# Advances and Challenges in Automated Interlinear Glossing

Michael Ginn

University of Colorado Boulder

Background

Shared Task

Robust Generalization

Multilingual Glossing

Future Work

# Background

## Interlinear Glossed Text

| | | | | | |
|---|---|---|---|---|---|
| **Ti-** | **j-** | **ya'** | **-tq** | **-a'** | **juntiir** |
| Inc- | E3s- | *give* | -Pl | -Enf | *everything* |

They give us everything.

IGT is a common format for language documentation

# Background

## Interlinear Glossed Text

**Transcription**

| | | | | | |
|---|---|---|---|---|---|
| **Ti-** | **j-** | **ya'** | **-tq** | **-a'** | **juntiir** |
| Inc- | E3s- | *give* | -Pl | -Enf | *everything* |

They give us everything.

# Background

## Interlinear Glossed Text

| | | | | | | |
|---|---|---|---|---|---|---|
| | **Ti-** | **j-** | **ya'** | **-tq** | **-a'** | **juntiir** |
| **Glosses** | Inc- | E3s- | *give* | -Pl | -Enf | *everything* |

They give us everything.

# Background

## Interlinear Glossed Text

| | | | | | |
|---|---|---|---|---|---|
| **Ti-** | **j-** | **ya'** | **-tq** | **-a'** | **juntiir** |
| Inc- | E3s- | *give* | -Pl | -Enf | *everything* |

**Translation** They give us everything.

# IGT can be used for…

Language preservation

Linguistic research

Moeller et al. (2020); Bender et al. (2013)

Language technologies
(MT, tagging, parsers)

Zhou et al. (2019); Georgi (2016)

Maintaining a standardized format

Morphological segmentation

Stem translation

**Creating annotated corpora requires significant effort and cost**

Annotating novel phenomena

Re-glossing the same morphemes many times

Maintaining a standardized format

Morphological segmentation

Stem translation

## Automated tools can aid annotators with repetitive tasks

Annotating novel phenomena

Re-glossing the same morphemes many times

Maintaining a standardized format

Morphological segmentation

Stem translation

## Automated tools can aid annotators with repetitive tasks

Annotating novel phenomena

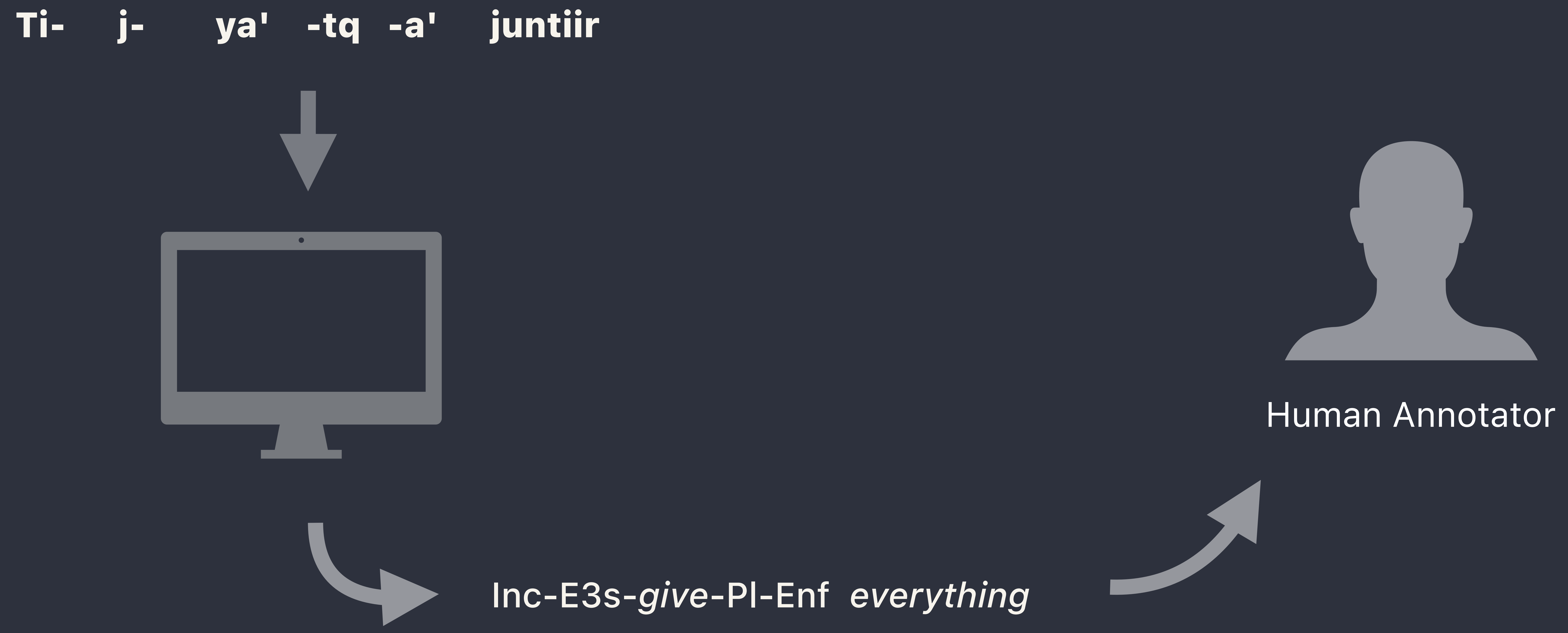
Re-glossing the same morphemes many times

# Background

## Interlinear Glossed Text

ML models can reduce annotator effort

Palmer & Baldridge (2009)



# Many approaches have been used to automate gloss prediction

MaxEnt Classifier

Palmer & Baldridge (2009)

Rule-Based Parsing

Bender et al. (2014)

CRFs

Moeller & Hulden (2018); McMillan-Major (2018)

RNNs

Moeller & Hulden (2018)

Transformers

Zhao et al. (2020)

How can we improve automated glossing systems?

Background

Shared Task

Robust Generalization

Multilingual Glossing

Future Work

# 2023 SIGMORPHON Shared Task

Ginn et al. (2023)

# 2023 SIGMORPHON Shared Task

- First public task for IGT glossing models
- Participants built systems for predicting glosses given transcriptions and (in some cases) translations

# 2023 SIGMORPHON Shared Task

los gato-s corr-en

↓

*the*-PL *cat*-PL *run*-3PL

**Open Track**

los gatos corren

↓

*the*-PL *cat*-PL *run*-3PL

**Closed Track**

# 2023 SIGMORPHON Shared Task

## Languages

**Arapaho**
175k tokens

**Gitksan**
1k tokens

**Lezgi**
9k tokens

**Natugu**
12k tokens

**Nyangbo**
11k tokens

**Tsez**
47k tokens

**Uspanteko**
45k tokens

# 2023 SIGMORPHON Shared Task

## Teams

| | |
|---|---|
| **COATES** | LSTM Encoder-Decoder |
| **LISNTeam** | Hybrid CRF-Neural |
| **SigMoreFun** | Multilingual Pretrained Transformers |
| **TeamSiggyMorph** | BiLSTM, ByT5 |
| **Tü-CL** | Straight-through gradient estimation, hard attention |

# 2023 SIGMORPHON Shared Task

## Results

Morpheme-Level Accuracy

| Submission | Arp | Ddo | Git | Lez | Ntu | Nyb | Usp | AVG | Complete? |
|---|---|---|---|---|---|---|---|---|---|
| TÜ-$CL_2$ | **78.47** | **73.95** | **11.72** | **62.10** | 56.32 | 85.24 | **70.05** | 62.55 | YES |
| TÜ-$CL_1$ | 76.56 | 70.29 | 9.26 | 62.03 | **56.38** | **86.74** | 60.42 | 60.24 | YES |
| TeamSiggyMorph$_1$ | - | 53.19 | - | 28.13 | 31.86 | 66.25 | 59.73 | 47.83 | |
| COATES$_1$ | 45.42 | 64.43 | 9.84 | 40.74 | 37.55 | 72.82 | 56.02 | 46.69 | YES |
| BASELINE | 44.19 | 51.23 | 8.54 | 41.62 | 18.17 | 14.22 | 57.24 | 33.60 | YES |

**Closed Track**

# 2023 SIGMORPHON Shared Task
## Results

| Morpheme-Level Accuracy | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Submission | Arp | Ddo | Git | Lez | Ntu | Nyb | Usp | AVG | Complete? |
| TÜ-CL$_2$ | **91.37** | **92.01** | 50.22 | **87.61** | 92.32 | **91.40** | **84.51** | 84.21 | YES |
| SigMoreFun$_2$ | 89.34 | 88.15 | **52.39** | 82.36 | 85.53 | 89.49 | 83.08 | 81.48 | YES |
| LISNTeam$_1$ | - | 91.39 | 50.80 | 87.17 | 92.60 | - | 82.42 | 80.88 | |
| TeamSiggyMorph$_2$ | - | 88.36 | 47.76 | 86.59 | 92.10 | 82.74 | 82.22 | 79.96 | |
| SigMoreFun$_1$ | 91.36 | 84.35 | 47.47 | 80.17 | 88.35 | 85.84 | 80.08 | 79.66 | YES |
| TÜ-CL$_1$ | 90.93 | 91.16 | 17.08 | 83.45 | 90.17 | 89.96 | 83.45 | 78.03 | YES |
| LISNTeam$_2$ | - | - | 51.09 | 86.52 | **92.77** | - | - | 76.79 | |
| BASELINE | 91.11 | 85.34 | 25.33 | 51.82 | 49.03 | 88.71 | 82.48 | 67.69 | YES |
| SigMoreFun$_4$ | 80.81 | 78.24 | 12.74 | 50.00 | 63.39 | 85.30 | 73.25 | 63.39 | YES |
| SigMoreFun$_3$ | 72.10 | 57.93 | 2.60 | 26.24 | 35.62 | 70.01 | 67.73 | 47.46 | YES |

Open Track

# 2023 SIGMORPHON Shared Task
## Observations

- Hard attention (Girrback, 2023) is highly effective at the joint segmentation and glossing task
  - Also provides an interpretable model
- Multilingual training (He et al., 2023) can provide benefits to low-resource languages

# What challenges remain with automated IGT systems?

Background

Shared Task

Robust Generalization

Multilingual Glossing

Future Work

# Robust Generalization

*Robust Generalization Strategies for Morpheme Glossing in an Endangered Language Documentation Setting.* Ginn and Palmer, 2023.

# Robust Generalization

- IGT corpora are often the product of a single documentation project
- Represent a limited domain of text (genre, speaker, etc)
- IGT models must generalize well to unseen texts for future documentation projects

# Robust Generalization

We **evaluate generalization** by splitting our dataset by **text genre**

Uspanteko corpus from Palmer et al. (2009)

12k lines

29 docs

Stories

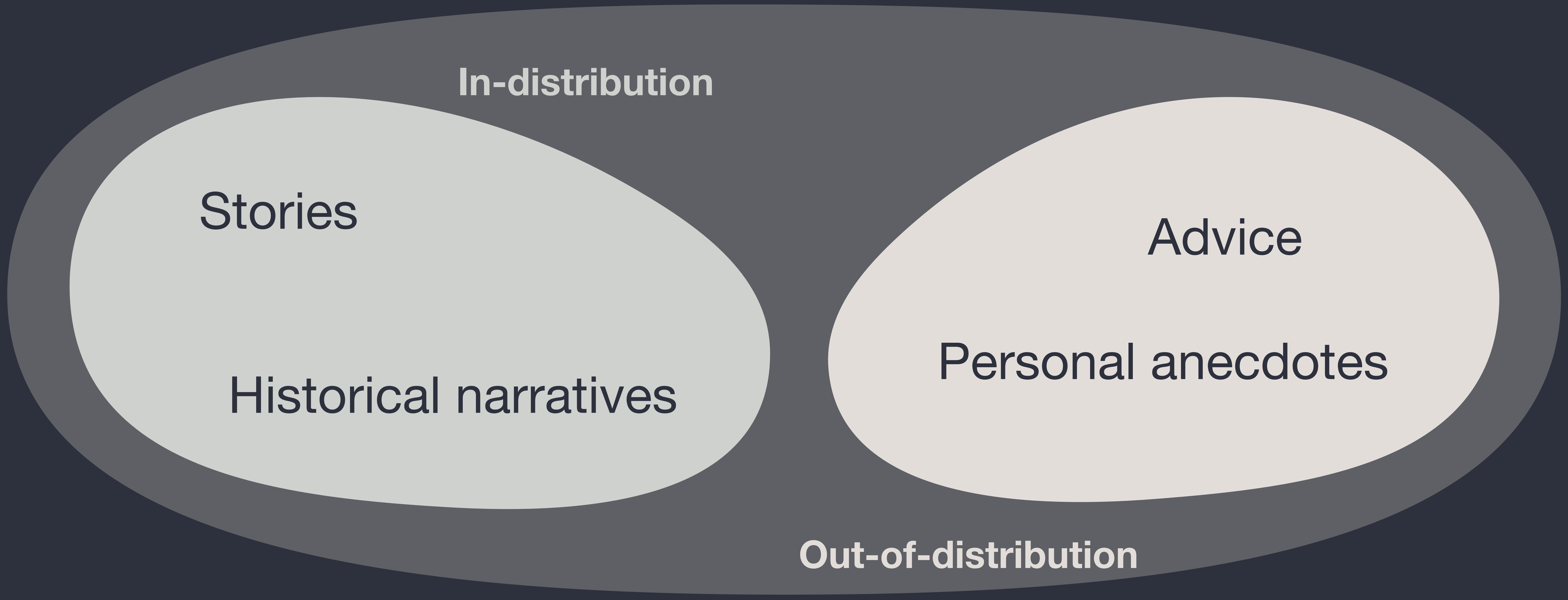
Advice

Historical narratives

Personal anecdotes

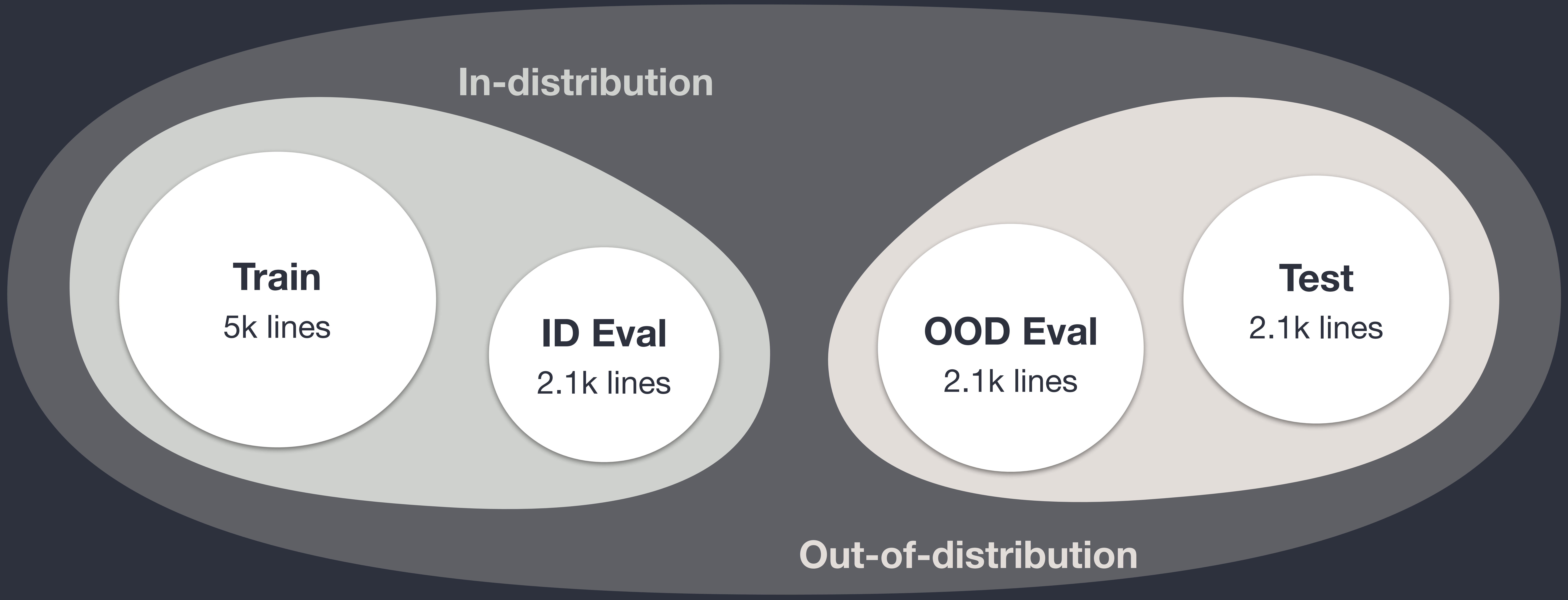
# Robust Generalization

We **evaluate generalization** by splitting our dataset by **text genre**



# Robust Generalization

ID data is used for **training** and **eval**,
and OOD is used for **eval** and **testing**



# Robust Generalization

**ID Eval**

2.1k lines

Perplexity: **77.8**

Accuracy: **84.5**

**OOD Eval**

2.1k lines

Perplexity: **94.0**

Accuracy: **74.6**

We demonstrate that the OOD data performs worse for **language modeling** and **gloss generation**.

**Evaluating generalization** is critical for robust IGT systems that can be used in **documentation projects**.

# Generalization Strategies

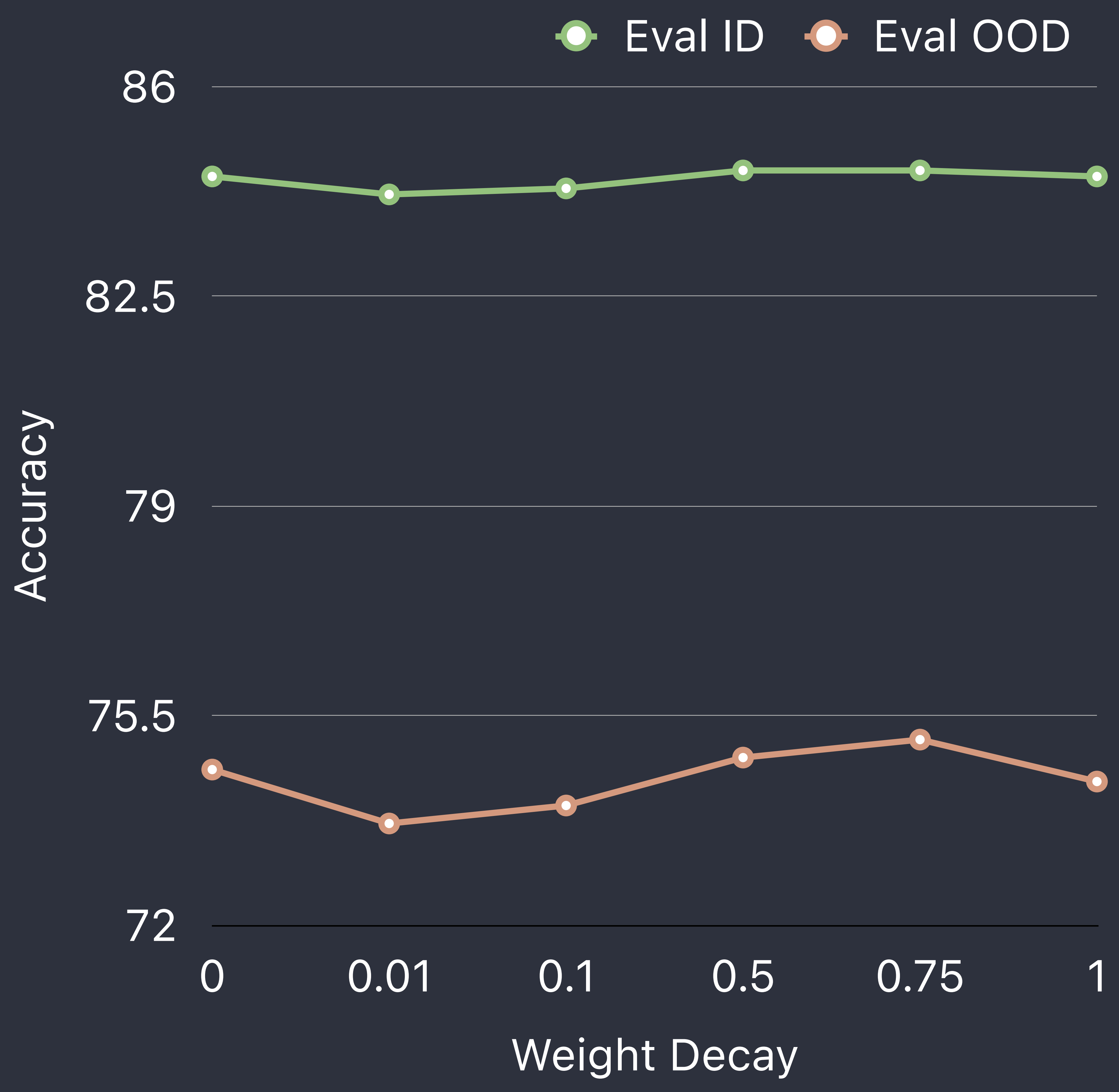
# Generalization Strategies

Weight Decay

Masked Language Modeling for OOV Tokens

Iterative Pseudo-Labeling

# Weight Decay




Higher weight decay helps **regularization** and **avoiding overfitting**.

# Generalization Strategies

Weight Decay

Masked Language Modeling for OOV Tokens

Iterative Pseudo-Labeling

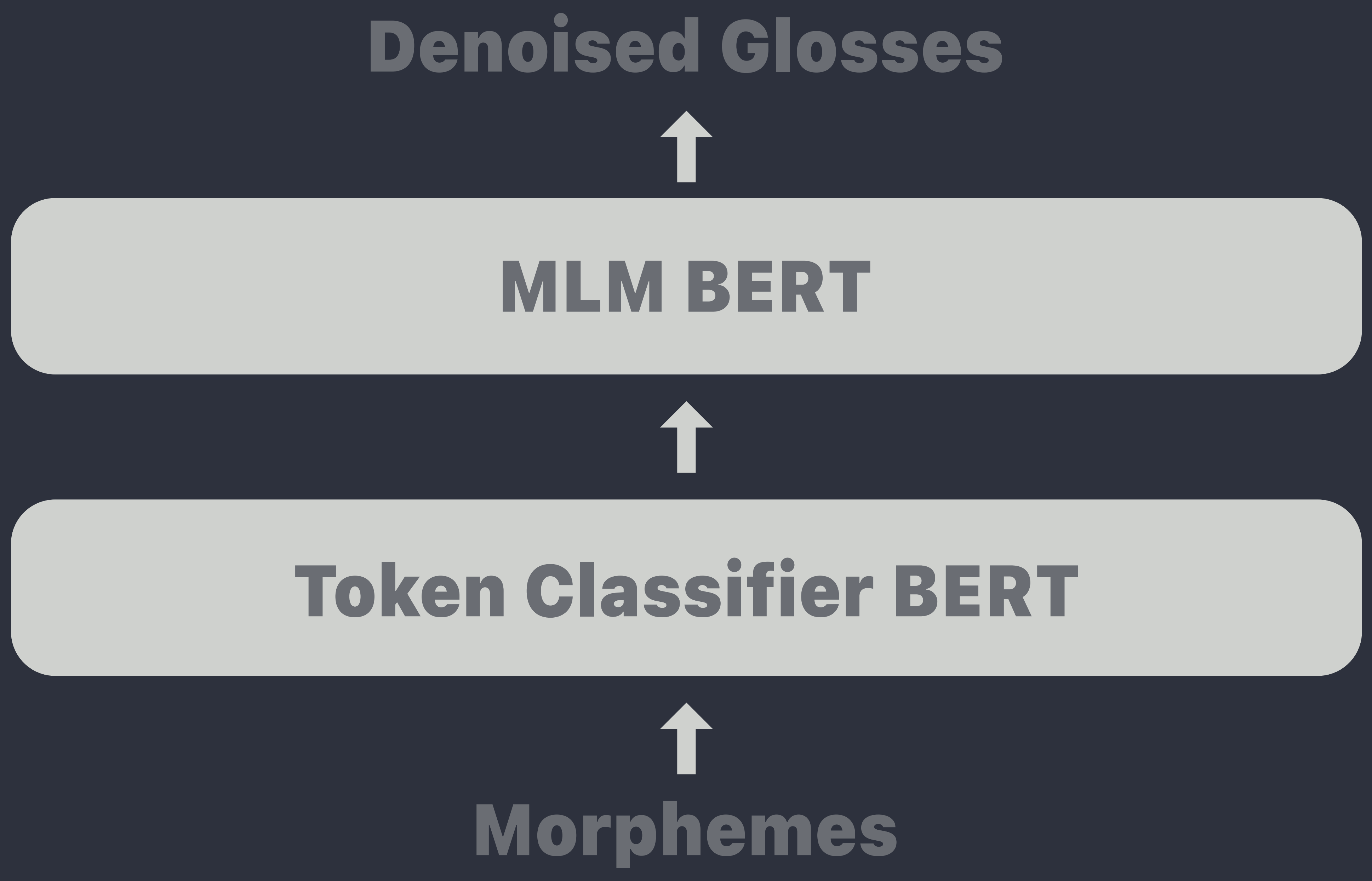
# Masked Language Modeling for OOV Tokens

- Out-of-vocabulary tokens are a greater cause of error in OOD texts
  - OOD: 6.2% vs ID: 3.0%
- Transformer glossing models may not handle OOV morphemes well
- We can often recover gloss from context

# Masked Language Modeling for OOV Tokens

We train a **masked language model** on gloss sequences and **apply it to the output** of the token classifier.

We achieve **limited improvement** (0.2%)



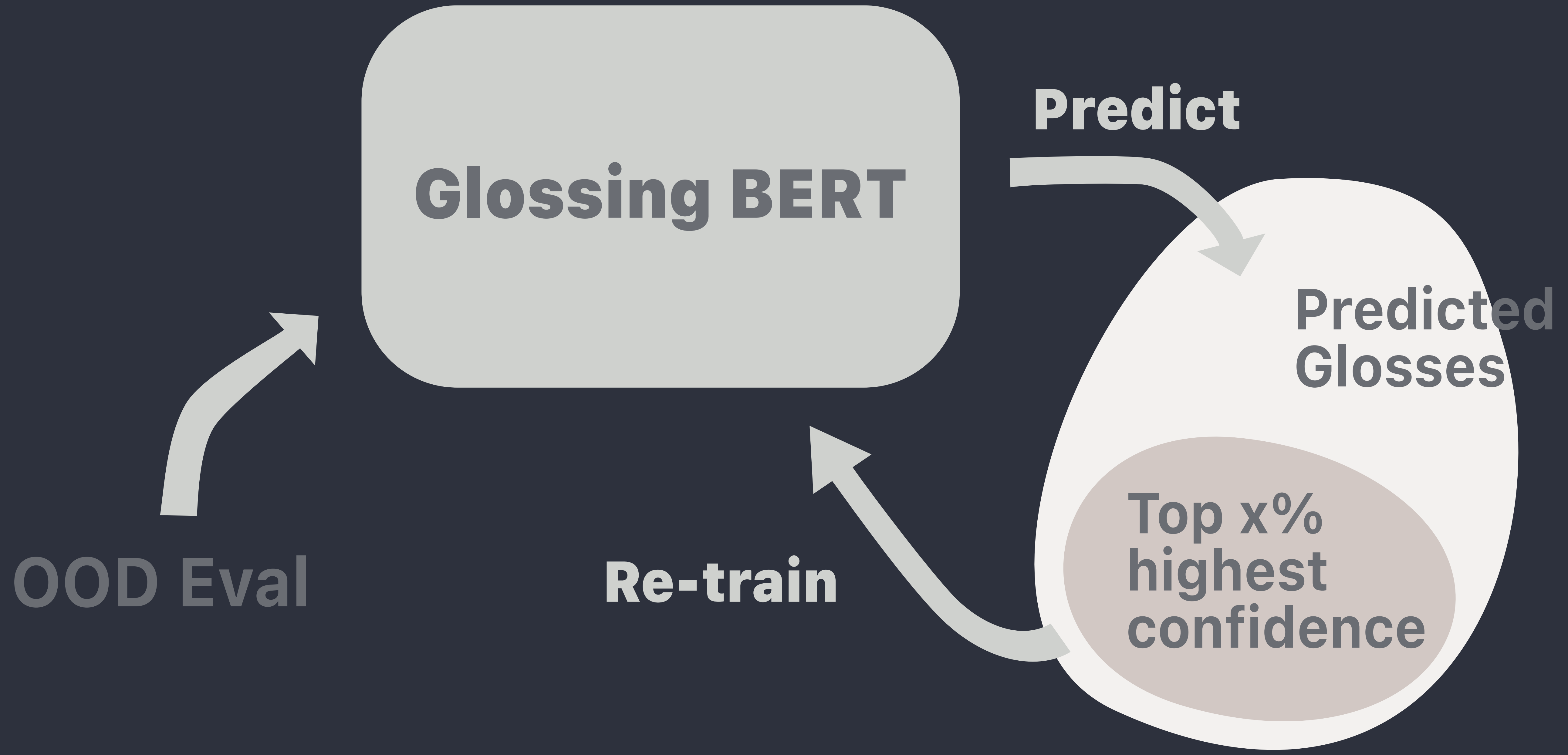
# Generalization Strategies

Weight Decay

Masked Language Modeling for OOV Tokens

Iterative Pseudo-Labeling

# Iterative Pseudo-Labeling

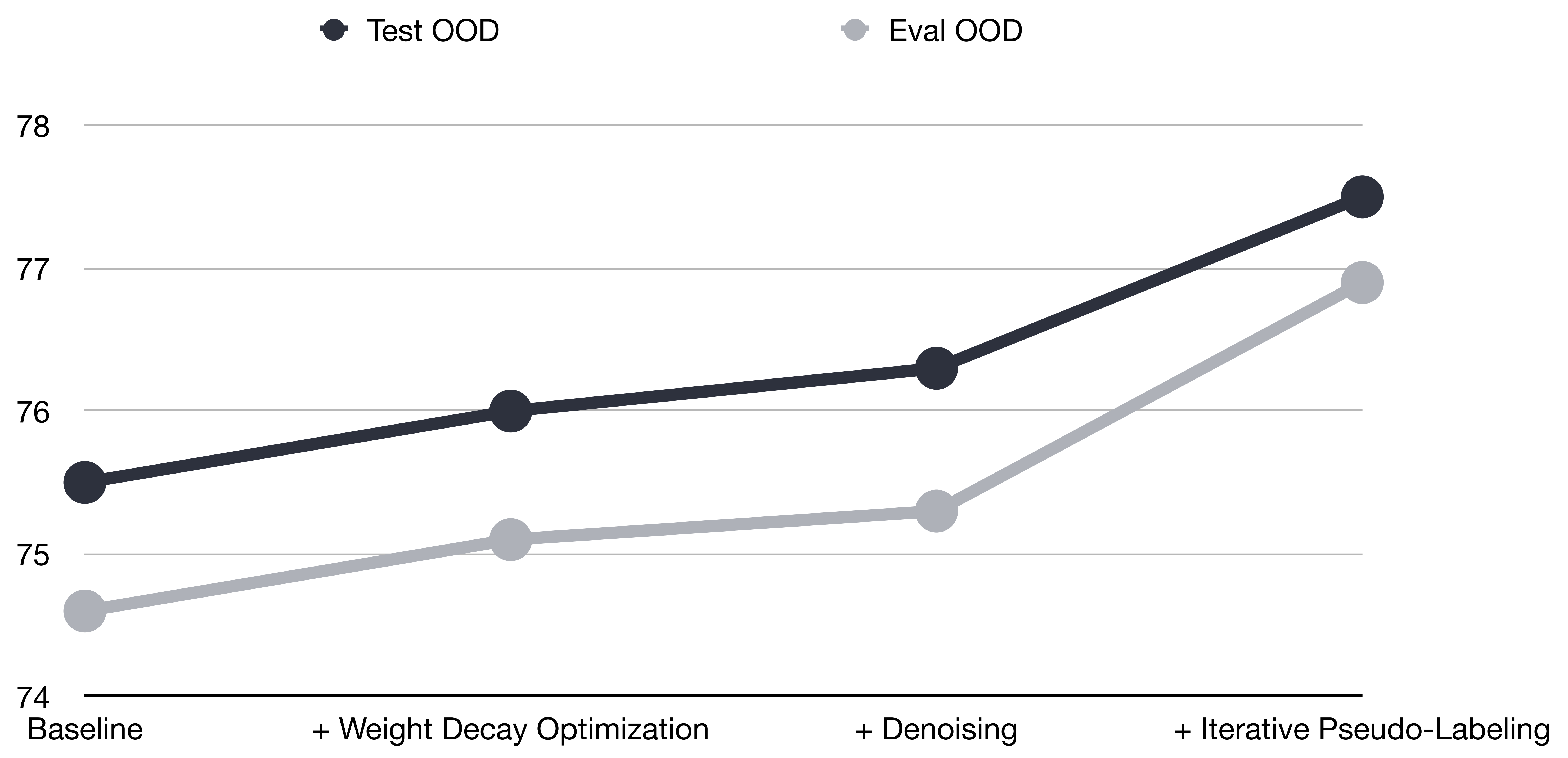



Use glossing model to do **inference on OOD data**

Select **top x% of predictions** by confidence and add to **training set**

**Repeat!**

# Results



Test OOD
Eval OOD
78
77
76
75
74
Baseline
+ Weight Decay Optimization
+ Denoising
+ Iterative Pseudo-Labeling

# Discussion

- Training strategies can improve robustness a limited amount
- Distributional shift remains a difficult problem for IGT models

Background

Shared Task

Robust Generalization

Multilingual Glossing

Future Work

# Multilingual Glossing

*GlossLM: Multilingual Pretraining for Low-Resource Interlinear Glossing.*
Ginn et al., 2024.

Can we leverage IGT across languages to improve automated glossing?

ByT5 pretrained on ODIN corpus

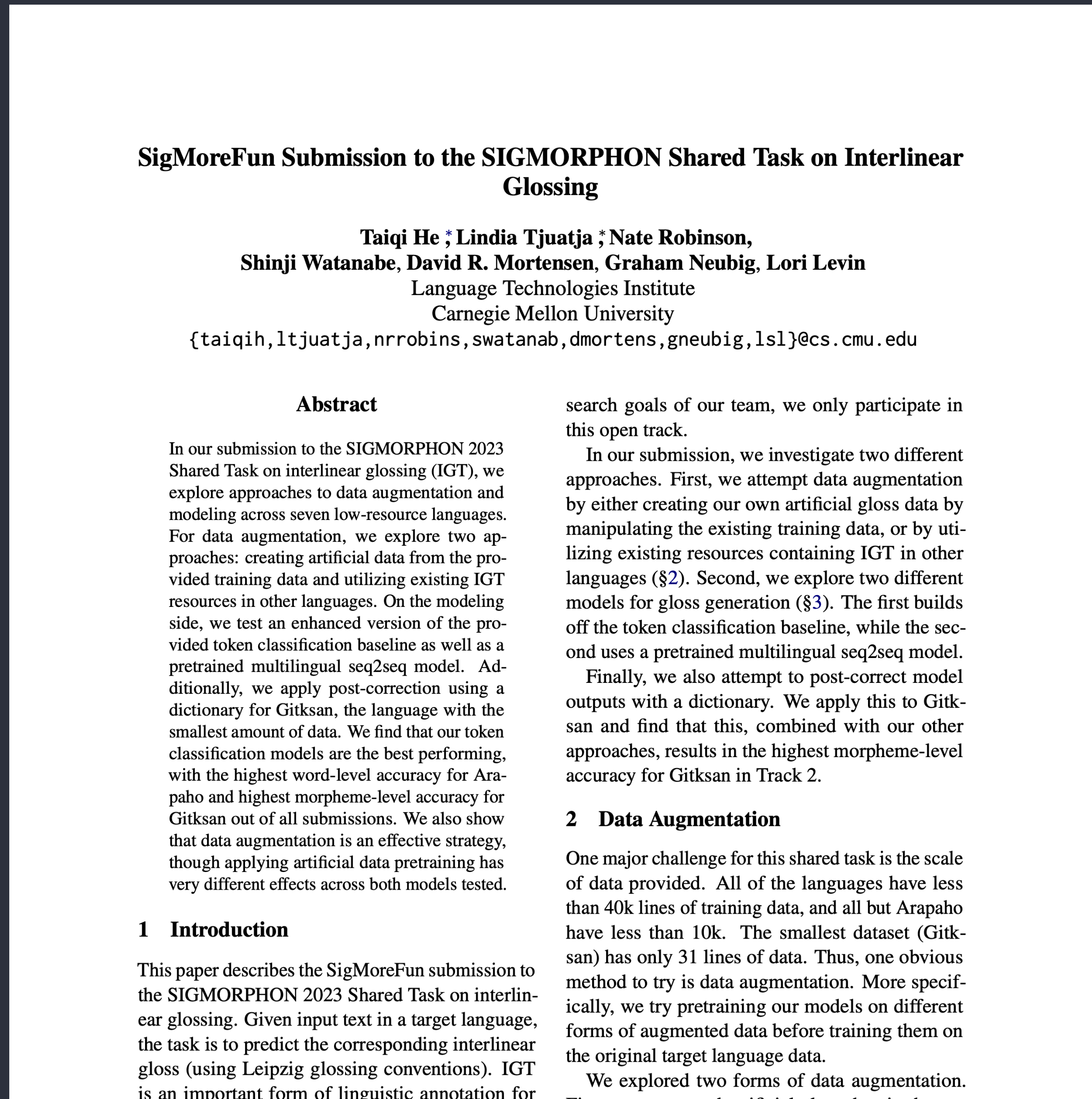
He et al. (2023)

CRF trained on IMTVault corpus

Okabe & Yvon (2024)

ByT5 pretrained on GlossLM corpus

Ginn et al. (2024)



**SigMoreFun Submission to the SIGMORPHON Shared Task on Interlinear Glossing**

**Taiqi He***, **Lindia Tjuatja***, **Nate Robinson**, **Shinji Watanabe**, **David R. Mortensen**, **Graham Neubig**, **Lori Levin**
Language Technologies Institute
Carnegie Mellon University
{taiqih,ltjuatja,nrrobins,swatanab,dmortens,gneubig,lsl}@cs.cmu.edu

**Abstract**

In our submission to the SIGMORPHON 2023 Shared Task on interlinear glossing (IGT), we explore approaches to data augmentation and modeling across seven low-resource languages. For data augmentation, we explore two approaches: creating artificial data from the provided training data and utilizing existing IGT resources in other languages. On the modeling side, we test an enhanced version of the provided token classification baseline as well as a pretrained multilingual seq2seq model. Additionally, we apply post-correction using a dictionary for Gitksan, the language with the smallest amount of data. We find that our token classification models are the best performing, with the highest word-level accuracy for Arapaho and highest morpheme-level accuracy for Gitksan out of all submissions. We also show that data augmentation is an effective strategy, though applying artificial data pretraining has very different effects across both models tested.

**1 Introduction**

This paper describes the SigMoreFun submission to the SIGMORPHON 2023 Shared Task on interlinear glossing. Given input text in a target language, the task is to predict the corresponding interlinear gloss (using Leipzig glossing conventions). IGT

search goals of our team, we only participate in this open track.

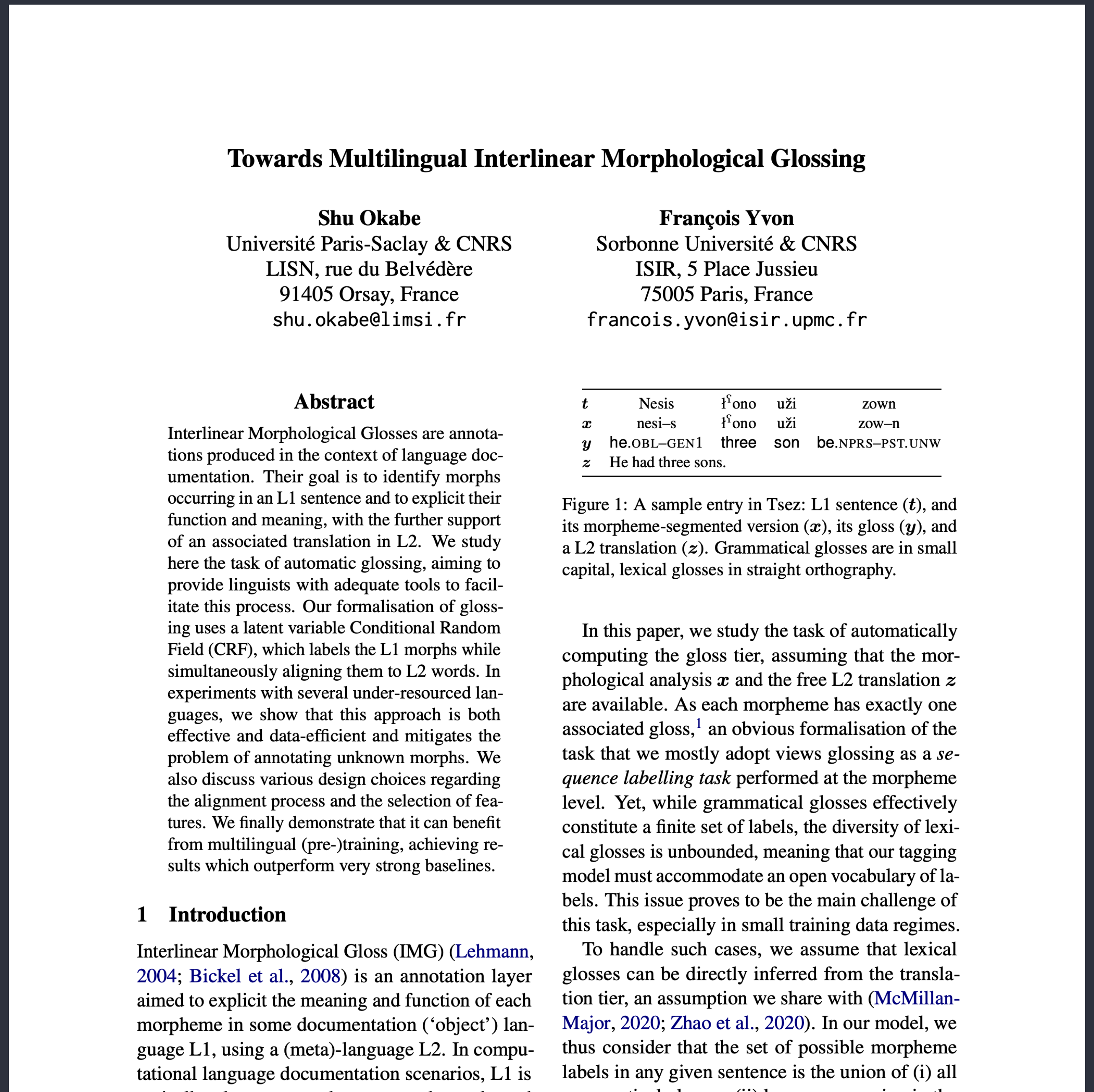
In our submission, we investigate two different approaches. First, we attempt data augmentation by either creating our own artificial gloss data by manipulating the existing training data, or by utilizing existing resources containing IGT in other languages (§2). Second, we explore two different models for gloss generation (§3). The first builds off the token classification baseline, while the second uses a pretrained multilingual seq2seq model.

Finally, we also attempt to post-correct model outputs with a dictionary. We apply this to Gitksan and find that this, combined with our other approaches, results in the highest morpheme-level accuracy for Gitksan in Track 2.

**2 Data Augmentation**

One major challenge for this shared task is the scale of data provided. All of the languages have less than 40k lines of training data, and all but Arapaho have less than 10k. The smallest dataset (Gitksan) has only 31 lines of data. Thus, one obvious method to try is data augmentation. More specifically, we try pretraining our models on different forms of augmented data before training them on the original target language data.

We explored two forms of data augmentation.



**Towards Multilingual Interlinear Morphological Glossing**

**Shu Okabe**
Université Paris-Saclay & CNRS
LISN, rue du Belvédère
91405 Orsay, France
shu.okabe@limsi.fr

**François Yvon**
Sorbonne Université & CNRS
ISIR, 5 Place Jussieu
75005 Paris, France
francois.yvon@isir.upmc.fr

**Abstract**

Interlinear Morphological Glosses are annotations produced in the context of language documentation. Their goal is to identify morphs occurring in an L1 sentence and to explicit their function and meaning, with the further support of an associated translation in L2. We study here the task of automatic glossing, aiming to provide linguists with adequate tools to facilitate this process. Our formalisation of glossing uses a latent variable Conditional Random Field (CRF), which labels the L1 morphs while simultaneously aligning them to L2 words. In experiments with several under-resourced languages, we show that this approach is both effective and data-efficient and mitigates the problem of annotating unknown morphs. We also discuss various design choices regarding the alignment process and the selection of features. We finally demonstrate that it can benefit from multilingual (pre-)training, achieving results which outperform very strong baselines.

**1 Introduction**

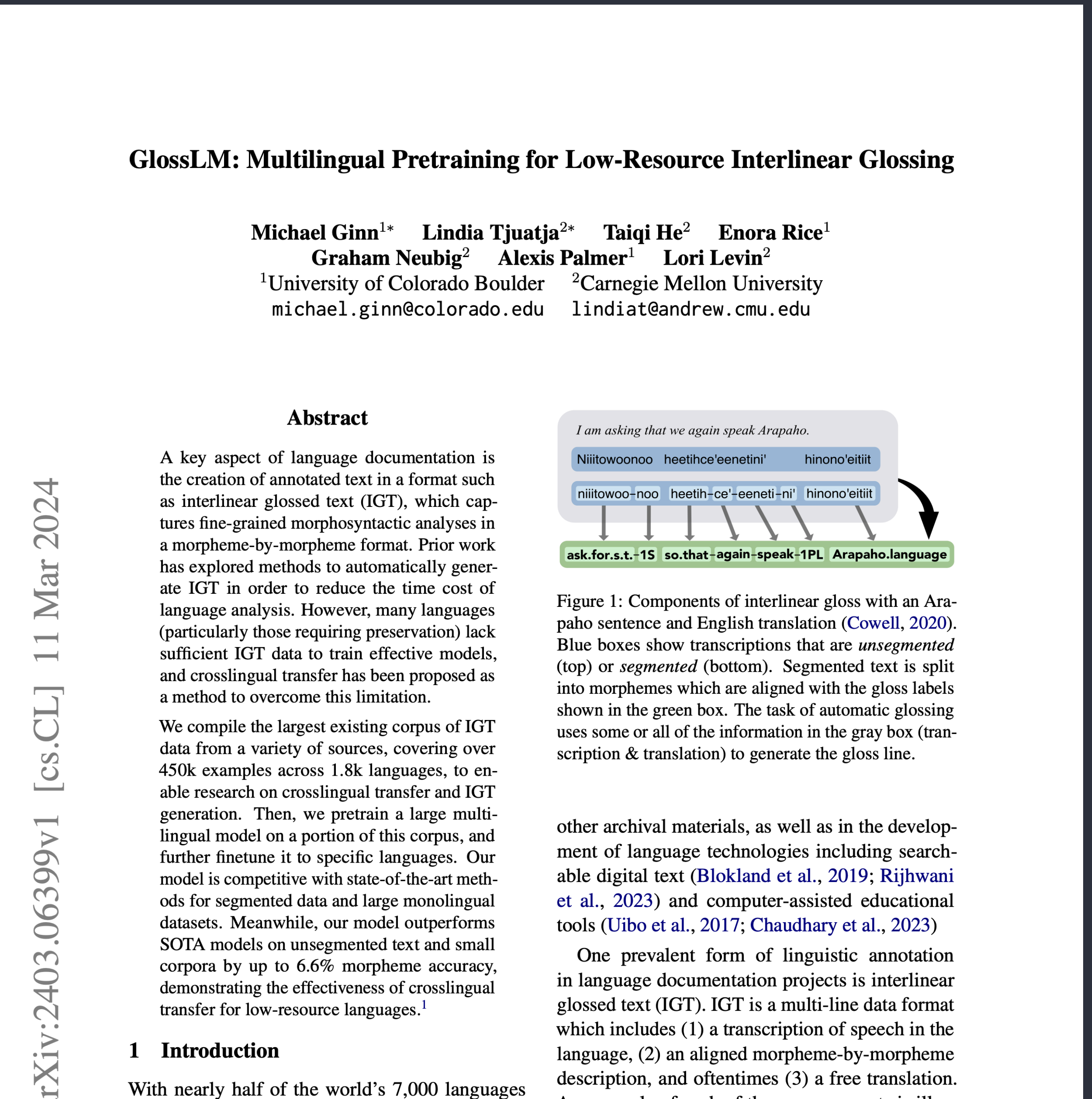
Interlinear Morphological Gloss (IMG) (Lehmann, 2004; Bickel et al., 2008) is an annotation layer aimed to explicit the meaning and function of each morpheme in some documentation ('object') language L1, using a (meta)-language L2. In computational language documentation scenarios, L1 is

| | | | | |
|---|---|---|---|---|
| $t$ | Nesis | ɬˁono | uži | zown |
| $x$ | nesi–s | ɬˁono | uži | zow–n |
| $y$ | he.OBL–GEN1 | three | son | be.NPRS–PST.UNW |
| $z$ | He had three sons. | | | |

Figure 1: A sample entry in Tsez: L1 sentence ($\boldsymbol{t}$), and its morpheme-segmented version ($\boldsymbol{x}$), its gloss ($\boldsymbol{y}$), and a L2 translation ($\boldsymbol{z}$). Grammatical glosses are in small capital, lexical glosses in straight orthography.

In this paper, we study the task of automatically computing the gloss tier, assuming that the morphological analysis $\boldsymbol{x}$ and the free L2 translation $\boldsymbol{z}$ are available. As each morpheme has exactly one associated gloss,[1] an obvious formalisation of the task that we mostly adopt views glossing as a *sequence labelling task* performed at the morpheme level. Yet, while grammatical glosses effectively constitute a finite set of labels, the diversity of lexical glosses is unbounded, meaning that our tagging model must accommodate an open vocabulary of labels. This issue proves to be the main challenge of this task, especially in small training data regimes.

To handle such cases, we assume that lexical glosses can be directly inferred from the translation tier, an assumption we share with (McMillan-Major, 2020; Zhao et al., 2020). In our model, we thus consider that the set of possible morpheme labels in any given sentence is the union of (i) all



**GlossLM: Multilingual Pretraining for Low-Resource Interlinear Glossing**

**Michael Ginn**[1*] **Lindia Tjuatja**[2*] **Taiqi He**[2] **Enora Rice**[1] **Graham Neubig**[2] **Alexis Palmer**[1] **Lori Levin**[2]
[1]University of Colorado Boulder [2]Carnegie Mellon University
michael.ginn@colorado.edu lindiat@andrew.cmu.edu

arXiv:2403.06399v1 [cs.CL] 11 Mar 2024

**Abstract**

A key aspect of language documentation is the creation of annotated text in a format such as interlinear glossed text (IGT), which captures fine-grained morphosyntactic analyses in a morpheme-by-morpheme format. Prior work has explored methods to automatically generate IGT in order to reduce the time cost of language analysis. However, many languages (particularly those requiring preservation) lack sufficient IGT data to train effective models, and crosslingual transfer has been proposed as a method to overcome this limitation.

We compile the largest existing corpus of IGT data from a variety of sources, covering over 450k examples across 1.8k languages, to enable research on crosslingual transfer and IGT generation. Then, we pretrain a large multilingual model on a portion of this corpus, and further finetune it to specific languages. Our model is competitive with state-of-the-art methods for segmented data and large monolingual datasets. Meanwhile, our model outperforms SOTA models on unsegmented text and small corpora by up to 6.6% morpheme accuracy, demonstrating the effectiveness of crosslingual transfer for low-resource languages.[1]

**1 Introduction**

With nearly half of the world's 7,000 languages

*I am asking that we again speak Arapaho.*
Niiitowoonoo heetihce'eenetini' hinono'eitiit
niiitowoo-noo heetih-ce'-eeneti-ni' hinono'eitiit
ask.for.s.t.-1S so.that-again-speak-1PL Arapaho.language

Figure 1: Components of interlinear gloss with an Arapaho sentence and English translation (Cowell, 2020). Blue boxes show transcriptions that are *unsegmented* (top) or *segmented* (bottom). Segmented text is split into morphemes which are aligned with the gloss labels shown in the green box. The task of automatic glossing uses some or all of the information in the gray box (transcription & translation) to generate the gloss line.

other archival materials, as well as in the development of language technologies including searchable digital text (Blokland et al., 2019; Rijhwani et al., 2023) and computer-assisted educational tools (Uibo et al., 2017; Chaudhary et al., 2023)

One prevalent form of linguistic annotation in language documentation projects is interlinear glossed text (IGT). IGT is a multi-line data format which includes (1) a transcription of speech in the language, (2) an aligned morpheme-by-morpheme description, and oftentimes (3) a free translation.

**IMTVault**
1.1k langs 80k rows
Nordhoff & Forkel (2023)

**APiCS**
76 langs 16k rows
Michaelis et al. (2013)

**ODIN**
936 langs 84k rows
Lewis & Xia (2010)

**GlossLM Corpus**

**UraTyp**
35 langs 1.7k rows
Norvik et al. (2022)

**SIGMORPHON**
7 langs 69k rows
Ginn et al. (2023)

**Guarani Corpus**
1 lang 803 rows
Zubizarreta (2023)

Standardized punctuation and formatting

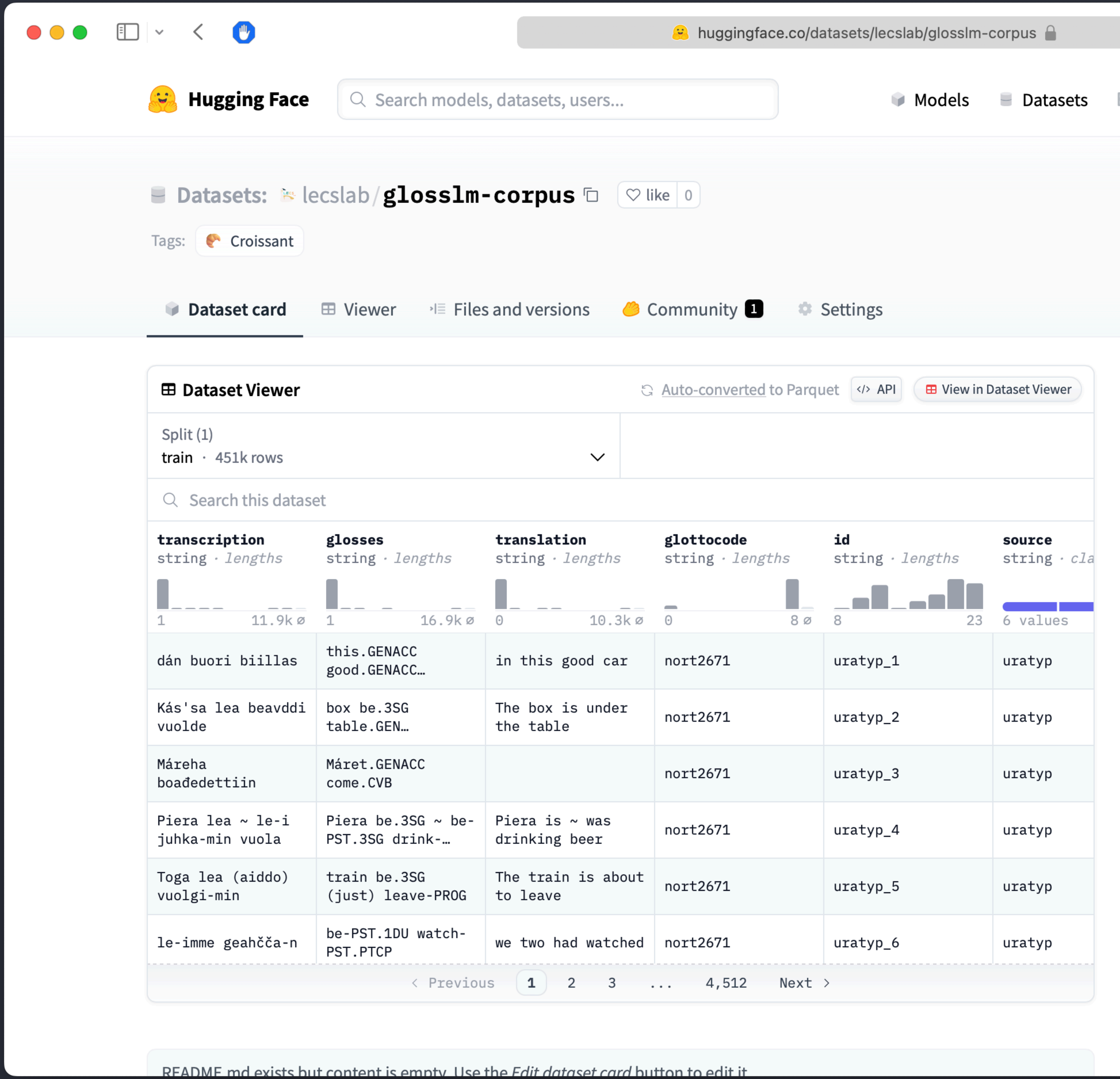
Filtering of low-quality rows

**GlossLM Corpus**

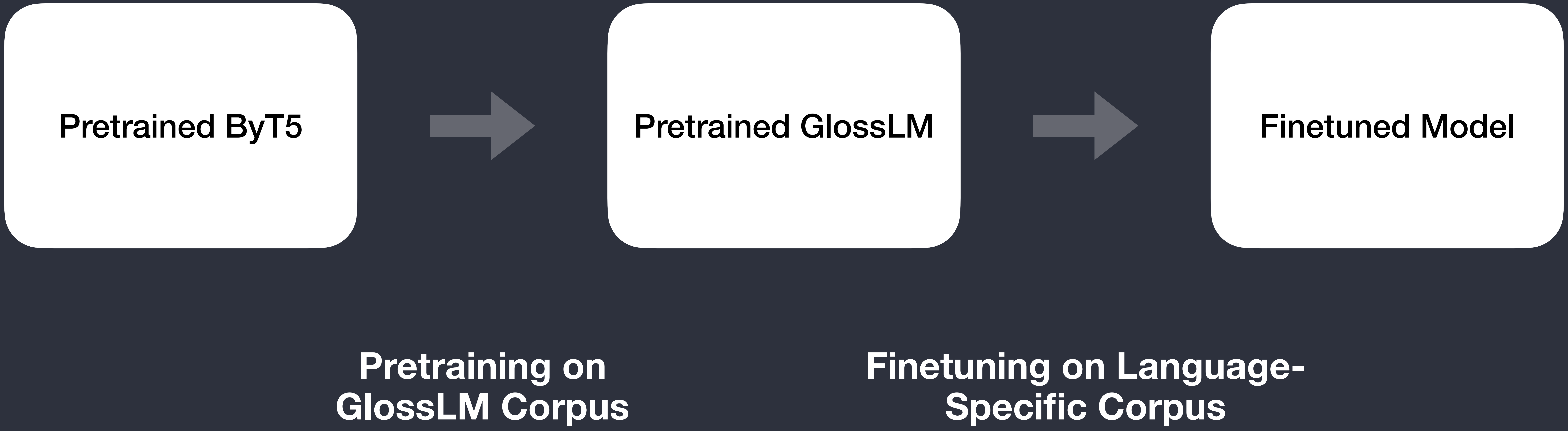
1.8k langs  451k rows

Translation language verification

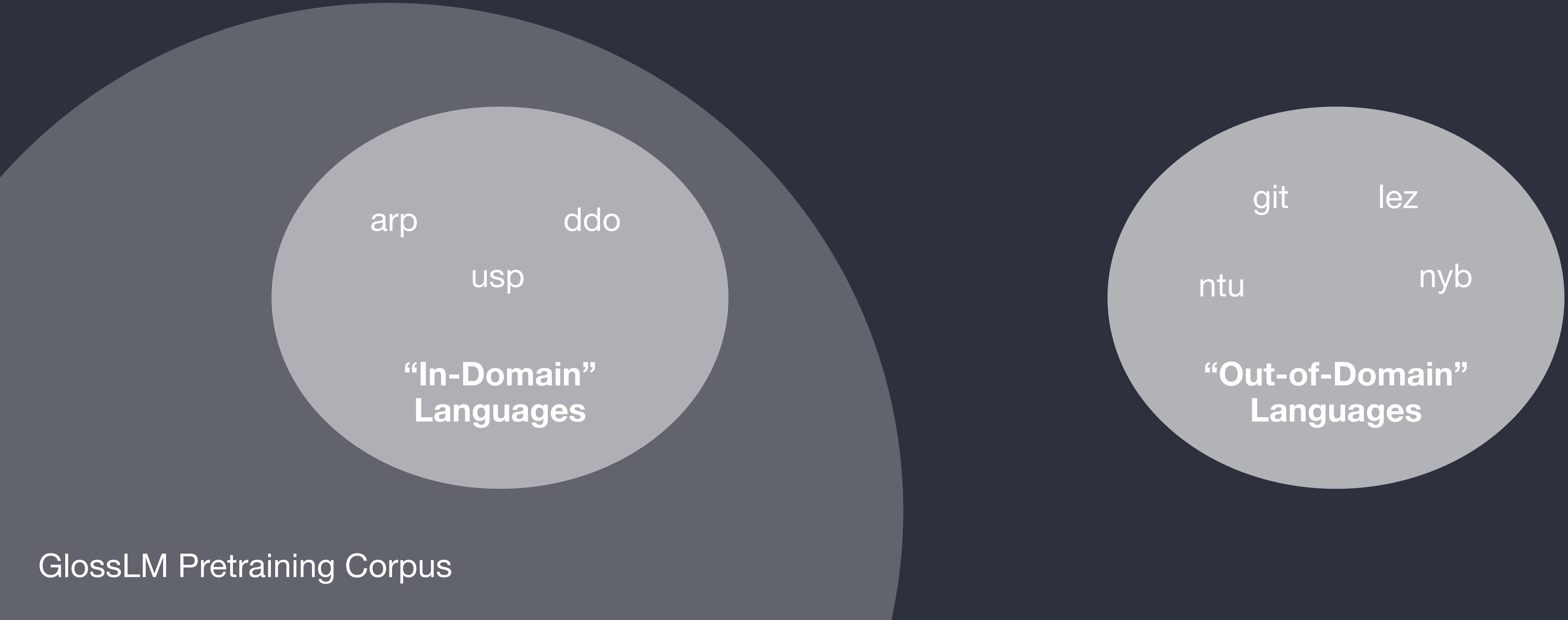
# GlossLM Corpus



# GlossLM Training

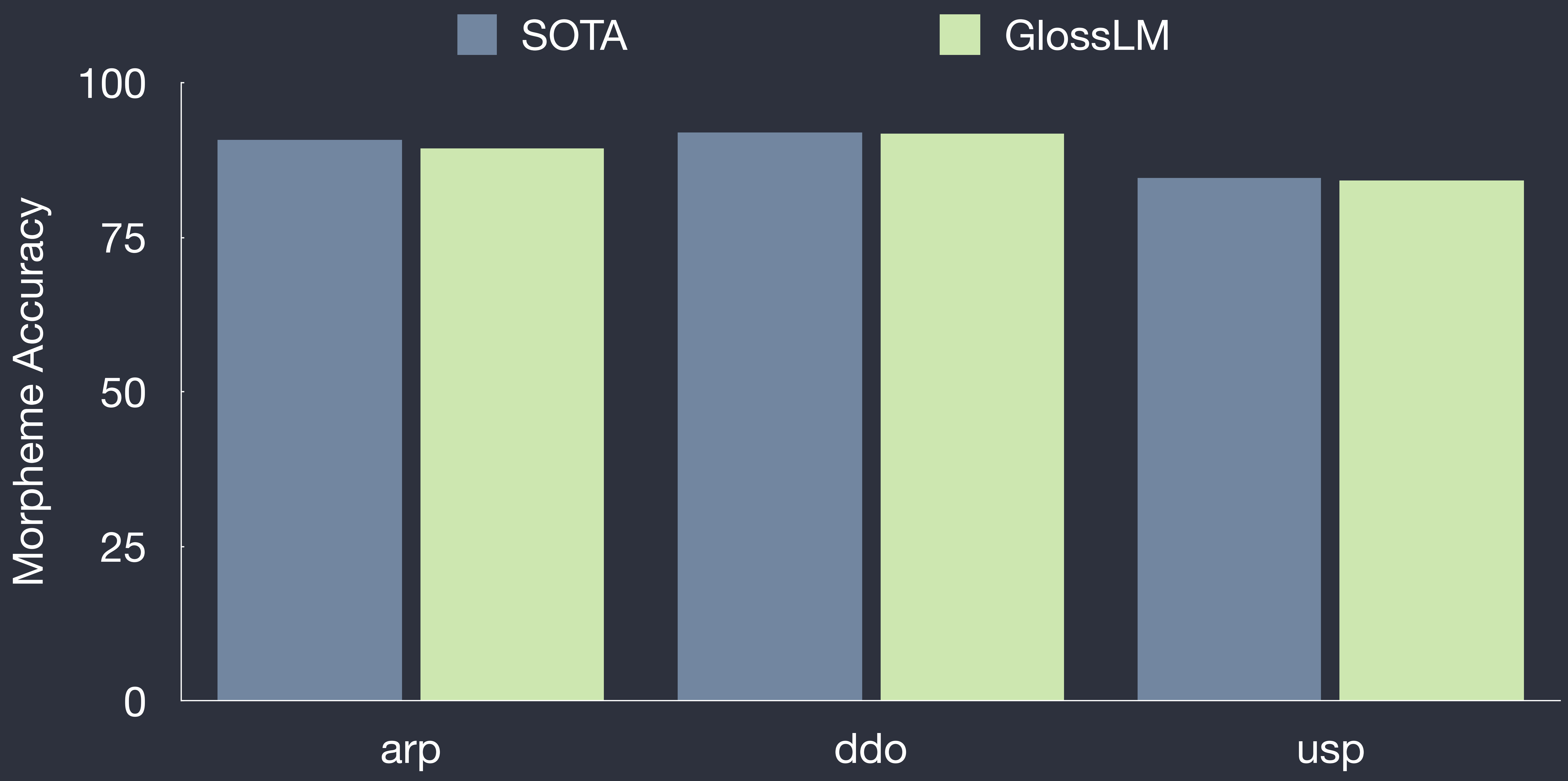


# Evaluation Languages



# How well does the pretrained model perform on seen languages?

* we focus on the unsegmented "closed-track" setting

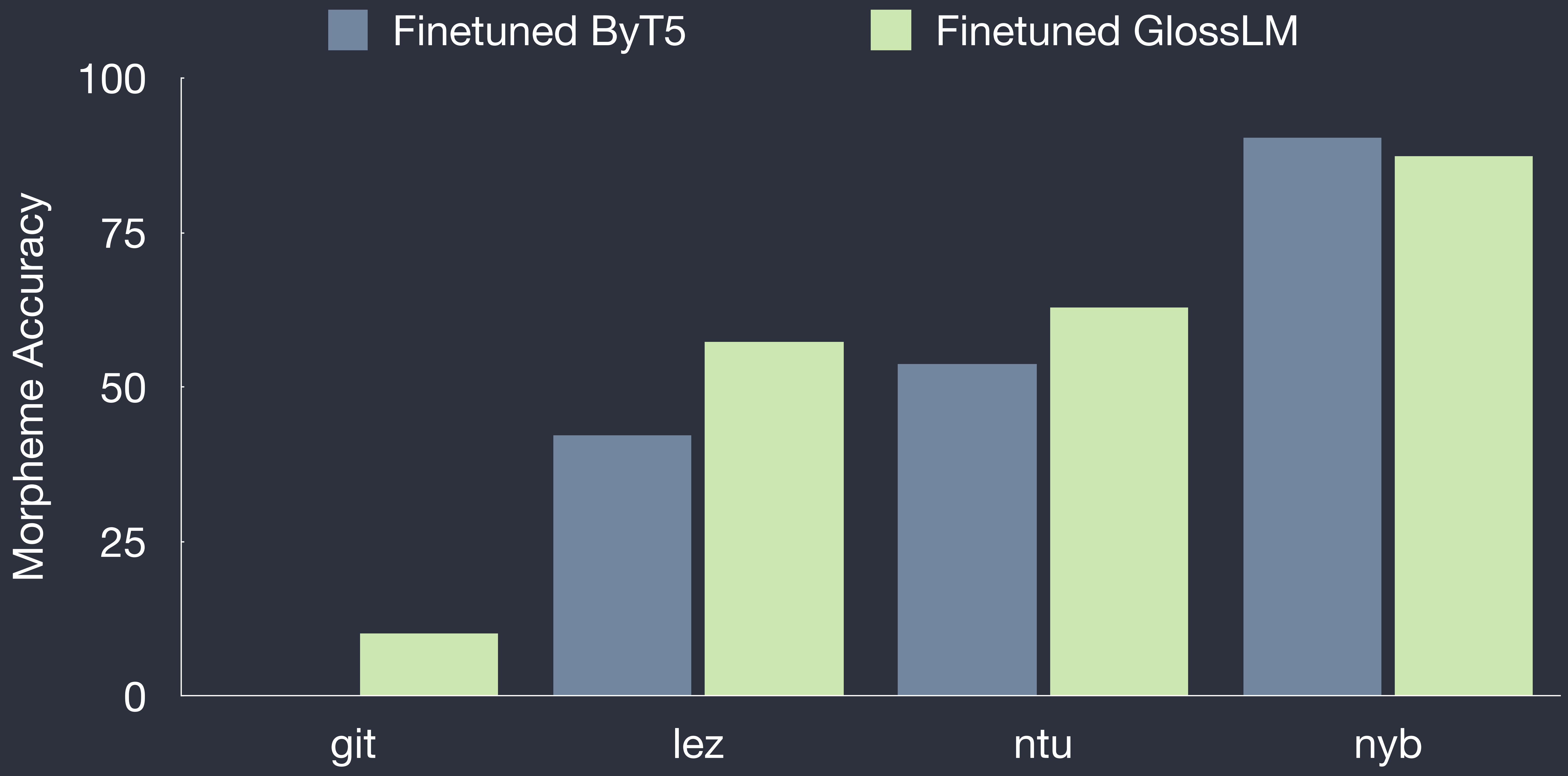


## How well does the pretrained model perform on seen languages?

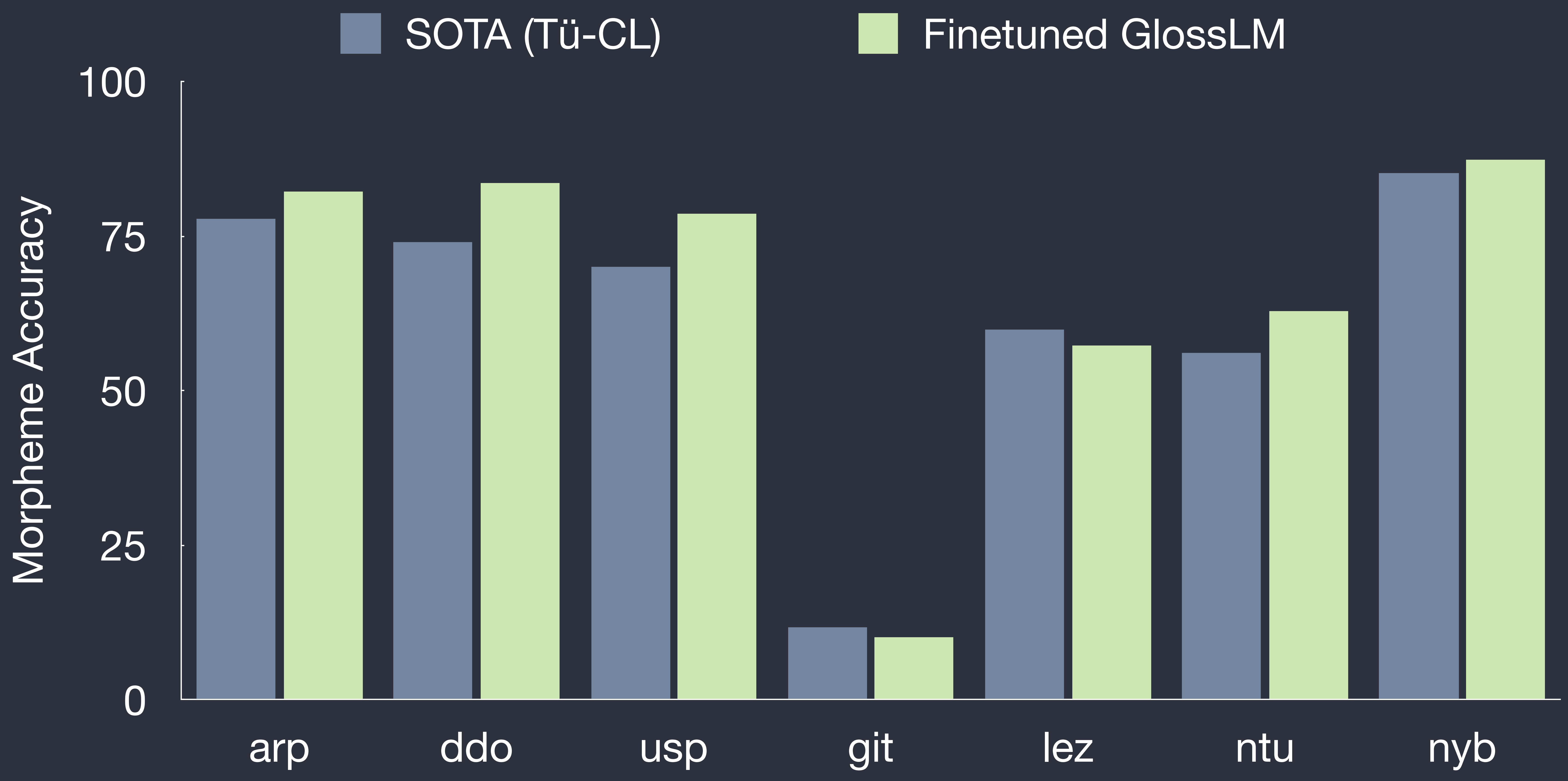
- Generally very close to SOTA
- Model does not seem to suffer from “curse of multilinguality”

- What about after finetuning?

# Does IGT pretraining help for finetuning models on new languages?



# How do **fine-tuned** GlossLM models compare to **SOTA**?



# Discussion

- Pretrained model is very competent
- Finetuned models are even better!

- Benefits from pretraining

Background

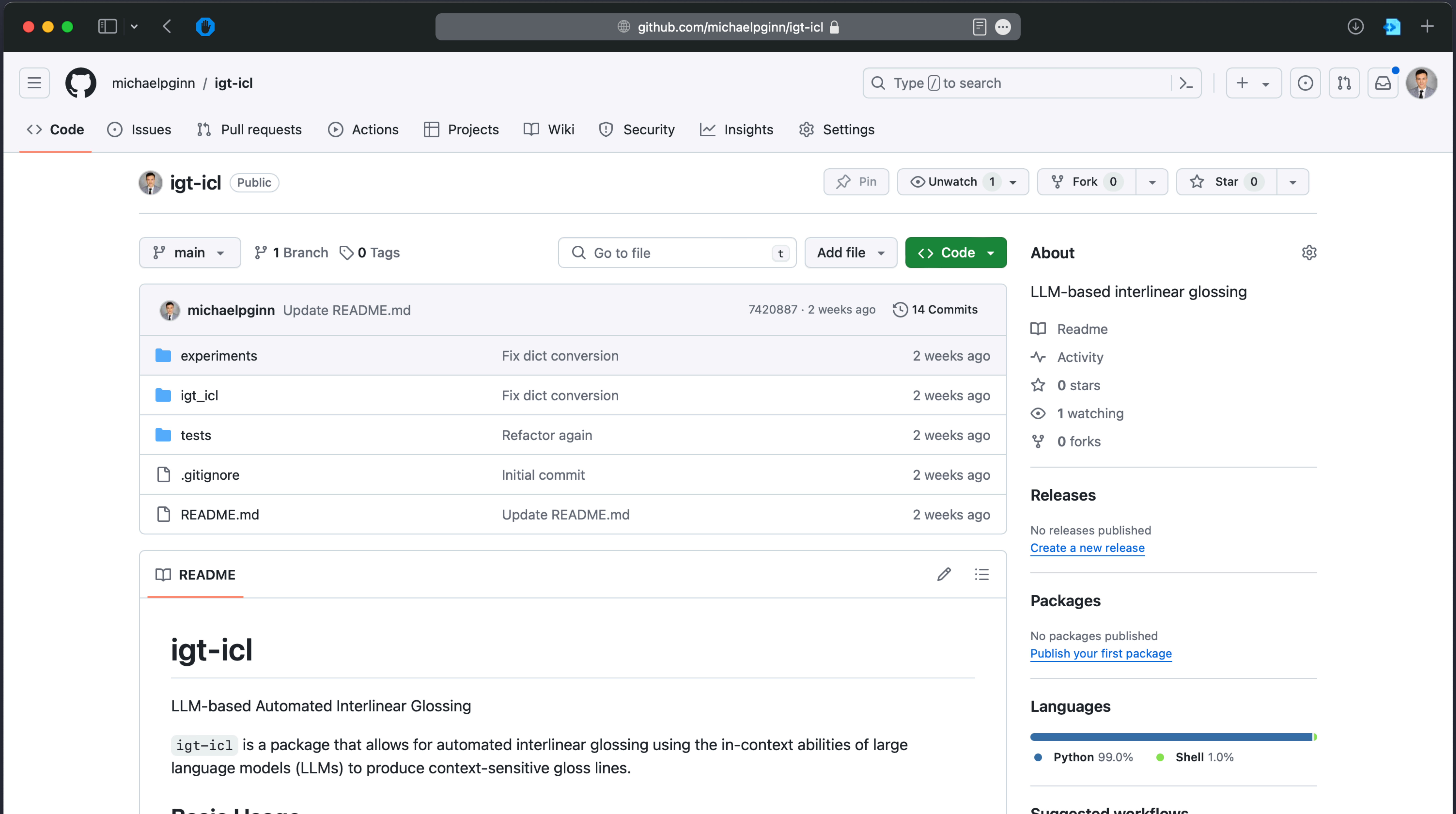
Shared Task

Robust Generalization

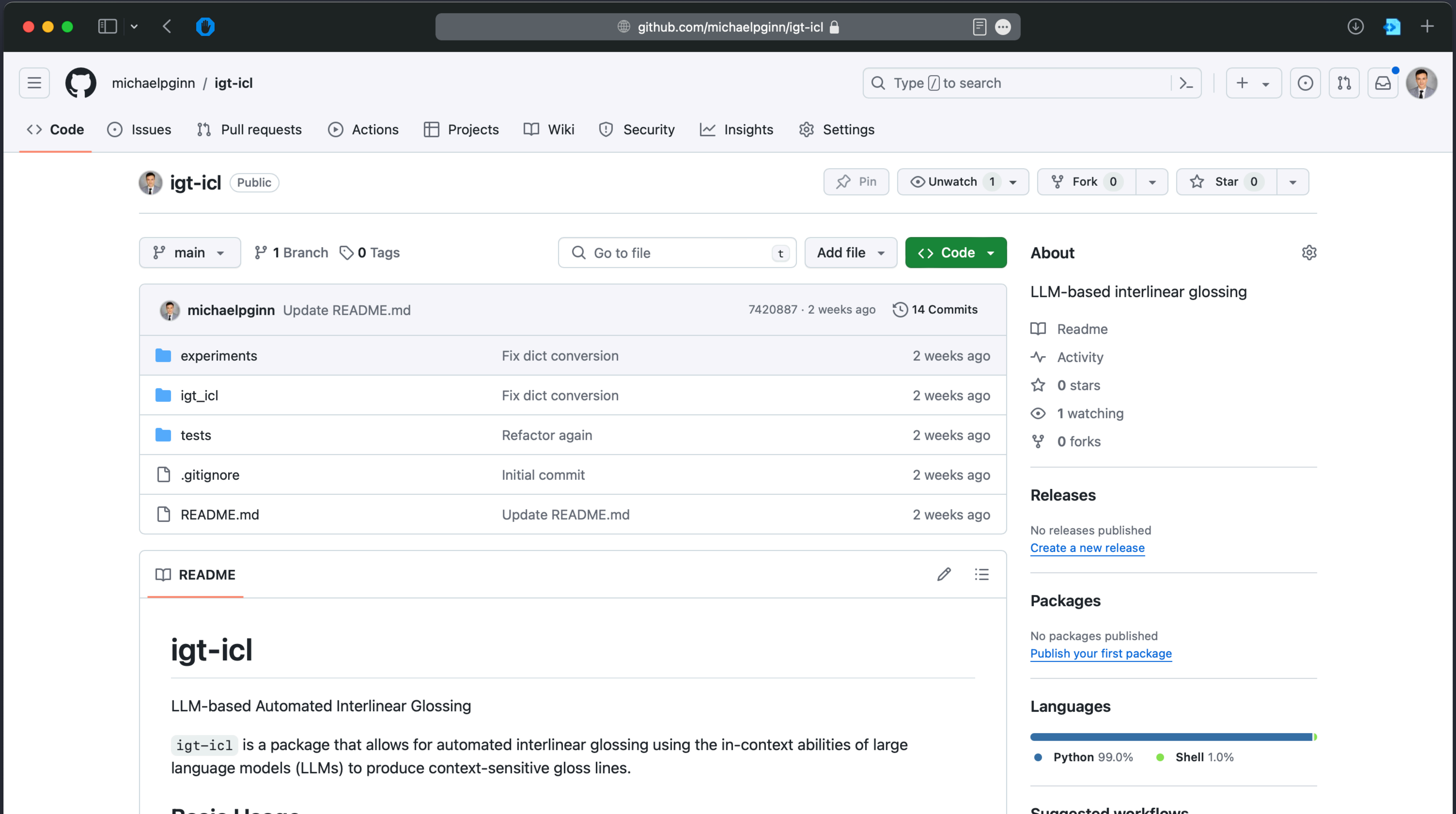
Multilingual Glossing

Future Work

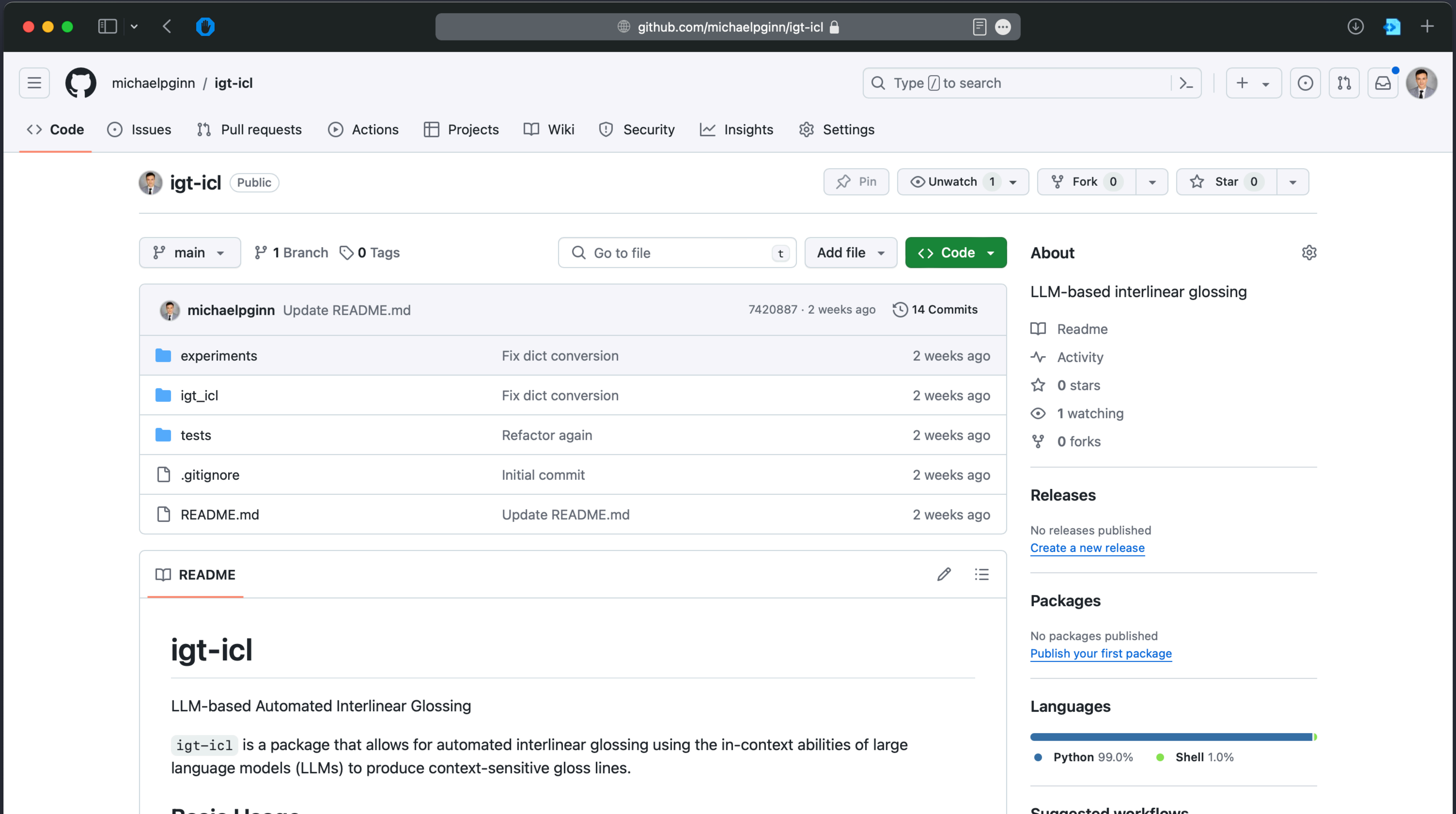
# Can LLM-based glossing systems be controllable?



# Can LLM-based glossing systems be cost-efficient?



# Can LLM-based glossing systems be cost-efficient?



# Summary

- Automated IGT Glossing models are becoming more capable with modern techniques
- IGT models must be robust to distributional shift for real-world usage
- IGT models can benefit from multilingual training

# Thank you!

This material is based upon work supported by the National Science Foundation under Grant No. 2149404, "CAREER: From One Language to Another". 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.