Hybrid Neural-LLM Pipeline for Morphological Glossing in Endangered Language Documentation: A Case Study of Jungar Tuvan

Siyu Liang[1], **Talant Mawkanuli**[2], **Gina-Anne Levow**[1]
[1] Department of Linguistics, University of Washington
[2] Department of Middle Eastern Languages and Cultures, University of Washington
{liangsy, tmawkan, levow}@uw.edu

Abstract

Interlinear glossed text (IGT) creation remains a major bottleneck in linguistic documentation and fieldwork, particularly for low-resource morphologically rich languages. We present a hybrid automatic glossing pipeline that combines neural sequence labeling with large language model (LLM) post-correction, evaluated on Jungar Tuvan, a low-resource Turkic language. Through systematic ablation studies, we show that retrieval-augmented prompting provides substantial gains over random example selection. We further find that morpheme dictionaries paradoxically hurt performance compared to providing no dictionary at all in most cases, and that performance scales approximately logarithmically with the number of few-shot examples. Most significantly, our two-stage pipeline combining a BiLSTM-CRF model with LLM post-correction yields substantial gains for most models, achieving meaningful reductions in annotation workload. Drawing on these findings, we establish concrete design principles for integrating structured prediction models with LLM reasoning in morphologically complex fieldwork contexts. These principles demonstrate that hybrid architectures offer a promising direction for computationally light solutions to automatic linguistic annotation in endangered language documentation.

1 Introduction

Interlinear glossed text (IGT) is essential for linguistic documentation and preservation, aligning language transcriptions with morpheme segmentation, glosses, and translations (Lehmann, 2004a). Despite its importance, IGT creation remains highly labor-intensive, creating bottlenecks in language documentation projects (Chelliah and Reuse, 2010). Recent advances in neural sequence models and large language models offer new possibilities for automated IGT generation, yet each approach has limitations: structured models lack flexibility and world knowledge, while LLMs struggle with consistency and require extensive in-context examples or expensive fine-tuning, all inaccessible in most real-life use cases.

We present a hybrid pipeline that combines a BiLSTM-CRF model for initial gloss prediction with LLM-based post-correction, evaluated on Jungar Tuvan (henceforth Tuvan), a morphologically complex Turkic language. Through systematic experiments across four LLMs and multiple design choices, we demonstrate that this two-stage approach substantially improves over the BiLSTM baseline for most models. Our ablation studies reveal key design principles: retrieval-augmented prompting significantly outperforms random example selection; morpheme dictionaries generally hurt performance for most models; and optimal few-shot parameters range from five to fifteen examples.

Our contributions are the following: (1) we present a hybrid architecture combining structured prediction with LLM reasoning for automatic glossing; (2) we carry out comprehensive ablation studies establishing design principles for retrieval strategies, glossary configurations, and few-shot scaling; (3) we provide comparative evaluation of model performance across generation versus correction tasks, providing evidence-based guidance in fieldwork contexts.

2 Related Work

2.1 IGT and Language Documentation

IGT serves as a standard format in field linguistics, encoding source language transcriptions, morphological segmentation, gloss labels, and free translations (Lehmann, 2004b; Comrie et al., 2008). For many endangered and low-resource languages, IGT represents the primary digitized documentation (Chelliah and Reuse, 2010; Hargus et al., 2020).

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics, pages 16–30
March 26, 2026. ©2026 Association for Computational Linguistics

Recent work has developed tools for IGT extraction from grammatical descriptions (Schenner and Nordhoff, 2016; Round et al., 2020; Nordhoff and Krämer, 2022) and multi-modal IGT generation from speech (He et al., 2024). The SIGMORPHON 2023 shared task on automatic IGT generation (Ginn et al., 2023) further stimulated interest in computational approaches to glossing, leading to subsequent work on large-scale IGT modeling and evaluation, including pretrained language models for glossing (Ginn et al., 2024b) and LLM-based prompting approaches for low-resource IGT generation (Elsner and Liu, 2025). While these efforts demonstrate steady progress on benchmark datasets, they have not yet displaced existing fieldwork practices; in most documentation projects, IGT creation remains largely manual, relying on tools such as ELAN and FLEx (Wittenburg et al., 2006; International, 2025). Traditional workflows also include rule-based morphological parsers and deterministic dictionary lookup within tools like FLEx, which supports semi-automatic glossing and lexicon building from texts; our approach is intended to complement rather than replace such methods.

2.2 Automatic Morphological Analysis and Glossing

Traditional approaches to automatic glossing employ structured prediction models. Sequence-to-sequence architectures have been applied to morphological segmentation (Ruzsics and Samardžić, 2017; Liu et al., 2021; Rice et al., 2024), while CRF-based models have proven effective for morphological tagging (Buys and Botha, 2016; Malaviya et al., 2018). BiLSTM-CRF architectures in particular balance local pattern recognition with global constraints (Ma and Hovy, 2016; Cotterell and Heigold, 2017), achieving strong performance on sequence labeling tasks in morphologically rich languages.

Recent work specifically targeting IGT generation has explored neural encoder-decoder models with translation data (Zhao et al., 2020), CRF-based approaches for low-resource scenarios (Barriga Martínez et al., 2021; Okabe and Yvon, 2023), and lightweight models using structured linguistic representations (Shandilya and Palmer, 2023). Moeller et al. (2020) demonstrate how IGT can support downstream morphological analysis tasks. However, these models require substantial annotated corpora and struggle with rare morphemes and novel combinations. The recent work by Rice et al. (2025) also identifies significant gaps between computational morphology research outputs and real-world language documentation needs, highlighting the importance of user-centered design.

2.3 LLMs for Linguistic Annotation

Large language models have shown promise for linguistic annotation tasks in low-resource settings. Recent work (Ginn et al., 2024a; Elsner and Liu, 2025) explore LLM-based gloss prediction and prompting strategies for IGT, demonstrating that prompt design and example selection substantially affect performance. Zhang et al. (2024) show that providing dictionaries and grammar sketches enables translation for unseen languages. Yang et al. (2025b) evaluate models on metalinguistic reasoning using reference grammars and IGT, though their benchmark relies on curated reference grammar data without the full fieldwork contexts. LLM post-correction and refinement steps are also widely explored across NLP tasks as cascaded or post-editing stages (Zouhar et al., 2021; Izacard et al., 2023).

Among recent studies, few-shot prompting has emerged as a key technique for adapting LLMs to specialized tasks. Studies demonstrate that careful selection and presentation of in-context examples significantly impacts performance (Logan IV et al., 2022; Winata et al., 2021), with retrieval-based example selection often outperforming random selection (Stahl et al., 2024). However, LLMs face challenges in low-resource settings: they require extensive in-context examples (increasing inference cost), struggle with paradigmatic consistency, and lack the inductive biases of structured sequence models.

2.4 Hybrid and Multi-Stage Architectures

Hybrid approaches combining multiple model types have proven effective across NLP tasks. Retrieval-augmented generation (RAG) enhances LLMs by dynamically incorporating relevant examples (Lewis et al., 2021; Jiang et al., 2023), with recent work exploring specialized RAG architectures for domain adaptation (Siriwardhana et al., 2023; Yu, 2022). Cascaded architectures leverage specialized models for different subtasks (Izacard et al., 2023), while post-processing steps that refine outputs using external knowledge have shown consistent gains (Zouhar et al., 2021). Our work extends these ideas to morphological annotation, proposing a two-stage pipeline where a BiLSTM-

CRF provides initial structure and an LLM refines predictions through contextual inference and consistency checks.

3 Data

3.1 Language and Corpus

Tuvan is a Turkic language spoken in the Republic of Tuva of the Russian Federation, Mongolia, and the Xinjiang Uyghur Autonomous Region of China, with approximately 280,000 speakers across these regions (Harrison and Anderson, 2002). The present study focuses on the variety of Jungar Tuvan, spoken in the Altay region of Xinjiang, China. We treat Jungar Tuvan as a low-resource variety used in documentation contexts; we do not adjudicate its formal endangerment status in this paper. Jungar Tuvan shares the core typological properties of Tuvan—canonical agglutinative morphology, extensive case marking (nominative, accusative, genitive, dative, locative, ablative, comitative), complex aspectual systems, and productive derivational morphology—while also exhibiting vowel harmony and consonant alternations that create allomorphic variation (Mawkanuli, 1999, 2005).

Our corpus comprises 895 IGT-annotated sentences drawn from data collected during fieldwork in Xinjiang, China from the 1987 to 1995. The data span 40 recording sessions across conversational registers and narratives. All IGT annotations were produced manually following a consistent project-specific schema informed by typological conventions (Lehmann, 2004b; Comrie et al., 2008). Glosses distinguish lexical items (e.g., *money*, *give*) from grammatical morphemes (e.g., DAT, 1SG, PRS).

Table 1 summarizes corpus statistics. The tagset comprises 240 unique gramamtical morpheme labels (e.g., 1SG, PST), and 1258 unique content word glosses.

Metric	#
Total sentences	895
Narratives	40
Unique grammatical morphemes	240
Unique content morphemes	1258
Average words per sentence	8.38 (± 5.71)
Average morphemes per word	1.69 (± 0.30)

Table 1: Corpus statistics for the Tuvan fieldwork dataset.

Example 1 illustrates a representative IGT instance from the corpus, demonstrating the alignment structure our models must learn.

(1) jïlgï-nan iyi joon bar
horse-ABL two big EXIST
"(We) have two big horses."

3.2 Data Split

We perform a train-test split at the document level, allocating approximately 85% of sentences (760) to training and 15% (135) to testing. To avoid information leakage, no segments of the same narrative appear in both splits, ensuring that models cannot exploit discourse-level or speaker-specific patterns from related utterances. We further verify the absence of near-duplicate sentences using character-level TF-IDF (term frequency–inverse document frequency) cosine similarity with a threshold of 0.95.

4 Methodology

4.1 Task Formalization

We frame glossing as a structured prediction problem: given a hyphen-segmented Tuvan utterance $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ where each x_i represents a morpheme, produce a parallel sequence of gloss labels $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ where each y_i is drawn from a tagset. We assume gold morpheme boundaries and do not use translations in the glossing model; segmentation and glossing are treated as separate steps. We evaluate using token-level accuracy, defined as:

$$\text{Accuracy} = \frac{1}{N}\sum_{i=1}^{N}[\hat{y}_i = y_i] \quad (1)$$

where N is the total number of morphemes in the test set, $\hat{y}_i$ is the predicted gloss, and y_i is the reference gloss. This metric directly reflects annotation workload reduction: higher accuracy means fewer manual corrections required.

4.2 BiLSTM-CRF Model

Our baseline employs a two-layer bidirectional LSTM with CRF decoding, widely used for sequence labeling (Lample et al., 2016; Ma and Hovy, 2016; Huang et al., 2015). The model uses 100-dimensional character-level embeddings (randomly initialized and trained from scratch), 128-dimensional hidden layers, and learns to predict gloss labels for segmented morphemes. We

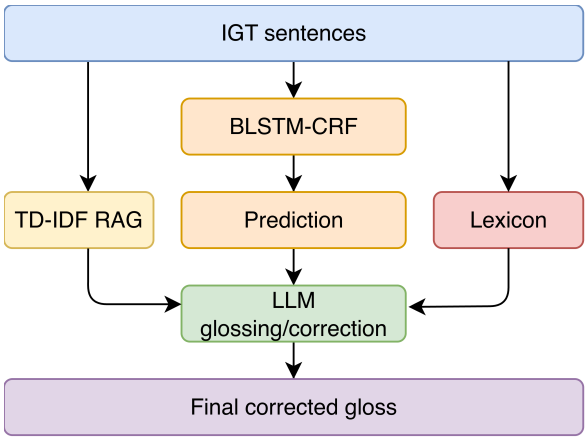

Figure 1: Hybrid pipeline combining BiLSTM-CRF structured prediction with LLM post-correction using retrieval-augmented prompting.

train for up to 100 epochs with early stopping (patience=10) on validation loss. This architecture captures local morphological patterns but cannot leverage broader linguistic knowledge or rare paradigms not well-represented in training data.

4.3 LLM Configuration and Prompting

We evaluate four LLMs: `deepseek-v3.2-exp` (DeepSeek-AI et al., 2025), `qwen3-max` (Yang et al., 2025a), `gpt-4o-mini` (OpenAI et al., 2024), and `gemma-3-27b-it` (Kamath et al., 2025). All models use greedy decoding with temperature zero for deterministic outputs. Prompts follow a consistent template: natural language instruction, k retrieved examples showing morpheme segmentation and gloss pairs, optional morpheme dictionary, and the test input. For the hybrid pipeline, prompts additionally include the BiLSTM prediction as a hypothesis to correct, framed as a "rough initial attempt" requiring verification. Complete prompt templates for all experiments are provided in Appendix A. Figure 1 illustrates our hybrid pipeline architecture.

4.4 Experimental Design

We conduct several experiments testing both LLM generation and hybrid correction. In Experiment 1 (Retrieval vs. Random Selection), we compare character-level TF-IDF (term frequency–inverse document frequency) cosine similarity-based retrieval against uniform random sampling for selecting three in-context examples to explore the effect of similarity-based retrieval. Retrieval operates at the sentence level: for each test sentence, we retrieve the most similar training sentences based on their Tuvan source text representations. Experiment 2 (N-Shot Scaling) varies example count from 1 to 20 with RAG and no glossary, mapping the accuracy-cost tradeoff for RAG LLM generation.

Experiment 3 (Glossary Ablation) tests four glossary configurations—none, top-100 most frequent morphemes, all grammatical morphemes, and the entire 1,498-pair dictionary—all using three-shot RAG to reveal whether partial dictionaries help or hinder performance. The glossary is provided to the LLM within the prompt as a plain-text key:value list (Appendix A), rather than as a deterministic lookup table. Finally, Experiment 4 (Hybrid Pipeline) evaluates BiLSTM plus LLM correction with varying n-shot counts including a zero-shot condition that tests whether LLMs can correct predictions without in-context examples. These ablations correspond to fieldwork-relevant choices about retrieval quality, example budget, and availability of lexical resources. Detailed prompt templates for RAG LLM generation (Experiments 1–3) and hybrid correction (Experiment 4) are provided in Appendix A.

5 Results

5.1 Baseline: BiLSTM-CRF Performance

Our BiLSTM-CRF baseline achieves 0.474 token accuracy on the test set with training data of 760 sentences. The model learns frequent morphological patterns (case markers, possessives, tense/aspect) but struggles with infrequent lexical morphemes and combinations of grammatical morphemes not attested in the training data. Error analysis reveals that 0.38 of errors involve lexical items appearing fewer than 5 times in training, and 0.24 involve grammatical morphemes in novel combinations.

5.2 Experiment 1: Retrieval vs. Random Selection

Table 2 shows the effect of retrieval enhancement across all four LLMs using 3-shot prompting without glossary.

Retrieval-augmented generation provides meaningful improvements across all models: +0.388 for `deepseek-v3.2-exp`, +0.319 for `qwen3-max`, +0.293 for `gpt-4o-mini`, and +0.276 for `gemma-3-27b-it`. `deepseek-v3.2-exp` achieves the highest absolute accuracy with RAG (0.506), while all models show substantial gains from retrieval. The consistent gains across

Model	Random	RAG
deepseek-v3.2-exp	0.118	**0.506**
qwen3-max	0.062	0.381
gpt-4o-mini	0.103	0.396
gemma-3-27b-it	0.068	0.344

Table 2: Retrieval-augmented prompting (RAG) vs. random example selection across four LLMs (3-shot, no glossary). All models show meaningful improvement with RAG.

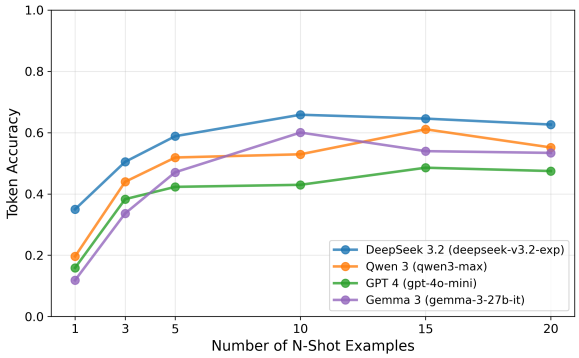

Figure 2: Experiment 2: n-shot scaling curves for RAG LLM generation. Performance scales approximately logarithmically with example count, plateauing around n=10–15 for most models. The BiLSTM baseline (0.474) is provided in the text for reference.

architectures demonstrate that similarity-based example selection is beneficial for morphological glossing tasks.

5.3 Experiment 2: N-Shot Scaling

Using the same TF-IDF-based retrieval from Experiment 1, we vary the number of retrieved examples from 1 to 20 without providing any glossary. Figure 2 shows performance scaling with example count (see Table 5 in Appendix B for detailed values).

Performance scales approximately logarithmically with example count. `deepseek-v3.2-exp` peaks at n=10 (0.658), then slightly declines at n=15 (0.646) and n=20 (0.626), while `qwen3-max` shows continued gains up to n=15 (0.611). `gpt-4o-mini` peaks at n=15 (0.486), and `gemma-3-27b-it` achieves 0.600 at n=10 before declining to 0.534 at n=20. The consistent pattern across models indicates diminishing marginal returns beyond 10 to 15 examples, with some models showing degradation at higher values. This decline may reflect either model saturation (distraction from excessive context) or retrieval quality degradation (less similar examples as the pool expands). We use character-level TF-IDF cosine similarity for retrieval; alternative similarity measures such as edit distance or contextualized embeddings might exhibit different scaling behavior, though we leave this investigation to future work. These results suggest practical operating points around n=5 to 10 for cost-sensitive applications and n=10 to 15 for maximum accuracy.

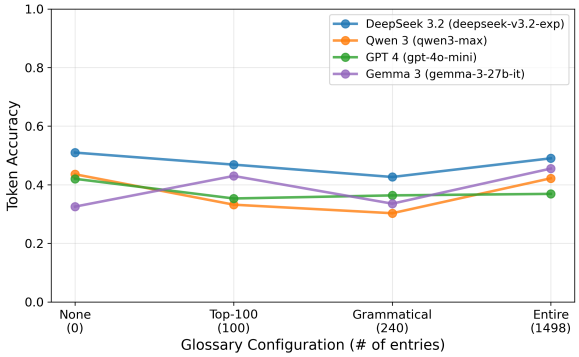

Figure 3: Experiment 3: glossary ablation results. Partial glossaries (Top-100, Grammatical) hurt performance compared to no glossary, while complete glossaries show modest gains. The negative effect suggests models are usually distracted by morphological information. The BiLSTM baseline (0.474) is provided in the text for reference.

5.4 Experiment 3: Glossary Ablation

To assess the impact of morpheme dictionaries on performance, we test four glossary configurations using 3-shot RAG: None (no dictionary provided), Top-100 (the 100 most frequent morpheme-gloss pairs), Grammatical (all 240 grammatical morpheme-gloss pairs), and Entire (the complete 1,498-pair dictionary including both grammatical and lexical morphemes). Figure 3 shows the results (see Table 4 in Appendix B for detailed values).

Counter-intuitively, providing morpheme dictionaries generally hurts performance. Partial glossaries consistently degrade accuracy across all models: Top-100 causes drops ranging from 0.041 to 0.104, while Grammatical shows similar or worse declines. Even the complete 1,498-pair dictionary fails to help for most models, with `deepseek-v3.2-exp`, `qwen3-max`, and `gpt-4o-mini` all performing worse with the entire glossary than with none (losses of 0.019, 0.014, and 0.051 respectively). Only `gemma-3-27b-it` benefits from dictionary information, achieving 0.455 with the entire glossary (+0.130 over None).

We lack direct evidence about whether models

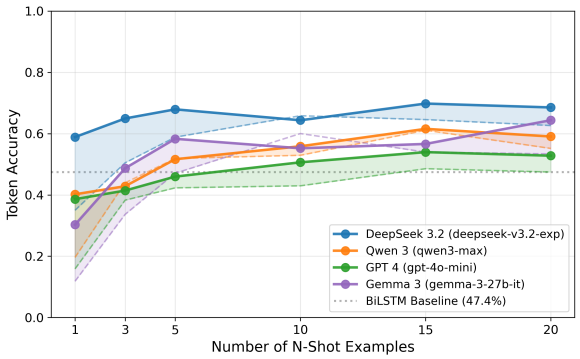

Figure 4: Experiment 4: hybrid pipeline improvement over RAG LLM generation. Solid lines show hybrid accuracy (BiLSTM + LLM correction), dashed lines show pure N-Shot baseline from Experiment 2, and shaded areas indicate improvement. The hybrid approach consistently improves performance across all four models, particularly in low-shot scenarios (n=1–5).

actually consult dictionary entries versus simply becoming distracted by additional prompt material. The degradation could reflect either inappropriate reliance on dictionary lookups for context-dependent glossing decisions, or models' difficulty balancing multiple information sources (retrieved examples, dictionaries, and morphological patterns). The no-glossary condition forces models to extract morphological structure from aligned examples, which appears more effective than dictionary consultation for most architectures, though we cannot definitively isolate the causal mechanism without fine-grained analysis of model predictions.

5.5 Experiment 4: Hybrid Pipeline Performance

We then evaluate the hybrid pipeline using the same TF-IDF retrieval as in Experiments 1 and 2, but providing the BiLSTM-CRF predictions as initial hypotheses for LLM correction. We test with n=1, 3, 5, 10, 15, 20 retrieved examples, all without glossaries. Figure 4 shows improvement from the hybrid pipeline over RAG LLM generation across varying n-shot counts for all four models.

The hybrid pipeline substantially outperforms pure BiLSTM predictions across all models. `gemma-3-27b-it` achieves 0.644 at n=20, a +0.170 gain over the BiLSTM baseline (0.474) and +0.110 improvement over its pure generation performance (0.534 at n=20). `deepseek-v3.2-exp` reaches 0.698 at n=15 (+0.224 over BiLSTM, +0.052 over pure generation), `qwen3-max` achieves 0.616 (+0.142 over BiLSTM, +0.005 over pure generation), and `gpt-4o-mini` reaches 0.540 (+0.066 over BiLSTM, +0.054 over pure generation), demonstrating that the hybrid approach benefits all models across performance tiers.

The improvements are particularly substantial in low-shot scenarios: at n=1, `gemma-3-27b-it` shows +0.185 improvement over pure generation (0.118 $\rightarrow$ 0.303), `deepseek-v3.2-exp` improves by +0.058, and `gpt-4o-mini` by +0.063. This demonstrates that the BiLSTM provides valuable structural guidance when few examples are available. As n increases, the gains diminish but remain consistent across all models, with `gemma-3-27b-it` showing +0.110 improvement even at n=20.

5.6 Error Analysis

To understand where improvements originate, we analyze errors by morpheme type. Comparing BiLSTM baseline against the best-performing hybrid configuration (`deepseek-v3.2-exp` with $n = 10$ RAG examples), we find asymmetric improvements.

The BiLSTM baseline achieves 0.923 accuracy on 168 grammatical morphemes but only 0.479 on 213 lexical morphemes. This confirms that structured models learn morphological paradigms effectively but struggle with lexical gaps. The hybrid pipeline improves lexical accuracy to 0.682 (+0.204 absolute gain) while grammatical accuracy drops slightly to 0.866 (-0.057). The substantial lexical gains outweigh minor grammatical losses, explaining the overall hybrid improvement.

When we stratify by training frequency, the pattern becomes clearer. We categorize each test morpheme by its frequency in the training data: infrequent (1 to 5 occurrences), common (6 to 20), and frequent (over 20). Table 3 shows accuracy by frequency bin.

Frequency	Count	BiLSTM	Hybrid
Infrequent (1 to 5)	69	0.029	0.426
Common (6 to 20)	58	0.448	0.724
Frequent (over 20)	86	0.860	0.859

Table 3: Accuracy by training frequency for lexical morphemes. Hybrid improvements concentrate in infrequent morphemes (+0.397), with no test morphemes completely unseen in training.

Infrequent morphemes show dramatic improvement: BiLSTM achieves only 0.029 accuracy while the hybrid reaches 0.426 (+0.397). Common mor-

phemes improve from 0.448 to 0.724 (+0.276). Frequent morphemes remain stable around 0.86 for both approaches. Notably, no test morphemes are completely unseen in this split. We did not enforce this property; with a document-level split and a small evaluation set, all test morphemes appear at least once in training. This demonstrates that hybrid gains stem from contextual inference on low-frequency items rather than handling zero-shot vocabulary, and that the baseline already captures frequent grammatical markers effectively when sufficient training examples exist.

5.7 Key Findings Summary

Our experiments reveal several important patterns for LLM-assisted morphological glossing. Retrieval-augmented prompting proves essential across all models, with similarity-based example selection dramatically outperforming random selection. Perhaps most surprisingly, providing morpheme dictionaries generally hurts performance: partial dictionaries universally degrade accuracy, while even complete dictionaries fail to help most models (only one of four shows gains). Performance scales approximately logarithmically with the number of in-context examples, typically peaking around ten to fifteen examples before plateauing or declining.

The hybrid architecture combining BiLSTM predictions with LLM correction improves performance across all tested models. These gains are particularly pronounced in low-shot scenarios where few examples are available, suggesting that the structured model provides valuable guidance when in-context learning is limited. Even without any examples, LLMs can successfully identify and correct many errors in structured predictions, indicating inherent morphological reasoning capabilities.

6 Discussion

6.1 The Case for Hybrid Architectures

Our results demonstrate that combining structured prediction with LLM reasoning yields consistent improvements across all tested models. The BiLSTM-CRF captures frequent morphological patterns from limited training data but struggles with rare morphemes and novel combinations. RAG LLM generation leverages broader linguistic knowledge and few-shot generalization but varies widely in accuracy depending on model choice.

The hybrid pipeline combines these complementary strengths through a division of labor revealed by error analysis. Structured models are more accurate on grammatical morphemes, which likely reflects their high frequency and regularity in the training data, while LLMs excel at contextual inference for infrequent lexical items, leveraging retrieved examples to handle vocabulary gaps. Hybrid improvements concentrate in low-frequency morphemes where BiLSTM lacks sufficient training signal, while both approaches perform similarly on frequent items. This asymmetry explains why the two-stage architecture succeeds: each component addresses the other's primary weakness. BiLSTM predictions provide structural guidance (particularly valuable in low-shot scenarios) while LLMs refine these predictions using patterns from retrieved examples. Our prompt design frames the BiLSTM output as a fallible hypothesis rather than an authoritative baseline, encouraging critical evaluation while providing useful structural constraints.

While we do not claim that BiLSTM-CRF represents the optimal base model for this task, our results suggest a broader principle: combining any trainable structured predictor with retrieval-augmented LLM post-correction can yield gains over either approach alone. This synergy proves universally beneficial across all performance tiers, with particularly strong gains when in-context learning is limited by few examples or budget constraints.

6.2 The Role of Morpheme Dictionaries

Perhaps our most surprising finding is that providing morpheme dictionaries generally hurts performance compared to providing no dictionary at all. Partial glossaries universally degrade accuracy, while even complete dictionaries fail to help most models. Only one model (`gemma-3-27b-it`) shows substantial gains from the complete dictionary, while three others perform worse with it than without.

These patterns suggest issues with how models integrate dictionary information, though we lack direct analysis of whether models actually consult dictionary entries. The degradation could stem from information overload, inappropriate dictionary consultation, or suboptimal prompt structure. Our approach provides dictionaries as simple unstructured key-value lists; alternative strategies merit investigation, such as organizing entries by morphological class or presenting dictionaries in structured formats.

Importantly, prompt structure can dramatically affect model behavior, and we have not systematically explored alternative formulations. Our findings should therefore be interpreted as demonstrating the ineffectiveness of our specific prompt design for dictionary integration, rather than fundamental limitations of dictionary use in morphological glossing. Future work should conduct systematic prompt engineering studies to identify more effective strategies for incorporating lexical resources.

6.3 Practical Implications for Fieldwork

Our findings suggest several considerations for practitioners working with LLM-assisted IGT annotation, though these patterns may vary across different languages, models, and documentation contexts. In our experiments, similarity-based retrieval consistently outperformed random example selection across all tested models, suggesting that investment in retrieval infrastructure may be worthwhile. The hybrid approach combining structured models with LLM post-correction showed gains across all configurations we tested, indicating potential value in this two-stage strategy.

For morpheme dictionaries, our results suggest caution: simple key-value presentations tended to hurt performance for most models in this setting, though alternative presentation strategies remain unexplored. Regarding few-shot example count, we observed approximately logarithmic scaling with diminishing returns beyond 10–15 examples, though optimal operating points likely depend on task complexity, model capabilities, and cost constraints. These patterns emerged from our specific experimental setup with Tuvan and four general-purpose LLMs; practitioners should validate these findings against their own languages and workflows, as model capabilities and architectural designs continue to evolve rapidly.

6.4 Ethical Considerations

Our experiments use data collected with informed consent under agreements restricting raw material sharing. We report only aggregate statistics and anonymized examples to protect speaker privacy. Speaker communities are not monolithic, and Tuvan speakers are distributed across multiple countries with different sociopolitical contexts. We do not claim community-wide consent; we follow the data agreements and consult the relevant fieldwork partners. Given the cross-border context, we avoid releasing raw data and refrain from identifying individuals or locations. Commercial APIs raise concerns about training data provenance (Bender et al., 2021; Sainz et al., 2023), and even at achieved accuracy levels, uncritical adoption risks introducing errors into the linguistic record. Decisions about automation should be made collaboratively with stakeholders, balancing efficiency gains against concerns about data control and quality.

6.5 Future Directions

Key directions include cross-linguistic evaluation on diverse morphological systems (polysynthetic, templatic, tonal) to test whether our design principles generalize beyond Turkic agglutination. Joint segmentation and glossing would address the limitation that our pipeline assumes gold boundaries. Parameter-efficient fine-tuning could improve performance with minimal language-specific data. Interactive interfaces with confidence estimation would help annotators prioritize review of uncertain predictions, and paradigm-level evaluation would better assess morphological generalization beyond token-level accuracy.

7 Conclusion

We present a hybrid automatic glossing pipeline combining BiLSTM-CRF structured prediction with LLM post-correction, evaluated on Tuvan fieldwork data across four LLMs. The hybrid approach consistently improves performance across all tested models, with particularly strong gains in low-shot scenarios where structural guidance from the BiLSTM proves most valuable. Error analysis reveals that improvements concentrate in lexical morphemes, especially rare vocabulary items, while BiLSTM already captures grammatical paradigms effectively. This demonstrates complementary strengths rather than simple performance stacking.

Our ablation studies reveal key design principles: retrieval-augmented prompting provides substantial gains; morpheme dictionaries generally hurt performance for most models; performance scales logarithmically with examples, plateauing around ten to fifteen; and hybrid correction benefits all models universally. These findings challenge conventional assumptions about prompt engineering, showing that more information is not always better.

While the utility of these accuracy levels for practical annotation workflows remains an open

question dependent on community priorities and annotation contexts, the substantial error rates necessitate careful human oversight. Model outputs should be treated as hypotheses requiring expert validation rather than authoritative annotations.

8 Limitations

Our evaluation focuses on a single language (Tuvan) from one language family (Turkic), leaving generalization to other morphological systems untested. The patterns we observe may not hold for polysynthetic, templatic, or non-concatenative morphology. Additionally, our test set is small, reflecting realistic annotation costs but introducing sampling variance that limits statistical precision.

Our evaluation metrics capture only token-level accuracy rather than higher-order properties like paradigm consistency, morphophonological regularity, or alignment with community language priorities. The pipeline also assumes gold morpheme boundaries and does not address the segmentation problem, which itself requires substantial linguistic expertise in fieldwork contexts.

The prompt engineering component of our study lacks systematic exploration. We tested only one prompt structure for each experiment, leaving unexplored how alternative formulations might affect dictionary effectiveness, example integration, or instruction following. Prompt design choices (information ordering, instruction phrasing, formatting conventions, and the balance of different knowledge sources) can dramatically impact model behavior, yet we lack the controlled comparisons needed to isolate their effects. Our glossary ablation findings in particular should be interpreted as showing that our specific prompt design failed to effectively leverage dictionary information, rather than demonstrating that dictionaries cannot help morphological glossing. We also lack fine-grained analysis of whether models actually consult dictionary entries or become distracted by additional prompt content.

Commercial API access limits transparency about training data and potential contamination from existing Tuvan linguistic resources. We use general-purpose LLMs rather than specialized models like GlossLM (Ginn et al., 2024b), which are explicitly trained on IGT data and may achieve higher absolute accuracy. However, our focus on establishing design principles (retrieval strategies, glossary configurations, hybrid architectures) likely generalizes across model types. Moreover, general-purpose LLMs offer practical advantages for fieldwork: no local infrastructure requirements, operation in few-shot regimes with minimal language-specific data, and accessibility to linguists without machine learning expertise. The tradeoff between specialized performance and practical accessibility merits further investigation.

References

Diego Barriga Martínez, Victor Mijangos, and Ximena Gutierrez-Vasques. 2021. Automatic Interlinear Glossing for Otomi language. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 34–43, Online. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Jan Buys and Jan A. Botha. 2016. Cross-Lingual Morphological Tagging for Low-Resource Languages. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1954–1964, Berlin, Germany. Association for Computational Linguistics.

Shobhana L. Chelliah and Willem J. de Reuse. 2010. *Handbook of Descriptive Linguistic Fieldwork*. Springer Science & Business Media. Google-Books-ID: d1FfE30hZ7EC.

Bernard Comrie, Martin Haspelmath, and Bickel Balthasar. 2008. The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morphene Glosses. Google-Books-ID: e8B7AQAACAAJ.

Ryan Cotterell and Georg Heigold. 2017. Cross-lingual Character-Level Neural Morphological Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 748–759, Copenhagen, Denmark. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui

Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint*. ArXiv:2501.12948 [cs].

Micha Elsner and David Liu. 2025. Prompt and circumstance":" A word-by-word LLM prompting approach to interlinear glossing for low-resource languages. In *Proceedings of the 22nd SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics*, pages 1–14, Albuquerque, New Mexico, USA. 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.

Sharon Hargus, Olga M. Semenova, and Siri G. Tuttle. 2020. Glossing Dene Languages. Publisher: Alaska Native Language Center.

K. David Harrison and Gregory D. S. Anderson. 2002. *A Grammar of Tuvan*. Scientific Consulting Services International. Google-Books-ID: RXnhAQAACAAJ.

Taiqi He, Kwanghee Choi, Lindia Tjuatja, Nathaniel Robinson, Jiatong Shi, Shinji Watanabe, Graham Neubig, David Mortensen, and Lori Levin. 2024. Wav2Gloss: Generating Interlinear Glossed Text from Speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–582, Bangkok, Thailand. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv preprint*. ArXiv:1508.01991 [cs].

SIL International. 2025. FieldWorks Language Explorer.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research*, 24(251):1–43.

Zhengbao Jiang, Frank Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active Retrieval Augmented Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992, Singapore. Association for Computational Linguistics.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry,

Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, C. J. Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju-yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. Gemma 3 Technical Report. *arXiv preprint*. ArXiv:2503.19786 [cs].

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural Architectures for Named Entity Recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

Christian Lehmann. 2004a. Data in linguistics. 21(3-4):175–210. Publisher: De Gruyter Mouton Section: The Linguistic Review.

Christian Lehmann. 2004b. Interlinear morphemic glossing. In Geert Booij, Christian Lehmann, Joachim Mugdan, Stavros Skopeteas, and Wolfgang Kesselheim, editors, *Morphologie*, pages 1834–1857. De Gruyter.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*. ArXiv:2005.11401 [cs].

Zoey Liu, Robert Jimerson, and Emily Prud'hommeaux. 2021. Morphological Segmentation for Seneca. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 90–101, Online. Association for Computational Linguistics.

Robert Logan IV, Ivana Balazevic, Eric Wallace, Fabio Petroni, Sameer Singh, and Sebastian Riedel. 2022. Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2824–2835, Dublin, Ireland. Association for Computational Linguistics.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Chaitanya Malaviya, Matthew R. Gormley, and Graham Neubig. 2018. Neural Factor Graph Models for Cross-lingual Morphological Tagging. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2653–2663, Melbourne, Australia. Association for Computational Linguistics.

Talant Mawkanuli. 1999. *The phonology and morphology of Jungar Tuva*. Ph.D., Indiana University, United States – Indiana. ISBN: 9780599667938.

Talant Mawkanuli. 2005. *Jungar Tuvan Texts*. Uralic and Altaic Series. Indiana University, Bloomington.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Kann, and Mans Hulden. 2020. IGT2P: From Interlinear Glossed Texts to Paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262, Online. Association for Computational Linguistics.

Sebastian Nordhoff and Thomas Krämer. 2022. IMT-Vault: 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.

Shu Okabe and François Yvon. 2023. Towards Multilingual Interlinear Morphological Glossing. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5958–5971, Singapore. Association for Computational Linguistics.

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, C. J. 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. *arXiv preprint*. ArXiv:2303.08774 [cs].

Enora Rice, Ali Marashian, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2024. TAMS: Translation-Assisted Morphological Segmentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6752–6765, Bangkok, Thailand. Association for Computational Linguistics.

Enora Rice, Katharina von der Wense, and Alexis Palmer. 2025. Interdisciplinary Research in Conversation: A Case Study in Computational Morphology for Language Documentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11284–11296, Suzhou, China. Association for Computational Linguistics.

Erich Round, Mark Ellison, Jayden Macklin-Cordes, and Sacha Beniamine. 2020. Automated Parsing of Interlinear Glossed Text from Page Images of Grammatical Descriptions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2878–2883, Marseille, France. European Language Resources Association.

Tatyana Ruzsics and Tanja Samardžić. 2017. Neu-

ral Sequence-to-sequence Learning of Internal Word Structure. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 184–194, Vancouver, Canada. Association for Computational Linguistics.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP Evaluation in trouble: On the Need to Measure LLM Data Contamination for each Benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Mathias Schenner and Sebastian Nordhoff. 2016. Extracting Interlinear Glossed Text from LaTeX Documents. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4044–4048, Portorož, Slovenia. European Language Resources Association (ELRA).

Bhargav Shandilya and Alexis Palmer. 2023. Lightweight morpheme labeling in context: Using structured linguistic representations to support linguistic analysis for the language documentation context. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 78–92, Toronto, Canada. Association for Computational Linguistics.

Shamane Siriwardhana, Rivindu Weerasekera, Elliott Wen, Tharindu Kaluarachchi, Rajib Rana, and Suranga Nanayakkara. 2023. Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering. *Transactions of the Association for Computational Linguistics*, 11:1–17. Place: Cambridge, MA Publisher: MIT Press.

Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring LLM Prompting Strategies for Joint Essay Scoring and Feedback Generation. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 283–298, Mexico City, Mexico. Association for Computational Linguistics.

Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. 2021. Language Models are Few-shot Multilingual Learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 1–15, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN : a professional framework for multimodality research. pages 1556–1559.

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. 2025a. Qwen3 Technical Report. *arXiv preprint*. ArXiv:2505.09388 [cs].

Changbing Yang, Franklin Ma, Freda Shi, and Jian Zhu. 2025b. 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.

Wenhao Yu. 2022. Retrieval-augmented Generation across Heterogeneous Knowledge. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 52–58, Hybrid: Seattle, Washington + Online. 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.

Vilém Zouhar, Martin Popel, Ondřej Bojar, and Aleš Tamchyna. 2021. Neural Machine Translation Quality and Post-Editing Performance. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10204–10214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Experimental Prompt Templates

This appendix provides the complete prompt templates used across all five experiments. All prompts follow a consistent structure with a system message defining the linguistic expert role, followed by task-specific instructions and formatting requirements.

A.1 System Message (All Experiments)

All experiments use the following system message:

> *You are a linguistic expert specializing in morpheme-by-morpheme glossing for an unknown language.*

A.2 Experiments 1–3: RAG LLM generation

Experiments 1 (Retrieval vs. Random), 2 (Glossary Ablation), and 3 (N-Shot Scaling) use RAG LLM generation with the following template:

> **User message:**
>
> Here are some examples of sentences with morpheme boundaries (marked by hyphens) and their glosses:
>
> *[For each example i in 1..k:]*
>
> Example *i*:
>
> Segmented: *[morpheme-segmented source text]*
>
> Gloss: *[corresponding gloss sequence]*
>
> *[If glossary provided:]*
>
> You are also given a morpheme dictionary mapping morphemes to their English glosses. For morphemes in the dictionary, use the provided gloss. For others, infer from context. Some morphemes may have multiple translations; choose the most appropriate for this context.
>
> Morpheme dictionary: *[morpheme1: gloss1, morpheme2: gloss2, ...]*
>
> *[End glossary section]*
>
> Please gloss this sentence:
>
> Segmented: *[test sentence with morpheme boundaries]*
>
> Output the gloss with the same structure (spaces between words, hyphens between morphemes). Enclose your gloss in ###. Example: ###word1-MORPH1 word2-MORPH2###

Experiment-specific variations:

- **Experiment 1:** Uses 3 examples, no glossary. Compares TF-IDF retrieval (RAG) against random sampling.
- **Experiment 2:** Uses RAG, no glossary. Varies $k \in \{1, 3, 5, 10, 15, 20\}$ examples.
- **Experiment 3:** Uses 3 RAG examples. Tests four glossary sizes: none (0 pairs), top-100 (100 pairs), grammatical (240 pairs), entire (1,498 pairs).

A.3 Experiment 4: Hybrid Pipeline (BiLSTM + LLM)

Experiment 4 uses BiLSTM-CRF predictions as initial hypotheses for LLM correction, with the following template:

> **User message:**
>
> You will be given:
>
> 1. A rough initial glossing attempt from a statistical model (may contain errors)
> 2. Some example sentences with their correct glosses
> 3. A morpheme dictionary
>
> Your task is to produce the correct gloss, using all available information.
>
> Here are some example sentences with correct glosses:
>
> *[For each example i in 1..k (or 0 for zero-shot):]*
>
> Example *i*:
>
> Segmented: *[morpheme-segmented source text]*
>
> Gloss: *[corresponding gloss sequence]*
>
> *[If glossary provided:]*
>
> You also have access to a morpheme dictionary: *[morpheme1: gloss1, morpheme2: gloss2, ...]*
>
> *[End glossary section]*
>
> Now, please gloss this sentence:
>
> Segmented: *[test sentence with morpheme boundaries]*
>
> Initial attempt (from statistical model): *[BiLSTM-CRF prediction]*
>
> This initial attempt may contain errors. Use the examples, dictionary, and linguistic patterns to produce the correct gloss. Maintain the same structure (spaces between words, hyphens between morphemes).
>
> IMPORTANT: Output ONLY the gloss wrapped in ###. Do not explain your

reasoning. Example format: ###word1-MORPH1 word2-MORPH2###

Design rationale: The hybrid prompt presents the BiLSTM prediction as a "rough initial attempt" rather than an authoritative baseline, encouraging the LLM to critically evaluate and correct errors. Clean gold examples (not error-correction pairs) provide paradigmatic context, while the statistical prediction narrows the search space by proposing plausible morpheme boundaries and candidate glosses.

A.4 Output Extraction

All experiments extract model outputs by locating text between ### delimiters. If delimiters are absent (indicating non-compliance), the entire output string is used as the predicted gloss. This extraction method proved robust across all four LLM providers.

B Detailed Results Tables

This section provides detailed numerical results for experiments presented as figures in the main text.

Glossary	deepseek	qwen3	gpt-4o	gemma-3
None	**0.510**	**0.436**	**0.420**	0.325
Top-100	0.469	0.332	0.353	0.430
Grammatical	0.426	0.303	0.364	0.335
Entire	0.490	0.422	0.369	**0.455**

Table 4: Glossary ablation study across four LLMs (3-shot RAG). Partial glossaries consistently degrade performance, while complete dictionaries fail to help most models. Model names abbreviated: deepseek = `deepseek-v3.2-exp`, qwen3 = `qwen3-max`, gpt-4o = `gpt-4o-mini`, gemma-3 = `gemma-3-27b-it`.

N	deepseek	qwen3	gpt-4o	gemma-3
1	0.350	0.196	0.159	0.118
3	0.505	0.439	0.383	0.336
5	0.588	0.519	0.423	0.471
10	**0.658**	0.529	0.430	**0.600**
15	0.646	**0.611**	**0.486**	0.540
20	0.626	0.552	0.475	0.534

Table 5: N-shot scaling for RAG LLM generation (RAG, no glossary). Performance improves logarithmically, with diminishing returns beyond n=10–15. Model names abbreviated as in Table 4.