learners and community members to understand, unless they have specialized training. Learners need explanations of meanings in plain-but-precise English. For instance, glossing a Michif verb as `SUBJCONJ-PRS-stop-laugh at someone-VTA.3SG-INV-1PL.EXCL`[1] would be useful to a linguist but incomprehensible to your average language-learner; what they really need to understand what it means is something like "if/when he/she stopped laughing at us (not including you)." The bottleneck is therefore not producing more glosses, but producing readable explanations that preserve the language's morphosyntactic distinctions while matching local pedagogical preferences.

Because the output domain of these analyzers expands exponentially with the number of morphemes covered, manual translation is impractical. For example, the system in Kazantseva et al. (2018) currently generates approximately 2.5 million forms, a scale that necessitated the use of an ad-hoc rule-based conversion script rather than human translation. However, the rule-based approach is still laborious and brittle: any change to the glossing "language", which happens frequently during the development of a morphological analyzer, will break the system and require a rewrite.[2]

[1] `SUBJCONJ`: introduces a subjunctive conjunct clause; `PRS`: present tense; `VTA.3SG`: transitive verb with a third-person singular animate agent; `INV`: inverse; `1PL.EXCL`: first-person plural exclusive patient ("we" not including the addressee).

[2] The authors of Kazantseva et al. (2018) informed us, p.c., that writing the system to turn glosses into readable English text took 3–6 person-months of work, and required constant revision as the system grew to encompass additional paradigms.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 9018–9029
11–16 May 2026.

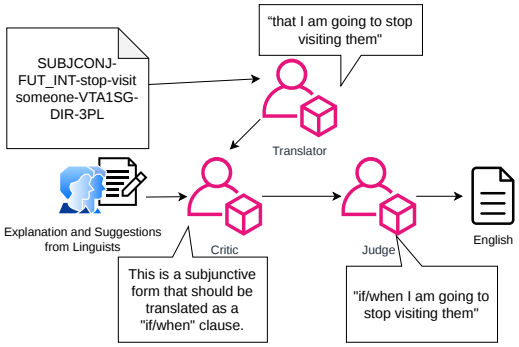

Figure 1: The pipeline of GLOSSMATE.

At the same time, Large language models (LLMs) offer a practical and more robust alternative to hand-engineered systems. Rather than relying on fixed rules, LLMs can map structured glosses and human-written explanations of grammatical gloss tags[3] directly to natural language, leveraging both symbolic cues in the gloss line and their generative capabilities.

In this paper, we propose GLOSSMATE (Multi-Agent Translator Explanations for Glosses), a human-in-the-loop pipeline that integrates linguists' expertise to guide and refine gloss-to-text translation (as shown in Figure 1) via LLMs. Our contributions are as follows:

1. We introduce a human-in-the-loop, multi-agent critique-and-judge pipeline for gloss-to-text translation that combines linguist input with in-context learning and structured critique. The application of a multi-agent workflow provides a more dynamic and adaptable approach to problem-solving that mirrors collaborative human review. We show that LLMs handle diverse gloss types and languages effectively, and that coupling them with expert feedback yields higher-quality translations.

2. We evaluate targeted few-shot selection strategies (choice and number of examples) to test how retrieval-augmented prompts improve translation quality.

3. We conduct practical case studies on three Indigenous languages spoken in Canada, demonstrating the effectiveness and feasibility of GLOSSMATE in real documentation settings.

[3]Morpheme glosses are commonly divided into *lexical* and *grammatical* glosses. In a label like `walk-PST`, `walk` is the lexical gloss (the core semantic unit), while `PST` is a grammatical gloss (typically uppercase) indicating a morphosyntactic category such as tense, aspect, or case.

2. The Challenge

Translating a gloss to English (or another high-resource language) might seem trivial, if one is only thinking of simple examples like `1SG-love-2SG` → "I love you". However, in morphologically complex languages, the task is actually quite delicate, especially when the morphemes encode morphosyntactic and semantic distinctions that English does not normally make. How should we best express the difference between inclusive and exclusive first-persons, in obviation, or in gender and tense distinctions that English does not make?

There is no *one* best way to translate these; rather, it comes down to the purpose of the system and the needs/preferences of its users. The "target language" here is not exactly English, but a systematic idiolect used for expressing the semantic distinctions of another language.

Likewise, the "source language", the gloss line, is an idiolect of its own. While linguists share broad glossing conventions like the Leipzig glossing rules[4], or the Unimorph abbreviation standard (Batsuren et al., 2022), they differ in the details: what is expressed in the gloss, what abbreviations are used, how far morpheme clusters are broken down, etc. Even for the same language, these glossing conventions are often specific to the document or system. Beyond shared assumptions about the syntax of glosses and abbreviation standards, there is no universal cross-linguistic representation for glosses, nor could there be; morphological analysis is always done towards some specific research or pedagogical purpose.

While the relationships between these "languages (source and target)" can be complex, they are at least *formulaic*, and it is possible to write an ad-hoc rule-based translator between them. However, as mentioned earlier, this approach is both time-consuming and fragile. Even minor modifications to the glossing conventions, such as adding a new affix or correcting existing errors, can render the previous implementation unusable. Moreover, the effort invested in developing such systems cannot be easily scaled or transferred across languages or projects.

Meanwhile, a statistical or neural machine translation approach runs into the problem of training data. These two "languages (source and target)" are often idiosyncratic to specific researchers/teachers/systems, and there is rarely training data of any scale, except that which the project team can make themselves.

Thus, the transition from training task-specific sequence-to-sequence models to leveraging large, general-purpose pretrained LLMs presents a

[4]https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

promising new direction. LLMs not only excel at generating fluent English from structured inputs but have also likely encountered linguistic glosses from a wide range of languages during pretraining. As a result, they may be able to perform gloss-to-English translation effectively in a zero-shot setting or with minimal few-shot in-context examples. Moreover, unlike rule-based pipelines, they are more resilient to variation: a change in the glossing conventions will not necessarily "break" the existing pipeline.

The challenge in using LLMs is that their zero-shot performance is not adequately aligned with two key needs. For one, there are many language-specific aspects of interpreting a gloss that one could not know just from the gloss itself; e.g., in experiential verbs, does the subject represent the experiencer or the stimulus? Neither a human nor LLM could simply guess this without familiarity with the language or its language family. The other need is to express things in line with the users'/schools' preferences for English renderings. Again, this is not something one could know *a priori*. For both of these, it is necessary to furnish the translator with examples and/or guidance, and much of what follows investigates means of doing so.

3. Related Work

In-context Learning for Low-Resource Languages In-context learning (ICL) has recently been established as a powerful alternative for low-resource data tasks in LLMs. By directly learning from ICL demonstrations provided in the prompt, the model can make inference without updating parameters or fine-tuning. Previous studies have shown that carefully selected few-shot exemplars can improve model generalization even when training data is limited or domain-specific (Cahyawijaya et al., 2024).

However, these systems will still struggle with many fieldwork documented languages because many rare or unseen words existing in those languages are underrepresented in pretraining corpora. To mitigate this challenge, researchers have begun integrating linguistic resources into ICL prompts, such as explanations from dictionaries and grammatical descriptions. They can help enrich contextual understanding and reduce hallucinations (Lu et al., 2024a; Aycock et al., 2025; Zhang et al., 2024; Yang et al., 2024b). Other work incorporates morphological analyzers to decompose input sentences into morphemes, improving compositional generalization and translation fidelity (Moisio et al., 2023; Ismayilzada et al., 2025).

These strategies are particularly valuable for linguistic documentation tasks, where examples tend to be morphologically rich and structurally complex. Based on prior approaches, our work combines human-authored gloss explanations in the provided ICL prompts to capture linguistic structure more effectively and improve translation quality.

Multi-agent Reasoning in LLMs Beyond text generation, LLMs have demonstrated capabilities in critique, evaluation, and reasoning. A growing body of work shows that LLM-as-a-judge approaches have shown strong correlation with human annotators across diverse natural language generation tasks (Bavaresco et al., 2025; Liu et al., 2023; Zheng et al., 2023).

Building upon this foundation, several recent studies have further demonstrated the effectiveness of multi-agent architectures for complex reasoning tasks. Multi-agent systems, where multiple LLM instances collaborate through structured interaction protocols, have shown particular promise in domains requiring diverse perspectives or specialized expertise (Wang et al., 2025; Li et al., 2023). Such approaches have been successfully applied to tasks ranging from code review and mathematical reasoning to factual verification (Tao et al., 2024; Chen, 2025; Du et al., 2024).

More than treating the LLM as an autonomous judge, our approach positions it as a collaborative assistant that reasons over explicit linguistic annotations and receives iterative feedback from domain experts. This design choice aligns with recent findings in meta-linguistic reasoning (Bean et al., 2024; Ginn et al., 2024; Yang et al., 2025), which demonstrate that while LLMs exhibit promising capabilities in reasoning over explicit linguistic descriptions and formal representations, they also benefit from structured feedback, expert supervision, and access to domain-specific knowledge bases.

4. Language and Data Preparation

We conduct experiments on two main data sources: (1) SIGMORPHON 2023 Shared Task datasets (Ginn et al., 2023), which provide diverse, annotated sentence-level IGT examples across multiple typologically distinct languages, and (2) verb conjugator datasets (Kazantseva et al., 2018; Lu et al., 2024b; Davis et al., 2021), used with permissions from their authors and language communities. These datasets provide dense mappings between inflected verb forms, glosses, and (in some cases) English translations.

The SIGMORPHON Shared Task datasets (Ginn et al., 2023) are fully open-source, allowing our experiments to be replicated and ensuring transparency in evaluation. In contrast, the conjugator datasets are not publicly available due to community data-protection policies (refer to Section 9 for details), but they serve as valuable controlled

Language	Family	Train	Dev	Test
Arapaho	Algonquian	39,501	4,938	4,892
Gitksan	Tsimshian	31	42	37
Lezgi	Nakh-Daghestanian	701	88	87
Natügu	Austronesian	791	99	99
Tsez	Nakh-Daghestanian	3,558	445	445

Table 1: Size of the SIGMORPHON shared task datasets (# samples) Ginn et al. (2023).

testbeds for assessing morphological generalization.

Together, we test our pipeline under two types of scenarios: One is naturalistic documentation settings (where glosses and translations come from human-annotated corpora) and the other one is controlled morphological settings (where glosses are generated systematically by conjugators).

4.1. SIGMORPHON Shared Task Data

The datasets released for the SIGMORPHON 2023 Shared Task on Interlinear Glossing (Ginn et al., 2023) include seven low-resource languages: Arapaho, Gitksan, Lezgi, Natügu, Nyangbo, Tsez, and Uspanteko. In our experiments, we exclude Nyangbo due to the absence of translation data, and Uspanteko because its translations are provided in Spanish rather than English, leaving five languages for analysis. These languages exhibit considerable typological diversity. For instance, Arapaho is highly agglutinative and polysynthetic, whereas Gitksan ranges from analytic to moderately synthetic and is not polysynthetic. For the SIGMORPHON datasets, we use the official training splits for constructing in-context examples and the test splits for evaluation. Details for the 5 languages we used here are shown in Table 1.

4.2. Verb Conjugator Datasets

The conjugator datasets (Kazantseva et al., 2018; Lu et al., 2024b; Davis et al., 2021) are derived from three community-built morphological tools for Kanyen'kéha (also known as Mohawk), Michif, and Oneida. Each provides machine-readable mappings between linguistic features (such as person, number, gender, direction, aspect) and corresponding inflected verb forms, accompanied by glosses and English translations (if applicable).

4.2.1. Kanyen'kéha Conjugator

The Kanyen'kéha dataset originates from Kawennón:nis: the Wordmaker for Kanyen'kéha (Kazantseva et al., 2018). It implements a symbolic finite-state model of Kanyen'kéha verbal morphology, designed to support language learners through automatic generation of conjugated verb forms. A sample of a single entry is shown in Figure 2.


```
"input": {
    "root": "7nahkwaya7k-r",
    "agent": "2-dl",
    "patient": "3-sg-n",
    "option": "command"
},
"output": [
    [0, "seni", ["agent", "pronoun"],
        "2-dl > 3-sg-n", ""],
    [1, "'nahkw\u00e1:ya'k", ["root", "
        root"], "verb", ""],
    [2, "", ["aspect", "command"], "IMP
        ", ""],
    [2, "", [], "", "(You two) beat a
        drum !"]
]
```


Figure 2: The figure shows a Kanyen'kéha example. Each generated entry includes the surface verb, its morpheme segmentation, gloss labels, and English translation, yielding structured data for our controlled experiments.

Kanyen'kéha is a polysynthetic Iroquoian language characterized by complex bound pronoun systems. It encodes both agent and patient features, and by extensive use of prepronominal prefixes and aspectual suffixes.

The Kanyen'kéha dataset contains 248,796 instances. Given the large size of the whole dataset, we simulate a sparse data scenario while preserving linguistic diversity. We draw compact yet representative subsets of 2,000 items for in-context learning retrieval and 2,000 items for testing. We aim to ensure that the dataset is "fair and representative", by which we mean that the sampled sets maintain the overall mix of morphological types present in the full corpus rather than over-selecting any single verb or inflectional class. We generate the dataset as follows:

Each example includes an `input.root` field (e.g., `7nahkwaya7k-r`). We parse this string at the final hyphen to obtain a verb stem and a class (here, verb stem $=$ `7nahkwaya7k`, class $=$ `r`[5]). We treat each $(\text{verb stem}, \text{class})$ pair as a bucket and assign every example to its corresponding bucket. We first load the full dataset and perform a single global shuffle to mitigate ordering bias (e.g., near-duplicates or topical clustering). Within each bucket, we also shuffle indices to avoid repeatedly selecting the same items across runs. We then allocate the sample proportionally: if a bucket b comprises a fraction p_b of the corpus, a sample of size N (e.g., $N = 2000$) targets approximately $p_b N$ items from b. Final allocations are rounded so that totals sum to exactly N, never overdraw any small bucket (allocations are capped by bucket size), and use minimal, seed-controlled randomness to break

[5]There are in total three classes in the Kanyen'kéha conjugator dataset: Class b stands for passive verb, class p stands for transitive verb, and class r stands for active verb.

```
"text": "ee-doo-paahpihitaahk",
"time": "PRS",
"gloss": "CONJPRSgolaugh at
    someoneVTA1PL.INCLINV2SG",
"context": "CONJ",
"preverb": "doo-",
"preverb_base": "doo-",
"preverb_translation": "go",
"verb": "paahpihVTA",
"verb_base": "paahpiheew",
"verb_translation": "laugh at someone",
"type": "VTA",
"verb_type": "VTA",
"subject": "VTA1PL.INCL",
"object": "VTA2SG",
"mood": "IMP",
"vta_type": "1",
"direction": "INV",
"subject_translation": "we and you",
"subject_base": "kiyanaw",
"object_translation": "you",
"object_base": "kiya"
```


Figure 3: This figure presents a sample of Michif. The conjugator encodes person/number agreement, transitivity type (e.g., `VTA`, `VAI`), and direct/inverse marking, and it produces detailed gloss fields (e.g., `VTA1PL.INCLINV2SG`). The resource primarily provides structured gloss-level information and feature encodings rather than authoritative sentence-level translations.

ties.

4.2.2. Michif conjugator

The Michif conjugator follows the system described by Davis et al. (2021), ported subsequently to the Gramble platform (Littell et al., 2024). Michif is a mixed Cree–French language exhibiting Algonquian-style verbal morphology alongside French-origin lexical roots. A Michif example is shown in Figure 3.

The Michif collection comprises 131,588 instances. Unlike the Kanyen'kéha dataset, these entries do *not* include a gold English translation line. To create a subset, we sample 100 entries for in-context learning examples and 100 test examples using similar proportional, bucketed protocol as for Kanyen'kéha: For Michif, we use each item's `verb` as the basis for defining a type. Then items are grouped into corresponding buckets, and a fixed-size sample is allocated proportionally. In addition, as no gold English translations are supplied by the Michif conjugator, human linguists provide reference English translations for all 200 sampled items.

4.2.3. Oneida conjugator

The Oneida data are derived from the conjugation engine presented by Lu et al. (2024b), developed in collaboration with the Oneida Nation of the Thames. Like Kanyen'kéha, Oneida is an Iroquoian language


```
"text": "takyunyʌ·ni'he'",
"object_translation": ">I(IMP)",
"object_base": "1SG",
"object": "1SG",
"subject_base": "2PL",
"subject": "2PL",
"subject_translation": "you_all>",
"root_base": "unyʌ·ni'he'",
"root": "unyʌ·ni'he'",
"root_translation": "_make(s)
    _something_for",
"subj_number": "PL",
"subj_gender": "n/a",
"obj_number": "SG",
"obj_gender": "n/a",
"subj_person": "2",
"subj_inclusivity": "n/a",
"obj_person": "1",
"obj_inclusivity": "n/a"
```


Figure 4: The figure shows an Oneida example. Each entry encodes person/number agreement and root translation.

with rich polysynthesis and dual-person marking on verbs. The conjugator generates morphologically complex verb forms annotated with morpheme segmentation. An example is shown in Figure 4.

The Oneida collection comprises 8,401 instances. Same as the case of Michif, no gold English translations are supplied by the conjugator, so human linguists provide reference English translations for the sampled items: We sample 100 in-context learning examples and 100 test examples using the same proportional, bucketed protocol as for Kanyen'kéha and Michif.

5. GlossMATE Pipeline

Our GLOSSMATE pipeline operates through the following three sequential stages:

1. Retrieval-Based In-Context Construction For a given test gloss, GLOSSMATE first constructs an In-Context Learning (ICL) prompt [6] by retrieving relevant examples from the training corpus. The objective is to provide the translation model (TRANSLATOR Agent) with semantically and morphologically aligned exemplars that mirror the target instance. We start the experiment with three retrieval strategies based on the SIGMORPHON Shared Task training splits: (i) Random sampling, which serves as a control baseline; (ii) Gloss-overlap retrieval, which selects examples that share the greatest number of overlapping gloss tokens with the test instance; and (iii) Distinctive morpheme retrieval: To identify informative grammatical cues for in-context examples, we first compute a *distinctiveness score* for each morpheme in the train-

[6] We provide our code and complete prompt template in our github repo: https://github.com/changbingY/GlossMATE.git.

ing set using a TF–IDF-inspired scheme. Let $D = \{d_1, \ldots, d_N\}$ denote the set of glossed examples and $V = \{m_1, \ldots, m_M\}$ the set of morphemes. For each morpheme m_i, we define:

$$\text{Distinct}(m_i) = \frac{\sum_{j=1}^{N} f(m_i, d_j)}{\sum_{k=1}^{M} \sum_{j=1}^{N} f(m_k, d_j)} \times \frac{N}{\text{DF}(m_i)},$$

where $f(m_i, d_j)$ is the frequency of m_i in example d_j, and $\text{DF}(m_i)$ is the number of examples in which m_i appears. This scoring balances term frequency (how representative a morpheme is) and inverse document frequency (how rare it is across examples), thereby highlighting morphemes that are frequent yet not ubiquitous.

Besides strategies of selecting ICL examples, we also evaluate models' sensitivity to context size: We vary the number of in-context examples from 1 to 5 and examine its effect on performance. Because the data structure of the verb conjugator dataset (Kazantseva et al., 2018; Lu et al., 2024b; Davis et al., 2021) differs from that of the SIGMORPHON shared task dataset, the analysis of in-context example count and sampling strategy is conducted solely on the SIGMORPHON dataset (Ginn et al., 2023). We then apply random sampling and gloss-overlap selections for the conjugator datasets. We use the best-performing configuration for the subsequent experiments.

2. Multi-Agent Critique and Refinement After the initial translation is produced by the primary translation agent (TRANSLATOR), we introduce a structured CRITIQUE stage designed to evaluate and refine the model output through explicit reasoning. This stage is implemented under a review-oriented prompt template[7] that enforces diagnosis and provides corrective suggestions for the translation.

Each critique prompt begins with a fixed instruction header defining the model's role as a meticulous translation reviewer and corrector. The reviewer receives the full input context used during generation, including the in-context examples, glossed sentence, along with the model's previous translation. We ask the model to return feedback identifying potential omissions, morphological mismatches, or stylistic inconsistencies in the translation. To enhance interpretability, the prompts for all critic agents also include **linguistically annotated gloss-tag explanations** that define the function of grammatical markers (e.g., tense, aspect, clusivity, inverse alignment) provided by fieldwork linguists or community members.

For three Indigenous languages (Kanyen'kéha, Michif, and Oneida), we additionally extend this step to a **human-in-the-loop critique mode**. Instead of relying solely on model-generated critiques, trained linguist annotators provide structured feedback following the same evaluation categories. This setting enables us to benchmark the interaction between automated and expert-guided critique, as detailed in Section 7.1.

3. Adjudication and Final Selection In the final stage, a JUDGE agent converts the critique into a single, publishable decision (the final translated sentence). Concretely, for each reviewed item, the JUDGE Agent receives (i) the full original context (ICL exemplars, gloss line, and gloss-tag explanations), (ii) the candidate translation produced before critique, and (iii) the prior critique.

6. Experiment Setup

We conduct experiments using three publicly available large language models (LLMs) of comparable scale but distinct architectural families: Qwen2.5-7B (Yang et al., 2024a), Gemma 3-12B (Team et al., 2025), and LLaMA3-8B (Dubey et al., 2024). All models are instruction-tuned and accessed via open-source platforms (the Hugging Face Hub). Inference is performed using the vLLM engine (Kwon et al., 2023). All experiments are executed on NVIDIA A6000 Ada GPUs. Details of model parameters are as follows: The temperature was set to 0.7 and the top-p parameter is 0.9. The maximum generation length was limited to 4096 tokens, and we use a repetition penalty of 1.1.

To reduce variance from random initialization, we use identical parameters across all language splits and repeat each experiment three times, reporting the averaged results. We further compute macro-averaged scores across languages to assess overall model performance.

For evaluation, we adopt three complementary metrics: BLEU, chrF, and Cosine Similarity between Sentence-BERT (Reimers and Gurevych, 2019) embeddings.

7. Results and Analysis

Our main results are shown in Table 3, Figure 5, and Figure 6.

In-context gains depend on model architecture and retrieval strategy rather than the sheer number of examples. Figure 5 illustrates BLEU, chrF, and Cosine similarity scores as a function of the number of in-context examples among SIGMORPHON Shared Task Datasets. While all models outperform their zero-shot baselines, the degree and consistency of improvement vary across models. The Gemma3-12B model exhibits a steady upward trend across all metrics, suggesting that larger models can effectively digest increasing contextual

[7] Full prompt template is available at: https://github.com/changbingY/GlossMATE.git.

evidence. In contrast, Qwen2.5-7B and LLaMA3-8B display non-monotonic behavior: performance often improves from one to three examples but plateaus or declines thereafter. These findings indicate that while adding in-context examples can improve performance, the effect is model-dependent.

Across retrieval strategies, both distinctive and overlap selections consistently yield higher scores than random sampling, confirming that semantically or structurally relevant exemplars are more beneficial than arbitrary ones. The overlap strategy usually outperforms the distinctive approach.

Multi-agent reasoning further enhances translation quality, though gains vary by model and language. Table 3 summarizes results across eight typologically diverse languages under three experimental settings: base generation (G), generation with in-context learning (G+ICL), and generation with both in-context learning and multi-agent critique (G+ICL+MA). Overall, introducing multi-agent feedback yields consistent improvements over both base and ICL-only configurations, with notable boosts in BLEU, chrF, and Cosine similarity for most languages.

Qwen2.5-7B exhibits the most stable and substantial gains under the multi-agent setting, achieving particularly large improvements for Lezgi and Tsez, where BLEU increases by up to 5–10 points beyond the ICL setting. In contrast, Gemma3-12B and LLaMA3-8B show more variable trends: the addition of multi-agent critique sometimes leads to modest or uneven improvements.

Providing linguistic explanations enhances model accuracy and semantic consistency across all architectures. Figure 6 compares models' performance with and without linguistically annotated gloss-tag explanations appended to the inputs. Across all three metrics, the inclusion of explanations yields clear improvements for each model. Gemma3-12B achieves the highest absolute scores, confirming its stronger capacity to internalize structured linguistic reasoning. Qwen2.5-7B and LLaMA3-8B also show steady gains, with BLEU increasing by 2–3 points and similar trends in chrF and semantic similarity.

These results suggest that explicit linguistically annotated gloss-tag explanations help models better interpret gloss–translation correspondences, improving both surface alignment and meaning preservation. The consistency of gains across architectures implies that even mid-sized LLMs can benefit from structured linguistic supervision.

7.1. Human-in-the-Loop Experiment

To assess whether expert feedback can efficiently improve already-usable model drafts, we conduct a human-in-the-loop experiment with trained linguists familiar with the target Indigenous languages. Importantly, our goal is not to replace expert translation, but to evaluate an assistive post-editing workflow: the relevant comparison is editing model outputs versus writing learner-facing explanations from scratch.

In this setup, each linguist reviews the output produced by the TRANSLATOR agent before the automated CRITIC and JUDGE stages. Annotators provide structured critiques targeting (i) adequacy with respect to the gloss (e.g., person/number, clusivity, argument roles, inverse/direct marking), (ii) grammatical alignment, and (iii) learner-facing naturalness and style. Across languages, annotators required approximately one hour per 100 items (about 36 seconds per item), reflecting a lightweight review process focused on correcting systematic adequacy mismatches rather than producing full translations.

We use these critiques as additional feedback exemplars in the prompt, and we further refine exemplar selection by retrieving critique examples jointly based on similarity in both the gloss and the model output. This strategy aims to surface recurring error patterns (e.g., consistent pronoun or role confusions) so that limited expert time is concentrated on the highest-impact issues.

Table 2 shows that incorporating human critiques generally yields additional improvements over the fully automatic pipeline, though gains are not uniform across languages. We emphasize that in the conjugator datasets, each instance is typically a single inflected verb form rather than a full sentence, which makes surface-overlap metrics such as BLEU/chrF brittle: correcting a single feature (e.g., inclusive vs. exclusive "we", or a direct/inverse role) may change only one or two tokens while substantially improving adequacy.

It is important to note that, in language documentation contexts, even low-frequency errors and fabrications can be unacceptable in pedagogical materials. From this perspective, human-in-the-loop review plays an important role as a quality-control mechanism: it helps ensure correctness and user-preferred phrasing once the model is already near-adequate, where post-editing can be fast while still preventing consequential mistakes.

8. Conclusion and Future Work

In this work, we presented GLOSSMATE, a human-in-the-loop, multi-agent framework for translating interlinear glossed text into fluent English. By combining linguist-provided gloss explanations, retrieval-augmented in-context learning, and a critique-and-judge workflow, GLOSSMATE achieves more accurate and interpretable translations across diverse low-resource languages. Our experiments show

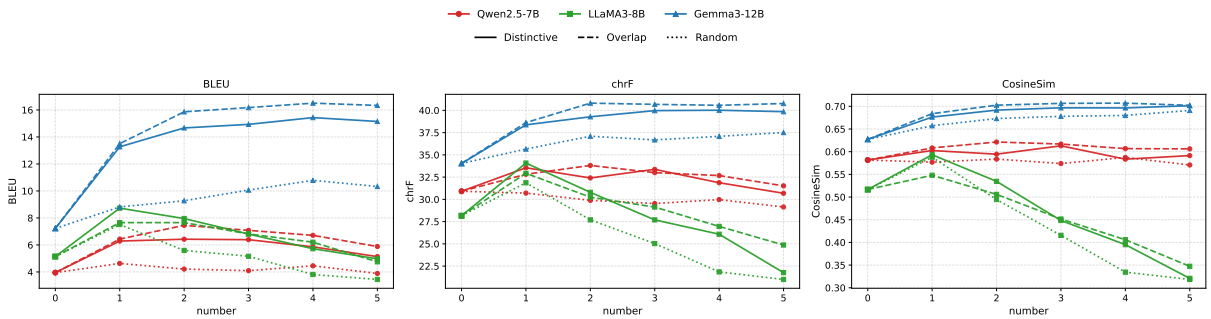

Figure 5: This figure presents the performance of three model architectures (Qwen2.5-7B, LLaMA3-8B, and Gemma3-12B) under different in-context learning (ICL) example extraction strategies for the SIGMORPHON Shared Datasets: Distinctive, Overlap, and Random, and varying numbers of ICL examples. The baseline denotes zero-shot prompting.

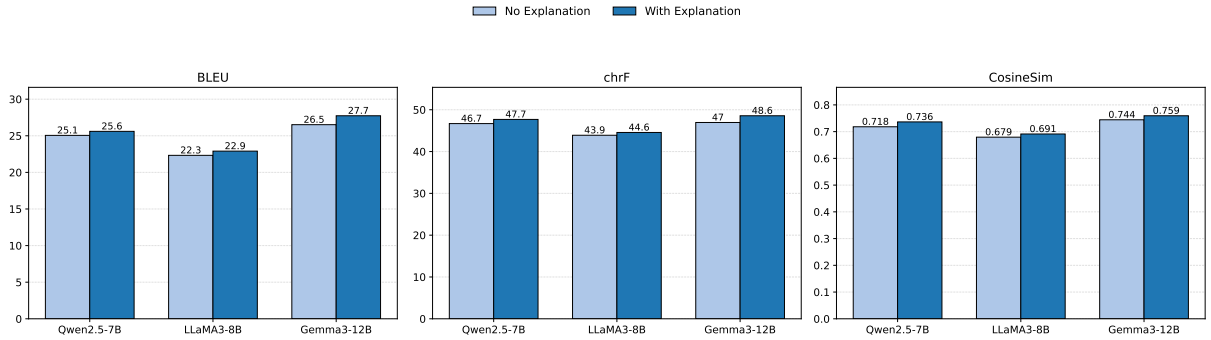

Figure 6: This figure presents the comparison between the with and without gloss explanation settings across all languages for three model types (Qwen2.5-7B, LLaMA3-8B, and Gemma3-12B). We calculate the average score across eight languages.

Language	Metric	G+ICL+MA	G+ICL+MA(+Human)
Kanyen'kéha	BLEU	31.29	**32.47**
	chrF	48.67	**50.32**
	CosineSim	0.762	**0.772**
Michif	BLEU	65.34	**67.32**
	chrF	79.02	**80.34**
	CosineSim	0.912	**0.913**
Oneida	BLEU	53.95	**53.96**
	chrF	**65.95**	65.92
	CosineSim	0.899	0.899

Table 2: Performance of Gemma3-12B across three Indigenous languages under two experimental settings: G+ICL+MA (Generation + In-Context Learning + Multi-Agent) and G+ICL+MA (+Human), where the critique stage includes human-in-the-loop feedback from trained linguists.

that structured linguistic supervision and multi-agent reasoning consistently enhance translation quality and robustness to gloss variation. Our work also highlights the promise of collaborative LLM systems that integrate expert knowledge to support language documentation and revitalization.

In future work, we plan to complement surface-overlap metrics with more semantic and structure-aware evaluation. One promising direction is a round-trip check: translate a gloss into English and then condition an LLM to recover the gloss (or a normalized set of features) from the English, comparing the reconstructed analysis to the original. This "back-translation" style evaluation may better capture whether crucial morphosyntactic distinctions (e.g., clusivity, obviation, inverse, aspect) are preserved, even when multiple English renderings are acceptable and BLEU/chrF are brittle for short verb-only outputs.

9. Limitations

While we recognize the importance of open data for reproducibility and future research, the verb conjugator datasets used in this study cannot be released publicly. These datasets are derived from community-built morphological tools originally designed for language learning, not for large-scale data dissemination. Public release poses potential risks of misuse or unintended exposure of culturally sensitive material. We respect the communities' preferences and data-governance policies, and we plan to explore pathways for broader accessibility

		Qwen2.5-7B			LLaMA3-8B			Gemma3-12B		
Language	**Metric**	**G**	**G+ICL**	**G+ICL+MA**	**G**	**G+ICL**	**G+ICL+MA**	**G**	**G+ICL**	**G+ICL+MA**
Arapaho	BLEU	3.45	5.21	6.75	3.29	9.14	5.18	8.27	**12.77**	11.43
	chrF	25.83	28.15	29.76	22.71	30.26	27.91	30.66	35.32	**35.98**
	CosineSim	0.555	0.595	0.610	0.488	0.616	0.574	0.627	**0.689**	0.679
Tsez	BLEU	4.98	5.95	10.12	3.88	8.03	5.31	7.65	**14.75**	13.67
	chrF	35.55	36.69	40.51	22.67	33.85	32.59	37.59	**42.66**	42.23
	CosineSim	0.678	0.703	0.740	0.356	0.627	0.645	0.705	0.774	**0.776**
Natügu	BLEU	2.93	7.34	9.57	6.91	9.64	7.65	6.54	13.60	**14.75**
	chrF	29.42	33.57	33.87	30.19	33.23	32.56	31.01	**37.88**	32.79
	CosineSim	0.552	0.658	0.668	0.593	0.621	0.602	0.598	**0.731**	0.708
Lezgi	BLEU	4.78	12.18	23.13	5.97	13.76	13.40	9.33	**34.11**	22.44
	chrF	30.35	36.66	47.94	30.59	37.24	38.47	35.18	**50.50**	42.66
	CosineSim	0.576	0.615	**0.749**	0.586	0.536	0.606	0.620	0.707	0.679
Gitksan	BLEU	3.61	6.79	6.82	5.66	5.61	5.89	4.30	7.26	**8.98**
	chrF	33.40	34.43	37.23	34.65	32.42	35.45	35.71	38.07	**41.16**
	CosineSim	0.544	0.551	0.585	0.558	0.518	0.584	0.583	0.636	**0.657**
Kanyen'kéha	BLEU	7.86	28.81	29.69	6.79	20.10	30.35	7.54	29.56	**31.29**
	chrF	27.00	46.75	47.95	27.67	35.17	45.44	29.76	40.74	**48.67**
	CosineSim	0.61	0.740	**0.763**	0.600	0.690	0.733	0.625	0.716	0.762
Michif	BLEU	2.06	65.99	66.00	3.72	62.35	63.23	4.67	**67.54**	65.34
	chrF	16.46	78.34	78.34	15.86	77.85	78.01	16.98	78.89	**79.02**
	CosineSim	0.303	0.901	0.901	0.312	0.894	0.899	0.325	0.912	**0.912**
Oneida	BLEU	12.19	50.38	52.75	12.96	51.23	52.21	14.64	53.10	**53.95**
	chrF	36.51	65.05	65.89	36.98	65.42	66.10	37.60	**66.30**	65.95
	CosineSim	0.644	0.865	0.872	0.652	0.871	0.885	0.666	0.879	**0.899**

Table 3: Evaluation across all languages with metrics as rows and models/settings as columns. G= basic gloss setting; ICL = In-context learning; MA = Multi-Agent.

through consent-based and ethically aligned collaborations in the future.

10. Acknowledgments

With deep appreciation, we extend our gratitude to the members of the Oneida Nation of the Thames, the Prairies to Woodlands Indigenous Language Revitalization Circle, and Onkwawenna Kentyohkwa for their support and collaboration. We are also sincerely grateful to our colleagues at NRC: Akwiratékha' Martin, Anna Kazantseva, Sowmya Vajjala, Jackie Lo, and Rebecca Knowles, for their valuable insights, thoughtful feedback, and continued support throughout this work. We would also like to thank Garrett Nicolai from UBC for his valuable suggestions on this project. We are indebted as well to all those who contributed to this project or supported us during this journey. Without them, this work would not have been possible.

11. Bibliographical References

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. 2025. Can llms really learn to translate a low-resource language from one grammar book? In *ICLR*.

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastasopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóǧa, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh

Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. UniMorph 4.0: Universal Morphology. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, et al. 2025. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. In *The 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*.

Andrew Michael Bean, Simeon Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, Ryan Andrew Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

Yujia Chen. 2025. Autoreview: An llm-based multi-agent system for security issue-oriented code review. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, pages 1022–1024.

Fineen Davis, Eddie Antonio Santos, and Heather Souter. 2021. On the computational modelling of michif verbal morphology. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. 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.

Leander Girrbach. 2023. Tü-CL at SIGMORPHON 2023: Straight-through gradient estimation for hard attention. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 171–185, Toronto, Canada. Association for Computational Linguistics.

Atticus Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of plains cree verbs. *Morphology*, 27.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, Bhuwan Dhingra, Antoine Bosselut, Duygu Ataman, and Lonneke Van Der Plas. 2025. Evaluating morphological compositional generalization in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1270–1305, Albuquerque, New Mexico. Association for Computational Linguistics.

Anna Kazantseva, Owennatekha Brian Maracle, Ronkwe'tiyóhstha Josiah Maracle, and Aidan Pine. 2018. Kawennón:nis: the wordmaker for Kanyen'kéha. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 53–64, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language

model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Qintong Li, Leyang Cui, Lingpeng Kong, and Wei Bi. 2023. Collaborative evaluation: Exploring the synergy of large language models and humans for open-ended generation evaluation.

Patrick Littell, Darlene Stewart, Fineen Davis, Aidan Pine, and Roland Kuhn. 2024. Gramble: A tabular programming language for collaborative linguistic modeling. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7913–7925, Torino, Italia. ELRA and ICCL.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024a. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Yanfei Lu, Patrick Littell, and Keren Rice. 2024b. Empowering Oneida language revitalization: Development of an Oneida verb conjugator. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5757–5767, Torino, Italia. ELRA and ICCL.

Jeffrey Micher. 2017. Improving coverage of an Inuktitut morphological analyzer using a segmental recurrent neural network. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 101–106, Honolulu. Association for Computational Linguistics.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Anssi Moisio, Mathias Creutz, and Mikko Kurimo. 2023. Evaluating morphological generalisation in machine translation by distribution-based compositionality assessment. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 738–751, Tórshavn, Faroe Islands. University of Tartu Library.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Wei Tao, Yucheng Zhou, Yanlin Wang, Wenqiang Zhang, Hongyu Zhang, and Yu Cheng. 2024. Magis: Llm-based multi-agent framework for github issue resolution. *Advances in Neural Information Processing Systems*, 37:51963–51993.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. 2025. Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems. *arXiv preprint arXiv:2502.11098*.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.

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, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025. LingGym: How far are LLMs from thinking like field linguists? In *Proceedings of the*

2025 Conference on Empirical Methods in Natural Language Processing, pages 1314–1340, Suzhou, China. Association for Computational Linguistics.

Changbing Yang, Garrett Nicolai, and Miikka Silfverberg. 2024b. Multiple sources are better than one: Incorporating external knowledge in low-resource glossing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4537–4552, Miami, Florida, USA. Association for Computational Linguistics.

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

Xingyuan Zhao, Satoru Ozaki, Antonios Anastasopoulos, Graham Neubig, and Lori Levin. 2020. Automatic interlinear glossing for under-resourced languages leveraging translations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5397–5408, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.