When is Multilinguality a Curse? Language Modeling for 350 Languages

Tyler Chang and Catherine Arnett

UC San Diego

EleutherAI

tachang@ucsd.edu

catherine@eleuther.ai

Collaborators

James Michaelov

Benjamin Bergen

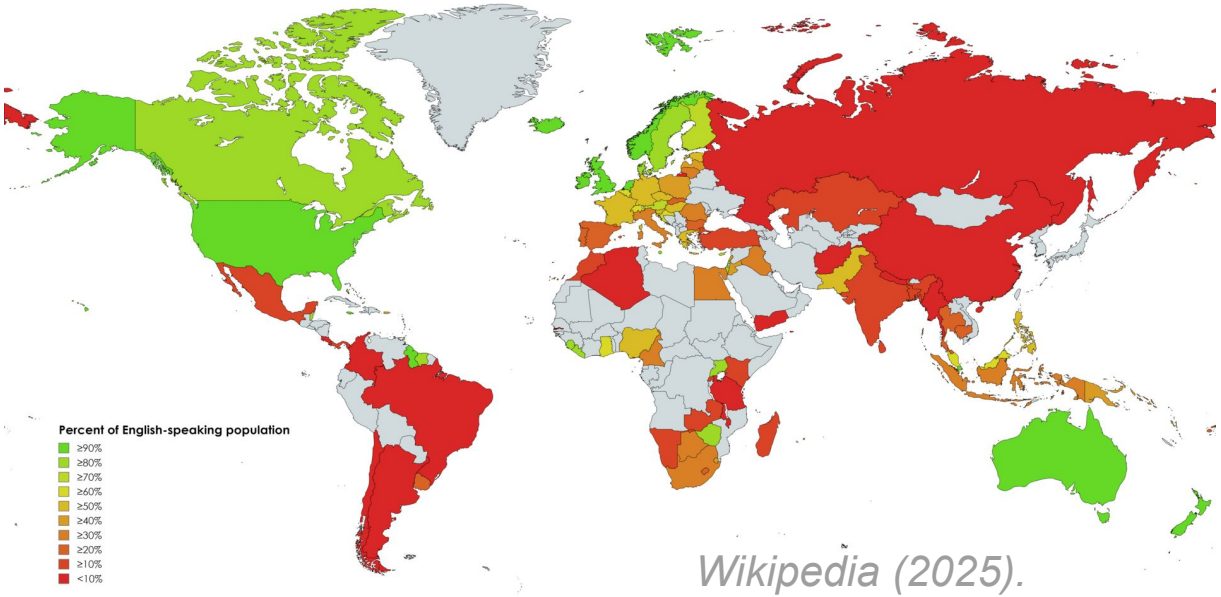
Zhuowen Tu

Massachusetts Institute of Technology

UC San Diego

English is only spoken as a native language (L1) by about 5% of the world's population, and as a second language by about 15%.

Ethnologue (2024)



Wikipedia (2025).

However, language models often focus primarily on English, or a small set of priority languages.

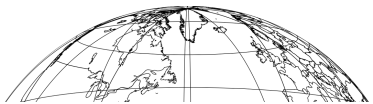
- GPT-3: **93%** English data. *Brown et al. (2020).*
- PaLM 1: **78%** English data. *Chowdhery et al. (2022).*
- Llama 3: **92%** English data. *Llama Team (2024).*

Even specifically multilingual models:

- XGLM: **77%** data in the top 10 languages. *Lin et al. (2022).*
- BLOOM: **96%** data in the top 10 languages. *Le Scao et al. (2022).*

Multilingual language models are the standard for NLP research in non-English languages.

E.g. Conneau et al. (2020), Lin et al. (2022), Le Scao (2022), Imani et al. (2023), Lin et al., (2024), Bandarkar et al. (2024).

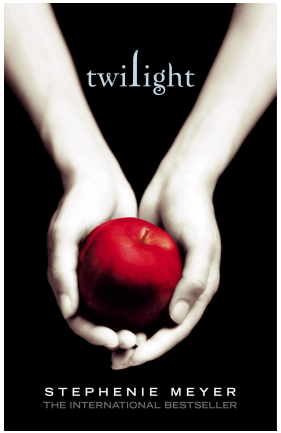


*** Dedicated models for low-resource languages are often not available. ***

However, multilingual language models are large and expensive to train.

Using small models:

- **More manipulations** of training data, languages, model size, etc.
- **More accessible** for researchers working with less compute and resources.
- For many low-resource languages, **data is the limiting factor**, not model size.
    - E.g. 5MB: a little more than the *Twilight* series.



However, multilingual language models are large and expensive to train.

- Smaller models are useful for proofs of concept and refining hypotheses. Promising results can be replicated at larger scales.

When Is Multilinguality a Curse? Language Modeling for 250 High- and Low-Resource Languages

Tyler A. Chang[a,c] **Catherine Arnett**[b] **Zhuowen Tu**[a] **Benjamin K. Bergen**[a]
[a]Department of Cognitive Science
[b]Department of Linguistics
[c]Halıcıoğlu Data Science Institute
University of California San Diego
{tachang,ccarnett,ztu,bkbergen}@ucsd.edu

Abstract

Multilingual language models are widely used to extend NLP systems to low-resource languages. However, concrete evidence for the effects of multilinguality on language modeling performance in individual languages remains scarce. Here, we pre-train over 10,000 monolingual and multilingual language models for over 250 languages, including multiple language families that are under-studied in NLP. We assess how language modeling performance in each language varies as a function of (1) monolingual dataset size, (2) added multilingual dataset size, (3) linguistic similarity of the added languages, and (4) model size (up to 45M parameters). We find that in moderation, adding multilingual data improves low-resource language modeling performance, similar to increasing low-resource dataset sizes by up to 33%. Improvements depend on the syntactic similarity of the added multilingual data, with marginal additional effects of vocabulary overlap. However, high-resource languages consistently perform worse in multilingual pre-training scenarios. As dataset sizes increase, adding multilingual data begins to hurt performance for both low-resource and high-resource languages, likely due to limited model capacity (the "curse of multilinguality"). These results suggest that massively multilingual pre-training may not be optimal for any languages involved, but that more targeted models can significantly improve performance.

et al., 2022; Imani et al., 2023), cross-lingual transfer learning (Pires et al., 2019; Conneau et al., 2020a), and multilingual text generation (Lin et al., 2022; Scao et al., 2022). However, while multilingual language models produce strong results across many languages, multilingual pre-training work almost exclusively focuses on pre-training a small number of models with some fixed distribution over languages (e.g. mBERT, XLM-R, XGLM, and BLOOM; Devlin et al., 2019; Conneau et al., 2020a; Blevins et al., 2022; Lin et al., 2022; Scao et al., 2022). This distribution over languages typically favors high-resource languages spoken in regions with high economic influence (Bender, 2011; Joshi et al., 2020).

Thus, it is largely unknown how different pre-training language distributions, such as different quantities of multilingual data or different selections of languages, affect multilingual language model performance in different languages. Multilingual models have been studied extensively during inference and fine-tuning (Pires et al., 2019; Conneau et al., 2020b; Karthikeyan et al., 2020; Winata et al., 2021; Chai et al., 2022; Alabi et al., 2022; Guarasci et al., 2022; Winata et al., 2022; Wu et al., 2022; Eronen et al., 2023), but these studies generally rely on the same sets of pre-trained models. For pre-training, there is mixed evidence for the benefits of multilingual vs. monolingual data (Conneau et al., 2020a; Wu and Dredze, 2020; Pyysalo et al., 2021; §2). As multilingual language

(EMNLP 2024)

Effects of multilingual data on individual language performance.

There is previous evidence for a **curse of multilinguality**, where training on too many languages hurts performance.

Multilinguality ❌

E.g. Conneau et al. (2020), Pyysalo et al. (2021).

However, multilingual data can **improve performance** for low-resource languages.

Multilinguality ✅

Wu & Dredze (2020), Adebara et al. (2023), Imani et al. (2023).

How do different training language distributions affect language modeling performance?

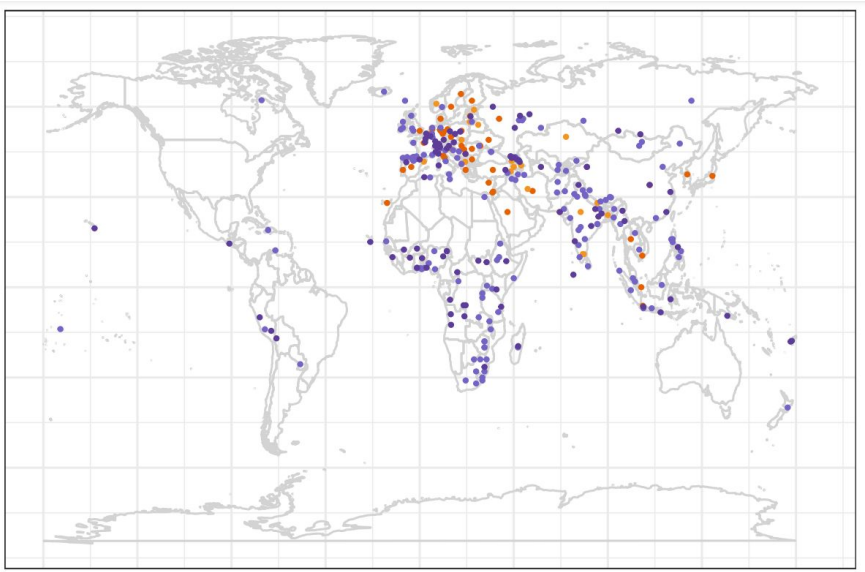
Run a controlled study with 10,000 small GPT2 models systematically varying:

- **Target language data** quantity: 1M, 10M, 100M, 1B tokens.
- **Added multilingual data** quantity: 10M, 100M, 1B added tokens.
- **Linguistic similarity** of added multilingual data:
  - Using a combined measure of lexical, syntactic, and geographic similarity, we select either the 10 most similar or 10 least similar languages to the target language.
- **Model size**: 8M (tiny), 19M (mini), and 45M (small) parameters.

Dataset

Combined 24 multilingual text datasets:

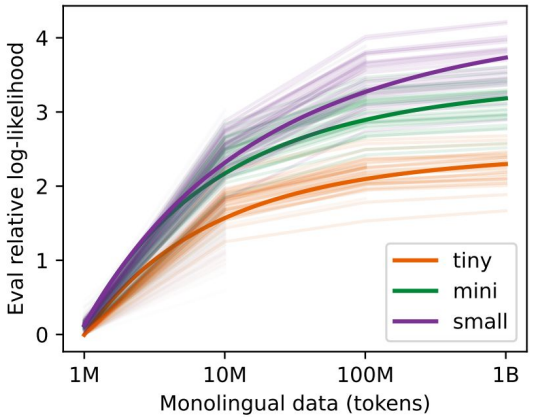
- **252 languages**, ranging from 1.5M to 1B tokens each.
- Languages from 5 continents, 29 language families, and 30 writing systems.

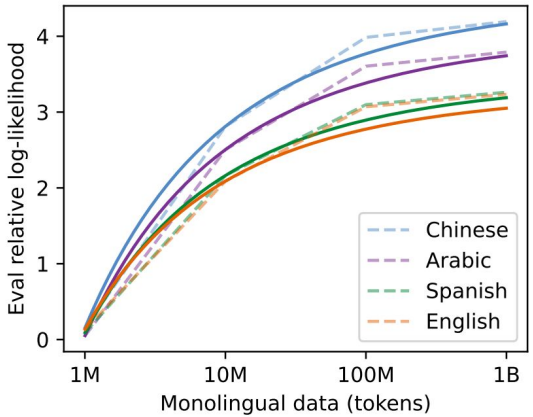


Quantifying model performance

To measure multilingual language model performance, we train monolingual baseline models, and we fit curves to **estimate the number of training tokens** based on model perplexity.

→ This serves as a measure of model performance in a language.





Recap:

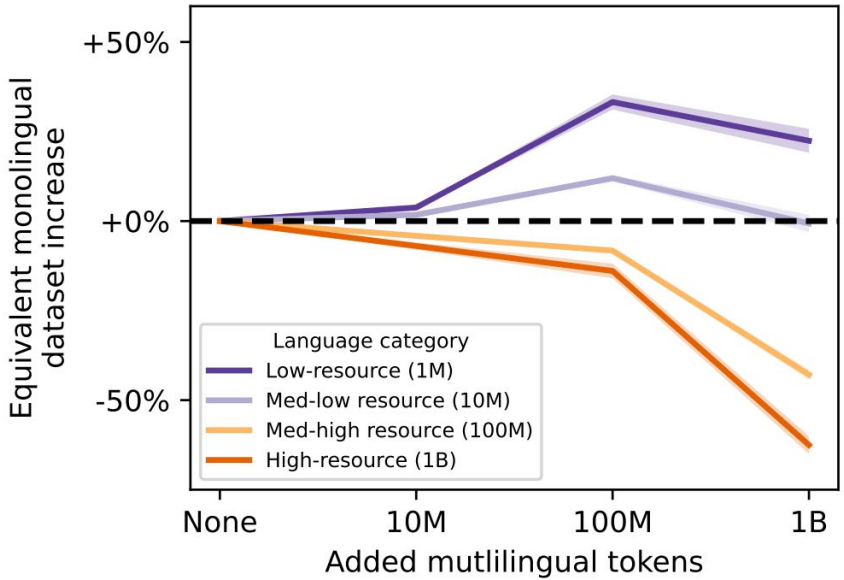
Run a controlled study with small GPT2 models systematically varying:

- **Target language data** quantity.
  - i.e. "low-resource" to "high-resource".
- **Added multilingual data** quantity.
- **Linguistic similarity** of added multilingual data.
- **Model size**: here, we focus on ~45M parameter models.

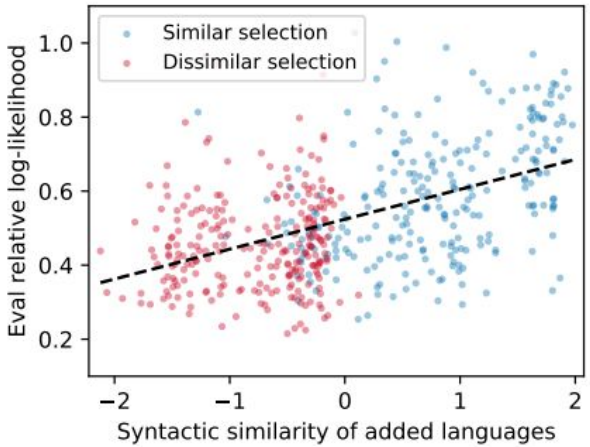
Measure performance using the **analogous monolingual tokens** it would take to achieve the same performance in the target language.

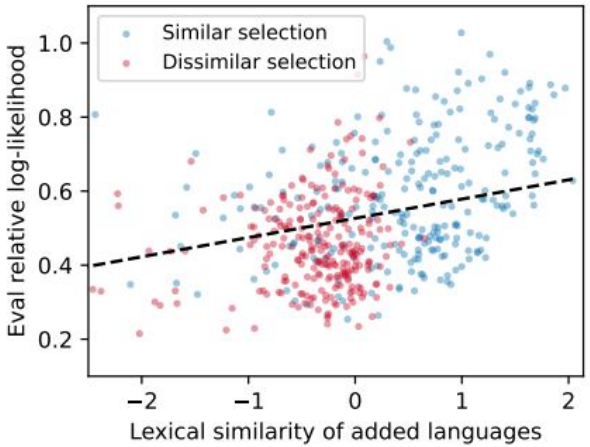
Multilingual data improves low-resource language modeling performance **analogous to increasing dataset sizes up to 33%**.

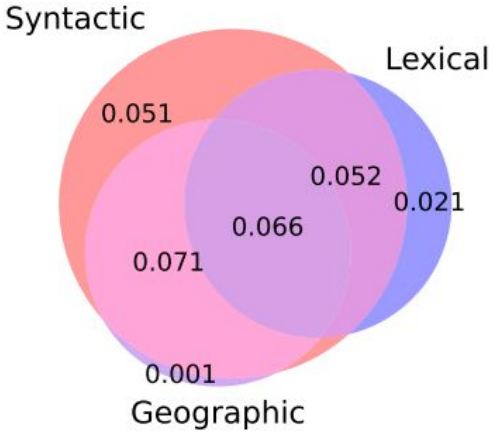
High-resource languages consistently perform worse in multilingual settings.



Improvements for low-resource languages depend on the **similarity of languages** in the added multilingual data.







Multilinguality is not universally good or bad.

- We should re-consider large, multilingual models as the default.
  - Sometimes monolingual models perform better.
- Monolingual baselines
  - Monolingual models do not exist for many languages.
  - But we have the tools to build them!

Goldfish: Monolingual Language Models for 350 Languages

Tyler A. Chang[1,2], **Catherine Arnett**[3], **Zhuowen Tu**[1], **Benjamin K. Bergen**[1]
[1]Department of Cognitive Science
[2]Halıcıoğlu Data Science Institute
[3]Department of Linguistics
University of California San Diego
{tachang, ccarnett, ztu, bkbergen}@ucsd.edu

Abstract

For many low-resource languages, the only available language models are large multilingual models trained on many languages simultaneously. However, using FLORES perplexity as a metric, we find that these models perform worse than bigrams for many languages (e.g. 24% of languages in XGLM 4.5B; 43% in BLOOM 7.1B). To facilitate research that focuses on low-resource languages, we pre-train and release Goldfish, a suite of monolingual autoregressive Transformer language models up to 125M parameters for 350 languages. The Goldfish reach lower FLORES perplexities than BLOOM, XGLM, and MaLA-500 on 98 of 204 FLORES languages, despite each Goldfish model being over 10× smaller. However, the Goldfish significantly underperform larger multilingual models on reasoning benchmarks, suggesting that for low-resource languages, multilinguality primarily improves general reasoning abilities rather than basic text generation. We release models trained on 5MB (350 languages), 10MB (288 languages), 100MB (166 languages), and 1GB (83 languages) of text data where available. The Goldfish models are available as baselines, fine-tuning sources, or augmentations to existing models in low-resource NLP research, and they are further useful for crosslinguistic studies requiring maximally comparable models across languages.

it contributes to model under-performance in low-resource languages (Wu and Dredze, 2020; Blasi et al., 2022). These barriers to research in low-resource languages are likely to exacerbate existing inequities across language communities in NLP research (Bender, 2011; Joshi et al., 2020).

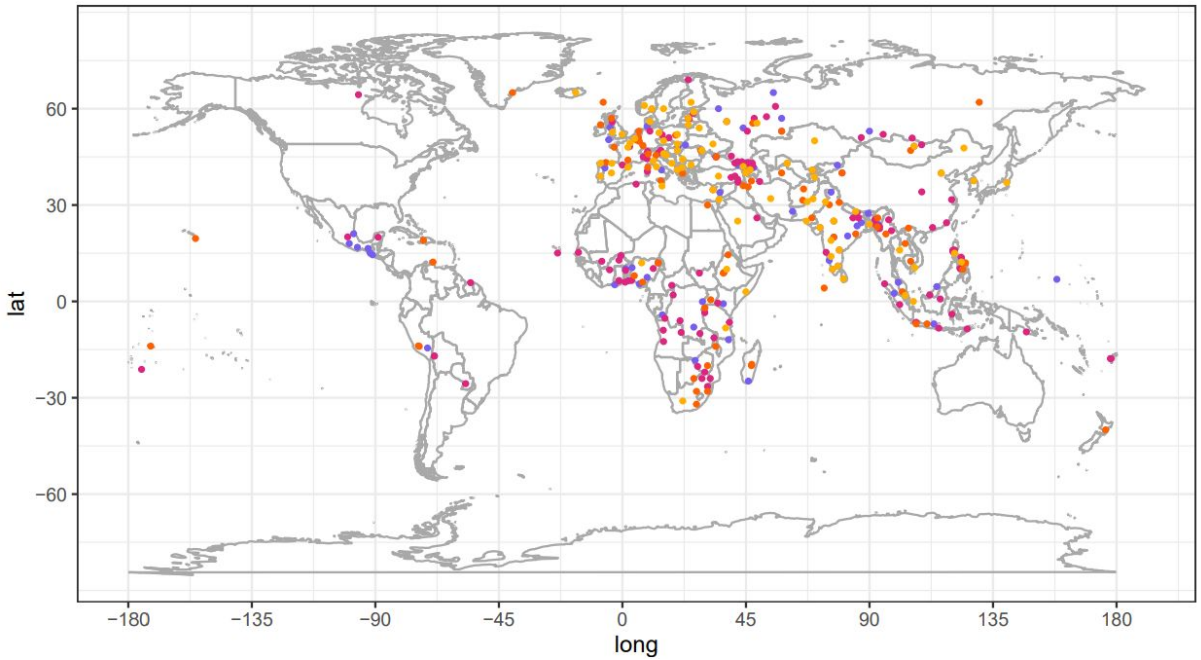
To address this lack of available models, we introduce Goldfish, a suite of over 1000 monolingual language models for 350 diverse languages.[1] The models reach lower perplexities than XGLM (Lin et al., 2022), BLOOM 7.1B (Scao et al., 2022), and MaLA-500 (Lin et al., 2024) on 98 out of 204 FLORES languages, despite each Goldfish model being over 10× smaller. The Goldfish also outperform simple bigram models, which are surprisingly competitive with larger models for low-resource languages (e.g. lower perplexities than BLOOM 7.1B on 43% of its languages; §4). However, despite better perplexities, the Goldfish underperform larger multilingual models on reasoning benchmarks, suggesting that multilingual pre-training may benefit abstract reasoning capabilities over more basic grammatical text generation (§5).

Finally, to enable comparisons across languages, we release monolingual models trained on comparable dataset sizes for all languages: 5MB, 10MB, 100MB, and 1GB when available, after accounting for the fact that languages require different numbers of UTF-8 bytes to encode comparable con-

(preprint)

Goldfish

1154 monolingual language models trained comparably for 350 languages.



Data size
5MB
10MB
100MB
1GB

Goldfish

Data: 24 sources from before + Glot500 + MADLAD-400 + NLLB.

- Deduplicated, ISO 639-3 language codes, ISO 15924 script codes.
- 5MB, 10MB, 100MB, 1000MB per language, after accounting for byte premiums.

Language	Language (ISO 639-3)	Script (ISO 15924)	Byte Premium	Scaled MB	Tokens	Dataset Proportions (OSCAR, NLLB, MADLAD-400, Glot500, Other)
Afrikaans	afr	latn	1.04	1000.00	239682048	
Amharic	amh	ethi	1.72	1000.00	211767808	
Standard Arabic	arb	arab	1.47	1000.00	196197376	
Azerbaijani	aze	latn	1.30	1000.00	233091584	
Belarusian	bel	cyrl	2.01	1000.00	254138368	
Bengali	ben	beng	2.43	1000.00	194737152	

...

Goldfish

Architecture: decoder-only GPT2, 125M parameters each.

Monolingual tokenizers: 50K vocab size each.

350 languages covering five continents, 28 top-level language families, and 32 scripts (writing systems).

Goldfish

Data size	Model output	
5MB	“Goldfish are a few years of the most of the most of the most...”	x350 languages
10MB	“Goldfish are a great way to the best way to the best way...”	x288 languages
100MB	“Goldfish are a great way to get your fish in the wild.”	x166 languages
1GB	“Goldfish are a species of fish that are found in the sea.”	x83 languages

Sequence-level FLORES perplexity eval

	Bigrams	XGLM 4.5B	XGLM 7.5B	BLOOM 7.1B	MaLA-500 10B
Bigrams		24 / 102	0 / 30	20 / 46	11 / 175
Goldfish (ours)	**202** / 202	**60** / 102	2 / 30	**32** / 46	**111** / 175

Win rates

Note: evaluated with perplexity, XGLM 4.5B and BLOOM 7.1B perform **worse than bigrams** on 24% and 43% of languages respectively.

Goldfish outperform XGLM 4.5B, BLOOM 7.1B, and MaLA-500 10B on **over 50% of languages.**

MultiBLiMP (Jumelet et al., 2025)

- Subject-verb agreement task for 101 languages.
- Goldfish has higher performance across languages, especially for low-resource languages.

- **Goldfish models outperform Llama3 8B for 39 out of 70 languages.**

			Resources			Languages				
Model	Size	Variant	Low	Mid	High	GF	Aya	EU	Eng	All
Llama3	8B	base	77.2	90.3	96.2	89.4	95.2	92.7	99.4	86.9
	8B	it	75.7	88.9	95.6	88.3	94.6	91.4	99.4	85.6
	8B	tülu3	74.7	88.1	95.3	87.6	94.0	90.9	98.6	84.9
	70B	base	**81.1**	93.9	97.8	92.6	97.1	95.5	99.0	**90.2**
Aya	8B	it	68.3	82.3	96.3	83.3	95.7	88.0	99.0	80.4
	32B	it	75.7	89.4	97.7	89.0	97.3	92.8	98.4	86.4
Gemma3	4B	base	71.1	93.0	96.4	89.2	95.3	94.6	98.2	85.8
	4B	it	63.9	82.7	88.2	80.0	86.6	85.1	93.4	77.1
	12B	base	75.8	95.4	97.7	92.0	96.8	96.6	99.0	88.8
	12B	it	69.5	89.4	92.6	85.9	91.4	90.9	95.6	82.9
	27B	base	78.3	**96.3**	**98.0**	93.2	**97.4**	**97.1**	98.6	**90.2**
	27B	it	73.1	92.3	94.4	88.9	93.7	93.3	96.0	85.8
OLMo2	32B	base	74.4	84.6	92.5	85.1	90.8	87.8	**99.5**	82.7
	32B	it	72.5	83.6	91.9	84.0	90.0	86.9	99.1	81.5
EuroLLM	9B	base	72.6	92.0	95.7	88.9	94.9	96.7	99.4	85.8
GPT2	1.6B	xl	62.8	61.2	68.4	63.5	66.0	64.0	98.3	63.4
Goldfish	500M	base	88.0	95.6	95.9	**93.8**	95.2	95.8	96.4	93.8

But reasoning performance is poor.

	# Langs	Chance	Goldfish	XGLM 4.5B	XGLM 7.5B	BLOOM 7.1B	MaLA-500 10B
Belebele	121	25.0	28.2	30.1	**30.6**	30.2	**30.6**
XCOPA	11	50.0	54.9	57.9	**60.6**	56.9	55.6
XStoryCloze	10	50.0	52.5	57.1	**59.9**	58.2	55.7

Zero-shot, select highest probability completion.

The large multilingual models may be better at leveraging **cross-lingual transfer for reasoning capabilities**, but results are still close to chance.

- Multilinguality may benefit reasoning capabilities, but monolinguality is better for grammatical text generation.

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Posts | Docs | Enterprise

Gold Fish

goldfish-models

Follow

14 followers · 0 following

AI & ML interests

multilingual NLP, low-resource languages

Organizations

Collections 4

1000MB Goldfish
Goldfish trained on 1000MB of data after byte premium scaling.

- goldfish-models/eng_latn_1000mb — Text Generation • Updated Aug 26, 2024 • 235 • 1
- goldfish-models/afr_latn_1000mb — Text Generation • Updated Aug 26, 2024 • 6
- goldfish-models/amh_ethi_1000mb — Text Generation • Updated Aug 26, 2024 • 7
- goldfish-models/arb_arab_1000mb

100MB Goldfish
Goldfish trained on 100MB of data after byte premium scaling

- goldfish-models/afr_latn_100mb — Text Generation • Updated Aug 26, 2024 • 4
- goldfish-models/als_latn_100mb — Text Generation • Updated Aug 26, 2024 • 7
- goldfish-models/amh_ethi_100mb — Text Generation • Updated Aug 26, 2024 • 8
- goldfish-models/arb_arab_100mb

Expand 4 collections

Models 1154

Sort: Recently updated

- goldfish-models/zho_hans_1000mb — Text Generation • Updated Aug 26, 2024 • 5
- goldfish-models/uzb_latn_1000mb — Text Generation • Updated Aug 26, 2024 • 7
- goldfish-models/vie_latn_1000mb — Text Generation • Updated Aug 26, 2024 • 162
- goldfish-models/ukr_cyrl_1000mb — Text Generation • Updated Aug 26, 2024 • 14
- goldfish-models/urd_arab_1000mb — Text Generation • Updated Aug 26, 2024 • 9
- goldfish-models/tgl_latn_1000mb — Text Generation • Updated Aug 26, 2024 • 6

What are these small monolingual models good for?

- Multilingual models have different performance for different languages
- Why?
  - Different amounts of training data
  - Languages are differentially difficult for language models to learn
    → **morphological type**

Why do language models perform worse for morphologically complex languages?

Catherine Arnett
Department of Linguistics
University of California San Diego
ccarnett@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
University of California San Diego
bkbergen@ucsd.edu

Abstract

Language models perform differently across languages. It has been previously suggested that morphological typology may explain some of this variability (Cotterell et al., 2018). We replicate previous analyses and find additional new evidence for a performance gap between agglutinative and fusional languages, where fusional languages, such as English, tend to have better language modeling performance than morphologically more complex languages like Turkish. We then propose and test three possible causes for this performance gap: morphological alignment of tokenizers, tokenization quality, and disparities in dataset sizes and measurement. To test the morphological alignment hypothesis, we present MorphScore, a tokenizer evaluation metric, and supporting datasets for 22 languages. We find some evidence that tokenization quality explains the performance gap, but none for the role of morphological alignment. Instead we find that the performance gap is most reduced when training datasets are of equivalent size across language types, but only when scaled according to the so-called "byte-premium"—the different encoding efficiencies of different languages and orthographies. These results suggest that languages of particular morphological types are not intrinsically advantaged or disadvantaged in language modeling. Differences in performance can be attributed to disparities in dataset size. These findings bear on ongoing efforts to improve performance for low-performing and under-resourced languages.

small number of high-resource languages remains extremely poor (Joshi et al., 2020; Ranathunga and de Silva, 2022; Søgaard, 2022; Atari et al., 2023; Ramesh et al., 2023). This has been attributed to a lack of research on non-English languages (Blasi et al., 2022), a lack of training data, and the possibility that evaluations are skewed towards high-resource languages (Choudhury, 2023).

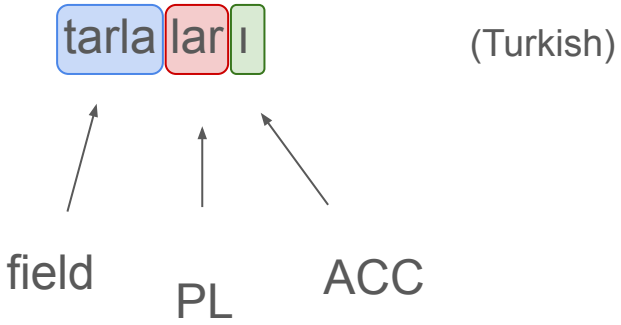
Beyond these systemic biases, it's also possible that certain linguistic features lead to higher or lower language modeling performance. Specifically, it has been proposed that languages with more complex morphology are harder to model (Cotterell et al., 2018; Park et al., 2021). Languages with more inflectional classes are morphologically more complex, and thus harder to predict. This can be described in terms of enumerative complexity (Ackerman and Malouf, 2013).

Greater morphological complexity may lead to worse language model performance, as morphologically rich languages tend to have a large number of very infrequent word forms produced by combinations of morphemes, which leads to data sparsity (Shin and You, 2009; Bender, 2011; Botev et al., 2022). This claim finds empirical support in Gerz et al. (2018a), who demonstrated over a sample of 50 languages that morphologically rich (agglutinative) languages performed worse than less morphologically rich (fusional) languages. In the current work (§3), we replicate this analysis and extend it to much larger transformer models, both in monolingual and multilingual settings. We, too, find a

(COLING 2025)

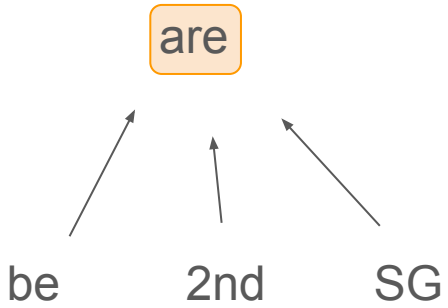
Morphological Type

- Agglutinative languages
  - use different morphemes to represent each feature



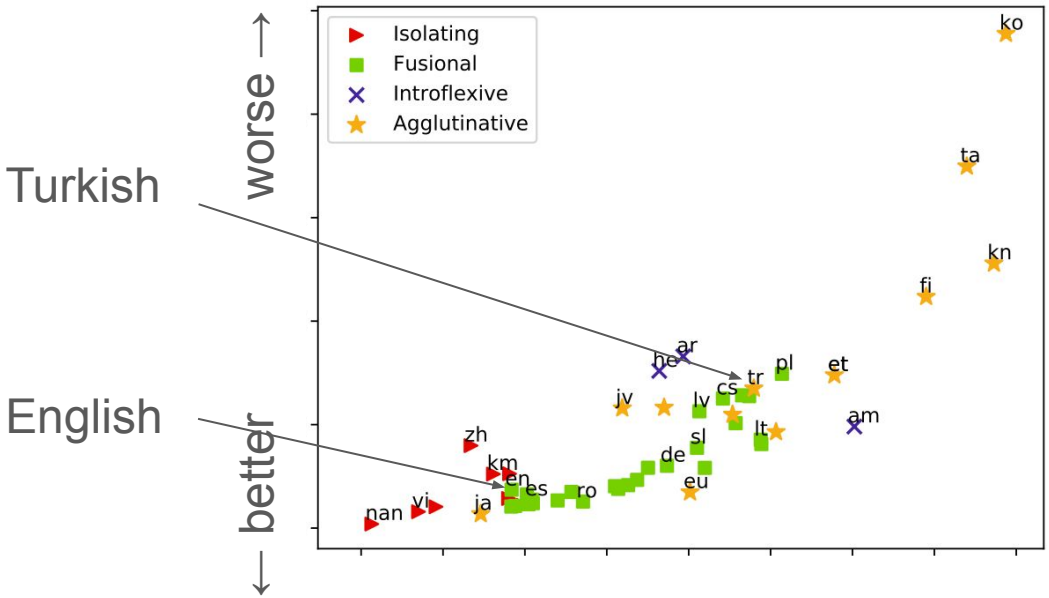
(Turkish)

- Fusional languages
  - multiple morpho-syntactic features in a single morpheme



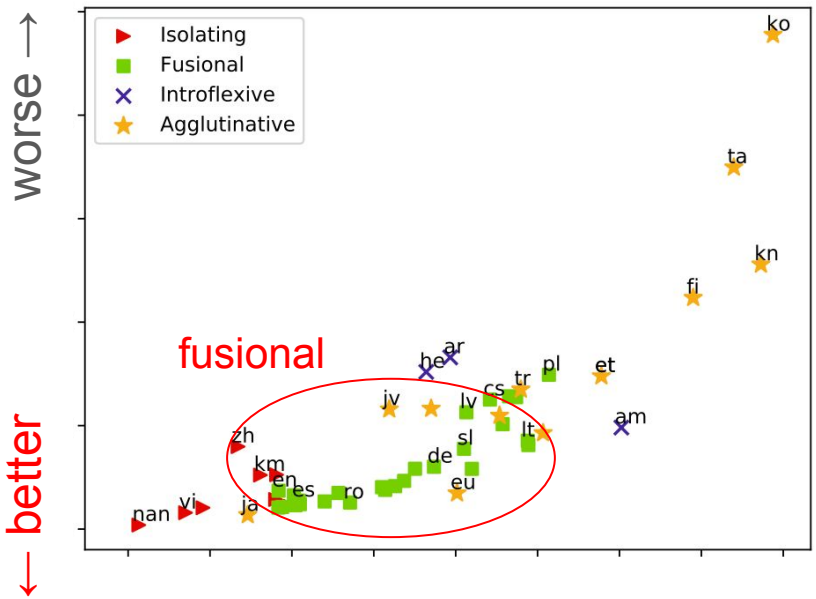
(English)

Correlation between Typology and Performance



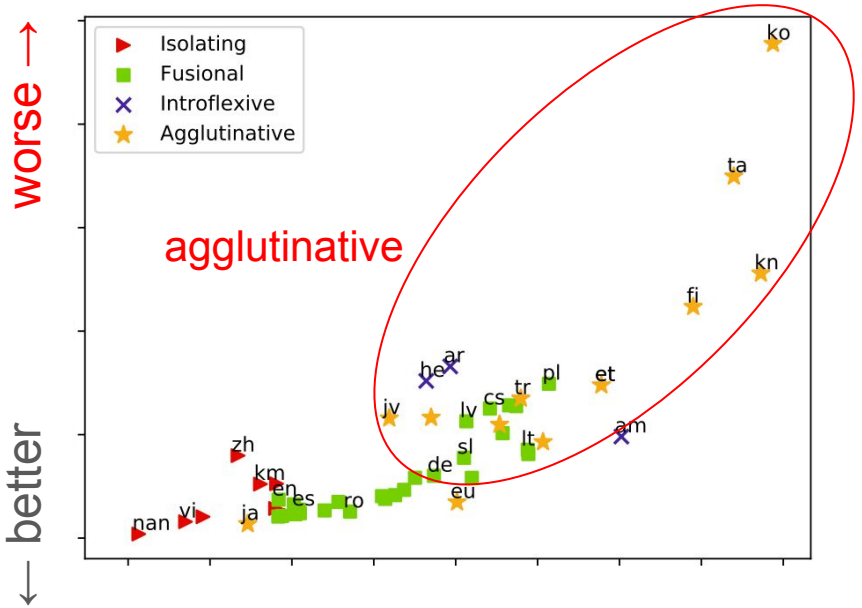
Gerz et al. (2018)

Correlation between Typology and Performance



Gerz et al. (2018)

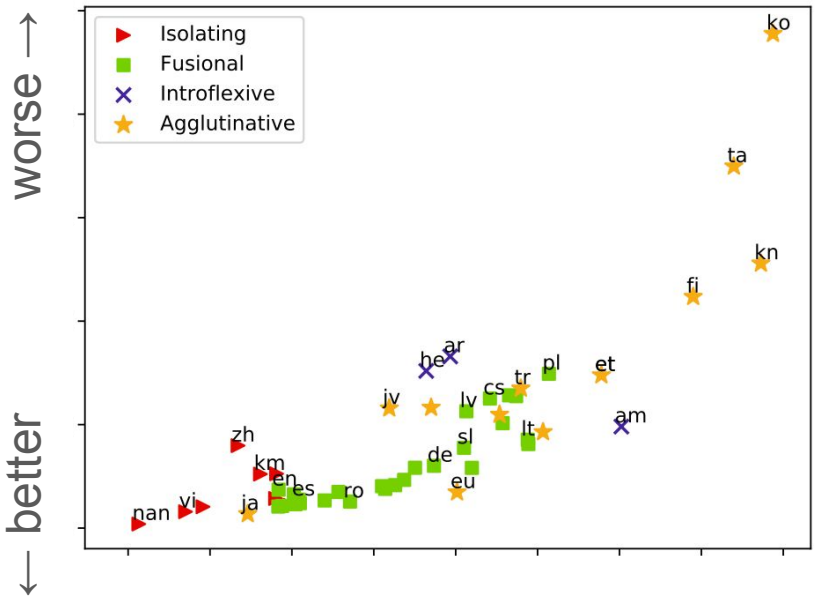
Correlation between Typology and Performance



Gerz et al. (2018)

Correlation between Typology and Performance

- Fusional languages have systematically better performance than agglutinative languages

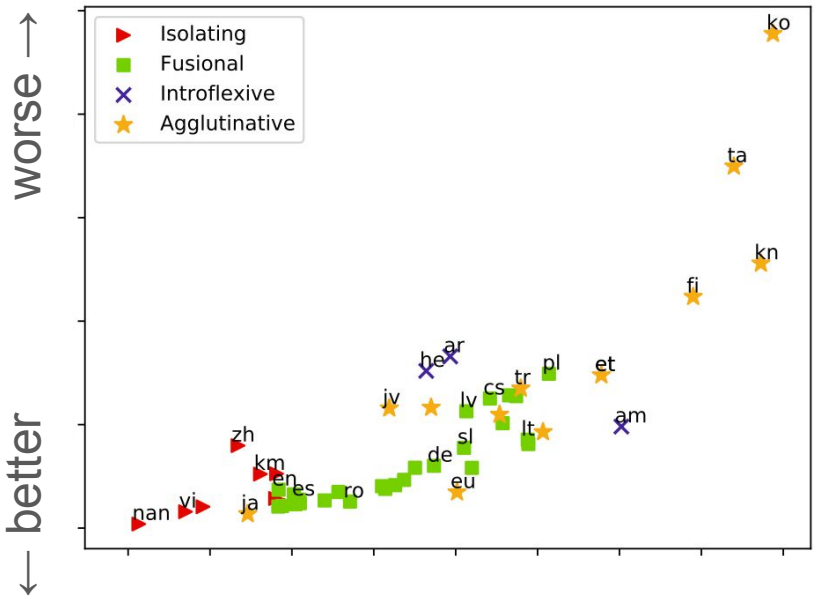


Gerz et al. (2018)

Correlation between Typology and Performance

Limitations:

- Did not control for number of training tokens
  - Even after re-running the analysis with training tokens taken as a predictor, there is still a significant effect of morphological type, where fusional languages show better performance than agglutinative languages.



Gerz et al. (2018)

Replicating Performance Gap

- Measure model performance
  - Models: XGLM, BLOOM, mT0, MaLA-500, and LLaMA 2
    (Lin et al., 2022; Scao et al., 2022; Muennighoff et al., 2023; Lin et al., 2024; Touvron et al., 2023)
  - Tasks: XStoryCloze, XCOPA, XNLI, Wikipedia, XWinograd, and SIB-200
    (Lin et al., 2022; Ponti et al., 2020; Conneau et al., 2018; Guo et al., 2020; Muennighoff et al., 2023; Adelani et al., 2024)
- Control for
  - Amount of training data
  - Language family
  - Model
  - Task

There is still a significant effect of morphological type

Testing Goldfish

- If these effects are explained by differences in training data
  - Goldfish should show no effect
- If morphological type affects learnability
  - Goldfish should show the same effects

Results

- Controlling for dataset size and taking into account byte premiums, Goldfish models **do not** show differences according to morphological typology
- Languages aren’t necessarily harder or easier to learn because of their grammar/structure
    - This is only possible to learn with comparable monolingual models

The Blessings of Multilinguality

- Multilingual models leverage crosslingual transfer for improved performance
  - Shared knowledge and representations across languages
- When and how does crosslingual transfer happen?

On the Acquisition of Shared Grammatical Representations in Bilingual Language Models

Catherine Arnett[a,b] **Tyler A. Chang**[c] **James A. Michaelov**[c,d,e] **Benjamin K. Bergen**[c]
[a]Department of Linguistics, UCSD [b]EleutherAI
[c]Department of Cognitive Science, UCSD
[d]Department of Brain and Cognitive Sciences, MIT [e]MIT Libraries CREOS
catherine@eleuther.ai, tachang@ucsd.edu, jamic@mit.edu, bkbergen@ucsd.edu

Abstract

Crosslingual transfer is crucial to contemporary language models' multilingual capabilities, but how it occurs is not well understood. We ask what happens to a monolingual language model when it begins to be trained on a second language. Specifically, we train small bilingual models for which we control the amount of data for each language and the order of language exposure. To find evidence of shared multilingual representations, we turn to structural priming, a method used to study grammatical representations in humans. We first replicate previous crosslingual structural priming results and find that after controlling for training data quantity and language exposure, there are asymmetrical effects across language pairs and directions. We argue that this asymmetry may shape hypotheses about human structural priming effects. We also find that structural priming effects are less robust for less similar language pairs, highlighting potential limitations of crosslingual transfer learning and shared representations for typologically diverse languages.

🤗 B-GPT models ✿ code and data

model would assign a higher probability to a prepositional object (PO) dative sentence (e.g. "*the chef gives a hat to the swimmer*") following another PO sentence than following a double object (DO) dative sentence (e.g. "*the chef gives the swimmer a hat*"; sentences from Schoonbaert et al., 2007). In crosslingual structural priming, targets that share a grammatical construction with the prime are more likely, even if the two sentences are in different languages (Figure 1).

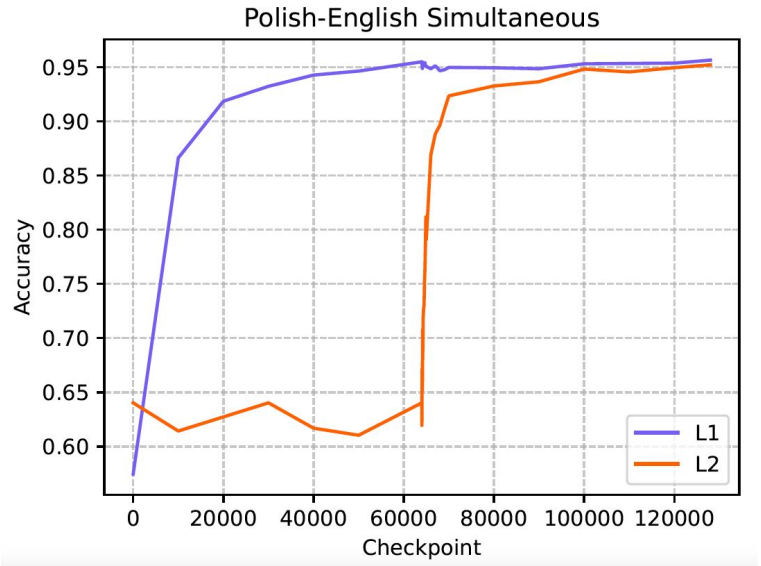
Human experiments demonstrate robust structural priming effects in a wide variety of languages (Bock, 1986; see Pickering and Ferreira (2008) for review) and have been used to argue that bilinguals have shared grammatical representations for their languages. Structural priming has previously been used to study the structural representations learned by language models (Prasad et al., 2019; Sinclair et al., 2022; Frank, 2021; Li et al., 2022; Choi and Park, 2022; Michaelov et al., 2023; Jumelet et al., 2024; Zhou et al., 2024). Because the grammatical structure is primed rather than a specific semantic meaning, Sinclair et al. (2022) argue that structural priming effects provide evidence for abstract

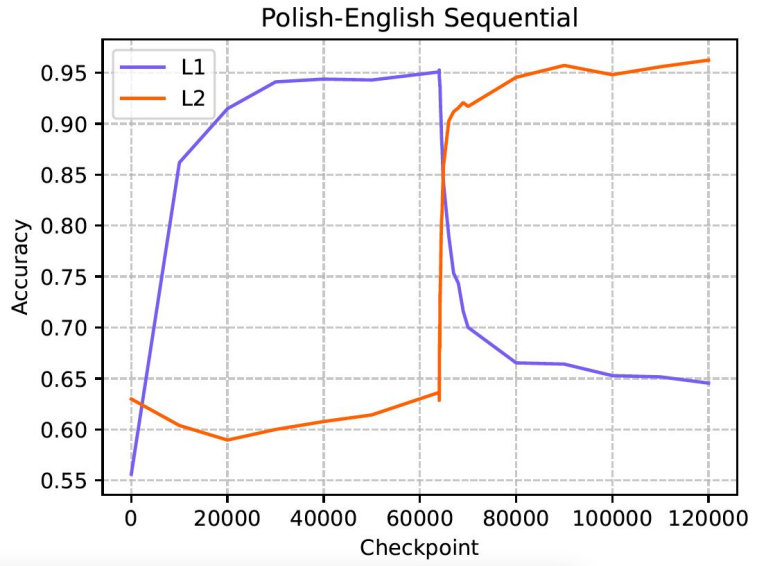
(ACL 2025)

B-GPT Models

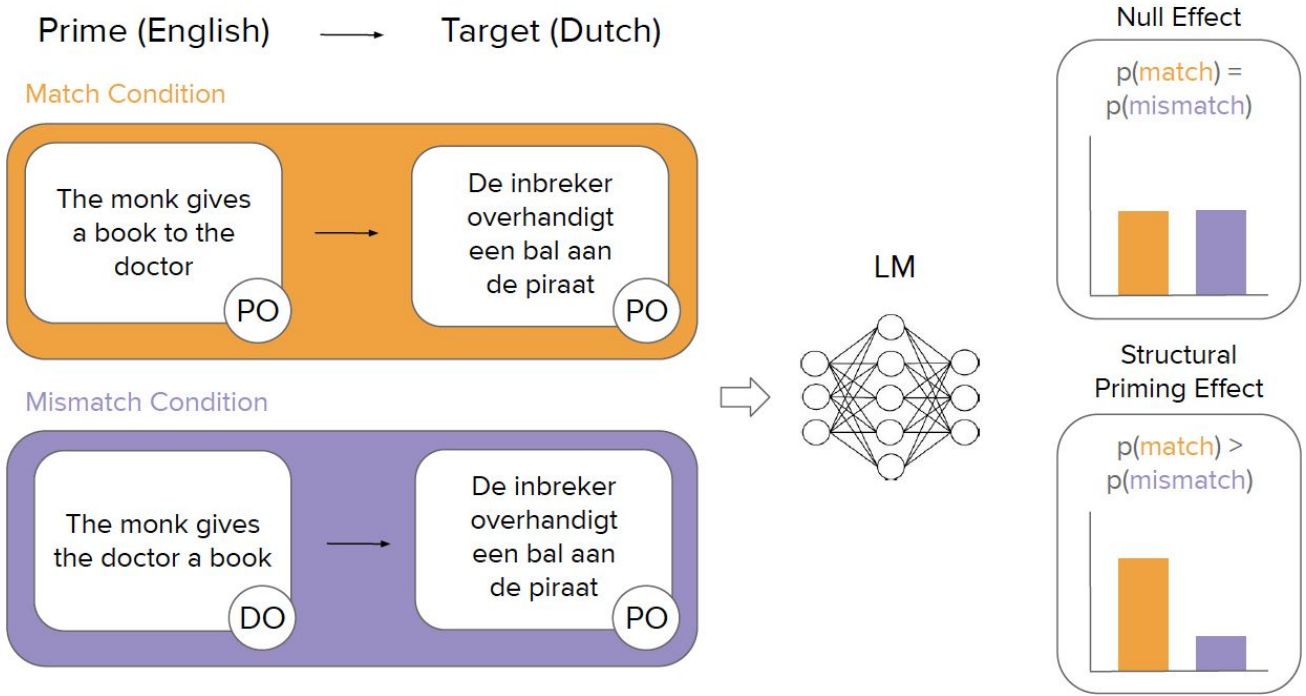
- GPT-2 base-size models (124M parameters)
  - Dutch, Spanish, Greek, and Polish
- Trained bilingually
  - First half of training: first language data only
  - Second half of training: exposed to second language
- Manipulate:
  - Order of language exposure:
    - English-Dutch vs. Dutch-English
  - Language mixing condition:
    - Simultaneous (50%-50% mix)
    - Sequential (100% L2)

MultiBLiMP

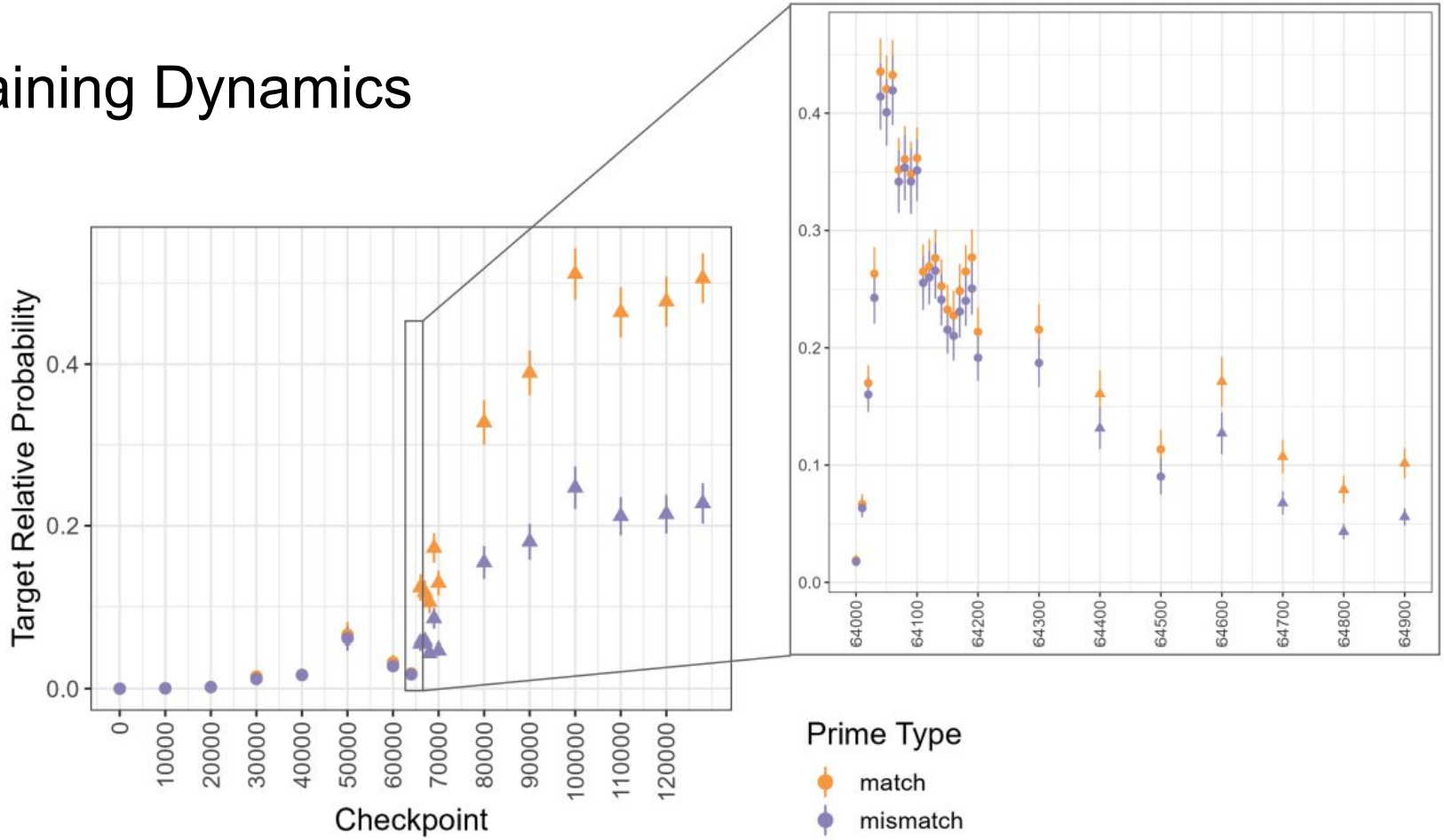

Polish-English Simultaneous
Accuracy
Checkpoint
L1
L2
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0
20000
40000
60000
80000
100000
120000


Polish-English Sequential
Accuracy
Checkpoint
L1
L2
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0
20000
40000
60000
80000
100000
120000

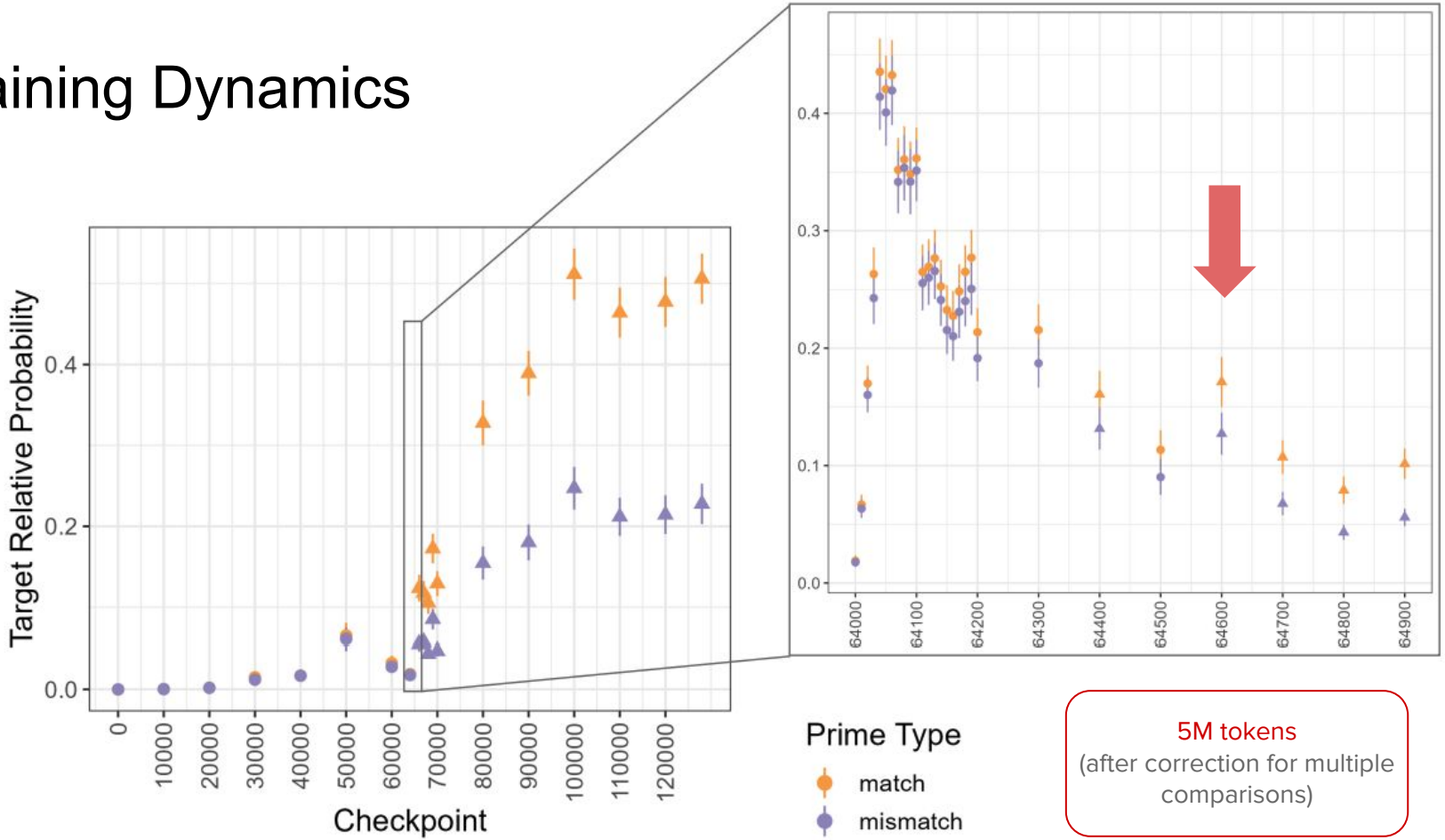
Crosslingual Structural Priming



Training Dynamics

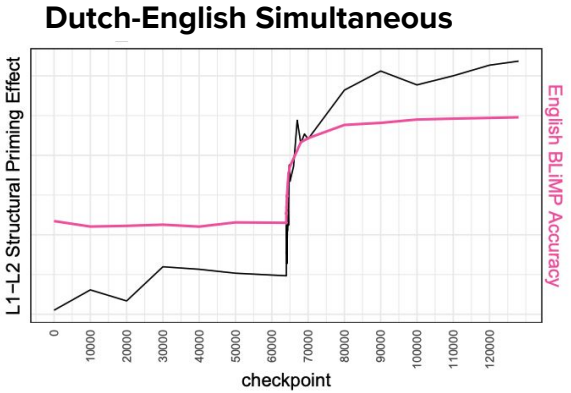


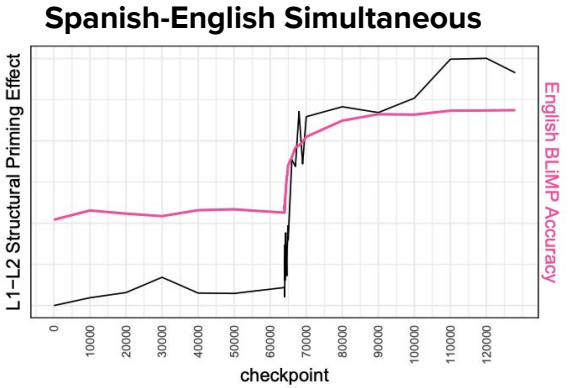
Training Dynamics

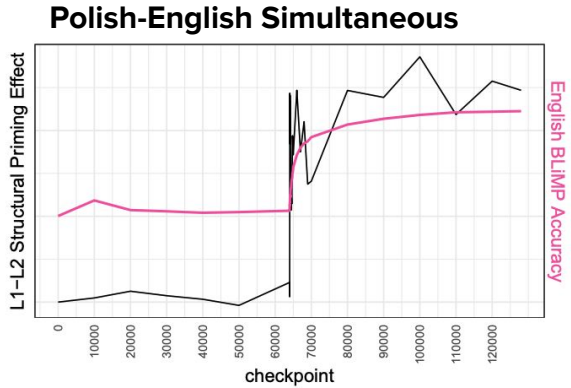


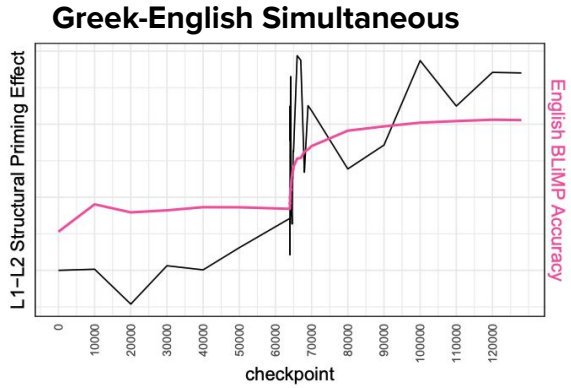
5M tokens
(after correction for multiple comparisons)

Training Dynamics









Shared Grammatical Representations

- Evidence of shared representations only appear after exposure to L2
- Model learns shared representations of L2 very quickly

Future Work

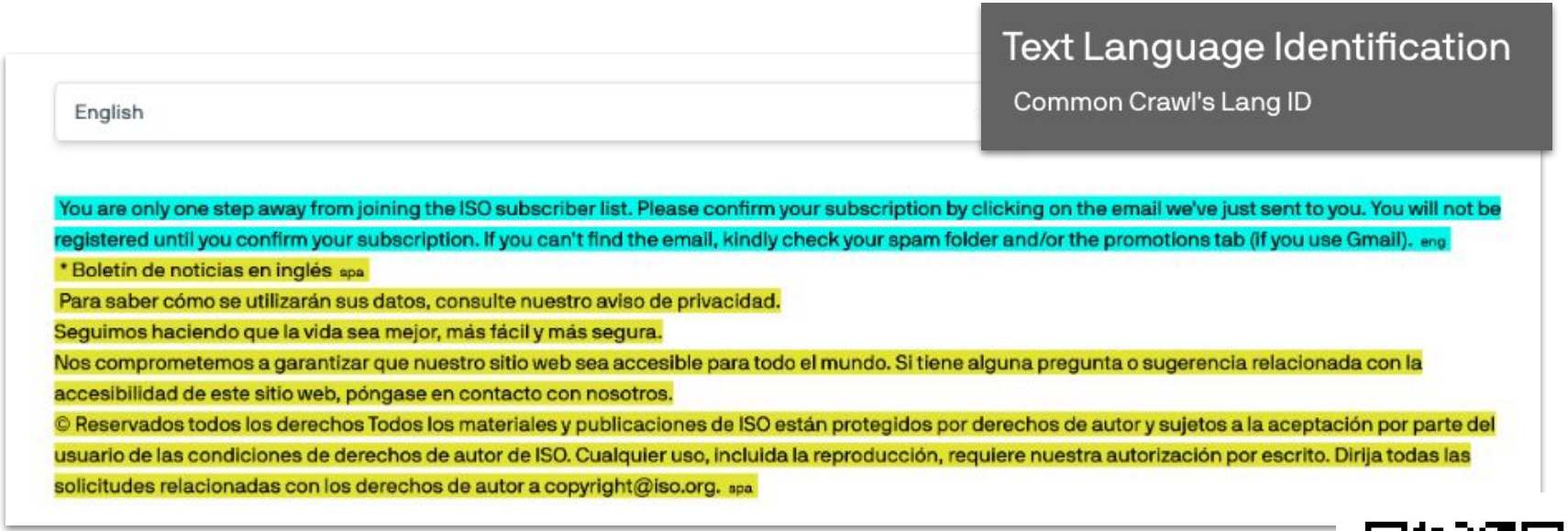
- Mechanistic Interpretability
  - What is the nature of these representations?
  - Where in the model are they located?
- Small models are easier and cheaper to interpret

Conclusion

- When is multilinguality a curse?
  - For high-resource languages, multilingual data may decrease performance
  - English-centric models may not always be the best option
  - We need more monolingual baselines
- When is multilinguality a blessing?
  - When crosslingual transfer benefits performance
- Small models
  - Monolingual models help understand how morphological typology impacts language model learning
  - Bilingual models provide insights into crosslingual transfer
- Limiting factors
  - Training data
  - Evaluations

Shared Task: Language Identification

Shared Task: Language Identification

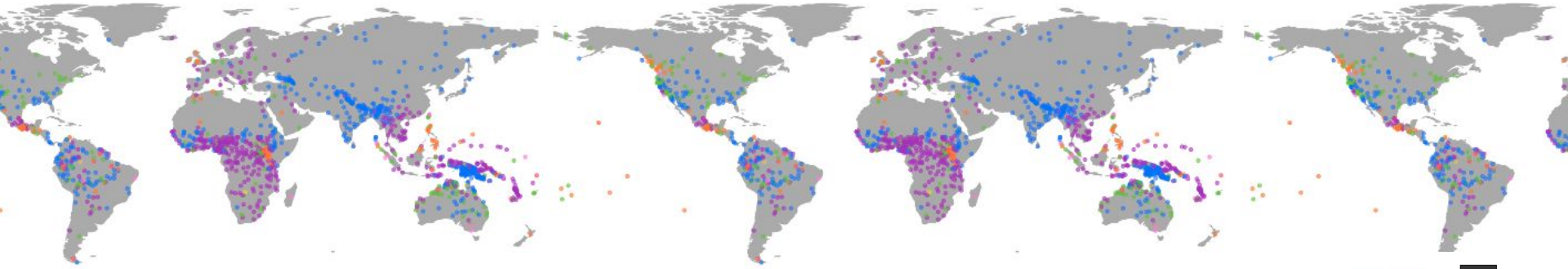


Shared task: https://wmdqs.org/shared-task/

Shared Task: Crowd-Sourcing Multilingual Evaluation

5TH MULTILINGUAL REPRESENTATION LEARNING (MRL) WORKSHOP 2025

CO-LOCATED WITH EMNLP IN SUZHOU, CHINA NOVEMBER 5th - 9th 2025



SIGTYP

https://sigtyp.github.io/ws2025-mrl.html

Thank you!

catherine@eleuther.ai
tachang@ucsd.edu

@catherinearnett
@tylerachang