Multilinguality Does not Make Sense: Investigating Factors Behind Zero-Shot Cross-Lingual Transfer in Sense-Aware Tasks

Roksana Goworek[1] **Haim Dubossarsky**[1,2,3]
[1] Queen Mary University of London
[2] The Alan Turing Institute
[3] University of Cambridge
{r.goworek, h.dubossarsky}@qmul.ac.uk

Abstract

Cross-lingual transfer allows models to perform tasks in languages unseen during training and is often assumed to benefit from increased multilinguality. In this work, we challenge this assumption in the context of two underexplored, sense-aware tasks: polysemy disambiguation and lexical semantic change. Through a large-scale analysis across 28 languages, we show that multilingual training is neither necessary nor inherently beneficial for effective transfer. Instead, we find that confounding factors, such as fine-tuning data composition and evaluation artifacts, can better account for the perceived advantages of multilinguality. Our findings call for more rigorous evaluations in multilingual NLP, and more nuanced and sensible choice of models for transfer. We release fine-tuned models and benchmarks to support further research, with implications extending to low-resource and typologically diverse languages.

1 Introduction

Cross-lingual transfer enables multilingual pretrained models to leverage knowledge acquired in one language to perform tasks in another (e.g., Wu and Dredze, 2019; Ponti et al., 2018). This ability underpins many of today's advances in multilingual NLP and has been evaluated across a wide range of tasks - from syntactic parsing and POS tagging to more complex, semantics-driven tasks like question answering, language inference, and paraphrasing (see Philippy et al. (2023) for a review).

Polysemy disambiguation, despite being foundational to linguistic meaning and a long-standing challenge in NLP (Navigli, 2009; Bevilacqua et al., 2021), has received relatively little attention in transfer learning research. Polysemy poses unique challenges to transfer due to its inherently language-specific nature (Rzymski et al., 2020). For example, while the English word mole denotes 'a small burrowing mammal' and 'a skin blemish,' its Hindi counterpart तिल (til) refers to the latter sense but also means 'sesame seed'. Such non-aligned senses across languages complicate direct transfer, making it an ideal testbed for evaluating the true limits of cross-lingual generalization.

A common assumption is that multilinguality itself, set by a model's exposure to multiple languages, is key for cross-lingual transfer. While it is evident that models must support both source and target languages, the extent to which broader multilinguality facilitates transfer remains unclear and understudied. Does access to more languages intrinsically improve transfer, or do other factors, such as language similarity or training setup, drive the observed gains?

In this work, we investigate the role of multilinguality in cross-lingual transfer for two underexplored, **sense-aware tasks**: polysemy disambiguation and lexical semantic change detection. We examine whether multilingual fine-tuning is truly essential for successful transfer, or whether previously reported benefits stem from confounding factors. Our results show that multilinguality is neither necessary nor intrinsically beneficial in these settings, challenging prevailing assumptions about the mechanisms underlying multilingual transfer.

While our findings may seem narrowly focused, they illuminate broader issues central to NLP, with implications for low-resource and domain-specific settings for which relying on broader multilingual models and corpora is not an option, as well as the wider research community. We release the best-performing models to support further analysis.[1]

2 Related Work

Cross-lingual transfer studies have expanded from tasks like syntactic parsing, POS tagging, and semantic classification (Wu and Dredze, 2019;

[1] Best-performing MONO and MULTI models are available in the collection Multilinguality Does not Make Sense.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35004–35029
November 4-9, 2025 ©2025 Association for Computational Linguistics

Choenni et al., 2023a; Pires et al., 2019; de Vries et al., 2022) to more complex ones such as NER, NLI (Dolicki and Spanakis, 2021; Srinivasan et al., 2021), and more recently QA, paraphrasing, and sentiment analysis (Lauscher et al., 2020; Ahuja et al., 2023; Choenni et al., 2023b; Wang et al., 2023). Most studies rely on pretrained multilingual models like XLM-R and mBERT.

Research also focused on factors affecting transfer, including linguistic similarity (Lauscher et al., 2020), pretraining corpus size and diversity (Srinivasan et al., 2021; Ahuja et al., 2023), lexical overlap (Patil et al., 2022; de Vries et al., 2022), and model architecture (K et al., 2020). Language selection also varied, from high-resource (Choenni et al., 2023b) to extremely low-resource languages tackled with continuous pretraining (Ebrahimi and Kann, 2021; Imani et al., 2023).

Despite extensive work, the role of multilinguality itself is rarely tested directly and is often assumed to be beneficial. A notable exception is Shaham et al. (2024), who compare PaLM 2 fine-tuned on monolingual versus multilingual data for instruction tuning, and find that modest multilingual exposure aids transfer, while too much can degrade it. However, their analysis is limited to a single model and task, leaving open questions about pretraining and generalizability. Choenni et al. (2023b) show that source language examples can influence target predictions, but without explicitly controlling for multilinguality, offering only indirect evidence. Chang et al. (2024) directly manipulated multilinguality by pretraining over 10,000 models across 250 languages, but only evaluated using perplexity with no standard tasks, making comparisons difficult and replication impractical.

Notably underrepresented in transfer research are sense-aware tasks, such as polysemy disambiguation and lexical semantic change (LSC), despite the central role polysemy has in NLP. Unlike sentiment analysis or other tasks amenable to meaning-preserving translations, polysemy exhibits substantial language-specific variation (Rzymski et al., 2020), making it particularly suitable for rigorous evaluation of zero-shot multilingual transfer ability.

Exceptions for polysemy are few. Raganato et al. (2020) reported transfer to 12 languages, but used only English as a source language, while Dairkee and Dubossarsky (2024) found near-zero transfer to Hindi, raising questions about the feasibility of zero-shot transfer in low-resource settings. In contrast, Goworek et al. (2025) recently showed notable zero-shot transfer from English to 10 low-resource languages, highlighting the need for further investigation.

With regard to LSC, Arefyev et al. (2021) showed that training on polysemy disambiguation generalizes to semantic change detection, linking the two tasks. Cassotti et al. (2023) followed up and fine-tuned models on multilingual polysemy datasets, achieving state-of-the-art results (Schlechtweg et al., 2020) and strong transfer to unseen languages (Periti and Tahmasebi, 2024).

None of these works studied multilinguality itself. A notable exception is Berend (2022), who explicitly examined the role of multilinguality in word sense disambiguation transfer. However, their study investigated the role of multilinguality only at the pretraining stage, rather than during fine-tuning. This is a less practical perspective given the ubiquitous use of multilingual models in NLP today, where many languages lack high quality monolingual models.

Overall, while multilingual transfer has been widely studied, the direct contribution of multilinguality in fine-tuning remains underexplored, especially in lexically focused tasks like polysemy and LSC. This work addresses that gap by systematically testing the role of multilinguality in sense-aware transfer. By manipulating multilingual conditions, we clarify when and how multilinguality supports cross-lingual generalization, helping to resolve prior conflicting findings and inform future research on transfer learning.

3 Methods

In this study, we set to isolate multilinguality from confounds, enabling a clear assessment of its independent impact on zero-shot transfer in polysemy disambiguation and LSC tasks. We conduct a large-scale evaluation of zero-shot and full-shot cross-lingual transfer performance across 28 languages, focusing on the unique contribution of multilingual training. To achieve this, we implement an experimental framework that systematically controls for potential confounds, ensuring that observed effects are attributed to multilinguality rather than artifacts of training data size or pretraining exposure.

3.1 Tasks

The Word in Context (WiC) formulation of polysemy disambiguation is used. By pairing two sentences with the same polysemous word Pilehvar and Camacho-Collados (2019) transformed this task into a binary classification problem, where a target word appears either in the *same sense* or in a *different sense* as per the example below:

> The couple went for a **date** last night.
> He marked this **date** on my calendar.

This reformulation removed the need for a sense-label per word, which is language-specific, making the task more suitable for cross-lingual transfer.

For LSC, we compare the models' representations of words occurring in natural sentences across two corpora from different time periods. The underlying assumption, which is the basis of all distributional semantics, is that changes in words' meaning are reflected in measurable changes to their representations over time (Periti and Montanelli, 2024).

3.2 Data

MCL, XL and Hindi datasets, which together span 18 languages, were used for training and evaluation, while AM^{2}iCO and LSCD (LSC Detection), spanning 14 and 7 languages, respectively, were used only for evaluation.[2] As some languages have only development and test data, we followed Cassotti et al. (2023) who sampled 75% of the development data of each language to enable training on these languages (keeping the remaining 25% for setting hyperparameters), using their exact train-dev-test splits. All WiC datasets are class-balanced, setting the chance baseline at 50%. German, French and English are overrepresented in the dataset relative to other languages (see Figure 4).

MCL by Martelli et al. (2021) spans English, Arabic, French, Russian and Chinese, constructed by annotating sentences from native corpora: BabelNet (Navigli and Ponzetto, 2010), the United Nations Parallel Corpus (Ziemski et al., 2016) and Wikipedia, with inter-annotator agreement of 0.95 and 0.9 on English and Russian, respectively.

XL by Raganato et al. (2020) used WordNet of Bulgarian, Chinese, Croatian, Danish, Dutch, English, Estonian, Japanese, Korean and Farsi, filtering out fine-grained senses. French, German and Italian used Wiktionary. The reported mean human accuracy was 80%, and varied between 74% for German, 87% for Danish and 97% for Farsi.

Hindi WiC by Dairkee and Dubossarsky (2024) consists of 12,000 sentence pairs, constructed from a sense-annotated Hindi Corpus (Singh and Siddiqui, 2016) of 60 polysemous nouns.

AM^{2}iCO by Liu et al. (2021) has English paired with 14 target languages. Compiled from Wikipedia Dumps of each language by selecting words with at least two different Wikipedia pages that show ambiguity in both the target language and English. Overall human accuracy was 90.6%, with an inter-annotator agreement of 88.4%.

Lexical Semantic Change Detection (LSCD) covers seven languages from different sources: English, German, Latin, and Swedish from Schlechtweg et al. (2020), Spanish from Zamora-Reina et al. (2022), Chinese from Chen et al. (2023), and Norwegian from Kutuzov et al. (2022).[3] For each language and target word, an equal number of sentences were sampled from corpus 1 (historical) and corpus 2 (modern).

3.3 Multilingual Base Models

To ensure robustness, we use five multilingual models that differ in their language coverage, and in the pretraining proportions of these languages.

XLM-R-large (Conneau et al., 2020) and **mBERT** (Devlin et al., 2019) are pretrained on 100 and 104 languages, respectively, covering all languages in our study, which enables us to train and evaluate on all languages in zero- and full-shot transfer. Both models are commonly used in multilingual research. XLM-R was extensively used in the context of WiC and LSC (Raganato et al., 2020; Cassotti et al., 2023; Dairkee and Dubossarsky, 2024), providing a strong baseline for comparison.

BLOOM (BigScience Large Open-science Open-access Multilingual Language Model) by Le Scao et al. (2023) is pretrained on 46 languages, with a focus on low-resource languages, 10 of which overlap with those used in our study.

LLaMA3-8B-Instruct (Grattafiori et al., 2024; AI@Meta, 2024) is a decoder-only instruction-tuned language model with 8 billion parameters, significantly larger than the other models. Pretraining data is not public, and is assumed to include major NLP datasets (Sainz et al., 2023), like WiC.

MuRIL (Multilingual Representations for Indian Languages) by Khanuja et al. (2021) is pretrained on 16 Indian languages + English, which, along with Hindi, are the only languages overlapping with our study. It was used to test an edge case of zero-shot transfer (see §4.2).

[2] AM^{2}iCO uses English as a pivot language therefore is biased toward it; LSCD is not in the WiC format.

[3] Norwegian$_1$, Norwegian$_2$, are two corpus pairs, compar-

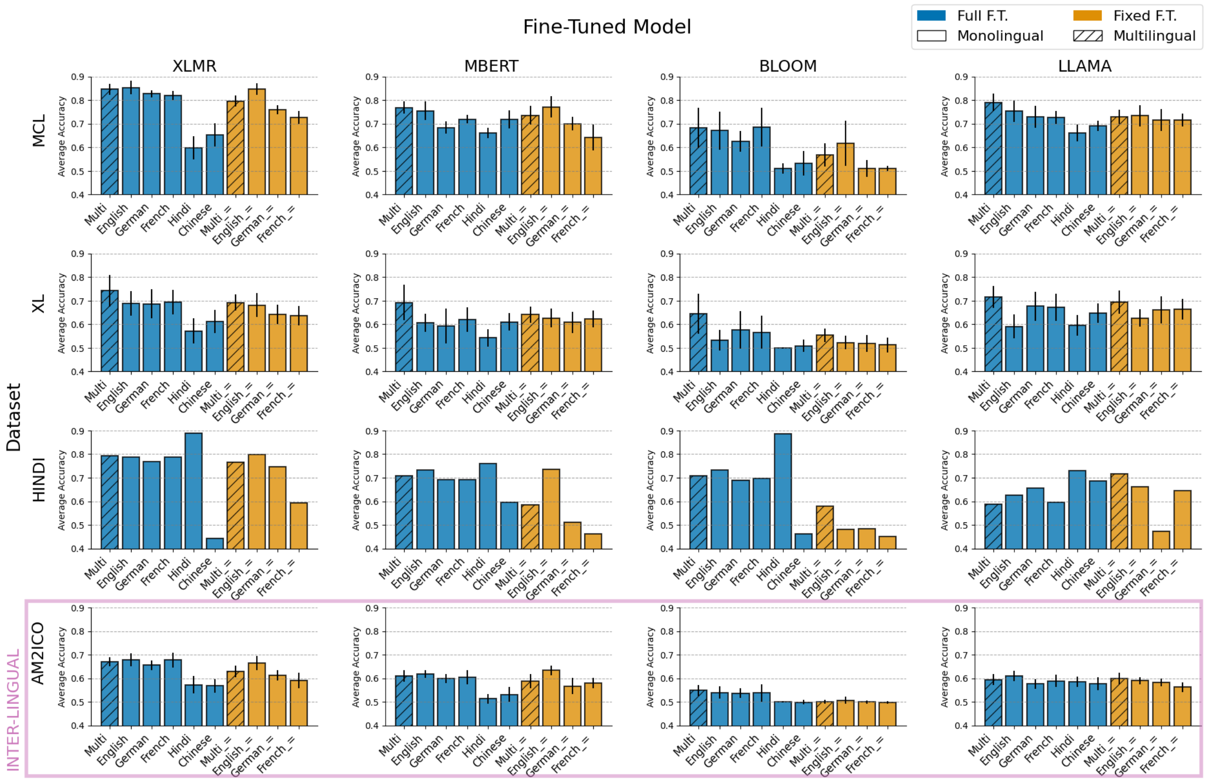

Figure 1: Mean accuracies and SD (bars) for multilingual and monolingual models on WiC datasets using different pretrained models. Colors indicate whether fine-tuning was done on all data or on its sampled portion (Hindi and Chinese appear only in the former due to their smaller data which did not allow subsampling). Hindi, as a single language dataset, does not have SD. For detailed results see Appendix H.

3.4 Model Training and Testing

We fine-tune models on different combinations of the data, creating three conditions: (1) **MONO**lingual models, trained on a single language. Only the 5 largest languages were considered; (2) **MULTI**lingual models, trained on all languages; and (3) **MULTI**$_{\otimes}$lingual models, trained on all but one held-out language. We also compare these models while controlling for the total amount of fine-tuning in the FIXED F.T. condition, denoted by **MODEL**$_{=}$. By systematically comparing performance across conditions while controlling for fine-tuning size and pretraining exposure, we isolate the effect of multilinguality on cross-lingual transfer in WiC and LSC tasks.

Fixed Fine-Tuning To control for the different training sizes across languages, fixed-size versions of datasets were created for the MULTI, ENGLISH, FRENCH and GERMAN datasets by randomly subsampling 8,750 examples from their training sets.[4]

Finetuning Details All encoder models were fine-tuned in the same way, selectively focusing on different subsets of the WiC datasets. Following Cassotti et al. (2023) and after a failed pilot study with cross-encoders, a Siamese bi-encoder was used to generate two distinct vector representations (embeddings) for the two usages of the target word in the two sentences. The model outputs the cosine distance between the output embeddings of the two inputs, and, in order to adapt it to the binary nature of WiC, a threshold is applied to decide if the words are classified as having the same sense. The model is trained to update its parameters (i.e., embeddings) to maximize this distance when the target appears in different meanings (label 0) and minimize it when the meanings are the same in the two sentences (label 1), by minimizing contrastive loss. During training, as well as inference, special tokens, <t> and </t>, are placed around the target word in each sentence to signal what word the model should focus on.

To preserve comparability and avoid cherry-picking, we fixed a single hyperparameter configuration and did not perform per-model/per-language

ing the words from four time periods.

[4] Other languages had too little data to subsample.

tuning or multi-seed sweeps, which may understate best-case performance and limit our characterization of variance across runs. For more discussion of this decision, see Appendix E.[5]

To verify that performance is significantly improved by fine-tuning rather than being the result of inherent sense-distinction ability of contextualized embeddings, we conduct no-fine-tuning experiments with frozen pretrained models. The results corroborate this and can be found in Appendix I. For details of LLaMA's fine-tuning (same data, different procedure), see Appendix L.1.

Testing To evaluate models on WiC, a threshold for each model is determined by maximizing accuracy on the validation set of the training language. This threshold collapses the cosine distance between the output embeddings of the two inputs into a binary label. For LSC, we follow Cassotti et al. (2023) and output a fixed-size vector for the target word in each sentence. Taking all vectors for each word across the two time-bins, a change score is computed using APD and PRT measures (Kutuzov and Giulianelli, 2020), and evaluated against gold-label change scores using Spearman correlation, as standard in LSC (Periti and Tahmasebi, 2024).

4 Experiments and Evaluation

4.1 Multilinguality Effects and Confounds

To rigorously assess the role of multilinguality in zero-shot transfer, we conduct three types of controlled comparisons, ensuring that fine-tuning size is held constant and distinguishing between full-shot and true zero-shot transfer scenarios. We provide a comprehensive evaluation of four multilingual base-models (§3.3) on 28 languages, first examining whether multilingual training offers an advantage and then systematically accounting for key confounding factors. We summarize the average accuracies of the models in Figure 1 and present the true zero-shot transfer comparison in Table 2. Due to space constraints, except for Figure 1, we show and discuss XLM-R results in the main text, only briefly discussing other models while providing their full results in Appendix H.

Testing for Multilingual Advantage We contrast between MONO models, which are fine-tuned on a single language, and the MULTI model, trained on all available data in the MCL and XL WiC datasets. If multilinguality is indeed critical, then MULTI should consistently outperform MONO models, which by definition do not have access to information outside of their own language beyond their pretraining stage. Our results show a mixed pattern, with many MONO models on-par or even better than MULTI on the MCL, AM^{2}iCO and HINDI datasets, and a multilinguality advantage on the XL dataset (blue bars of the FULL F.T condition in Figure 1). LLaMA is the exception to the rule, though its results are 5%-10% points lower than XLM-R. We attribute the low transfer of Hindi and Chinese MONO models to their smaller dataset sizes, and subsequently tested whether dataset size is a potential confound in our analyses.

Dataset Size Confound MULTI is fine-tuned on an order of magnitude more data than any other model (see Figure 4), which gives it an unfair advantage over all monolingual models. We therefore fix the size of the training datasets across all models (§3.4) to enable a fair comparison, and repeat the same analysis under these controlled conditions. The results in Figure 1, and especially the comparison between FULL F.T and FIXED F.T (blue and orange bars, respectively) which are summarized in Table 1, show that the performance of the MULTI model drops much more than MONO English when training size is controlled, perhaps due to a larger relative drop in training size, making MONO English the best model across most datasets. Similar drops are observed for mBERT, BLOOM, and LLaMA (Tables 7-9), which for the latter largely diminished the advantage it had in FULL F.T. Despite its drop in the FIXED F.T condition, LLaMA seems to benefit from multilingual fine-tuning more than other models. We attribute this to its nature as a generative model and pretraining uncertainty which may include data contamination unwanted for our experiments (Sainz et al., 2023). Additionally, LLaMA relies on English prompts with explicit task instructions, which showed improved performance, particularly for HINDI and CHINESE, where training data is more limited. Overall, this contrast (blue and orange) undermines the notion that training on a multilingual dataset inherently improves transfer. Instead, the advantage originally observed for MULTI in the Full F.T may largely be due to more training data (122.4k examples for MULTI cf. 54.7k for German, 46.1k for French, and 15.1k for English; see Table 4). This finding does not imply that training on more examples is not a good strategy to improve transfer, which it clearly

[5] Training hyperparameters and model sizes are in Table 6.

is, only that attributing the performance gains to multilinguality is flawed.

Error Analysis If multilinguality equips models with novel sense-understanding, then not only should its performance improve, but also its errors should differ from MONO models that lack exposure to such linguistic diversity. Instead, Figure 6 shows greater overlap of errors between MULTI and MONO models, which even increases for FIXED F.T, supporting a notion that MULTI is simply another MONO model (see formulation of alignment measures in Appendix F). This provides converging evidence that supports previous findings, further undermining the importance of multilinguality.

Dataset / Model	MCL	XL	Hindi	Am^{2}iCO
MULTI	-5.0	-5.0	-2.7	-4.0
ENGLISH	-0.6	-0.6	1.2	-1.5
GERMAN	-6.8	-4.4	-2.0	-4.3
FRENCH	-9.5	-5.5	-19.6	-8.7

Table 1: Percentage change in average accuracy of XLM-R per dataset (Fixed F.T. - Full F.T.)

Language	German	French		English		Hindi	Chinese	
Dataset	XL	MCL	XL	MCL	XL	Hindi	MCL	XL
Z.S. MULTI$_{\otimes=}$	67.6	79.0	65.3	**82.9**	**68.7**	76.6	78.3	72.2
GERMAN$_=$	74.1	77.0	63.8	78.3	60.6	74.7	75.0	67.5
FRENCH$_=$	68.6	73.7	65.5	77.0	63.0	59.3	71.2	66.9
ENGLISH$_=$	**69.3**	**84.1**	**65.6**	89.2	68.9	**79.9**	**78.4**	**76.2**
HINDI$_=$	62.1	53.8	59.7	66.0	56.8	88.9	60.7	66.4
CHINESE$_=$	62.3	61.3	58.1	68.1	61.3	44.2	70.8	69.4

Table 2: XLM-R (fixed F.T.) **zero-shot** results for MONO and MULTI models. For each of the 5 column languages, the Z.S. MULTI$_{\otimes=}$ row reports the model's performance when it is trained multilingually with that language held out. Rows *lang*$_=$ report monolingual models trained on *lang*, evaluated on each column test language. Grey cells indicate full-shot conditions (target language included in fine-tuning).

The Z.S. MULTI$_{\otimes=}$ row presents the results of five different multi$_{\otimes=}$ models (each trained on all available languages excluding the target language, specified by the column, respectively). The {lang}$_=$ rows show the results of the specified monolingual model on the test set specified by the column. Grey cells mark full-shot conditions for monolingual models.

Full-shot Exposure Artifact We currently measure transfer by conflating zero- and full-shot, when the target language is present in training, together. Since MULTI is trained on all the languages it is later evaluated on (except Hindi), it is effectively evaluated under full-shot conditions. In contrast, MONO models are only full-shot with respect to their training language, and zero-shot for all the rest. Thus, comparing transfer between MULTI and MONO models is unfair, as full-shot learning is expected to be much better than zero-shot transfer.

To address this, we train five different MULTI$_\otimes$ models, excluding one language at a time on which that specific model is later evaluated on, while still fixing the amount of fine-tuning data as before, and compare them to MONO models. We focus on the 5 languages with the most training data. If multilinguality is truly of merit, then MULTI models, that were trained on 14 languages overall (minus the held-out language they are evaluated on), should outperform MONO models.

Table 2 shows that MONO English outperforms zero-shot MULTI$_\otimes$ on all languages, and the only case where zero-shot MULTI$_\otimes$ outperforms MONO models is English, where the ENGLISH model is full-shot, and thus not considered in this evaluation (grayed). This further disproves the assumed benefit of multilingual training, as prior advantage in performance that was originally associated with multilinguality stemmed, at least in part, from mixing full-shot with zero-shot transfer conditions, which only MULTI possessed. Similar results were obtained for mBERT, BLOOM and even LLaMA, although with GERMAN as the best zero-shot model (see Tables 10-12). Interestingly, the strong zero-shot performance of MULTI models on English could be related to the prevalence of English in the base models' pretraining data.

4.2 Underlying Factors of Successful Transfer

Correlation with Model's Pretraining Size Our analyses reject multilinguality as a key driver of zero-shot transfer, showing its effects stem from training data size - a confound, albeit a beneficial one. Yet, even after controlling for this, transfer results varied considerably across target languages.

Before their fine-tuning, multilingual models undergo pretraining on large-scale datasets spanning multiple languages. The prevalence of a target language within a model's pretraining corpus, or "pretraining size", may influence the model's ability to represent, process and transfer to that language.

To test this hypothesis, and understand what drives variation in transfer results, we computed Pearson correlations between languages' log-transformed pretraining sizes in models for which pretraining sizes were available (XLM-R, mBERT, BLOOM) and their corresponding accuracies on the four WiC datasets and on LSCD.[6]

Figure 2 shows strong correlations between a language's pretraining size and its accuracy as a target language. We attribute BLOOM's poor correlations to its unusual pretraining language distribution that focuses on low-resource languages. The lack of correlation in XL remains unclear and requires further investigation (see §5 for discussion).

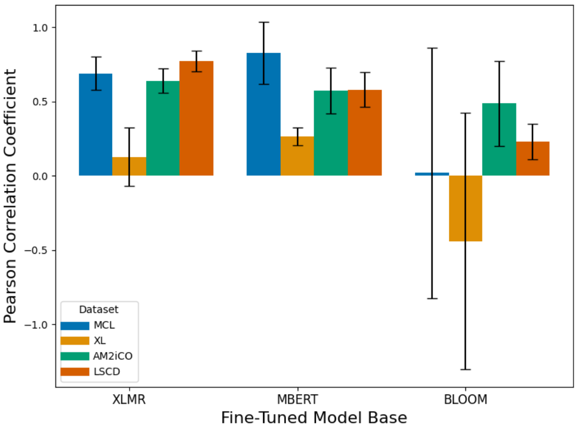

Figure 2: Mean correlations between languages pretraining sizes and zero-shot performance of MONO models.

Linguistic Similarity We tested the link between linguistic similarity and transferability using syntactic similarity scores (Littell et al., 2017), which quantify similarity based on shared syntactic features, as is standard in NLP (Philippy et al., 2023). Pearson correlations between similarity scores and model accuracies were computed across target language pairs. Syntactic similarity showed some correlation with zero-shot performance, the effect was weaker and less consistent than that of pretraining size (see Appendix 5). Moreover, we found that syntactic similarity is highly correlated with pretraining size, further undermining its contribution.

Zero-shot Transfer from Unknown Languages While the correlation analysis provides useful insights, it remains slightly inconclusive. A key limitation is that we cannot pretrain models to directly manipulate this variable for a controlled experiment. We also lack sufficient fine-tuning data for languages that were less prominent during pretraining, making it difficult to determine whether poor model performance stems from limited fine-tuning or limited exposure during pretraining.

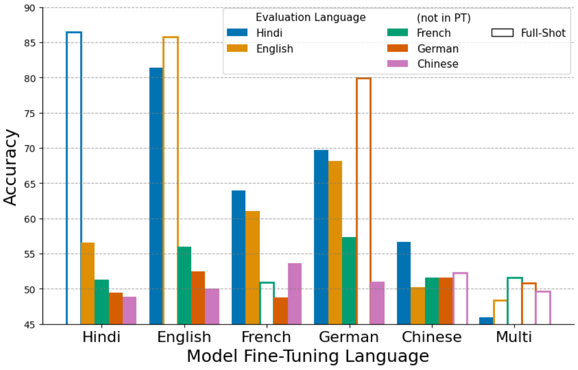

Figure 3: MuRIL accuracy scores. "(not in PT)" are languages absent from MuRIL's pretraining.

To counter this, we analyzed transfer in an edge case scenario. We used MuRIL (§3.3), a model pretrained exclusively on 16 Indian languages and English. Evaluating zero-shot transfer performance of models that are fine-tuned on languages entirely absent from MuRIL's pretraining corpus can offer insight on the role of pretraining size in transfer.

Figure 3 shows graded transfer effects associated with the pretraining size, supporting the correlation analysis above. Hindi, being the dominant language in MuRIL (despite less data than English, it benefits from exposure to 16 Indic languages, including 2–4 with the same script and 9 with lexical overlap), enjoys the largest zero-shot transfer regardless of the fine-tuning language (tallest blue bars across all different models, not considering the full-shot condition in white), followed by English (yellow bars). Second, we observe substantial transfer from French, German and even Chinese on Hindi, languages MuRIL was not pretrained on (see §5 for further discussion). Third, MuRIL was able to learn and perform well on German, despite it being absent from its pretraining stage.

We find that the MULTI model, despite being trained predominantly on German, French, English, and Chinese, failed to learn any meaningful sense disambiguation, let alone transfer them to Hindi. This highlights a stark contrast: while MuRIL trained on monolingual data can effectively learn the task from languages it was not pretrained on, the multilingual data appears too noisy for effective transfer in this extreme zero-shot setting.

4.3 Lexical Semantic Change Detection

[6] Pretraining sizes are provided in Appendix 13. For BLOOM, we removed languages not in its pretraining.

[6] For Norwegian, two corpora pairs are available – the period 1929–1965 paired with 1970–2013, reported as NO_1, as well as 1980–1990 with 2012–2019, reported as NO_2.

LSCD Language ↓	XL-LEXEME	MULTI	GERMAN	FRENCH	ENGLISH	HINDI	CHINESE
English	.757	.703	.737	.681	**.772**	.436	.673
German	**.873**	.863	.841	.867	.844	.635	.641
Swedish	.754	**.801**	.754	.618	.724	.480	.489
Latin	-.035	.117	**.161**	.136	.135	-.177	.091
Spanish	.665	.696	.670	.664	**.711**	.354	.383
Chinese	.734	.652	.649	.499	**.737**	.524	.593
Norwegian$_1$	.668	.729	.638	.697	**.777**	.525	.400
Norwegian$_2$	.634	**.655**	.604	.580	.645	.433	.439
Average	.631	.652	.632	.593	**.668**	.401	.464

Table 3: Spearman correlations of XLM-R models' APD scores with graded semantic change scores across LSCD tasks in different languages. Best scores are in bold, scores within 0.05 of the best are underlined.

We evaluate XL-LEXEME (Cassotti et al., 2023), trained on multilingual and inter-lingual data (all WiC datasets plus AM^{2}iCO), against the same MONO models from earlier analyses. As in WiC, MULTI is matched or outperformed in all but two languages by MONO English or German. XL-LEXEME outperforms only on German, but all well-trained MONO models perform well on this task. Notably, MONO English outperforms all other models on average and in 4 out of 8 test languages. Results correlate strongly with pretraining sizes in all three models (Figure 2), echoing WiC patterns.

Zero-shot transfer performs on par with full-shot. On German, MONO FRENCH and ENGLISH outperform the GERMAN model, and ENGLISH is most effective on Spanish, Chinese, and Norwegian. PRT confirms these trends, with ENGLISH showing the highest correlation. Across mBERT and BLOOM, MONO models often match or outperform MULTI models (Tables 26-28).[7]

Overall, our results show that the best results in LSC are obtained by monolingual models, dismissing multilinguality as an important factor in cross-lingual transfer also on this task.

5 Discussion

Disentangling the effects of multilinguality from confounding factors such as data quality and pretraining exposure is inherently challenging. Rather than providing a single definitive explanation, we identify consistent trends and highlight effects that reflect the nuanced nature of cross-lingual transfer.

Quantity over Multilinguality Our results show that the amount of fine-tuning data matters more than the number of languages included during training, as FULL F.T. MONO models trained on more data typically outperform the FIXED F.T. MULTI model. In both fine-tuning conditions, the MULTI model only partially outperforms monolingual models, with the ENGLISH model often matching or exceeding it, even in the FULL F.T. setting. Taken together, this pattern of results challenges the assumption that multilingual training inherently improves zero-shot transfer; rather, data quantity drives most gains. Notably, Berend (2022), who also tested the role of multilinguality in word sense disambiguation but in the pretraining stage, reached the same conclusions. Our results not only complement those of Berend but are also more practical as multilingual models are in much wider use than monolingual ones in transfer scenarios, where the latter are actually lacking for many languages.

Quality Beyond Quantity The frequent outperformance of the MONO English model on most tasks could point to data quality as an important factor. Certainly, the quality of training data can impact performance, and consequently, transfer to other languages. English may enjoy datasets of higher quality, often attributed to their curation, annotation protocols, sense granularity and text quality (Philippy et al., 2023). However, assessing or normalizing data quality across languages for polysemy disambiguation is nearly impossible without costly native-speaker annotations. Thus, this factor remains difficult to study directly.

Other Underlying Factors ENGLISH's strong performance may be tied to its dominance in pretraining data which is supported by its strong overall correlations reported in §4.2. The contrasting results from MuRIL, where transfer was strongest to Hindi when Hindi had a pretraining size advantage over English, further underscore this point and provide an informative exception.

Our results show that both pretraining size and language similarity are tied to transfer, but the former has a stronger correlation. The influence of pretraining size on cross-lingual transfer is well established (Lauscher et al., 2020; Srinivasan et al., 2021; Ahuja et al., 2023), as the correlation between language similarity and transfer performance (Wu and Dredze, 2019; Pires et al., 2019; K et al., 2020). However, while both variables clearly influence transfer, the effect seems to depend on the spe-

[7] LLaMA models were not evaluated on LSCD as prompting generative models with hundreds of usages to produce a scalar 'change score' is not practically feasible.

cific task on which transfer is evaluated: pretraining size correlates more with semantic tasks (e.g., NLI, QA), while linguistic similarity benefits syntactic tasks (e.g., POS tagging, dependency parsing) (Lauscher et al., 2020; Philippy et al., 2023).

We suggest that sense-aware tasks, such as polysemy disambiguation and semantic change, should follow the semantic route. Given the language-specific nature of word senses, syntactic similarity offers limited benefit in the transfer of knowledge. Our findings confirm this: transfer performance in these tasks depends more on pretraining size than on linguistic similarity. The rationale of this interpretation is rooted in the type and relevance of linguistic *knowledge* that is being transferred (Rajaee and Monz, 2024; Goldman et al., 2025). We propose that syntactic tasks rely more heavily on structural similarities across languages, and thus benefit from linguistic similarity. As syntactic patterns diverge across typologically distant languages, their utility in transfer diminishes. In contrast, semantic tasks are more dependent on the diversity and scale of pretraining data, which exposes models to a wider range of meaning representations.

Lexical Semantic Change The results on LSCD align with the conclusions observed in the WiC tasks: multilinguality does not necessarily lead to improved performance, and monolingual models, particularly ENGLISH, often outperform multilingual ones. This holds even when the test language is included in the multilingual setup (and, of course, absent from the monolingual model), as seen in the cases of Spanish, Chinese, and Norwegian, where the English model outperformed both multilingual models. These findings reinforce the paper's central claim: multilinguality is neither necessary nor inherently beneficial for effective transfer, and monolingual models can surpass multilingual ones even under comparison conditions that favor the latter. In such cases, the multilingual model benefits from full-shot exposure to the test language, while the monolingual model is evaluated in a zero-shot setting—yet still performs better. These results further emphasize the role of dataset quality in transfer, suggesting that the perceived advantages of multilinguality may be confounded by differences in training data quality.

As lexical semantic change gains popularity and more languages are being studied using computational methods, it is important not to perpetuate the misconception that multilingual models are always the best solution—not even when the target language is included in the multilingual training setup. This conclusion is further supported by Baes et al. (2025), who recently showed that the state-of-the-art model for LSCD—the multilingual XL-LEXEME by Cassotti et al. (2023) also included in our comparisons—is not the best choice across different semantic change scenarios. Their findings call for a more nuanced approach to model selection—one that accounts for the specific conditions of the linguistic inquiry being studied, whether it is a different language or type of change.

Large Models and Architectures Notably, larger models like LLaMA, pretrained on much more data, underperform compared to smaller models like XLM-R. This corroborates our main finding that increasing the amount of training data alone – whether in terms of language coverage or volume – does not guarantee better transfer. However, architectural differences (encoder vs. decoder), hyperparameter tuning, and adaptation methods likely also influence performance and require further investigation. Additionally, Berend (2022) suggests that multilingual models may not even be necessary for effective transfer, highlighting that monolingual pretrained models can achieve strong cross-lingual generalization on sense-aware tasks when combined with appropriate adaptation techniques.

6 Conclusion

This study is the first to directly investigate the role of multilinguality in transfer for sense-aware tasks, polysemy and lexical semantic change. Our results indicate that the improved performance typically attributed to multilinguality largely stems from confounding factors such as training size, rather than linguistic diversity. Indeed, training on all available languages increases data quantity, inadvertently benefiting transfer, but this effect should not be mistakenly credited to multilingual diversity itself.

While further research is needed to determine whether these patterns generalize across other tasks, they may underline broader implications for multilingual NLP research that traditionally favored multilingual quantity over data quality. As expanding datasets across multiple languages is costly, and sometimes not practical for some tasks and languages, future research should prioritize understanding the characteristics of training regimes that facilitate transfer in order to optimize training resources for effective cross-lingual transfer.

7 Limitations

Zero-shot transfer enables models to perform tasks in languages without task-specific training data. However, its effectiveness depends on the pretrained model's ability to capture the linguistic properties of both the source and target languages, as well as the availability of sufficient, high-quality data in the fine-tuning language. Our study is constrained by this limitation, as we were only able to train models on a limited set of languages with highly imbalanced data distributions. As in broader cross-lingual transfer research, high-resource languages dominate, while lower-resource languages often lack sufficient data for effective fine-tuning.

We evaluated zero-shot transfer on two tasks: WiC and LSCD. Both use the same bi-encoder architecture and cosine similarity to produce their outputs. While these results provide useful insights, further research is needed to validate the findings across other tasks and model architectures.

Our evaluation of LLaMA is limited to the WiC task, and several factors constrain its interpretability. First, its pretraining data is unknown so we could not assess the correlation of performance with pretraining size. Second, LLaMA's pretraining could include the datasets used in our training or evaluation. And lastly, adapting WiC to LLaMA's generative format (see Appendix L.1) introduced further limitations, such as the lack of a threshold to control sense discrimination, and a cue about the task for LLaMA, which encoder-based models lacked. Considering these, LLaMA is not directly comparable to the other models. Future work is needed to better understand the impact of these architectural differences. Due to computational resource constraints, we were unable to include other larger-scale models (8B+ parameters) or mixture-of-experts (MoE) variants in our experiments. Despite this, our study remains broadly relevant as encoder-based models remain crucial for tasks like classification, regression, and ranking in retrieval systems, where their efficiency and lower resource demands make them well-suited. Additionally, on-device performance requires fast inference and low memory usage. While recent hardware and quantization techniques enable some larger models to run locally on powerful devices, encoder models remain preferred for many real-world applications that require fast, cost-effective, and scalable solutions, especially in specialized or low-resource settings.

Further investigation into the factors influencing zero-shot performance, particularly in low-resource languages, is essential for improving cross-lingual transfer and democratizing access to language technologies.

8 Ethical Considerations

This study investigates factors influencing successful zero-shot transfer, with the goal of informing the NLP community on how to develop more effective methods for speakers of low-resource languages. We evaluate models on tasks in a total of 28 languages and provide details on the languages and dataset sources used to create the datasets.

We do not foresee any direct ethical risks arising from our findings. Rather, this work promotes more responsible resource allocation by encouraging a shift away from continual dataset creation in favor of improving cross-lingual transfer techniques and understanding what defines high-quality datasets. However, it is important to acknowledge that zero-shot methods, while beneficial, may still introduce biases due to disparities in pretraining data, potentially disadvantaging underrepresented languages. Additionally, reliance on transfer from high-resource languages may reinforce linguistic hierarchies, where certain languages disproportionately influence model behavior.

Future work should continue to critically assess the impact of cross-lingual transfer on linguistic diversity and ensure that improvements in NLP benefit a wide range of language communities equitably.

9 Acknowledgments

We are grateful to Francesco Periti and Dominik Schlechtweg for their invaluable feedback and insightful advice on this manuscript, and to Mahmud Akhter and Pierluigi Cassotti for their support.

This work has in part been funded by the research program Change is Key! supported by Riksbankens Jubileumsfond (under reference number M21-0021).

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

AI@Meta. 2024. Llama 3 model card.

Nikolay Arefyev, Daniil Homskiy, Maksim Fedoseev, Adis Davletov, Vitaly Protasov, and Alexander Panchenko. 2021. Deepmistake: Which senses are hard to distinguish for a wordincontext model. In *Computational Linguistics and Intellectual Technologies - Papers from the Annual International Conference 'Dialogue' 2021*.

Naomi Baes, Raphael Merx, Nick Haslam, Ekaterina Vylomova, and Haim Dubossarsky. 2025. LSC-eval: A general framework to evaluate methods for assessing dimensions of lexical semantic change using LLM-generated synthetic data. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 10905–10939, Vienna, Austria. Association for Computational Linguistics.

Gábor Berend. 2022. Combating the curse of multilinguality in cross-lingual WSD by aligning sparse contextualized word representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2459–2471, Seattle, United States. Association for Computational Linguistics.

Michele Bevilacqua, Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. Recent trends in word sense disambiguation: A survey. In *International Joint Conference on Artificial Intelligence*, pages 4330–4338. International Joint Conference on Artificial Intelligence, Inc.

Pierluigi Cassotti, Lucia Siciliani, Marco DeGemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. Xl-lexeme: Wic pretrained model for cross-lingual lexical semantic change. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1577–1585.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Ben Bergen. 2024. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4074–4096, Miami, Florida, USA. Association for Computational Linguistics.

Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokić, and Chu-Ren Huang. 2023. Chiwug: A graph-based evaluation dataset for chinese lexical semantic change detection. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 93–99.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023a. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pages 613–641.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023b. How do languages influence each other? studying cross-lingual data sharing during LM fine-tuning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13244–13257, Singapore. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Farheen Dairkee and Haim Dubossarsky. 2024. Strengthening the wic: New polysemy dataset in hindi and lack of cross lingual transfer. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15341–15349.

Wietse de Vries, Martijn Wieling, and Malvina Nissim. 2022. Make the best of cross-lingual transfer: Evidence from POS tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Błażej Dolicki and Gerasimos Spanakis. 2021. Analysing the impact of linguistic features on cross-lingual transfer. *arXiv preprint arXiv:2105.05975*.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, et al. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *arXiv preprint arXiv:2502.21228*.

Roksana Goworek, Harpal Singh Karlcut, Hamza Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Satkirti Purighella, Paridhi Gupta, Muhinyia Ndegwa, Bao Khanh Tran, and Haim Dubossarsky. 2025. SenWiCh: Sense-annotation of low-resource languages for WiC using hybrid methods. In *Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 61–74, Vienna, Austria. Association for Computational Linguistics.

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

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. 2020. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, Shruti Gupta, Subhash Chandra Bose Gali, Vish Subramanian, and Partha Talukdar. 2021. Muril: Multilingual representations for indian languages. *Preprint*, arXiv:2103.10730.

Andrey Kutuzov and Mario Giulianelli. 2020. UiO-UvA at SemEval-2020 task 1: Contextualised embeddings for lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 126–134, Barcelona (online). International Committee for Computational Linguistics.

Andrey Kutuzov, Samia Touileb, Petter Mæhlum, Tita Enstad, and Alexandra Wittemann. 2022. NorDiaChange: Diachronic semantic change dataset for Norwegian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2563–2572, Marseille, France. European Language Resources Association.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Qianchu Liu, Edoardo Maria Ponti, Diana McCarthy, Ivan Vulić, and Anna Korhonen. 2021. AM2iCo: Evaluating word meaning in context across low-resource languages with adversarial examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7151–7162, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, Roberto Navigli, et al. 2021. Semeval-2021 task 2: Multilingual and cross-lingual word-in-context disambiguation (mcl-wic). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)*, 41(2):1–69.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden. Association for Computational Linguistics.

Vaidehi Patil, Partha Talukdar, and Sunita Sarawagi. 2022. Overlap-based vocabulary generation improves cross-lingual transfer among related languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–233, Dublin, Ireland. Association for Computational Linguistics.

Francesco Periti and Stefano Montanelli. 2024. Lexical semantic change through large language models: a survey. *ACM Comput. Surv.*, 56(11).

Francesco Periti and Nina Tahmasebi. 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4262–4282, Mexico City, Mexico. Association for Computational Linguistics.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings* *of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Ivan Vulić, Goran Glavaš, Nikola Mrkšić, and Anna Korhonen. 2018. Adversarial propagation and zero-shot cross-lingual transfer of word vector specialization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 282–293, Brussels, Belgium. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. Xl-wic: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206.

Sara Rajaee and Christof Monz. 2024. Analyzing the evaluation of cross-lingual knowledge transfer in multilingual language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2895–2914, St. Julian's, Malta. Association for Computational Linguistics.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):13.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1–23, Barcelona (online). International Committee for Computational Linguistics.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Satyendr Singh and Tanveer J Siddiqui. 2016. Sense annotated hindi corpus. In *2016 International Conference on Asian Language Processing (IALP)*, pages 22–25. IEEE.

Anirudh Srinivasan, Sunayana Sitaram, Tanuja Ganu, Sandipan Dandapat, Kalika Bali, and Monojit Choudhury. 2021. Predicting the performance of multilingual nlp models. *arXiv preprint arXiv:2110.08875*.

Fei Wang, Kuan-Hao Huang, Kai-Wei Chang, and Muhao Chen. 2023. Self-augmentation improves zero-shot cross-lingual transfer. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–9, Nusa Dua, Bali. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 833–844, Hong Kong, China. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Dominik Schlechtweg. 2022. LSCDiscovery: A shared task on semantic change discovery and detection in Spanish. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 149–164, Dublin, Ireland. Association for Computational Linguistics.

Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3530–3534, Portorož, Slovenia. European Language Resources Association (ELRA).

A Model Training Details

Model Type	Model Name	Training Size
MULTI	MULTI	122.4k
MONO	GERMAN	54.7k
	FRENCH	46.1k
	ENGLISH	15.1k
	HINDI	7k
	CHINESE	2.5k

Table 4: Fine-tuning sizes.

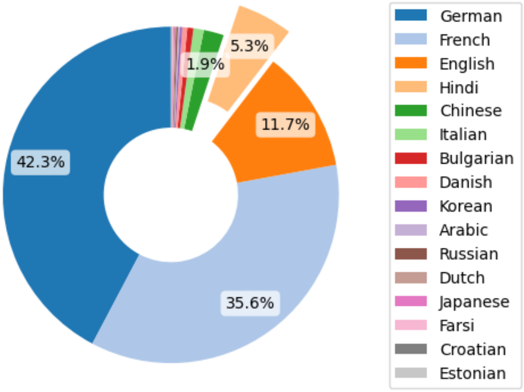

Figure 4: Proportions of languages in WiC datasets used for training. MULTI is trained on all of them except Hindi.

Model	Number of Parameters
XLM-R	560M
BLOOM	560M
mBERT	178M
LLaMA3-8B-Instruct	8B
MuRIL	506M

Table 5: Number of parameters for each selected pre-trained model.

Hyperparameter	Value
Hidden activation	`gelu`
Hidden dropout probability	0.1
Hidden size	1024
Initializer range	0.02
Intermediate size	4096
Layer norm epsilon	1×10^{-5}
Max position embeddings	514
Number of attention heads	16
Number of hidden layers	24
Position embedding type	Absolute
Vocabulary size	250004
Learning rate	1×10^{-5}
Weight decay	0.0
Max sequence length (λ)	128

Table 6: Fine-tuning hyperparameters used in our experiments.

B Percentage Change Between Full F.T. - Fixed F.T.

Dataset Model	MCL	XL	Hindi	Am^{2}iCO
MULTI	-3.3	-5.1	-12.3	-2.4
ENGLISH	1.6	2.1	0.1	1.5
GERMAN	1.6	1.5	-18.2	-3.0
FRENCH	-7.7	0.3	23.1	-2.8

Table 7: Percentage change in average accuracy of mBERT per dataset (Fixed F.T. - Full F.T.)

Dataset Model	MCL	XL	Hindi	Am^{2}iCO
MULTI	-11.5	-9.1	-13.0	-4.8
ENGLISH	-5.3	-0.9	-24.9	-3.2
GERMAN	-11.5	-5.8	-20.4	-3.7
FRENCH	-17.6	-5.4	-24.5	-4.0

Table 8: Percentage change in average accuracy of BLOOM per dataset (Fixed F.T. - Full F.T.)

Dataset Model	MCL	XL	Hindi	Am^{2}iCO
MULTI	-5.9	-2.2	12.8	0.5
ENGLISH	-1.9	3.5	3.6	-2.1
GERMAN	-1.4	-1.5	-18.3	0.7
FRENCH	-1.1	-0.8	4.8	-2.6

Table 9: Percentage change in average accuracy of LLaMA per dataset (Fixed F.T. - Full F.T.)

C Linguistic Similarity

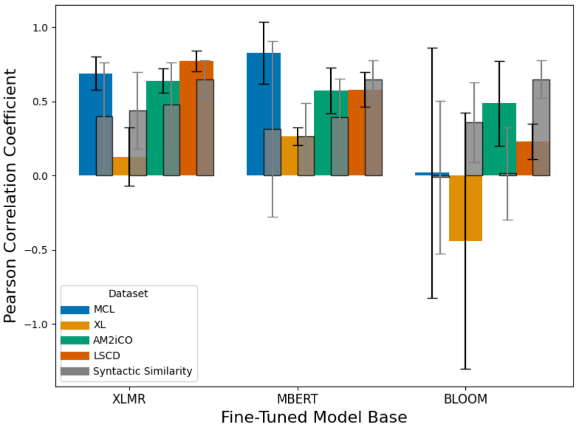

Figure 5: Mean correlations of syntactic similarity (between the fine-tuning language of MONO models and the target language) and zero-shot performance on the target languages. Superimposed on the correlation reported in Figure 2) for comparison to correlations with pretraining sizes.

D Zero-Shot Comparison of mBERT and BLOOM Models

Language	German	French	French	English	English	Hindi	Chinese	Chinese
Dataset	XL	MCL	XL	MCL	XL	Hindi	MCL	XL
Z.S. MULTI$_{\otimes =}$	62.9	74.9	62.7	69.9	**64.6**	58.4	65.8	64.5
GERMAN$_{=}$	73.9	72.2	61.3	**73.6**	56.3	51.0	67.7	63.9
FRENCH$_{=}$	**66.2**	66.0	65.9	72.5	62.8	46.2	58.2	60.4
ENGLISH$_{=}$	65.6	**78.3**	**62.8**	84.3	63.9	**73.4**	**73.0**	**68.1**
HINDI$_{=}$	55.7	67.0	56.4	69.6	55.9	76.1	64.1	60.7
CHINESE$_{=}$	61.9	72.7	63.4	78.1	61.4	59.7	70.8	68.0

Table 10: mBERT (Fixed F.T.) Zero-shot performance of MULTI models (Z.S. MULTI) and monolingual models not trained on the target language. Grey cells indicate full-shot scenarios. All trained on 8,750 examples.

Language	German	French	French	English	English	Hindi	Chinese	Chinese
Dataset	XL	MCL	XL	MCL	XL	Hindi	MCL	XL
Z.S. MULTI$_{\otimes =}$	48.6	54.1	52.9	**54.1**	**53.9**	**57.9**	50.0	50.1
GERMAN$_{=}$	61.3	49.4	50.7	49.7	51.0	48.4	50.0	50.0
FRENCH$_{=}$	50.3	52.3	60.4	51.1	53.1	45.1	50.0	50.0
ENGLISH$_{=}$	51.2	**67.4**	**57.1**	74.6	58.6	48.2	**55.6**	**52.8**
HINDI$_{=}$	49.7	50.0	50.1	50.0	49.9	88.6	50.0	50.0
CHINESE$_{=}$	49.8	50.0	50.1	50.0	50.0	46.2	54.5	59

Table 11: BLOOM (Fixed F.T.) Zero-shot performance of MULTI models (Z.S. MULTI) and MONO models not trained on the target language. Grey cells indicate full-shot scenarios. All trained on 8,750 examples.

Language	German	French	French	English	English	Hindi	Chinese	Chinese
Dataset	XL	MCL	XL	MCL	XL	Hindi	MCL	XL
Z.S. MULTI$_{\otimes =}$	66.7	**77.6**	67.4	74.1	68.4	**71.5**	72.7	68.9
GERMAN$_{=}$	76.6	77.1	**68.0**	**77.5**	64.9	47.2	**75.9**	**71.9**
FRENCH$_{=}$	68.6	70.5	65.8	77.1	67.8	64.5	74.5	71.6
ENGLISH$_{=}$	63.6	74.2	64.5	84.0	70.2	66.2	71.1	64.4
HINDI$_{=}$	59.8	66.4	56.8	72.7	64.1	72.9	69.4	64.7
CHINESE$_{=}$	64.0	70.9	62.1	70.7	**68.7**	68.5	71.5	70.2

Table 12: LLaMA (Fixed F.T.) Zero-shot performance of MULTI models (Z.S. MULTI) and MONO models not trained on the target language. Grey cells indicate full-shot scenarios. All trained on 8,750 examples.

E On the Lack of Extensive Hyperparameter Tuning

To keep cross-model and cross-language comparisons fair and interpretable, we deliberately avoided extensive per-model or per-language hyperparameter tuning. Instead, we adopted a single configuration taken from independent prior work and applied it uniformly across all models and settings, with an identical early-stopping rule. This choice limits researcher degrees of freedom and reduces the risk of inadvertently "tuning into" our research hypothesis, which can happen when many parameters are adjusted differently across conditions. While more aggressive tuning could raise the peak performance of individual systems, it would blur causal attribution and undermine the comparability that our study seeks to emphasize.

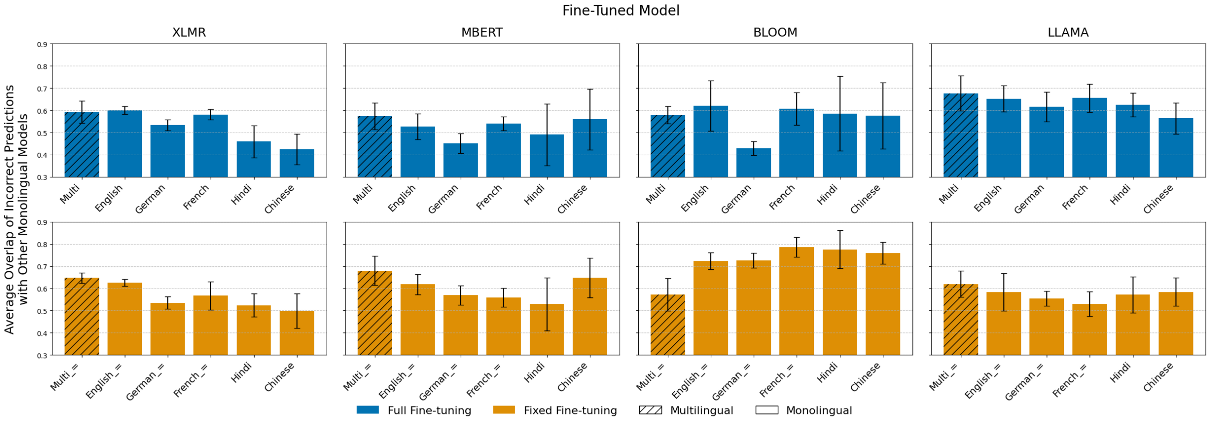

Figure 6: Average proportion of each model's incorrect predictions on the WiC task (across all test files) that are also made by monolingual models in the same fine-tuning condition, as defined in Equation F. Multilingual models are indicated with diagonal hatching. The same Hindi and Chinese models are included in both conditions due to their limited fine-tuning size.

F Error Analysis

We perform error analysis to further analyze how, and if multilingual models differ from their monolingual counterparts.

Let:

- M be the set of **monolingual models** in the same fine-tuning condition,
- m be the model of interest, where $m \in M \cup \{\textsc{multi}\}$,
- $\mathrm{Err}(m)$ be the set of incorrect predictions made by model m,
- $\mathrm{Align}(m, m') = \frac{|\mathrm{Err}(m) \cap \mathrm{Err}(m')|}{|\mathrm{Err}(m)|}$, the proportion of m's errors also made by model m'.

Then, the **average proportion of alignment** of model m with the monolingual models is given by:

$$\mathrm{AvgAlign}(m) = \begin{cases} \frac{1}{|M|-1} \sum_{\substack{m' \in M \\ m' \neq m}} \frac{|\mathrm{Err}(m) \cap \mathrm{Err}(m')|}{|\mathrm{Err}(m)|}, & \text{if } m \in M \\ \frac{1}{|M|} \sum_{m' \in M} \frac{|\mathrm{Err}(m) \cap \mathrm{Err}(m')|}{|\mathrm{Err}(m)|}, & \text{if } m = \textsc{multi} \end{cases}$$

G Pretraining Sizes of Languages Used in Our Analysis

ISO Code	Language	XLM-R	mBERT	BLOOM	MuRIL
AR	Arabic	3.37	0.72	4.26	-
BG	Bulgarian	4.07	0.24	-	-
BN	Bengali	2.24	0.12	2.91	-
DA	Danish	3.84	0.24	-	-
DE	German	4.21	1.66	-	-
EN	English	5.71	2.89	6.12	3.30
ES	Spanish	3.99	1.66	5.10	-
ET	Estonian	1.96	0.24	-	-
EU	Basque	1.10	0.12	1.16	-
FA	Persian	4.72	0.43	-	-
FI	Finnish	4.01	0.43	-	-
FR	French	4.06	1.66	5.27	-
HI	Hindi	3.05	0.12	3.18	1.95
HR	Croatian	3.07	0.24	-	-
ID	Indonesian	5.01	0.43	2.98	-
IT	Italian	3.44	1.14	-	-
JA	Japanese	4.25	1.14	-	-
KA	Georgian	2.31	0.12	-	-
KK	Kazakh	2.00	0.12	-	-
KO	Korean	4.01	0.43	-	-
LA	Latin	1.25	0.06	-	-
NL	Dutch	3.41	0.72	-	-
NO	Norwegian	3.91	0.43	-	-
RU	Russian	5.63	1.66	-	-
SV	Swedish	2.57	0.72	-	-
TR	Turkish	3.09	0.43	-	-
UR	Urdu	1.90	0.12	1.28	1.00
ZH	Simplified Chinese	3.87	1.14	5.50	-
ZH	Traditional Chinese	2.87	1.14	0.54	-

Table 13: Models' log-transformed pretraining sizes (originally in GB) of languages used in our analysis. Language proportions for LLaMA-3–8B-Instruct's pretraining corpus are not publicly available.

H All Results on Monolingual and Cross-Lingual WiC tasks

Type	Dataset →	MCL					XL													Hindi
	Model \ Language	AR	EN	FR	RU	ZH	BG	DA	DE	EN	ET	FA	FR	HR	IT	JA	KO	NL	ZH	HI
MULTI	MULTI	82.6	88.0	82.6	84.3	85.5	83.0	76.0	83.7	70.4	62.3	66.6	78.0	70.6	72.5	66.7	78.4	77.1	79.4	79.3
MONO	GERMAN	82.4	85.2	81.0	82.9	82.1	69.0	72.1	84.4	67.6	62.3	62.0	67.3	65.2	65.7	62.1	68.2	73.0	73.4	76.7
	FRENCH	80.8	85.5	81.3	81.4	81.3	67.5	73.9	71.0	67.9	60.0	67.9	78.1	65.7	69.3	63.0	66.7	76.3	73.8	78.8
	ENGLISH	82.4	90.2	84.6	85.1	84.4	70.8	73.8	67.8	70.6	60.0	62.8	65.9	70.8	64.2	62.9	73.5	74.5	76.2	78.7
	HINDI	56.1	66.0	53.8	62.5	60.7	61.7	61.0	62.1	56.8	50.3	53.8	59.7	48.5	52.2	51.9	58.2	61.2	66.4	88.9
	CHINESE	62.7	68.1	61.3	61.9	72.6	57.3	62.4	62.3	61.3	55.1	72.4	58.1	58.8	59.8	57.0	58.2	63.4	68.8	44.2
FIXED FINE-TUNING	MULTI$_=$	77.6	83.1	80.7	77.6	79.0	69.0	70.3	75.1	69.0	63.1	73.8	67.2	69.6	65.2	65.5	70.2	68.7	73.2	76.6
	GERMAN$_=$	75.9	78.3	77	73.2	75	67.1	65.8	74.1	60.6	57.4	60.6	63.8	64.5	63	61.5	61.8	66.9	67.5	74.7
	FRENCH$_=$	70.5	77	73.7	70.6	71.2	59	64.9	68.6	63	56.9	72	65.5	63.5	63.5	61.5	59.2	64.6	66.9	59.3
	ENGLISH$_=$	82.8	89.2	84.1	83.4	84.0	71.9	71.8	69.3	68.9	58.2	66.6	65.6	73.5	62.7	61.5	68.5	72.6	75.1	79.9
	HINDI$^*_=$	56.1	66.0	53.8	62.5	60.7	61.7	61.0	62.1	56.8	50.3	53.8	59.7	48.5	52.2	51.9	58.2	61.2	66.4	88.9
	CHINESE$^*_=$	62.7	68.1	61.3	61.9	72.6	57.3	62.4	62.3	61.3	55.1	72.4	58.1	58.8	59.8	57.0	58.2	63.4	68.8	44.2
Z.S. MULTI	GERMAN$_{\otimes=}$	75.9	82.9	77.7	77.9	78.4	68.2	68.3	67.6	68.1	59	69.2	67.6	69.1	66	63.6	67.4	67.4	71.6	79
	FRENCH$_{\otimes=}$	74.2	81.1	79	77.8	76.4	70.6	68.7	70.7	68.5	59.5	66.9	65.3	70.3	62.3	61.7	67.2	71	71.1	76.4
	ENGLISH$_{\otimes=}$	76.6	82.9	78.6	74.5	77.7	70.7	68	73.6	68.7	63.1	75.9	65.8	73	64.9	61.7	68.4	68	71.6	76.8
	CHINESE$_{\otimes=}$	80.1	83.3	81.2	77.3	78.3	70.7	69.8	71.9	67.9	58.5	66.5	66.8	70.8	66.6	62.5	68.6	70.4	72.2	79.9

Table 14: XLM-R models' accuracies on all monolingual WiC tasks. Full-shot conditions where the fine-tuning included the target language are highlighted in yellow.

Type	Dataset →	AM^{2}iCO													
	Model \ Language	AR	BN	DE	EU	FI	ID	JA	KA	KK	KO	RU	TR	UR	ZH
MULTI	MULTI	68.0	65.7	69.3	65.4	70.0	67.8	65.8	67.2	63.2	66.9	69.9	67.4	64.5	66.6
MONO	GERMAN	64.5	63.9	67.8	66.6	68.3	67.1	64.4	64.2	62.5	66.9	67.7	66.4	61.2	65.9
	FRENCH	67.7	68.0	70.0	66.7	72.5	68.2	68.3	68.9	61.8	67.9	72.4	68.2	61.5	66.5
	ENGLISH	67.4	66.1	69.9	66.2	73.0	68.7	68.5	66.5	61.8	68.9	71.8	68.7	66.0	67.6
	HINDI	58.4	51.3	60.6	55.0	60.1	54.8	61.5	52.3	52.2	61.5	60.3	58.0	55.8	59.7
	CHINESE	56.0	50.1	59.7	57.0	58.6	56.4	60.0	53.0	55.0	58.9	57.1	57.7	55.5	59.9
FIXED FINE-TUNING	MLCL$_=$	61.0	62.7	66.3	62.1	67.2	64.1	64.8	61.0	58.2	63.2	65.9	62.4	60.5	61.8
	GERMAN$_=$	60.2	57.6	64.6	61.9	63.6	61.2	61.6	63.2	60.2	59.4	64.4	61.1	57.5	61.3
	FRENCH$_=$	60.7	52.6	63.4	57.4	63.4	60.0	62.1	57.0	56.2	59.8	62.0	59.4	57.0	56.1
	ENGLISH$_=$	65.6	62.7	70.0	63.6	71.1	67.6	67.3	64.5	60.2	67.9	70.8	66.6	66.2	65.6
	HINDI$^*_=$	58.4	51.3	60.6	55.0	60.1	54.8	61.5	52.3	52.2	61.5	60.3	58.0	55.8	59.7
	CHINESE$^*_=$	56.0	50.1	59.7	57.0	58.6	56.4	60.0	53.0	55.0	58.9	57.1	57.7	55.5	59.9
Z.S. MULTI	GERMAN$_{\otimes=}$	68.2	68.0	70.9	68.7	75.3	70.4	69.0	70.3	64.8	69.8	73.8	68.2	65.8	66.6
	FRENCH$_{\otimes=}$	72.0	69.7	72.5	69.9	75.3	72.4	72.6	72.6	68.0	73.1	74.6	72.7	68.2	69.1
	ENGLISH$_{\otimes=}$	62.7	60.4	64.9	59.3	66.5	62.5	62.9	60.8	57.8	63.3	64.7	62.1	62.2	62.8
	CHINESE$_{\otimes=}$	71.1	72.3	72.3	68.1	76.9	70.1	69.9	70.4	68.0	72.5	74.4	72.2	68.8	67.5

Table 15: XLM-R models' accuracies on inter-lingual WiC tasks. Each language in the AM^{2}iCO dataset is paired with English, making the tasks inter-lingual.

Type	Dataset →	MCL					XL													Hindi
	Model \ Language	AR	EN	FR	RU	ZH	BG	DA	DE	EN	ET	FA	FR	HR	IT	JA	KO	NL	ZH	HI
MULTI	MULTI	75.1	81.4	76.2	76.7	74.8	79.5	65.9	84.8	64.1	56.2	64.4	74.6	68.6	72.0	61.2	67.1	69.7	72.2	70.7
MONO	GERMAN	65.1	72.5	68.4	68.2	67.6	55.6	59.3	82.9	57.6	55.4	59.1	58.3	57.1	57.1	54.6	54.2	57.5	61.6	69.3
	FRENCH	71.6	74.6	71.7	71.9	69.7	57.9	62.2	63.6	60.4	58.7	60.1	76.2	56.9	63.5	57.9	57.7	65.2	65.5	69.3
	ENGLISH	73.5	82.1	76.1	73.2	72.7	57.9	63.3	61.9	64.1	52.8	62.1	61.5	61.8	59.6	57.6	54.8	63.4	66.4	73.3
	HINDI	64.7	69.6	67	65.5	64.1	51.1	57.5	55.7	55.9	48.2	59.8	56.4	53.2	52.9	50.7	51	52	60.7	76.1
	CHINESE	68.7	78.1	72.7	69.1	70.8	56.4	63	61.9	61.4	56.7	65.9	63.4	60.5	60	56.9	55.9	63.6	68	59.7
FIXED FINE-TUNING	MULTI$_=$	71.5	79.3	76.1	70.6	70.2	63.9	62.1	72.4	64.2	60.3	66	65.4	66.7	63.7	60.9	58.8	63.6	65.7	58.4
	GERMAN$_=$	69.4	73.6	72.2	67.1	67.7	59.6	59.2	73.9	56.3	56.9	63.9	61.3	59.3	61.8	58.7	58.3	57.3	63.9	51.0
	FRENCH$_=$	61.4	72.5	66.0	62.9	58.2	61.2	62.8	66.2	62.8	57.2	67.5	65.9	64.5	62.8	55.7	58.3	64.1	60.4	46.2
	ENGLISH$_=$	74.6	84.3	78.3	75.4	73	62.5	62.3	65.6	63.9	54.1	64.6	62.8	67.9	60	58.7	59.4	64.8	68.1	73.4
	HINDI$^*_=$	64.7	69.6	67	65.5	64.1	51.1	57.5	55.7	55.9	48.2	59.8	56.4	53.2	52.9	50.7	51	52	60.7	76.1
	CHINESE$^*_=$	68.7	78.1	72.7	69.1	70.8	56.4	63	61.9	61.4	56.7	65.9	63.4	60.5	60	56.9	55.9	63.6	68	59.7
Z.S. MULTI	GERMAN$_{\otimes=}$	72.1	76.2	74.4	69.3	69.2	61.3	61.9	62.9	64.5	54.4	65.0	63.2	68.9	61.7	59.5	60.4	62.0	65.9	61.5
	FRENCH$_{\otimes=}$	65.5	74.2	74.9	70.0	67.5	63.0	61.5	68.2	64.1	56.2	65.8	62.7	66.7	60.5	57.9	58.5	60.5	63.3	64.7
	ENGLISH$_{\otimes=}$	57.3	69.9	63.8	62.2	63.8	61.2	59.1	69.4	64.6	56.7	63.2	64.2	60.0	62.7	59.0	60.1	64.9	63.2	43.8
	CHINESE$_{\otimes=}$	66.2	73.9	71.6	68.0	65.8	63.4	61.7	69.1	64.3	57.9	62.8	62.2	65.7	62.5	60.7	60.0	64.1	64.5	58.3

Table 16: mBERT models' accuracies on all monolingual WiC tasks. Full-shot conditions where the fine-tuning included the target language are highlighted in yellow.

Type	Dataset →	AM^{2}iCO													
	Model \ Language	AR	BN	DE	EU	FI	ID	JA	KA	KK	KO	RU	TR	UR	ZH
MULTI	MULTI	60.5	61.6	64.9	59.1	64.4	62.7	60.4	59.9	57.2	59.7	64.8	61.1	57.8	60.9
MONO	GERMAN	56.7	60.1	61.9	59.6	61.9	61.7	59.7	58.1	59.2	57.8	62.3	61.7	57.2	58.7
	FRENCH	58.5	57	62.4	61	64.4	64.1	62.6	58	55	59.6	64.2	62.4	57.5	60.7
	ENGLISH	59.8	60	63.8	60.6	63.9	64.7	61.5	61.2	59	62.4	63.9	61.6	61.5	61.5
	HINDI	51.2	50.9	57.2	50.1	50.8	52.2	49.9	50.2	49.2	50.8	53.8	50.8	51	50
	CHINESE	53	52.7	59.7	50.2	52.7	55.9	52.9	51	48.8	52.9	58.8	52.8	50.7	51.5
FIXED FINE-TUNING	MLCL$_=$	58.3	60.4	64.7	56.5	61.1	61.0	59.0	55.5	53.0	58.1	61.7	58.7	58.2	55.6
	GERMAN$_=$	55.8	55.9	63.5	55.1	59.1	60.6	56.6	54.6	51.0	57.3	60.0	57.8	53.5	53.4
	FRENCH$_=$	58.9	57.7	61.7	56.6	60.4	59.0	57.8	54.5	53.5	58.5	60.3	58.6	56.5	56.9
	ENGLISH$_=$	62.8	62.6	66.6	60.3	65.3	65.1	63.7	62	59.5	63.9	65.7	63.1	63	62.8
	HINDI$^*_=$	51.2	50.9	57.2	50.1	50.8	52.2	49.9	50.2	49.2	50.8	53.8	50.8	51	50
	CHINESE$^*_=$	53	52.7	59.7	50.2	52.7	55.9	52.9	51	48.8	52.9	58.8	52.8	50.7	51.5
Z.S. MULTI	GERMAN$_{\otimes=}$	62.3	63.1	66.4	61.9	68.2	63.8	64.5	63.0	60.0	64.6	65.8	65.2	61.2	65.6
	FRENCH$_{\otimes=}$	65.4	64.6	68.9	64.8	71.0	69.1	66.1	64.4	65.2	64.8	67.5	68.6	64.2	67.6
	ENGLISH$_{\otimes=}$	59.2	57.9	60.8	56.8	60.4	60.4	60.8	57.0	55.8	58.5	61.0	60.0	57.0	60.1
	CHINESE$_{\otimes=}$	63.3	66.0	69.0	62.8	68.6	65.1	64.0	64.9	60.2	65.7	65.8	67.1	59.5	65.5

Table 17: mBERT models' accuracies on inter-lingual WiC tasks.

Type	Dataset →	MCL					XL													Hindi
	Model \ Language	AR	EN	FR	RU	ZH	BG	DA	DE	EN	ET	FA	FR	HR	IT	JA	KO	NL	ZH	HI
MULTI	MULTI	71.4	73.9	65.8	54.8	75.6	74.0	56.1	80.8	62.3	58.7	55.1	72.3	54.2	66.7	57.5	69.3	59.5	71.0	70.9
MONO	GERMAN	62.9	66.1	62.3	55.2	66.1	51.8	54.9	82.9	58.4	55.9	54	57.5	51	53.9	56.4	56.2	55.5	61.2	68.8
	FRENCH	72.7	74.9	68.4	54.6	72.8	52.8	50.1	52.7	61	52.3	55.9	74.4	50	57.9	54.4	56.4	52.8	66.2	69.7
	ENGLISH	68.7	74.1	70.4	53.3	68.8	50.1	50.4	50.2	62.4	51	52.9	60.1	50.7	52	52.9	50.1	49.4	59.1	73.1
	HINDI	55	50	50	50	50	50	50	49.7	49.9	50	49.2	50.1	49.8	50	50	50.9	50	50	88.6
	CHINESE	61.8	50	50	50	54.5	49.5	50	49.8	50	49.5	53	50.1	49	50.2	50	52	49.9	59	46.2
FIXED FINE-TUNING	MULTI$_=$	64	59.2	54.9	50.7	55.2	53.3	50.7	61.7	53.9	56.2	52.2	57.3	53.4	58.6	53.5	56.8	56.3	55.7	57.9
	GERMAN$_=$	57.3	49.7	49.4	48.5	50.0	49.9	50.6	61.3	51.0	46.9	50.5	50.7	54.2	50.8	50.2	54.7	53.5	50.0	48.4
	FRENCH$_=$	52.2	51.1	52.3	50.0	50.0	49.7	49.6	50.3	53.1	49.0	52.2	60.4	48.3	53.7	50.1	50.5	49.7	50.0	45.1
	ENGLISH$_=$	60.8	74.6	67.4	50.5	55.6	50.7	51.5	51.2	58.6	50.3	50	57.1	50	54.6	51.1	52.7	49.2	52.8	48.2
	HINDI$^*_=$	55	50	50	50	50	50	50	49.7	49.9	50	49.2	50.1	49.8	50	50	50.9	50	50	88.6
	CHINESE$^*_=$	61.8	50	50	50	54.5	49.5	50	49.8	50	49.5	53	50.1	49	50.2	50	52	49.9	59	46.2
Z.S. MULTI	GERMAN$_{\otimes=}$	52.5	55.5	54.2	48.9	54.5	51.8	50.3	48.6	57.5	51.5	53.2	57.1	52.7	58.3	53.2	58.8	54.0	52.8	54.7
	FRENCH$_{\otimes=}$	55.9	52.6	50.7	51.8	52.9	47.5	51.2	55.7	55.9	53.8	49.4	52.9	52.7	58.4	53.2	56.0	50.5	52.9	48.1
	ENGLISH$_{\otimes=}$	62.8	54.1	54.3	52.9	57.1	50.9	51.7	58.3	53.9	56.2	50.7	56.7	49.5	54.2	53.4	57.7	52.6	56.0	44.4
	CHINESE$_{\otimes=}$	50.5	51.0	54.7	50.1	50.0	51.4	49.9	57.2	54.8	54.6	51.2	56.0	53.7	57.4	51.1	56.0	54.4	50.1	43.7

Table 18: BLOOM models' accuracies on all monolingual WiC tasks. Full-shot conditions where the fine-tuning included the target language are highlighted in yellow.

Type	Dataset →	AM^{2}iCO													
	Model \ Language	AR	BN	DE	EU	FI	ID	JA	KA	KK	KO	RU	TR	UR	ZH
MULTI	MULTI	58.0	56.6	55.2	54.0	52.3	57.4	56.2	52.4	53.2	53.6	51.8	52.8	55.0	59.6
MONO	GERMAN	55.2	55.6	52	54	52.6	53.4	55.5	51.8	57.2	49.3	53	53.3	54	55
	FRENCH	57.8	54.1	55.5	51.3	50.9	58.4	53.6	49.1	50.7	51.4	51.2	50.4	57.5	60.8
	ENGLISH	59.2	54.7	55.8	52.8	53.4	56.3	55.4	51.4	51.8	49.9	52.6	50.3	55.8	55.1
	HINDI	50	50	50	49	50	50.1	50.1	50.1	49.8	50.4	50.1	50	50	50.4
	CHINESE	50	50	50	47.5	49.9	50	51.4	50.1	49.5	49.8	50.9	49.9	50	48.2
FIXED FINE-TUNING	MULTI$_=$	49.9	50.0	51.0	50.1	50.3	51.0	51.1	50.1	49.8	49.9	49.8	48.2	50.0	49.4
	GERMAN$_=$	50.0	50.0	50.0	50.2	49.2	50.3	50.7	50.1	50.0	48.9	48.9	49.8	50.0	51.5
	FRENCH$_=$	50.0	50.0	50.6	49.2	49.4	50.0	49.6	50.0	50.2	48.5	49.2	49.9	50.0	49.5
	ENGLISH$_=$	55.5	50	52.3	50.7	51.5	50.4	50	50.1	50	49.7	50.1	49.5	50	49.9
	HINDI$^*_=$	50	50	50	49	50	50.1	50.1	50.1	49.8	50.4	50.1	50	50	50.4
	CHINESE$^*_=$	50	50	50	47.5	49.9	50	51.4	50.1	49.5	49.8	50.9	49.9	50	48.2
Z.S. MULTI	GERMAN$_{\otimes=}$	52.2	49.6	51.4	51.1	50.0	51.4	50.8	52.5	53.0	50.5	49.2	51.3	50.2	53.1
	FRENCH$_{\otimes=}$	51.0	51.4	51.5	52.7	50.0	52.7	52.8	53.3	49.2	51.1	50.8	52.5	51.5	49.0
	ENGLISH$_{\otimes=}$	50.1	50.0	49.8	50.0	51.4	50.7	50.4	50.0	50.0	49.9	50.0	50.4	50.0	47.5
	CHINESE$_{\otimes=}$	52.7	50.1	54.1	50.9	51.3	49.4	50.6	50.5	50.2	51.5	48.1	51.9	50.2	49.5

Table 19: BLOOM models' accuracies on inter-lingual WiC tasks.

Type	Dataset →	MCL					XL													Hindi
	Model \ Language	AR	EN	FR	RU	ZH	BG	DA	DE	EN	ET	FA	FR	HR	IT	JA	KO	NL	ZH	HI
MULTI	MULTI	71.8	85.6	76.5	78.1	80.2	69.5	67.9	77.9	71.6	62.1	79.0	74.2	69.4	70.9	66.9	73.7	75.1	72.8	58.7
MONO	GERMAN	71.5	78.9	77.0	75.9	77.9	64.3	68.9	80.0	70.2	53.3	69.0	69.0	67.6	64.4	63.6	65.6	71.2	72.4	65.5
	FRENCH	70.4	78.7	72.7	71.7	75.3	62.8	67.0	69.9	69.6	53.3	69.2	76.5	63.0	67.6	63.1	68.6	72.7	70.3	59.7
	ENGLISH	70.3	85.0	76.9	74.8	70.9	54.9	62.3	60.3	67.7	50.5	63.2	63.2	55.1	57.1	53.2	55.6	59.8	65.4	62.6
	HINDI	64.8	72.7	66.4	68.7	69.4	56.4	58.3	59.8	64.1	50.8	67.6	56.8	60.0	58.4	57.5	58.2	62.8	64.7	72.9
	CHINESE	69.3	70.7	70.9	71.5	71.5	64.8	63.7	64.0	68.7	56.7	70.9	62.1	67.2	61.7	60.1	66.4	65.4	70.2	68.5
FIXED FINE-TUNING	MULTI$_=$	73.4	79.8	74.9	73.2	73.2	66.5	65.2	75.6	71.0	62.1	80.1	70.6	70.6	67.4	64.0	69.0	68.8	71.5	71.5
	GERMAN$_=$	70.9	77.5	77.1	71.7	75.9	60.5	65.3	76.6	64.9	57.4	74.1	68.0	66.2	64.4	61.2	59.8	69.2	71.9	47.2
	FRENCH$_=$	70.7	77.1	70.5	69.8	74.5	65.4	63.4	68.6	67.8	60.5	76.9	65.8	64.5	63.3	63.6	68.4	63.7	71.6	64.5
	ENGLISH$_=$	68.5	84.0	74.2	74.7	71.1	62.7	62.4	63.6	70.2	55.1	63.2	64.5	59.8	61.0	58.6	62.8	65.9	64.4	66.2
	HINDI$^*_=$	64.8	72.7	66.4	68.7	69.4	56.4	58.3	59.8	64.1	50.8	67.6	56.8	60.0	58.4	57.5	58.2	62.8	64.7	72.9
	CHINESE$^*_=$	69.3	70.7	70.9	71.5	71.5	64.8	63.7	64.0	68.7	56.7	70.9	62.1	67.2	61.7	60.1	66.4	65.4	70.2	68.5
Z.S. MULTI	GERMAN$_{\otimes=}$	63.2	81.4	74.8	70.8	70.6	64.6	64.3	66.7	71.4	62.1	78.1	70.4	66.9	66.2	62.0	66.4	71.4	65.7	67.6
	FRENCH$_{\otimes=}$	69.3	78.9	77.6	74.1	76.6	65.7	63.6	74.2	72.2	63.1	77.5	67.4	67.6	65.5	64.6	69.6	69.7	70.4	61.6
	ENGLISH$_{\otimes=}$	61.6	74.1	69.9	67.2	69.6	63.2	64.2	76.3	68.4	58.5	73.9	69.4	63.5	66.4	62.1	68.7	69.6	67.6	56.7
	CHINESE$_{\otimes=}$	70.4	78.3	72.8	71.7	72.7	63.1	62.9	72.0	68.2	65.1	71.1	67.8	68.9	62.7	62.7	67.5	68.8	68.9	62.9

Table 20: LLaMA models' accuracies on all monolingual WiC tasks. Full-shot conditions where the fine-tuning included the target language are highlighted in yellow.

Type	Dataset →	AM^{2}iCO													
	Language Model	AR	BN	DE	EU	FI	ID	JA	KA	KK	KO	RU	TR	UR	ZH
MULTI	MULTI	61.1	59.7	63.5	58.1	60.0	59.8	59.7	57.4	56.0	61.7	60.3	58.6	55.2	61.5
MONO	GERMAN	59.9	55.0	60.6	55.2	57.7	57.3	60.0	55.0	56.8	60.2	58.3	58.3	54.5	57.3
	FRENCH	58.5	58.0	62.5	57.1	63.6	58.8	59.4	56.5	53.2	60.5	60.3	60.2	57.2	59.0
	ENGLISH	61.9	61.0	63.9	57.6	63.9	60.1	62.0	59.5	57.0	63.6	62.3	60.1	59.5	62.1
	HINDI	59.8	59.7	62.1	56.1	60.0	59.8	59.1	55.8	55.2	59.5	60.5	57.0	55.5	58.6
	CHINESE	59.0	56.4	62.2	55.4	59.4	60.1	58.8	53.2	53.5	56.3	59.8	59.3	55.5	58.5
FIXED FINE-TUNING	MULTI$_=$	61.6	58.9	64.6	58.5	63.8	61.1	60.2	58.2	55.0	59.0	59.5	60.0	58.0	61.3
	GERMAN$_=$	57.1	57.7	60.3	56.7	60.5	59.3	59.5	58.5	55.2	60.8	58.9	57.5	55.8	58.0
	FRENCH$_=$	56.3	57.6	58.3	57.5	59.4	58.6	55.9	52.9	52.0	55.3	56.9	56.4	56.0	55.4
	ENGLISH$_=$	60.1	59.1	61.1	56.5	60.7	58.6	58.1	60.3	56.0	59.2	60.4	56.9	58.5	59.6
	HINDI$^*_=$	59.8	59.7	62.1	56.1	60.0	59.8	59.1	55.8	55.2	59.5	60.5	57.0	55.5	58.6
	CHINESE$^*_=$	59.0	56.4	62.2	55.4	59.4	60.1	58.8	53.2	53.5	56.3	59.8	59.3	55.5	58.5
Z.S. MULTI	GERMAN$_{\otimes=}$	73.1	73.0	78.9	72.1	78.6	76.3	75.8	72.4	71.5	76.0	80.0	75.8	69.8	76.3
	FRENCH$_{\otimes=}$	71.3	68.7	72.5	69.6	73.8	71.1	70.1	69.7	68.5	72.2	74.1	71.6	67.5	70.8
	ENGLISH$_{\otimes=}$	57.7	56.0	59.3	57.2	59.9	59.0	58.9	56.3	55.2	59.9	58.7	57.3	58.2	59.7
	CHINESE$_{\otimes=}$	74.6	72.7	76.3	71.0	76.1	74.5	74.7	72.3	68.5	74.2	76.3	74.3	70.5	74.9

Table 21: LLaMA models' accuracies on inter-lingual WiC tasks.

Type	Dataset →	MCL					XL													Hindi
	Language Model	AR	EN	FR	RU	ZH	BG	DA	DE	EN	ET	FA	FR	HR	IT	JA	KO	NL	ZH	HI
MONO	HINDI	56.8	56.6	50.9	50.1	49.2	50.3	52.7	49.5	56.3	51.5	50.5	51.3	50.5	57.4	52.7	53.0	53.5	48.8	86.5
	ENGLISH	56.2	86.6	64.5	50.7	49.1	50.2	52.4	52.5	70.4	51.0	49.1	56.0	46.3	53.4	50.0	50.0	51.0	50.1	81.4
	FRENCH	52.7	61.2	56.7	50.3	50.0	52.8	56.1	48.8	58.7	54.9	51.5	50.9	50.7	54.2	50.8	59.6	53.8	54.0	64.0
	GERMAN	55.3	68.5	62.7	51.0	51.2	49.9	56.0	80.0	62.0	55.1	56.5	57.4	48.8	55.2	51.1	50.0	56.4	51.0	69.8
	CHINESE	50.0	50.2	52.8	50.9	54.3	49.9	51.4	51.6	51.7	52.1	51.9	51.6	53.2	55.4	50.6	58.4	51.5	52.1	56.6
MULTI	MULTI	48.3	48.4	49.9	50.2	49.0	50.7	50.1	50.9	48.8	49.0	46.8	51.6	50.5	51.2	52.7	51.5	47.2	49.8	45.9

Table 22: MuRIL models' accuracies on all monolingual WiC tasks. Full-shot conditions where the fine-tuning included the target language are highlighted in yellow.

Type	Dataset →	AM^{2}iCO													
	Language Model	AR	BN	DE	EU	FI	ID	JA	KA	KK	KO	RU	TR	UR	ZH
MONO	HINDI	49.6	50.3	50.8	49.5	52.1	48.5	47.5	49.9	49.0	50.8	50.6	50.2	50.7	50.0
	ENGLISH	53.7	71.7	65.3	55.9	57.3	59.8	49.8	49.0	49.0	49.3	51.7	54.8	70.2	48.7
	FRENCH	49.5	51.1	51.6	48.8	49.7	49.9	50.0	48.1	50.7	47.5	51.9	51.4	51.5	50.2
	GERMAN	52.6	57.6	56.9	54.8	54.4	56.2	51.1	52.2	51.8	50.2	53.7	52.5	53.8	51.1
	CHINESE	50.0	50.1	50.0	50.0	49.9	50.0	50.0	50.0	50.0	50.0	49.9	50.0	50.0	50.0
MULTI	MULTI	49.6	47.7	50.4	48.5	49.5	49.3	47.7	48.9	51.5	50.2	49.1	49.3	49.2	48.1

Table 23: MuRIL models' accuracies on inter-lingual WiC tasks.

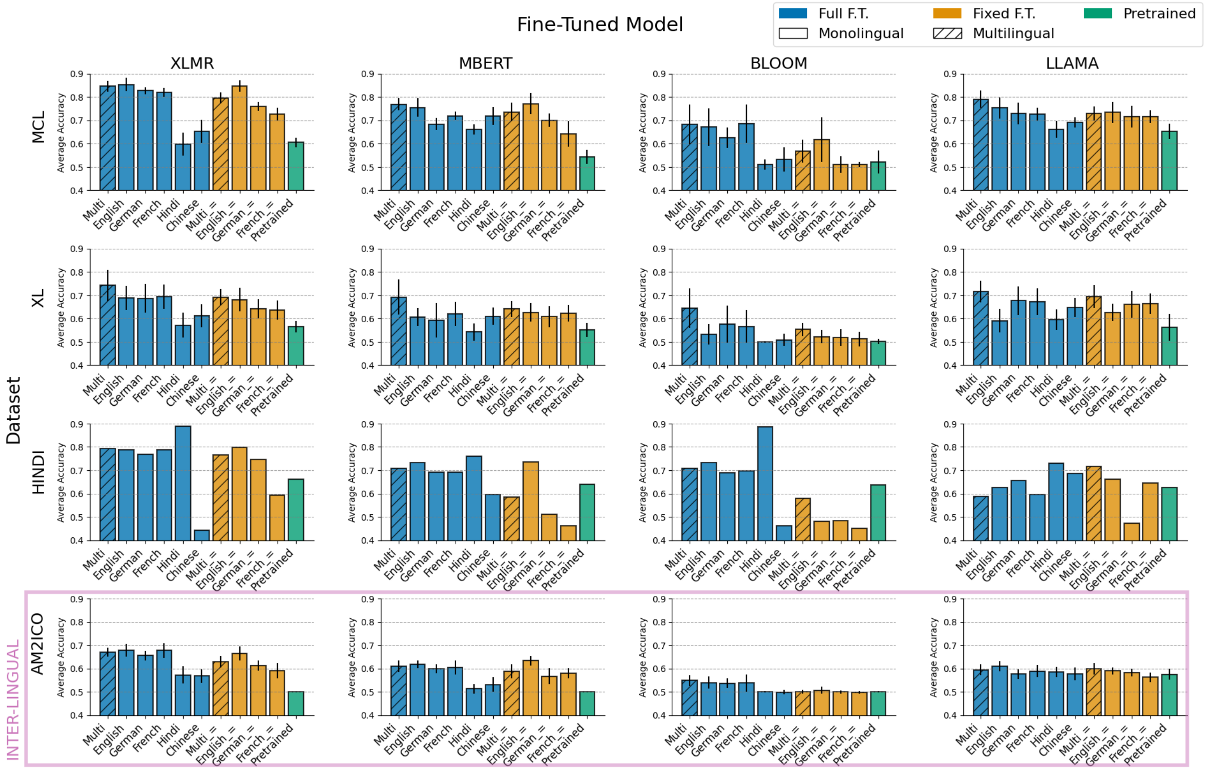

Figure 7: Mean accuracy with standard deviation for multilingual and monolingual models on WiC datasets, showing both pretrained (no-fine-tuning) and fine-tuned performance. Color encodes whether fine-tuning used the full training set, a subsampled portion, or whether the pretrained models were used off-the-shelf. Hindi and Chinese appear only in the full-data condition due to limited size. The Hindi dataset has no error bars because it is a single-language dataset.

I Performance of Pretrained Models without Fine-Tuning

It is logical to ask whether pretrained multilingual language models already separate word senses in a no-fine-tuning setting. Because contextual embeddings integrate sentence context via attention, one might expect embeddings of *mole* in "burrowing mammal" vs. "skin blemish" to diverge more than two occurrences of the same sense, even without fine-tuning. To probe this, we evaluate pretrained models without task-specific training by classifying sentence pairs using cosine distance between target-word embeddings, consistent with our main protocol.

Threshold selection and considerations. In our main experiment setup, each fine-tuned encoder uses a cosine-distance decision threshold set on dev data from its fine-tuning language (or multilingual dev for MULTI models). For encoders without fine-tuning, no model-specific dev data exists, leaving only flawed alternatives: (A) calibrate once on pooled multilingual dev data (introducing mixture bias), (B) tune separately on each test set (test leakage, inflated scores), or (C) calibrate per language (uses target-language supervision unavailable to the pretrained encoders). We adopt Option A as the least biased feasible choice. We sweep candidate thresholds on the pooled dev data to maximize accuracy and then fix that single global threshold for each encoder across all datasets. This avoids test leakage and keeps calibration constant, but the global threshold inevitably reflects mixture statistics and tailors to the average of all the data, and can favor high-resource languages. Accordingly, these numbers are a *sanity check* for learning rather than directly comparable performance estimates.

Results. Figure 7 shows that XLM-R, mBERT, and BLOOM consistently benefit from fine-tuning across nearly all conditions, with the few exceptions concentrated in low-data fine-tuning settings (e.g., Chinese or Fixed F.T.) when evaluated on Hindi. These gains indicate that fine-tuning substantially improves sense discrimination beyond what is present in the pretrained geometry.

For LLaMA, the dynamics differ slightly. The generative model is prompted (see Figure 9 for the prompt) rather than based on a cosine-distance threshold. Its zero-shot instruction-following yields a stronger baseline than the pretrained encoders on most tasks; accordingly, despite using the same WiC training data, its absolute gains from fine-tuning are smaller. This reflects an architectural/interface contrast—encoders benefit from shaping the embedding space and a calibrated decision threshold, whereas LLaMA already executes the task from the prompt and changes less with additional supervision.

We include full results for the no-fine-tuning baselines in Table 24. These baselines show that sense discrimination improves with fine-tuning; however, they are not intended for head-to-head comparison with fine-tuned models due to the necessarily different threshold calibration.

Dataset →	MCL					XL													Hindi
Language / Model	AR	EN	FR	RU	ZH	BG	DA	DE	EN	ET	FA	FR	HR	IT	JA	KO	NL	ZH	HI
XLM-R	62.2	62.2	61.6	58.7	58.0	58.5	60.6	56.1	56.1	53.1	53.9	53.1	57.0	54.8	56.9	55.7	58.7	60.0	66.1
mBERT	56.2	58.5	51.2	52.1	53.7	58.1	54.2	54.4	52.8	54.7	55.2	52.6	54.7	50.3	56.4	56.8	62.6	53.1	63.8
BLOOM	61.1	50.0	50.0	50.0	50.0	49.9	50.0	49.0	50.0	50.1	49.9	50.0	50.1	49.7	50.3	50.0	50.0	53.9	63.6
LLaMA	67.2	64.2	65.4	58.0	62.8	55.4	63.2	56.9	55.7	58.4	57.9	53.8	53.4	52.8	46.5	54.1	53.7	70.6	62.5
MuRIL	50.0	49.9	57.4	51.8	50.4	49.8	49.9	48.5	50.8	50.6	56.1	58.2	49.7	50.0	56.2	50.2	53.2	50.4	59.8

Table 24: Pretrained models' accuracies on all monolingual WiC tasks.

Dataset →	AM^{2}iCO													
Language / Model	AR	BN	DE	EU	FI	ID	JA	KA	KK	KO	RU	TR	UR	ZH
XLM-R	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	49.9	50.1	50.0	50.0
mBERT	50.0	50.0	50.1	50.1	50.0	50.0	50.0	50.0	50.0	50.0	49.9	49.9	50.0	50.0
BLOOM	50.0	50.0	50.0	50.8	49.9	50.1	50.2	50.1	49.8	49.9	49.7	49.9	50.0	49.8
LLaMA	58.3	55.7	61.6	55.3	57.3	59.3	58.7	54.7	53.2	57.9	57.7	56.8	59.0	60.3
MuRIL	50.5	49.7	50.3	50.0	49.8	50.6	49.7	50.6	50.0	49.6	49.8	50.1	52.2	50.0

Table 25: Pretrained models' accuracies on inter-lingual WiC tasks.

J Performance of XLM-R, mBERT and BLOOM-Based Models on LSCD Tasks

MuRIL models were excluded from this evaluation due to poor cross-lingual performance, resulting from a lack of pretraining on most of the target languages. LLaMA models were also not evaluated, as the LSCD task format is not well-suited to generative architectures.

Language	Metric	XL-LEXEME	MULTI	GERMAN	FRENCH	ENGLISH	HINDI	CHINESE
English	APD	.757	.703	.737	.681	**.772**	.436	.673
	PRT	.495	.492	.241	.337	**.535**	.363	.367
German	APD	**.873**	.863	.841	.867	.844	.635	.641
	PRT	.881	**.890**	.829	.831	.873	.755	.682
Swedish	APD	**.755**	.801	.754	.618	.724	.480	.489
	PRT	**.678**	.673	.522	.138	.627	.277	.332
Latin	APD	-.035	.117	**.161**	.136	.135	-.177	.091
	PRT	.467	.392	.445	.405	.429	**.512**	**.512**
Spanish	APD	.665	.696	.670	.664	**.711**	.354	.383
	PRT	.633	**.698**	.655	.649	.643	.355	.267
Chinese	APD	.734	.652	.649	.499	**.737**	.524	.593
	PRT	.702	**.708**	.623	.578	.684	.552	.432
Norwegian$_1$	APD	.668	.729	.638	.697	**.777**	.525	.400
	PRT	.769	.784	.730	.740	**.845**	.551	.435
Norwegian$_2$	APD	.634	**.655**	.604	.580	.645	.433	.439
	PRT	.532	.583	.557	.525	**.636**	.337	.396
Average	APD	.631	.652	.632	.593	**.668**	.401	.464
	PRT	.645	.652	.575	.525	**.659**	.463	.428

Table 26: Spearman correlations of XLM-R models' APD and PRT scores with graded semantic change scores across LSCD tasks, transposed to show languages as rows. Best scores are bolded; scores within 0.05 of the best are underlined.

Language	Metric	MULTI	GERMAN	FRENCH	ENGLISH	HINDI	CHINESE
English	APD	**.754**	.566	.551	.711	.506	.684
	PRT	**.470**	.397	.393	.434	.309	.328
German	APD	.760	.719	.772	**.810**	.551	.783
	PRT	.834	.759	**.846**	.845	.550	.705
Swedish	APD	.377	.473	.002	.437	.523	**.541**
	PRT	.082	.145	-.504	.266	.207	**.366**
Latin	APD	**.181**	-.092	-.058	-.158	.163	-.190
	PRT	.298	.339	.235	.038	**.462**	.412
Spanish	APD	.629	.500	.531	**.651**	.452	.536
	PRT	**.636**	.462	.546	.617	.332	.407
Chinese	APD	.654	.554	.386	**.668**	.651	.619
	PRT	**.690**	.623	.493	.686	.427	.529
Norwegian$_1$	APD	.596	.515	.561	.638	.631	**.700**
	PRT	.724	.698	.580	**.748**	.558	.601
Norwegian$_2$	APD	**.601**	.596	.464	**.604**	.412	.581
	PRT	.474	.527	.260	**.535**	.238	.421
Average	APD	**.569**	.479	.401	.545	.486	.532
	PRT	**.526**	.494	.356	.521	.385	.471

Table 27: Spearman correlations of mBERT models' APD and PRT scores with graded semantic change scores across LSCD tasks, transposed to show languages as rows. Best scores from fine-tuned models are bolded; scores within 0.05 of the best are underlined.

Language	Metric	MULTI	GERMAN	FRENCH	ENGLISH	HINDI	CHINESE
English	APD	**.689**	.463	.657	.656	.176	.291
	PRT	.469	.293	.459	**.487**	.215	.342
German	APD	**.634**	.606	.244	.410	.075	.144
	PRT	.720	**.723**	.474	.564	.274	.499
Swedish	APD	.278	**.399**	.303	.311	.276	.132
	PRT	.287	**.326**	.219	-.052	.134	-.041
Latin	APD	-.068	-.104	**.081**	-.197	-.230	-.173
	PRT	**.367**	.221	.323	.314	.125	.310
Spanish	APD	.501	.370	.474	**.599**	.326	.230
	PRT	.489	.372	.450	**.521**	.284	.178
Chinese	APD	.550	.467	.570	**.592**	.216	.299
	PRT	.546	.381	.506	**.568**	.395	.159
Norwegian$_1$	APD	.278	.289	.131	**.345**	-.002	.051
	PRT	.251	**.512**	-.015	.354	.240	.363
Norwegian$_2$	APD	**.395**	.193	.237	.089	-.109	-.046
	PRT	**.264**	.216	.261	.041	-.303	-.118
Average	APD	**.407**	.335	.337	.351	.091	.116
	PRT	**.424**	.380	.335	.350	.170	.211

Table 28: Spearman correlations of BLOOM models' APD and PRT scores with graded semantic change scores across LSCD tasks, transposed to show languages as rows. Best scores from fine-tuned models are bolded; scores within 0.05 of the best are underlined.

K LLaMA

K.1 LLaMA outperformed by XLM-R

Figure 8 shows that XLM-R generally outperforms LLaMA across most datasets under both full and fixed fine-tuning conditions, except when the models are trained on Hindi or Chinese. This can be attributed to LLaMA's instruction-following pretraining, which enables it to better generalize from limited examples, which is particularly valuable in low-resource settings. In contrast, XLM-R lacks such capabilities and must infer both the task and the language from sparse training data. This issue is exacerbated by the fact that LLaMA was provided with prompts, which were consistently in English due to tokenizer limitations, while XLM-R had no explicit indication of either the task or input language during training or evaluation except for the sparse training data.

While we include LLaMA in our evaluation to provide a comparison with recent large language models (LLMs) and instruction-tuned architectures, our results suggest that encoder-based models like XLM-R remain more effective for in-context polysemy disambiguation. Despite the generative capabilities of LLaMA, it underperforms on these tasks in most languages compared to XLM-R, particularly when ample training data is available. This highlights the continued relevance of encoder-based models, which demonstrate stronger task-specific performance and greater cross-lingual transfer.

L LLaMA Fine-Tuning Parameters

Fine-tuning details	
pre-trained LLMs	*Meta-Llama-3-8B-Instruct*
GPUs	NVIDIA A40 (48GB)
PEFT	LoRA
LoRA dropout	0.1
Weight decay	0.001
Learning rate	1e-4
LoRArank	128
LoRAalpha	256
Warmup ratio	0.05
Num train epochs	3
Gradient accumulation steps	4
Max seq. length	512
Batch size	8
Optimizer	paged_adamw_8bit
LoRA target modules	q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, down_proj

Table 29: Settings and parameters for fine-tuning Llama3Instruct.

L.1 Prompts and Response Parsing

We adapted the WiC task to a generative setting for LLaMA by using the prompt shown in Figure 9. This prompt format was used consistently during both training and evaluation. For training, the prompt was followed by the correct binary label ("1" or "0") as the target output. During evaluation, the model generated this label based on the prompt alone. We evaluated the pretrained model using several alternative prompt formulations on evaluation in all languages, ultimately selecting the one that yielded the highest overall performance.

To extract model predictions, we used the following simple rule-based parser:

```
if "1" in output and "0" not in output:
    return "1"
elif "0" in output and "1" not in output:
    return "0"
else:
    return None # counts as incorrect
```

In practice, nearly all outputs followed the expected format, with the vast majority consisting of a single "1" or "0".

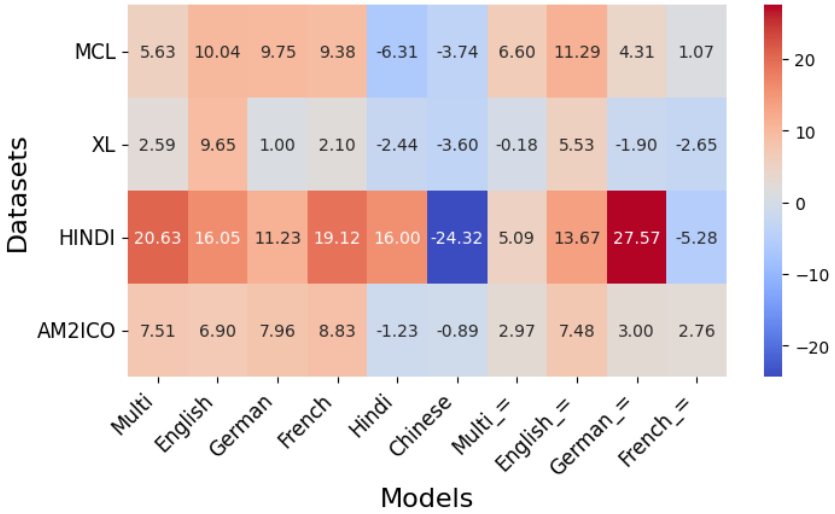

Figure 8: Differences between the per-dataset average accuracies of XLM-R based models and LLaMA-based models trained on the same data in both Full and Fixed Fine-Tuning conditions.

System message:

```
Act as an expert lexicographer: determine if a given word has the same sense in two sentences,
and respond with 1 if the sense is the same, or 0 if it is different.
```

User message:

```
Determine if "{target_word}" has the same meaning in these two sentences.
You MUST reply with ONLY:
1 – if the meaning is identical.
0 – if the meaning differs.

Sentence 1: "{sentence1}"
Sentence 2: "{sentence2}"

Answer (ONLY 1 or 0):
```

Figure 9: System and user prompt used to adapt the WiC task for LLaMA.