Massively Multilingual Joint Segmentation and Glossing

Michael Ginn[1] **Lindia Tjuatja**[2] **Enora Rice**[1] **Ali Marashian**[1]
Maria Valentini[1] **Jasmine Xu**[2] **Graham Neubig**[2] **Alexis Palmer**[1]
[1]University of Colorado Boulder [2]Carnegie Mellon University
michael.ginn@colorado.edu alexis.palmer@colorado.edu

Abstract

Automated interlinear gloss prediction with neural networks is a promising approach to accelerate language documentation efforts. However, while state-of-the-art models like GLOSSLM (Ginn et al., 2024b) achieve high scores on glossing benchmarks, user studies with linguists have found critical barriers to the usefulness of such models in real-world scenarios (Rice et al., 2025). In particular, existing models typically generate morpheme-level glosses but assign them to whole words without predicting the actual morpheme boundaries, making the predictions less interpretable and thus untrustworthy to human annotators.

We conduct the first study on neural models that **jointly predict interlinear glosses and the corresponding morphological segmentation** from raw text. We run experiments to determine the optimal way to train models that balance segmentation and glossing accuracy, as well as the alignment between the two tasks. We extend the training corpus of GLOSSLM and pretrain POLYGLOSS, a family of seq2seq multilingual models for joint segmentation and glossing that outperforms GLOSSLM on glossing and beats various open-source LLMs on segmentation, glossing, and alignment. In addition, we demonstrate that POLYGLOSS can be quickly adapted to a new dataset via low-rank adaptation.

1 Introduction

Nearly half of the world's 7,000 languages face extinction. For many speakers and linguists of these languages, **language documentation** has become an urgent goal. Documentation projects commonly involve the creation of interlinear glossed text (IGT), a dense annotation format combining morphological segmentation, tagging, and translation (Figure 1). Due to its structured format and common usage among linguists, IGT has proven useful for linguistic analysis (Bender et al., 2013; Zamaraeva, 2016; Moeller et al., 2020), language pedagogy (Alast and Baleghizadeh, 2024; Bonilla Carvajal, 2025), and development of language technology such as taggers (Georgi, 2016), searchable text databases (Blokland et al., 2019; Rijhwani et al., 2023), educational tools (Uibo et al., 2017; Chaudhary et al., 2023), and machine translation systems (Zhou et al., 2020; Ramos et al., 2025).

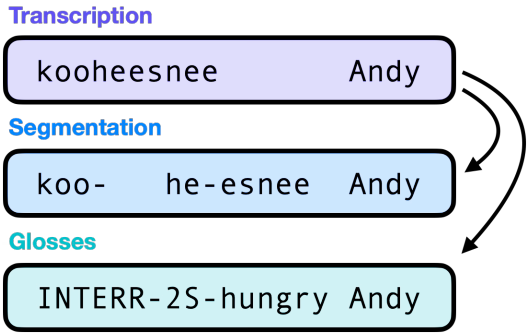

Figure 1: An interlinear glossed text example, showing the Arapaho for *"Are you hungry, Andy?"*. Our model predicts the segmentation and gloss line from the transcribed text.

Creating IGT is expensive, and a number of studies have proposed methods to automate IGT production with statistical and neural methods (McMillan-Major, 2020a; Zhao et al., 2020; Ginn et al., 2024a). In all of these studies, including the 2023 SIGMORPHON shared task (Ginn et al., 2023), the task is formulated as predicting the gloss line from the transcription or segmentation line. The former is more difficult (but also more useful), as it requires the model to infer morphological segmentation in addition to predicting glosses, and has been the primary focus of recent work.

Though state-of-the-art glossing models such as GLOSSLM (Ginn et al., 2024b) have achieved high accuracy across many languages, Rice et al. (2025) discovered several issues when using these models in a realistic documentation scenario:

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 35806–35820

1. First, documentary linguists typically perform **explicit morphological segmentation** before glossing each morpheme, so a model that produces glosses directly—without exposing the implicit segmentation—is confusing, less interpretable, and difficult to trust.

2. Second, the model produced very **inaccurate glosses** in two of the three languages studied, with the participants agreeing that correcting the predicted outputs would be more difficult than annotating from scratch, or using a simpler lookup-based method.

3. Third, the model often predicted gloss labels which were unfamiliar or unlike the glossing conventions preferred by the participants, and the existing system provided no way to **adapt its labels to the preferred conventions**.

In this work, we address these three concerns, building on the approach of GLOSSLM. We release an improved version of the GLOSSLM corpus, which adds 91k examples for a total of 341k examples, improves standardization and formatting, and ensures alignment between morphological segmentation and glosses. We train a multilingual model on the dataset for **joint segmentation and glossing**, optimizing for performance on both tasks, as well as **alignment between the two tasks**, outperforming various small LLMs and satisfying item 1. We show that per-language perplexity can roughly predict glossing accuracy for any language, addressing item 2 by enabling an automatic glossing system to avoid giving low-quality predictions, or to fall back to a simpler model. Finally, we show that POLYGLOSS can be rapidly adapted to small labeled datasets via low-rank adaptation, satisfying item 3. Unlike prior work that trains monolingual glossing models, we focus on creating **a single multilingual model that can be used out-of-the-box on many languages.** Our models and dataset are available on HuggingFace[1] and our code is on GitHub.[2]

2 Corpus

We create an enhanced version of the GLOSSLM corpus with significantly improved formatting. We consistently handle punctuation across all sources, ensuring that sentence-ending punctuation is surrounded by spaces while gloss-internal punctuation remains unchanged, as in the following Tsez example:

(1) Žeda kidbeqor kurno lel yayrno .
žeda-a kid-qor kur-n lel y-ayr-n
DEM1.IIPL.OBL-ERG girl-POSS.LAT throw-PFV.CVB wing II-lead-PST.UNW

Statistic	Count
Total examples	353,266
Unique languages	2,077
Train examples	340,251
Eval examples	6,148
Test examples	6,867
No glottocode	13,428
No metalang. glottocode	10,894
No segmentation	93,648
No translation	5,921
Misaligned	34,894

Table 1: POLYGLOSS corpus statistics

We fixed a number of source-specific formatting issues. For example, we noticed 4,882 instances in the Arapaho data where ",." was used inside glosses, and we confirmed with the original annotator that this was an error. Across sources, we identified instances where the morphological segmentation was **misaligned** with the glosses—that is, cases where there was a mismatch in either the number of words or the number of segments within a word (see §3.3 for more discussion of alignment). In these cases, if the segmentation field does not include any segmentation markers (and the gloss field does), we set the segmentation to blank. Otherwise, we keep the segmentation, but ensure the offending examples are within the training split, so as not to affect evaluation.

We also incorporate additional IGT data into the original GLOSSLM corpus. The largest of these is the Fieldwork dataset (He et al., 2024), which collects 80,461 IGT instances for 37 languages. We also update the IMTVault dataset (Nordhoff and Krämer, 2022) to the newest version (1.2), which includes IGT scraped from new linguistic publications, adding 39,741 examples. After removing 20,116 duplicates and filtering very low-quality examples,[3] we have **91,416 new unique examples** compared to the original GLOSSLM corpus. We introduced an auditing process on our dataset to quantify issues, and report full statistics in Table 1. We add two new languages as evaluation languages from the Fieldwork dataset: Hokkaido Ainu

[1] https://huggingface.co/collections/lecslab/polygloss

[2] https://github.com/lecs-lab/polygloss

[3] Examples which only contained punctuation or whitespace

(ainu1240) and Ruuli (ruul1235). Our dataset splits for all evaluation languages are reported in Table 2.

Language	Train	Eval	Test
Arapaho (arp)	36776	4687	4499
Tsez (ddo)	3626	444	442
Gitksan (git)	89	42	37
Uspanteko (usp)	8338	170	566
Ainu (ain)	6726	218	590
Lezgi (lez)	646	51	53
Natugu (ntu)	786	99	99
Nyangbo (nyb)	1221	225	248
Ruuli (ruc)	2158	212	333

Table 2: Number of examples for each evaluation language across train, eval, and test splits.

3 Evaluation

Since our model performs joint glossing and segmentation, we compute metrics for both, as well as an alignment score (§3.3) between the two.

3.1 Glossing

We compute a number of metrics for gloss prediction. Departing from prior work, we use **morpheme error rate** (c.f. word error rate as used in speech recognition) as our primary metric. While prior work (McMillan-Major, 2020a; Ginn et al., 2023) used morpheme-level accuracy as the primary metric, this assumes that the output has the correct number of morphological glosses. If there is a gloss inserted or deleted, all subsequent glosses will be counted as incorrect. Instead, we compute the morpheme error rate by first inserting [SEP] tokens between the glosses for each word and computing the edit distance, normalized to the length of the gold label sequence. The range is 0 or greater; a score higher than 1 is possible if the predicted sequence is longer than the gold sequence. In addition, we compute word and character error rates, BLEU scores (at all three levels of granularity), and morpheme and word-level accuracy.

3.2 Segmentation

We use standard metrics for evaluating segmentation. We primarily report the modified F1 score as defined in Mager et al. (2020), which computes precision based on morphemes in the predicted segmentation also occurring in the gold label, and vice versa for recall. We also compute character-level edit distance and whole-word accuracy.

3.3 Alignment

A key goal in this study is to predict morphological segmentations and glosses that are aligned with one another, making the gloss predictions more interpretable and trustworthy for a human annotator. To measure this, we propose a novel **alignment score**, which is computed based on predictions with no reference to the gold sequence.[4] First, the segmentation and gloss predictions are converted into *abstract sequences* that represent morphological structure. Ignoring punctuation, each morpheme sequence is converted to a single "x" character, and morpheme boundaries ("-" and "=") are left unchanged, as in the following example:

the cat-s ru-n ⇒ x x-x x-x
DET cat-PL run.SG ⇒ x x-x x

Then, the character-level edit distance is computed between the abstracted gloss and segmentation sequences, ranging from 0 to infinity. We normalize the edit distance by the length of the longer sequence,[5] and subtract from 1 to give a score in $[0, 1]$ where 1 is a perfect score. In this example, the alignment score is 0.78.

4 Model

4.1 Task Format

Using the PolyGloss corpus, we perform continued pretraining on a pretrained multilingual LLM for both segmentation and glossing. We train the model to predict glosses from both the segmented and unsegmented transcription, but we only evaluate on the latter, as it is the more difficult and realistic setting. We experiment with three different approaches for combining the two tasks and report results in 7.2.

Multitask Prediction In this setting, separate training examples are created for segmentation and for glossing. The examples are formatted as in the following Vera'a language example (replacing "glosses" with "segmentation" when appropriate):

Predict the glosses for the following text in Vera'a.
Text in Vera'a: o wōlēn 'ēqēk
Translation in English: Oh, over there is my garden
Glosses: INTERJ you.know-ZERO=ART garden-1SG

This setting is simple and allows for simultaneous inference of both glosses and segmentation. How-

[4] That is, a model could achieve a perfect alignment score while predicting incorrect glosses and segmentation.

[5] Unlike with the standard error rate, we don't know which sequence is correct if there is a mismatch.

ever, there is greater risk of misalignment, since the two tasks are trained separately and alignment is not enforced.

Concatenated Prediction Morphological segmentation and interlinear glosses are not in fact distinct tasks, as the latter depends inherently on the former. In this setting, we model this dependency by training the model to predict the segmentation followed by the glosses:

> Predict the morphological segmentation and glosses for the following text in Vera'a.
> Text in Vera'a: o wōlēn 'ēqēk
> Translation in English: Oh, over there is my garden
> Segmentation: o wōlē-0=n 'ēqē-k
> Glosses: INTERJ you.know-ZERO=ART garden-1SG

This introduces a natural dependency due to the causal training objective: while generating the gloss string, the model can attend to tokens in the segmentation. Of course, this is not a strict constraint, and carries the risk that a bad segmentation will affect the glosses as well. Since not all training examples have segmentation labels, we also create glossing examples in the multitask style shown in the previous section.

Interleaved Prediction While the concatenated setting trains the model with an implicit relationship between segments and glosses, it is still possible to generate misaligned predictions. In this setting, we introduce a hard constraint using an interleaved format that explicitly aligns segments and glosses. In this format, each gloss label is immediately followed by the corresponding morpheme in parentheses.

> Predict the glosses and morphological segmentation (in parentheses) for the following text in Vera'a.
> Text in Vera'a: o wōlēn 'ēqēk
> Translation in English: Oh, over there is my garden
> Output: INTERJ(o) you.know(wōlē)-ZERO(0)=ART(n) garden('ēqē)-1SG(k)

We hypothesized that this setting would have the best alignment, as any well-formed output should be perfectly aligned.

4.2 Base Model

Following (Ginn et al., 2024b), we perform continued pretraining on ByT5 (Xue et al., 2022), a byte-level encoder-decoder transformer language model based on the T5 architecture. By using byte-level tokenization, ByT5 avoids the issues that arise for rare languages with subword tokenizers, and has been shown to outperform T5 on multilingual glossing (He et al., 2023). We also experimented with finetuning an instruction-tuned decoder LLM, using Qwen 3 0.6B (Yang et al., 2025a), known to be a strong multilingual model. However, we saw poor results, discussed in Appendix A. We train one ByT5-based models for each task format in the preceding section, using the `byt5-base` checkpoint with 580M parameters.

4.3 Training

Due to the high cost of training runs, we do not perform extensive tuning. We train all models in bf16, using the AdamW optimizer with default parameters, a linear learning rate warmup for the first 3% of steps, cosine learning rate decay, and gradient clipping with max norm of 1. We train all models using 4 GH200 GPUs. For the ByT5-based models, we used a learning rate of 5E-5, batch size of 64, and 15 epochs. Evaluation uses beam search with default parameters and 2 beams. Parameters for the Qwen-based model are given in Appendix A.

5 Baselines

5.1 Multilingual Baselines

We compare our best model to the following multilingual models for glossing:

GLOSSLM We compare glossing performance to the pretrained GLOSSLM model, primarily to ensure that incorporating segmentation does not cause glossing performance to regress. As this model was not trained for segmentation, we cannot compare across all metrics. Additionally, the original GLOSSLM model did not include all of the segmented training data, so this is not a perfect head-to-head comparison.

In-Context Learning Following Ginn et al. (2024a), we use LLMs to predict both the segmentation and glosses in a single pass using the interleaved format. We retrieve ten similar examples (via chrF score) from the training set to provide in-context. We test three models: the **Qwen 3 0.6B model** with thinking (Yang et al., 2025a) and the **Cohere Aya Expanse 8B model** (Dang et al., 2024) and **Google Gemma 3 4B model** (Team, 2025) without thinking. These enable a direct comparison between our finetuned seq2seq models and ICL with LLMs of similar size.

5.2 Monolingual Baselines

In addition, we run preliminary experiments using monolingual models to understand the effect of different approaches to joint training and monolingual versus multilingual training. We use the following baselines, and provide comparisons in 7.1 and 7.2:

Finetuned ByT5 (separate) For each language in our test set, we train monolingual models on ByT5 for segmentation and glossing in the Multitask format, using **separate models for each task**. We use the same hyperparameters as our pretrained model, except we set a max 30 epochs and use early stopping with patience 5 to prevent overfitting. We do not perform extensive tuning.

Finetuned ByT5 (joint) In addition, we train monolingual models for **joint segmentation and glossing** using the Interleaved format and the same hyperparameters as the prior baseline.

Pipeline We train monolingual pipelines of two ByT5 models, where the first model predicts the morphological segmentation and the second model predicts glosses given segments. In this baseline, there is clear risk of error propagation, as incorrect segmentation likely results in incorrect glosses.

Hard Attention Transformer Following Girrbach (2023a), we train monolingual hard attention transformers using straight-through gradient estimation. The model is only trained on glosses, but generates an explicit latent segmentation based on the hard attention between glosses and characters in the input.

6 Results

We compare our best PolyGloss model (ByT5 architecture and Interleaved format) with multilingual baselines on the test set for glossing (Table 3), segmentation (Table 4), and alignment (Table 5). Results for all other metrics, such as BLEU score, morpheme accuracy, and word-level scores are available on our GitHub.

Overall, the multilingual PolyGloss model is state-of-the-art on glossing and very strong on segmentation. It outperforms GlossLM for glossing on the three overlapping evaluation languages, likely due to both better dataset preprocessing and improved training hyperparameters. It also far outperforms in-context learning approaches with LLMs that are orders of magnitude larger (220M vs 0.6B-8B). The smallest LLM (Qwen 0.6B) often struggles to conform to the desired format, resulting in misaligned outputs. The larger models (Gemma 4B and Aya 8B) have much higher alignment scores, but still struggle to perform glossing or segmentation accurately. More details on specific types of failures are described in Appendix C. While significantly larger LLMs might show better results, our model is clearly state-of-the-art given constraints on both model size and training budget.

6.1 Predicting Performance on Other Languages

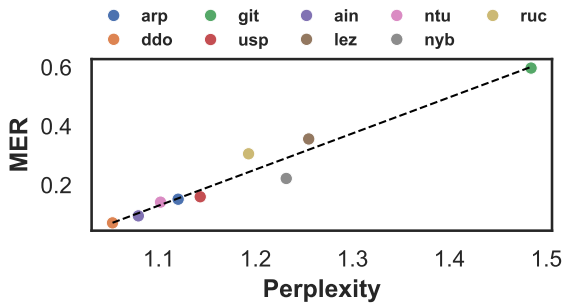

Figure 2: Relationship between validation set perplexity for a given language and glossing performance, as measured by morpheme error rate. There is a strong correlation ($r^2 = 0.951$), indicating that perplexity can be used as a rough predictor of glossing performance.

As identified in Rice et al. (2025), an issue with GlossLM was that there was no good way to predict glossing performance on a given language that is not one of our nine selected evaluation languages. For the PolyGloss model (ByT5, Interleaved), we compute per-language perplexity on our validation dataset and demonstrate that it has a strong correlation ($r^2 = 0.951$) with our target metric (Figure 2).

This provides a practical heuristic for the use of our model in real-world settings. A glossing software such as Plaid[6] could set an acceptable error rate threshold, and use the PolyGloss model to predict glosses only if the language's expected error rate is below that threshold. If not, then it will likely be a better user experience to either fall back to a simple method (such as predicting the highest-frequency gloss) or not showing predictions at all.

[6] https://www.langdoc.net/t/introducing-plaid/1250

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Qwen 3 0.6B (ICL)	0.868	0.904	0.919	0.730	0.773	0.895	0.877	0.706	0.883	0.839
Gemma 3 4B (ICL)	0.489	0.597	0.826	0.476	0.351	0.668	0.473	0.430	0.723	0.559
Aya Expanse 8B (ICL)	0.545	0.749	0.871	0.514	0.464	0.740	0.591	0.492	0.802	0.641
GLOSSLM	0.161	0.095	0.870*	0.163	0.909*	0.940*	0.893*	0.990*	0.731*	0.639*
POLYGLOSS (ByT5, multitask)	0.177	0.089	0.603	0.162	0.122	0.383	0.189	0.328	0.329	0.265
POLYGLOSS (ByT5, interleaved)	**0.152**	**0.072**	**0.597**	**0.160**	**0.095**	**0.357**	**0.142**	**0.222**	**0.306**	**0.234**

Table 3: Morpheme error rate (↓) for **glossing** on the held-out test set with multilingual models. For GLOSSLM, the only eval languages explicitly included in the pretraining corpus are `arp`, `ddo`, and `git`, so scores on other languages (marked with *) are very poor.

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Qwen 3 0.6B (ICL)	0.091	0.096	0.050	0.297	0.233	0.118	0.133	0.298	0.190	0.167
Gemma 3 4B (ICL)	0.505	0.310	0.159	0.508	0.575	0.338	0.435	0.617	0.345	0.421
Aya Expanse 8B (ICL)	0.457	0.249	0.132	0.509	0.569	0.302	0.337	0.465	0.324	0.371
POLYGLOSS (ByT5, multitask)	0.903	0.967	0.605	0.850	0.963	0.826	0.925	0.938	0.763	0.860
POLYGLOSS (ByT5, interleaved)	**0.910**	**0.972**	**0.595**	**0.855**	**0.963**	**0.782**	**0.935**	**0.959**	**0.782**	**0.862**

Table 4: Morpheme F1 (↑) for **segmentation** on the held-out test set with multilingual models.

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Qwen 0.6B (ICL)	0.342	0.669	0.488	0.744	0.716	0.711	0.697	0.788	0.792	0.661
Gemma 3 4B (ICL)	0.982	0.995	0.936	0.983	0.985	0.993	0.994	0.993	0.990	0.984
Aya Expanse 8B (ICL)	0.957	0.953	0.956	0.973	0.979	0.882	0.978	0.985	0.986	0.961
POLYGLOSS (ByT5, multitask)	0.984	0.995	0.919	0.990	0.982	0.973	0.985	0.988	0.941	0.973
POLYGLOSS (ByT5, interleaved)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

Table 5: **Alignment score** (↑) between predicted segmentation and glosses on held-out test set, multilingual models.

7 Ablations

Per-language breakdowns are provided for all ablations in Appendix B.

7.1 Effect of Joint Training

The best POLYGLOSS model is trained jointly on segmentation and glosses in a single training example. To disentangle the effect of this joint training approach, we compare three monolingual ByT5-based approaches: training for glossing and segmentation separately, jointly, and in a model pipeline (see subsection 5.2). We also evaluate the Girrbach (2023a) approach, which induces a latent segmentation using hard attention. We report the average scores in Figure 3.[7] The **Joint** setting is superior on all three metrics (and near-perfect on alignment), suggesting a harmonious relationship between the two training tasks. While the **Separate** setting is similar to the former on glossing and segmentation, its alignment score is significantly worse. This is unsurprising, as the two separate models for glossing and alignment are not guaranteed to produce the same errors, and thus may generate misaligned outputs for the same input. For example, the separated models make the following poorly aligned prediction for a Nyangbo sentence (alignment score of 0.882):

(2) vũnɔ gagãlĩ gɛ enu budzyuɖí yɛ
vũnɔ gagãlĩ gɛ e-nu bu-dzyuɖí yɛ
beverage well-well REL 3SG-be CM-strength 3SG

The misalignment occurs in the second word, which is predicted to be a single morpheme but two glosses. Meanwhile, the interleaved joint model predicts the perfectly aligned glosses and morphemes (which are also more accurate):

(3) vũnɔ gagãlĩ gɛ enu budzyuɖí yɛ
vũnɔ gagãlĩ gɛ e-nu bu-dzyuɖí yɛ
greet be_hard REL 3SG-be CM-dawadawa_tree FOC

For real-world usage, generating aligned outputs is critical in addition to achieving high accuracy.

[7] Error bars are large because of the variance across languages, and should not be used to determine significance.

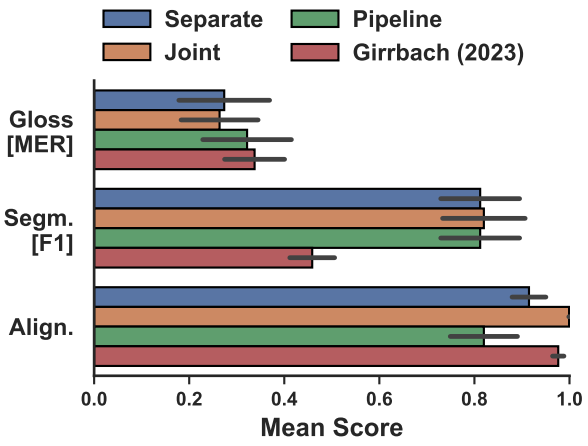

Figure 3: Scores for monolingual models using various approaches to multitask training. A lower glossing MER is better; higher is better for the other two metrics. Scores are averaged across nine languages and reported with standard error.

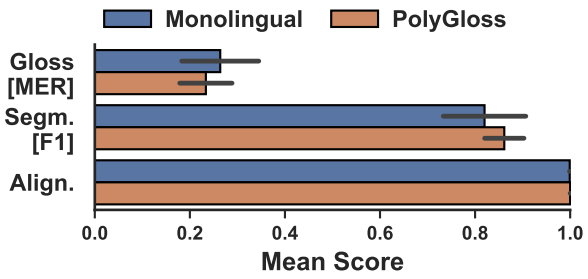

Figure 4: Ablation comparing monolingual joint models (using the interleaved format) and the multilingual PolyGloss using the same format. A lower glossing MER is better; higher is better for the other two metrics. Scores are averaged across nine languages and reported with standard error.

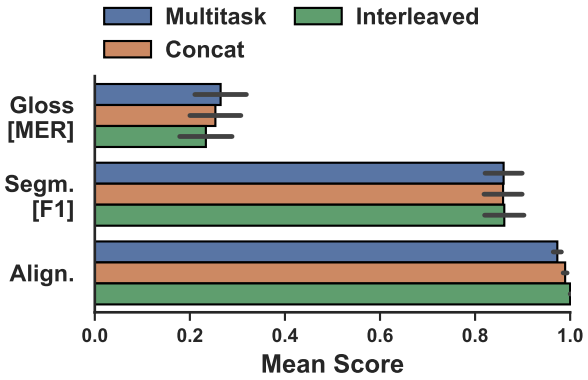

Figure 5: Scores for PolyGloss multilingual models using three different data formats. Scores are averaged across nine languages and reported with standard error.

The **Joint** approach is also clearly superior to the **Pipeline** setting for glossing and alignment (though segmentation is very similar). Because the pipeline is not trained end-to-end, error in the intermediate predictions is catastrophic for the gloss predictions, resulting in far higher error than the joint approach. Finally, the jointly trained model outperforms the hard attention approach of Girrbach (2023a) on segmentation, where the learned segmentations are not very accurate to the gold labels.

7.2 Effect of Multilingual Pretraining

We compare the monolingual joint models with the PolyGloss multilingual model (Figure 4). Both approaches use the interleaved format, enabling a direct comparison. We see that on average, the multilingual model outperforms the monolingual models on glossing and segmentation (and alignment is perfect for both), though the standard error is large. We observed that for the three languages with the most training data (`arp`, `usp`, and `ain`) the monolingual model is slightly superior on glossing. Meanwhile, for all other languages the multilingual model is far superior on glossing, demonstrating clear transfer learning benefits. Furthermore, we believe that serving a single multilingual glossing model is preferable to requiring linguists to train and host their own monolingual models (though we do propose an adaptation method in section 8).

7.3 Effect of Task Format

Results for the three ByT5-based PolyGloss models are very similar regardless of task format (Figure 5). The interleaved model is best overall on glossing (average MER 0.234), segmentation (average F1 0.862), and alignment (1.000). The multitask model is slightly worse at glossing, indicating potential benefits from explicitly conditioning the model on the morphological segmentation. As hypothesized, the multitask model struggles the most on alignment, indicating that its glossing and segmentation predictions are accurate but not necessarily aligned. The interleaved format results in perfect alignment thanks to its explicit constraint.

One concern with the interleaved and concatenated formats is that the expected output is roughly twice as long than the joint model, as it includes both the glosses and segmentation. Due to the autoregressive nature of generation, longer sequences can result in degraded accuracy. We test this by filtering the test set to only very long inputs; specifically, we only keep examples where the transcription length is in the 75th percentile for all examples of the same language. We compare glossing re-

sults for long inputs between the three formats in Figure 6. We observe that average performance is nearly identical on long inputs, thus suggesting there is no concern with the increased length of the interleaved and concatenated formats.

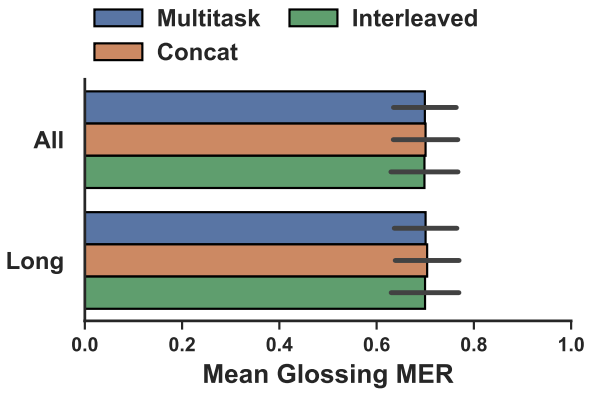

Figure 6: Glossing MER for PolyGloss multilingual models when evaluated either on all data, or only on long inputs (75th percentile in length per language). Full data in Table 10.

8 Adapting PolyGloss To New Data

Though we strived to collect as much IGT data as possible, it is inevitable that some languages will not be well-represented (or present at all) in the PolyGloss pretraining corpus. Furthermore, as glossing conventions vary between annotators, the glosses produced by PolyGloss may not match the desired schema. Ginn et al. (2024b) studied full-parameter finetuning for unseen languages, but we argue this is an unrealistic scenario for virtually all documentary linguists (due to both technical difficulty and compute requirements). Instead, we propose the use of low-rank adaptation (LoRA), which drastically reduces the computational cost of training and often requires less data (Hu et al., 2022). Given a small dataset of new glossed examples, it is feasible to run LoRA adaptation with limited computational resources. For example, a glossing software could periodically train an adapter overnight, enabling predictions in the linguist's target language without any additional effort.

We simulate a realistic annotation scenario for Vamale (which is never seen during pretraining), training LoRA adapters on increasingly large training datasets (increasing increments of 50 up to the full 380 examples). We use the Vamale data from Yang et al. (2025b) and train rank 8 adapters for 25 epochs, using a batch size of 32; on an A100, the largest training only took 12 minutes. We report MER scores when adapting the interleaved PolyGloss model and a ByT5 base model in Figure 7. The PolyGloss quickly adapts to the new language, while the ByT5-based model never improves.

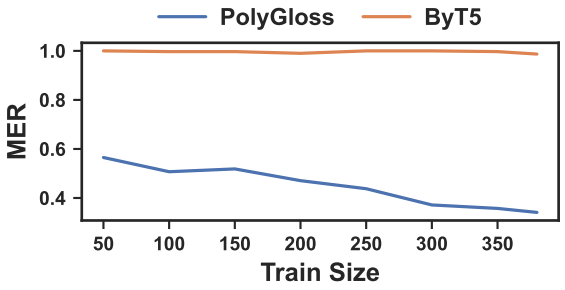

Figure 7: Morpheme error rate for Vamale when training LoRAs on the PolyGloss interleaved model and ByT5 with different size training sets

9 Alignment Score as a Reward Function

Our alignment score can also be used as a scalar reward function for reinforcement learning with verifiable rewards (RLVR, DeepSeek-AI and alia, 2025) when predicting a concatenated segmentation and glosses. We demonstrate this with a small pilot study using vanilla GRPO (Shao et al., 2024) to optimize the PolyGloss ByT5 Concatenated model on Gitksan. We use $\beta = 0.1$, $lr = 5E-5$, batch and group sizes both 8, and train for 50 epochs. We use a temperature of 0.6, top-p of 0.9, and repetition penalty of 1.05 when sampling.

We report results before and after tuning in Figure 8, observing that all three metrics improve slightly. This approach could be scaled to the full training dataset as an additional post-training step, providing a method to improve the model without any additional labeled data. Furthermore, RL could be used when adapting the model to a new language that lacks gold-labeled morphological seg-

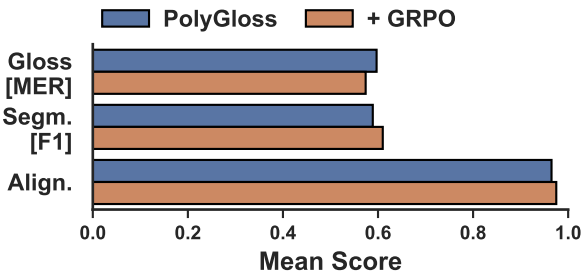

Figure 8: Scores before and after GRPO tuning for Gitksan.

mentations, since the alignment score will force the segmentation to (minimally) be aligned with the gold glosses.

10 Related Work

Research has explored automatic interlinear glossing using a variety of techniques including active learning (Palmer et al., 2010, 2009), conditional random fields (Moeller and Hulden, 2018; McMillan-Major, 2020b), neural models (Moeller and Hulden, 2018; Zhao et al., 2020), and large language models (Ginn et al., 2024a; Yang et al., 2024; Elsner and Liu, 2025; Shandilya and Palmer, 2025). Our work is directly inspired by He et al. (2023), a submission to the 2023 SIGMORPHON Shared Task on Interlinear Glossing (Ginn et al., 2023), and Ginn et al. (2024b), both of which used multilingual pretraining on glossed text.

Other work has focused on methods to automatically create IGT instances from other modalities, such as images of reference grammars (Round et al., 2020), LaTeX publications (Nordhoff and Krämer, 2022), and speech recordings (He et al., 2024). Recently, Aycock et al. and Yang et al. (2025b) proposed glossing as a method for testing LLMs' abilities to apply grammatical knowledge.

11 Conclusion

Tools for language documentation are only valuable when designed with the user—annotators and linguists—in mind. We address user feedback on automated glossing models and develop new multilingual models with major improvements over prior work. Most significantly, the POLYGLOSS models predict both morphological segments and interlinear glosses in a single forward pass, enabling more useful, interpretable, and trustworthy suggestions. We set a new state-of-the-art for glossing in several languages, as well as optimizing our model for segmentation accuracy and alignment between the two tasks. Finally, we offer practical recommendations for predicting performance and adapting the model to new data, keeping in mind computational constraints.

Going forward, we plan to work with developers of annotation software such as ELAN (The Language Archive, 2025) and FLEx (SIL GLobal, 2026) to integrate our model into real-world documentation workflows. One straightforward solution is to use our model to predict segmentation and gloss lines for samples with transcribed text, filter predictions according to model confidence, and allow the user to accept, reject, or edit the predictions.

Limitations

We do not compare against closed-sourced LLMs for several reasons. First, the training datasets for these models are opaque, and with much of the test data being available online, there is risk of contamination (of course, this may also be true for the Qwen and Cohere models as well). Second, endangered language data often bears considerations of data sovereignty, and language communities often do not want their data to be sent to a third-party provider. Our models are open-source and open-weights and can be finetuned locally.

For morphological segmentation, we do not differentiate between surface-level morphemes and underlying-form morphemes (and our dataset includes both). Thus, the task is not exactly identical across datasets, and there is no guarantee that the predicted segmentation may exactly match the input string.

We do not attempt to standardize the glosses in our dataset for two reasons. First, conventions and meanings vary greatly across annotators, and there is no way we could ensure the original intent of the annotator was preserved for the thousands of examples and languages in our dataset. Second, Ginn et al. (2024b) tried standardizing glosses to the UniMorph conventions (Batsuren et al., 2022) and found no evidence that it improved performance. Instead, our model outputs glosses according to the various conventions in its training data, and it can be adapted to a new convention via low-rank finetuning.

Ethical Considerations

Automatic approaches to interlinear glossing are intended to help accelerate the language documentation process and contribute to the fight against language death. However, there is risk of misuse, and these systems should not fully replace human annotators, which could result in erroneous documentation that hinders downstream applications. All data was taken from existing work and used in accordance with the original stakeholders' wishes. Finally, our work used a large amount of computational resources, which inevitably bears an environmental cost.

Acknowledgments

We thank Morris Alper for suggesting our "interleaved" format. Parts of this work were supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another." This work also used DeltaAI at NCSA through allocation CIS250116 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Seyede Faezeh Hosseini Alast and Sasan Baleghizadeh. 2024. The interplay of glossing with text difficulty and comprehension levels. *Language Teaching Research*, 28(3):1201–1230.

Seth Aycock, David Stap, Di Wu, Christof Monz, and Khalil Sima'an. Can llms really learn to translate a low-resource language from one grammar book? In *The Thirteenth International Conference on Learning Representations*.

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, and 76 others. 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.

Emily M. Bender, Michael Wayne Goodman, Joshua Crowgey, and Fei Xia. 2013. Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties. In *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 74–83, Sofia, Bulgaria. Association for Computational Linguistics.

Rogier Blokland, Niko Partanen, Michael Rießler, and Joshua Wilbur. 2019. Using computational approaches to integrate endangered language legacy data into documentation corpora: Past experiences and challenges ahead. In *Workshop on Computational Methods for Endangered Languages, Honolulu, Hawai'i, USA*, volume 2, pages 24–30.

Camilo Andrés Bonilla Carvajal. 2025. Interlinear translations reduce cognitive load on efl vocabulary acquisition. *Íkala, Revista de Lenguaje y Cultura*, 30(1).

Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos, and Graham Neubig. 2023. Teacher perception of automatically extracted grammar concepts for L2 language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3776–3793, Singapore. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepSeek-AI and alia. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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.

Ryan Alden Georgi. 2016. *From Aari to Zulu: massively multilingual creation of language tools using interlinear glossed text*. Ph.D. thesis.

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.

Leander Girrbach. 2023a. 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 151–165, Toronto, Canada. Association for Computational Linguistics.

Leander Girrbach. 2023b. 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.

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.

Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig, and Lori Levin. 2023. SigMoreFun submission to the SIGMORPHON shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 209–216, Toronto, Canada. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250, Online. Association for Computational Linguistics.

Angelina McMillan-Major. 2020a. Automating gloss generation in interlinear glossed text. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 355–366, New York, New York. Association for Computational Linguistics.

Angelina McMillan-Major. 2020b. Automating Gloss Generation in Interlinear Glossed Text. *Proceedings of the Society for Computation in Linguistics*, 3(1):338–349. Publisher: University of Mass Amherst.

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.

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.

Alexis Palmer, Taesun Moon, and Jason Baldridge. 2009. Evaluating automation strategies in language documentation. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 36–44.

Alexis Palmer, Taesun Moon, Jason Baldridge, Katrin Erk, Eric Campbell, and Telma Can. 2010. Computational strategies for reducing annotation effort in language documentation: A case study in creating interlinear texts for Uspanteko. *Linguistic Issues in Language Technology*, 3.

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

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.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-centric evaluation of OCR systems for kwak'wala. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 19–29, Remote. 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.

Bhargav Shandilya and Alexis Palmer. 2025. Boosting the capabilities of compact models in low-data contexts with large language models and retrieval-augmented generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7470–7483, Abu Dhabi, UAE. Association for Computational Linguistics.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

SIL GLobal. 2026. FLEx (fieldworks language explorer).

Gemma Team. 2025. Gemma 3.

The Language Archive. 2025. ELAN (version 7.0).

Heli Uibo, Jack Rueter, and Sulev Iva. 2017. Building and using language resources and infrastructure to develop e-learning programs for a minority language. *Proceedings of the Joint 6th Workshop on NLP for Computer Assisted Language Learning and 2nd Workshop on NLP for Research on Language Acquisition at NoDaLiDa*, 134:61–67.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *Preprint*, arXiv:2307.09702.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, and alia. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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.

Changbing Yang, Garrett Nicolai, and Miikka Silfverberg. 2024. 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.

Olga Zamaraeva. 2016. Inferring morphotactics from interlinear glossed text: Combining clustering and precision grammars. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 141–150, Berlin, Germany. 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.

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

A Base Model Ablation

In addition to the ByT5-based PolyGloss models, we also performed training on decoder-only instruction-tuned LLMs with significantly more parameters. We tried a number of hyperparameters through manually tuning, but were unable to match the ByT5 model performance. We report the best hyperparameters in Table 6.

	ByT5	Qwen
LR	5E-5	5E-5
Batch size	64	18
Epochs	15	15

Table 6: Hyperparameters for PolyGloss training.

We provide average scores in Figure 9 and full scores in Appendix B. Generally, none of the other base models converged to a decent loss, and the scores are unsurprisingly far worse than ByT5. We hypothesize a few possible explanations. First, the decoder-only models use subword tokenizers (as opposed to the byte-level tokenizer of ByT5), which can cause issues for rare languages—particularly for segmentation, where the expected output is the input string with morpheme boundaries inserted, a very difficult task using multi-character subword tokens. Second, these models are trained on much more data (with instruction tuning and reinforcement learning) than ByT5, making it more difficult to escape the local minimum during continued finetuning. Third, these models are much larger, and our dataset's size relative to the parameter count may result in high-variance or uninformative gradients, impeding training. Still, we expect with the right hyperparameters, these models should be able to at least match the accuracy of the ByT5 model, which could be explored by future work.

B Ablation Results

Full results for the monolingual ablations and task format ablations are given in Table 7, Table 8, and Table 9. We also provide evaluation metrics when

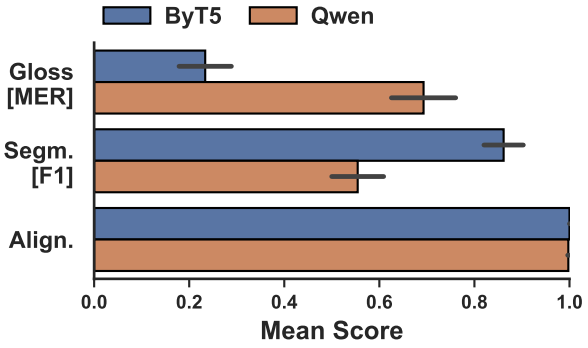

Figure 9: Scores for PolyGloss multilingual models using the Interleaved format and different base models. Scores are averaged across nine languages and reported with standard error.

evaluating PolyGloss only on long examples in Table 10.

C In-Context Learning Failures

For the smaller Qwen model, the majority of outputs are malformed, resulting in very poor scores. The larger models conformed to the format consistently, but still produced poor outputs. We observe the several types of failure within the Tsez evaluation data:

Hallucinated Transcription The most common failure case is that the model produces a segmentation that does not match the input transcription. For the Qwen model, the predicted segmentation was often wildly divergent, such as the following example in which the model hallucinates a word and omits two words from the input:

(4) Hič'č'a eɣenä eƛin:
xan-a eƛi-n
khan-ERG say-PST.UNW

For the larger models, this type of failure was also incredibly common, as in the following:

(5) Nedur yeda ukaynosi , kidbä šebi
nedur-a yeda-q r-oq-n kid-a šebi r-oq-n-ƛin
roqäƛin esirno .

DEM1.IPL.OBL-ERG food-POSS.ESS IV-happen-PST.UNW girl-ERG what IV-happen-PST.UNW-QUOT

Here, the segmentation is fairly close to the prediction transcription, but there are additional segments such as the first "-a" and "-q" inserted. In both cases, predicting an invalid segmentation naturally results in low-quality glosses. There is at least one character error in 79.9% of examples with the Qwen model, 98.0% of examples with Gemma, and 98.4% of examples with Aya. We also compute the character error rate between the predicted and gold transcriptions (ignoring segmentation markers), with a CER of 0.66 for Qwen, 0.46 for Gemma, and 0.38 for Aya. This suggests that the smaller the model, the more the transcription diverges. We suspect these errors occur due to tokenization, which our PolyGloss model avoids by using byte-level tokenization.

Bad Format For the Qwen model, the output often does not follow the interleaved format, missing the interleaved parentheses used to indicate glosses. This occurs in 17.4% of cases for Qwen, and never for the other two models. This could be addressed by trying different formats, providing more in-context examples, or using constrained decoding (Willard and Louf, 2023) to force the correct format.

Repetition Finally, the LLM occasionally gets stuck in an endless repetition, in 2.5% of cases for Qwen and never for the other models. This could potentially be addressed by tuning the repetition penalty.

D Use of AI Assistants

We used AI assistants via GitHub's Copilot to review pull requests. We otherwise did not use AI assistance at any point of the research study.

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Finetuned ByT5 (separate)	**0.092**	**0.059**	0.977	**0.116**	0.080	0.380	0.184	0.283	**0.292**	0.274
Finetuned ByT5 (joint)	0.128	0.064	0.841	0.170	**0.075**	0.382	0.144	0.225	0.346	0.264
Pipeline	0.140	0.076	0.972	0.173	0.096	0.454	0.238	0.314	0.433	0.322
Girrbach (2023b)	0.185	0.110	0.732	0.242	0.353	0.392	0.266	0.246	0.512	0.338
PolyGloss (ByT5, multitask)	0.177	0.089	0.603	0.162	0.122	0.383	0.189	0.328	0.329	0.265
PolyGloss (ByT5, concat)	0.171	0.080	**0.597**	0.165	0.108	**0.357**	0.180	0.310	0.315	0.254
PolyGloss (ByT5, interleaved)	0.152	0.072	**0.597**	0.160	0.095	**0.357**	**0.142**	**0.222**	0.306	**0.234**
PolyGloss (Qwen, interleaved)	0.453	0.626	0.902	0.418	0.480	0.872	0.739	0.912	0.835	0.693

Table 7: Morpheme error rate (↓) for **glossing** on the held-out test set. For GlossLM, the only languages explicitly included in the pretraining corpus are `arp`, `ddo`, and `git`, so scores on other languages (marked with *) are very poor.

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Finetuned ByT5 (separate)	0.921	0.975	0.191	0.851	0.972	0.795	0.930	0.957	0.717	0.812
Finetuned ByT5 (joint)	0.924	0.976	0.152	0.865	0.972	0.827	0.935	0.962	0.766	0.820
Pipeline	0.921	0.975	0.191	0.851	0.972	0.795	0.930	0.957	0.717	0.812
Girrbach (2023b)	0.531	0.447	0.241	0.548	0.348	0.349	0.441	0.730	0.494	0.459
Qwen 0.6B (ICL)		0.096		0.297	0.233	0.118	0.133	0.298	0.190	
Aya Expanse 8B (ICL)	0.457	0.249	0.132	0.509	0.569	0.302	0.337	0.465	0.324	0.371
PolyGloss (ByT5, multitask)	0.903	0.967	0.605	0.850	0.963	0.826	0.925	0.938	0.763	0.860
PolyGloss (ByT5, concat)	0.901	0.964	0.589	0.852	0.963	0.825	0.924	0.940	0.772	0.859
PolyGloss (ByT5, interleaved)	0.910	0.972	0.595	0.855	0.963	0.782	0.935	0.959	0.782	0.862
PolyGloss (Qwen, interleaved)	0.658	0.610	0.235	0.663	0.754	0.485	0.554	0.656	0.375	0.555

Table 8: Morpheme F1 (↑) for **segmentation** on the held-out test set.

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Finetuned ByT5 (separate)	0.980	0.988	0.650	0.977	0.979	0.885	0.907	0.963	0.904	0.915
Finetuned ByT5 (joint)	0.999	**1.000**	0.996	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	0.997	0.999
Pipeline	0.998	0.989	0.464	**1.000**	0.994	0.522	0.785	0.914	0.713	0.820
Girrbach (2023b)	0.998	0.998	**1.000**	1.000	0.895	0.985	0.976	0.994	0.941	0.976
PolyGloss (ByT5, multitask)	0.984	0.995	0.919	0.990	0.982	0.973	0.985	0.988	0.941	0.973
PolyGloss (ByT5, concat)	0.996	0.997	0.965	**1.000**	0.994	0.986	0.995	0.999	0.973	0.989
PolyGloss (ByT5, interleaved)	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
PolyGloss (Qwen, interleaved)	0.999	0.999	0.996	0.999	1.000	0.988	1.000	0.997	1.000	0.997

Table 9: **Alignment score** (↑) between morphological segmentation and predicted glosses on the held-out test set.

	arp	ddo	git	usp	ain	lez	ntu	nyb	ruc	*Avg.*
Glossing										
PolyGloss (Qwen, interleaved)	0.453	0.626	0.902	0.418	0.480	0.872	0.739	0.912	0.835	0.693
PolyGloss (ByT5, multitask, long only)	0.167	0.106	0.613	0.159	0.129	0.412	0.186	0.300	0.332	0.267
PolyGloss (ByT5, concat, long only)	0.161	0.099	0.628	0.169	0.115	0.409	0.191	0.293	0.320	0.265
PolyGloss (ByT5, interleaved, long only)	0.146	0.089	0.603	0.151	0.102	0.317	0.129	0.206	0.294	0.226
Segmentation										
PolyGloss (ByT5, multitask, long only)	0.919	0.959	0.620	0.829	0.967	0.816	0.919	0.945	0.800	0.864
PolyGloss (ByT5, concat, long only)	0.917	0.958	0.573	0.837	0.969	0.816	0.921	0.948	0.808	0.861
PolyGloss (ByT5, interleaved, long only)	0.924	0.965	0.628	0.849	0.961	0.791	0.943	0.966	0.825	0.872
Alignment										
PolyGloss (ByT5, multitask, long only)	0.987	0.995	0.903	0.985	0.981	0.972	0.975	0.992	0.951	0.971
PolyGloss (ByT5, concat, long only)	0.996	0.997	0.932	0.999	0.991	0.989	0.989	1.000	0.981	0.986
PolyGloss (ByT5, interleaved, long only)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Table 10: Glossing, segmentation, and alignment scores for PolyGloss models only evaluated on very long inputs (75th percentile in length per language).