Capacity Constraints and the Multilingual Penalty for Lexical Disambiguation

Anonymous ACL submission

Abstract

Multilingual language models (LMs) sometimes under-perform their monolingual counterparts, possibly due to *capacity limitations*. We quantify this "multilingual penalty" for lexical disambiguation—a task requiring precise semantic representations and contextualization mechanisms—using controlled datasets of human relatedness judgments for ambiguous words in both English and Spanish. Comparing monolingual and multilingual LMs from the same families, we find consistently reduced performance in multilingual LMs. We then explore three potential capacity constraints: representational (reduced embedding isotropy), attentional (reduced attention to disambiguating cues), and vocabulary-related (increased multi-token segmentation). Multilingual LMs show some evidence of all three limitations; moreover, these factors *statistically account for* the variance formerly attributed to a model's multilingual status. These findings suggest both that multilingual LMs do suffer from multiple capacity constraints, and that these constraints correlate with reduced disambiguation performance.

1 Introduction

In principle, training language models (LMs) on multiple languages should facilitate efficient cross-linguistic generalizations and widespread practical deployment. Yet as the number of languages on which a model is trained increases, multilingual models sometimes under-perform their monolingual counterparts (Conneau et al., 2020; Chang et al., 2024; Wang et al., 2020; Pfeiffer et al., 2022; Blevins et al., 2024), possibly due to *insufficient model capacity* (Chang et al., 2024). Here, we quantify the multilingual penalty associated with *lexical disambiguation*, and explore several distinct routes by which reduced capacity might manifest in multilingual LMs. We focus on lexical disambiguation in particular because it demands both precise semantic representations and well-developed mechanisms for disambiguation in context (Rivière and Trott, 2025); additionally, it has been extensively characterized across multiple disciplines, including NLP (Schlechtweg et al., 2018, 2025; Haber and Poesio, 2021; Trott and Bergen, 2021).

One potential source of reduced capacity is *representational*: notwithstanding cross-linguistic generalizations, multilingual LMs must compress more linguistic knowledge into the same number of dimensions, potentially reducing *within-language isotropy* (Ethayarajh, 2019). Because some multilingual LMs maintain language-specific representations (Chang et al., 2022), embeddings for words from a given language might occupy a narrower cone of vector-space (i.e., *anisotropy*) than in a monolingual model. While the downstream consequences of anisotropy are still under debate (Machina and Mercer, 2024; Godey et al., 2024; Rajaee and Pilehvar, 2022; Rudman et al., 2022), reduced isotropy could limit a model's ability to discriminate distinct word meanings in context (e.g., "tree *bark*" vs. "dog *bark*").

A related possibility is reduced *attentional* capacity. Attention heads likely play some role in contextualizing the meaning of target ambiguous words (Rivière and Trott, 2025). It is implausible that the same attention head could perform disambiguation across diverse languages—yet it is equally implausible that specialized attention heads could develop for each individual language. On net, this could limit a multilingual LM's ability to contextualize ambiguous words using attention mechanisms.

A third possibility is reduced *vocabulary* capacity. Multilingual LMs must cover far more words with a comparably sized vocabulary; consequently, more words will be segmented into multiple *subword tokens*, which in turn may not reflect meaningful morpheme boundaries (Arnett et al., 2024). This could be particularly problematic for disambiguation: when target words are split across to-

kens, extracting a coherent representation might be less straightforward. Indeed, there is some evidence that impaired performance in multilingual LMs can be attributed to challenges with tokenization (Rust et al., 2021).

In the current work, we evaluate disambiguation in both English and Spanish, using tightly controlled datasets of human judgments about ambiguous words in minimal pair contexts (Trott and Bergen, 2021; Rivière et al., 2025). We also use monolingual/multilingual "minimal pairs", i.e., from the same model family (e.g., BERT). In Section 2.2.1, we first quantify the multilingual penalty; we then quantify and explore the potential *correlates* of this penalty and ask which best accounts for the reduction in disambiguation performance (Section 2.2.5). All data and code required to reproduce the analyses in the current manuscript will be posted on GitHub when the anonymity period is over.

2 Current Work

2.1 Methods

2.1.1 Datasets

We used two datasets containing human relatedness judgments about ambiguous words across minimal pair contexts (e.g., "She liked the marinated lamb" vs. "She liked the friendly lamb"). RAW-C (Relatedness of Ambiguous Words—in Context) contained judgments for 672 English sentence pairs (Trott and Bergen, 2021), while SAW-C (Spanish Ambiguous Words—in Context) contained judgments for 812 Spanish sentence pairs (Rivière et al., 2025). These datasets were selected for their tight experiment control (ambiguous words were embedded in minimal pair contexts) and their *graded* human judgments about relatedness (consistent with recent work arguing for continuous representations of word meaning (Elman, 2009; Li and Joanisse, 2021; Trott et al., 2023; Li, 2024)). Both datasets had been anonymized by previous work.

2.1.2 Models

We assessed 24 unique model instances: 10 monolingual English models, 10 monolingual Spanish models, and 4 multilingual models. Our selection of models was guided by two primary factors: first, because the Spanish disambiguating cue occurred *after* the target word, we identified a range of *bidirectional models* (Rivière et al., 2025); and second, we selected model families that contained both monolingual English models and monolingual Spanish models, as well as multilingual variants. Model families included ALBERT/ALBETO (Lan et al., 2019; Cañete et al., 2020), BERT/BETO (Devlin et al., 2018; Cañete et al., 2020), DistilBERT/DistilBETO (Sahn et al., 2019; Cañete et al., 2020), and RoBERTa (Liu et al., 2019; Gutiérrez-Fandiño et al., 2022); all but one of these families (ALBERT) also included at least one multilingual variant, e.g., XLM-RoBERTa (Conneau et al., 2020), mBERT (Devlin et al., 2018), and DistilBERT (Sahn et al., 2019). A table listing all individual model instances (along with their number of parameters) can be found in Appendix A.

2.2 Results

2.2.1 Quantifying the "Multilingual Penalty"

As in past work (Trott and Bergen, 2021; Rivière et al., 2025), disambiguation performance was assessed by presenting each sentence pair from each dataset to a given model instance. We then calculated the cosine distance between the contextualized embeddings of the target word (e.g., "lamb") across all layers of that model. Finally, we regressed human relatedness judgments against cosine distance and used the resulting R^2 as a index of how successfully representations from that layer predicted relatedness. Multilingual models were evaluated using both datasets, and monolingual models were evaluated using only the dataset in the target language.

As depicted in Figure 2, models ranged considerably in their maximum performance, though none surpassed human inter-annotator agreement (Trott and Bergen, 2021; Rivière et al., 2025). We then estimated the "multilingual penalty" by regressing disambiguation performance from each layer of each model models (R^2) against several factors: Log Parameter Count, Language (English vs. Spanish), and Multilingual status (Yes vs. No); we also included random intercepts for each model. Performance improved for bigger models [$\beta = 0.09, SE = 0.03, p = 0.001$] and later layer depths [$\beta = 0.2, SE = 0.01, p < .001$]; multilingual models exhibited reduced performance even controlling for these other factors [$\beta = -0.16, SE = 0.04, p < .001$]. That is, representations from multilingual LMs were worse at predicting relatedness than were representations from equivalent layers of monolingual LMs. (Note that equivalent results were obtained using only

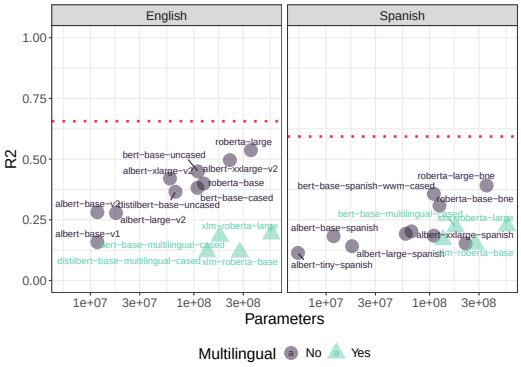


Figure 1: The best-performing layers of multilingual models generally exhibited reduced performance (as measured by R^2) compared to monolingual models of equivalent size.

maximal R^2 instead of the layer-wise measure; see Appendix B).

2.2.2 Decreased Representational Isotropy in Multilingual Models

As discussed above, there is no universally agreed-upon metric for evaluating the degree of *isotropy* in LM embeddings (Ethayarajh, 2019; Rudman et al., 2022). We focus here on Centered Isotropy, or CI, which is calculated by first centering and normalizing all embeddings for a sentence, then computing the average cosine distance between each pair of embeddings (see Appendix C for alternative metrics). For each layer of each model, we calculated the mean CI for each input sentence. We then fit a linear mixed model with mean CI as the dependent variable, fixed effects of Layer Depth, Multilingual status (and its interaction with Layer Depth), Language, and Log Parameter Count; and random intercepts for Target Word and LM. Multilingual LMs were associated with reduced mean CI overall $[\beta = -0.02, SE = 0.004, p < 0.001]$ (see also Figure 2a); moreover, this reduction was exacerbated at later layer depths $[\beta = -0.01, p < .001]$.

2.2.3 Decreased Attention to Disambiguating Cues in Multilingual Models

Disambiguation performance is likely linked to the degree of *attention* directed from an ambiguous word (e.g., "lamb") to potential disambiguating cues to the disambiguating cue (e.g., "marinated") (Rivière and Trott, 2025). For each attention head in each model, we calculated the attention score between the target word and the disambiguating cue. We then aggregated these scores by layer, computing both the average and maximum attention across all heads in a layer.

In the English dataset (RAW-C), we found no evidence that multilingual models exhibited reduced attention to disambiguating cues (for either metric). However, we did observe differences in maximal attention to disambiguating cues in the Spanish dataset (SAW-C), particularly in later layers of multilingual models (see Figure 2b). This was corroborated by the results of a statistical analysis regressing Max. Attention against Layer Depth, Multilingual status, Log Parameter Count, Language, an interaction between Language and Multilingual status, and random intercepts for model. The interaction effect was significant, with reduced attention for multilingual models tested in Spanish $\beta = 0.09, SE = 0.03, p = .01]$. (Note that the interaction was also significant, albeit smaller, when predicting mean attention.)

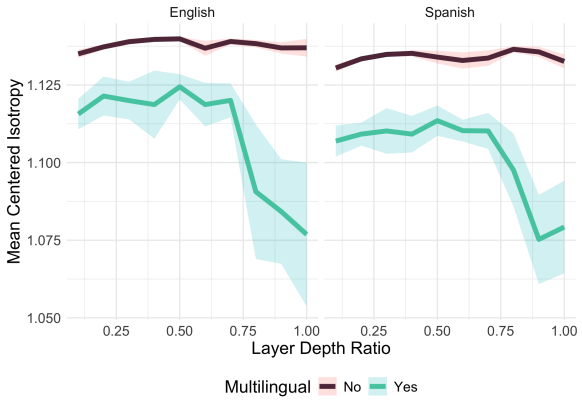
2.2.4 Increased Rate of Multi-Token Words for Multilingual Models

For both datasets, we counted the number of tokens corresponding to both the target word (e.g., "lamb") and disambiguating cue (e.g., "marinated") for each LM's tokenizer. We then built a series of linear mixed models predicting Number of Tokens (for the Target and Disambiguating Cue), with Multilingual status, Language, and Log Parameter Count as fixed effects, and random intercepts for Model, Target Word, and Sentence. As predicted, multilingual status was consistently associated with a higher number of tokens for both the target word $[\beta = 0.23, SE = 0.01, p < 0.001]$ and the disambiguating cue $[\beta = 0.43, SE = 0.04, p < .001]$; see also Figure 5.

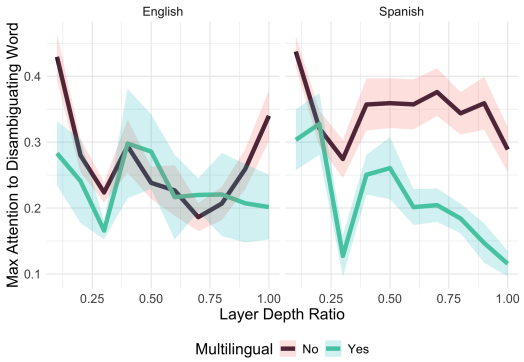
2.2.5 Which Factors Account for Reduced Disambiguation?

We then asked which factors accounted for the "multilingual penalty". We built a series of linear mixed models predicting layer-wise disambiguation performance from each factor—Cumulative Maximum Attention to the disambiguating cue across all layers up to layer ℓ; mean CI from ℓ; and the mean number of tokens in the target word—and assessed their fit using AIC (Akaike, 2003) (lower AIC corresponds to better fit). All models included baseline covariates (Log Parameter Count, Layer Depth) and random intercepts for model; AIC values were rescaled to this baseline.

As depicted in Figure 3, mean CI represented only a marginal improvement over the Baseline



(a) Multilingual models exhibited reduced embedding isotropy at equivalent layer depths.



(b) Attention heads in multilingual models directed less attention towards disambiguating cues in Spanish sentences than those in monolingual models.

Figure 2: Compared to their monolingual counterparts, multilingual models models showed evidence of reduced isotropy (left) and somewhat reduced attention to disambiguating cues (right).

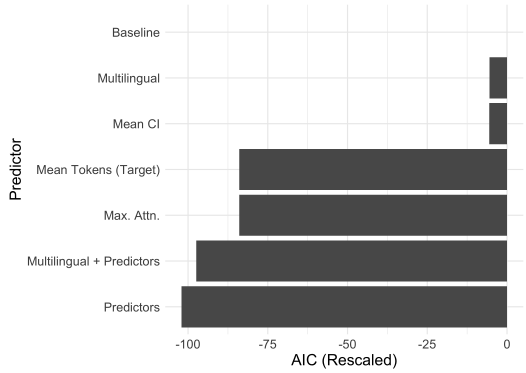


Figure 3: The AIC (scaled to the Baseline model) associated with linear mixed models predicting disambiguation performance from various factors (lower is better).

model—roughly equivalent to Multilingual status alone. Tokenization and attention yielded substantially better fit ($\Delta AIC > 78$). Crucially, when all three factors were combined, adding Multilingual status *hurt* model fit: the parameter was unnecessary because the other factors already captured its explanatory power. In the full model, increased isotropy predicted increased R^2 $[\beta = 1.2, SE = 0.27, p < .001]$ and multi-token words predicted decreased R^2 $[\beta = -0.58, SE = 0.13, p < .001]$. Counterintuitively, increased attention was associated with decreased performance $[\beta = -0.14, SE = 0.02, p < .001]$.

3 Discussion

We set out to quantify and explain the apparent penalty faced by multilingual LMs in disambiguation tasks. First, we confirmed that multilingual LMs consistently under-performed their monolingual counterparts on a disambiguation task in both English and Spanish (Section 2.2.1). Second, we found that multilingual LMs also displayed evidence of reduced isotropy (in both languages), reduced attentional capacity (in Spanish), and a higher rate of multi-token words (in both languages)—all potential *correlates* of a multilingual penalty. Third, we found that the combination of these factors statistically accounted for the variance explained by an LM's multilingual status (Section 2.2.5): that is, variance in isotropy, attention, and tokenization better accounted for variance in disambiguation performance than did a factor indicating whether an LM was multilingual. These results confirm that multilingual LMs do suffer from multiple kinds of *capacity limitations*, consistent with prior work (Chang et al., 2024); and moreover, that these *correlate* with reductions in disambiguation performance. Although this work is limited in scope and inferential power (see Section 4), it represents a proof-of-concept that at least in a subset of LMs, the multilingual penalty is correlated with measurable, relatively interpretable factors.

4 Limitations

A key limitation is *scope*: although the datasets benefited from tight experimental control, they were limited in size and covered only two languages (English and Spanish). Similarly, we relied on LMs from a restricted set of families, none of which are considered state-of-the-art; this limitation was driven in part by the need to rely on bidirectional

LMs (given that the ambiguous words in Spanish were always disambiguated by a word following the target), and by our aim to *match* multilingual LMs with monolingual LMs with similar training protocols and architectures. However, future work could investigate whether this multilingual penalty is observed in larger, state-of-the-art multilingual LMs across a variety of languages—and whether the same explanatory factors (i.e., reduced isotropy, reduced attention to disambiguating cues, and more multi-token words) consistently co-vary with multilingual status.

A related limitation is that our category of "multilingual" LM was quite coarse. To the extent that it exists, the multilingual penalty likely depends on the number and distribution of languages on which an LM is trained (Chang et al., 2024). Future work could also assess LMs trained on a small number of related languages—and also vary the relative balance of training volume across languages—to investigate how much the multilingual penalty (and its correlates) depends on the amount and type of multilinguality.

Certain findings were also surprising and raise questions about interpretation. For instance, we found a significant *negative* relationship between attention to disambiguating cues and overall disambiguation performance—the opposite of our predictions. One possibility is that better performance is actually driven by more distributed patterns of attention, perhaps reflecting greater redundancy; alternatively, our operationalization of "attention to the disambiguating cue" (i.e., the maximum attention from a given layer) could simply be flawed. This result also highlights the challenges of relying on attention maps in explanations of LM behavior (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019), particularly in the absence of causal intervention or a fine-grained analysis of training dynamics (Rivière and Trott, 2025).

This leads to a final limitation: notably, our analyses were *correlational* and do not establish a causal role for the factors we identified. It is possible that a single factor causally accounts for each of the other variables (e.g., perhaps differences in tokenization produce differences in measured isotropy or attention), or even that some unmeasured factor is truly responsible for the multilingual penalty. Future research could ask whether *causally intervening* on these factors—i.e., isotropy, attention to disambiguating cues, and tokenization—improves disambiguation performance.

5 Ethical Considerations

The primary ethical concerns relate to the inferential limitations discussed above: it is possible that limitations in our evaluation of multilingual LMs could lead to an underestimation (or overestimation) of their performance, which would have downstream effects on the relative risk of relying on such models.

References

Hirotugu Akaike. 2003. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Catherine Arnett, Pamela D. Rivière, Tyler A. Chang, and Sean Trott. 2024. Different tokenization schemes lead to comparable performance in Spanish number agreement. In *Proceedings of the 21st SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 32–38, Mexico City, Mexico. Association for Computational Linguistics.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. 2024. Breaking the curse of multilinguality with cross-lingual expert language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10822–10837, Miami, Florida, USA. Association for Computational Linguistics.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data.

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.

Tyler A. Chang, Zhuowen Tu, and Benjamin K. Bergen. 2022. The geometry of multilingual language model representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 119–136, Abu Dhabi, United Arab Emirates. 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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jeffrey L Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4):547–582.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. 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 55–65, Hong Kong, China. Association for Computational Linguistics.

Nathan Godey, Éric Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, St. Julian's, Malta. Association for Computational Linguistics.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Janosch Haber and Massimo Poesio. 2021. Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. 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 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942.

Jiangtian Li. 2024. Semantic minimalism and the continuous nature of polysemy. *Mind & Language*, 39(5):680–705.

Jiangtian Li and Marc F Joanisse. 2021. Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, 45(4):e12955.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Anemily Machina and Robert Mercer. 2024. Anisotropy is not inherent to transformers. 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 4892–4907, Mexico City, Mexico. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Sara Rajaee and Mohammad Taher Pilehvar. 2022. An isotropy analysis in the multilingual BERT embedding space. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1309–1316, Dublin, Ireland. Association for Computational Linguistics.

Pamela D Rivière, Anne L. Beatty-Martínez, and Sean Trott. 2025. Evaluating contextualized representations of (Spanish) ambiguous words: A new lexical resource and empirical analysis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8322–8338, Albuquerque, New Mexico. Association for Computational Linguistics.

Pamela D. Rivière and Sean Trott. 2025. Start making sense(s): A developmental probe of attention specialization using lexical ambiguity. *arXiv preprint arXiv:2511.21974*.

William Rudman, Nate Gillman, Taylor Rayne, and Carsten Eickhoff. 2022. IsoScore: Measuring the uniformity of embedding space utilization. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3325–3339, Dublin, Ireland. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. 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 3118–3135, Online. Association for Computational Linguistics.

Victor Sahn, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version

of BERT: smaller, faster, cheaper and lighter. In *Proceedings of Thirty-third Conference on Neural Information Processing Systems (NeurIPS2019)*.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. 2025. CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 33–47.

Dominik Schlechtweg, Sabine Schulte im Walde, and Stefanie Eckmann. 2018. Diachronic usage relatedness (DURel): A framework for the annotation of lexical semantic change. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 169–174, New Orleans, Louisiana. Association for Computational Linguistics.

Sean Trott and Benjamin Bergen. 2021. RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). 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 7077–7087, Online. Association for Computational Linguistics.

Sean Trott, Cameron Jones, Tyler Chang, James Michaelov, and Benjamin Bergen. 2023. Do large language models know what humans know? *Cognitive Science*, 47(7):e13309.

Zirui Wang, Zachary C. Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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 11–20, Hong Kong, China. Association for Computational Linguistics.

A Full list of models

Table 1: Language Models by Family

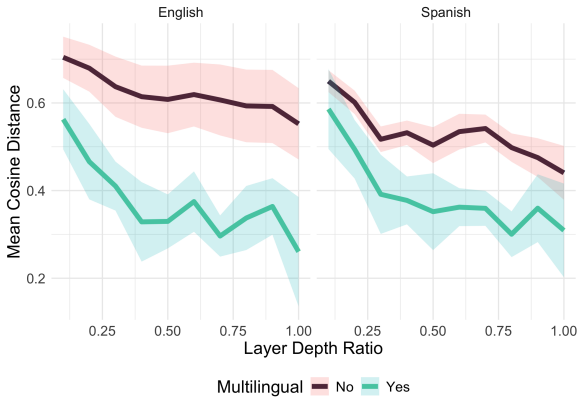
Model	Multi.	# Params
bert-base-cased	No	∼ 108M
bert-base-uncased	No	∼ 109M
bert-base-spanish-wwm-cased	No	∼ 110M
bert-base-spanish-wwm-uncased	No	∼ 110M
bert-base-multilingual-cased	Yes	∼ 178M
distilbert-base-uncased	No	∼ 66M
distilbert-base-spanish-uncased	No	∼ 67M
distilbert-base-multilingual-cased	Yes	∼ 135M
albert-tiny-spanish	No	∼ 5M
albert-base-v1	No	∼ 12M
albert-base-v2	No	∼ 12M
albert-base-spanish	No	∼ 12M
albert-large-v2	No	∼ 18M
albert-large-spanish	No	∼ 18M
albert-xlarge-v2	No	∼ 59M
albert-xlarge-spanish	No	∼ 59M
albert-xxlarge-v2	No	∼ 223M
albert-xxlarge-spanish	No	∼ 223M
roberta-base	No	∼ 125M
roberta-base-bne	No	∼ 125M
roberta-large	No	∼ 355M
roberta-large-bne	No	∼ 355M
xlm-roberta-base	Yes	∼ 278M
xlm-roberta-large	Yes	∼ 560M

B Analysis of maximal R^2

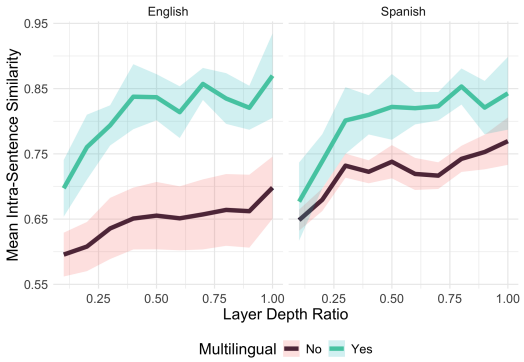
We carried out an additional analysis investigating which factors predicted *maximal* R^2. First, we replicated the analysis of the multilingual penalty. As in the main manuscript, we found that overall performance was higher for bigger models [$\beta = 0.15, SE = 0.03, p < .001$] and lower for models tested in Spanish [$\beta = -0.1, SE = 0.03, p = .002$]. Crucially, multilingual models exhibited reduced performance even controlling for these other factors [$\beta = -0.22, SE = 0.04, p < .001$]. That is, an LM's multilingual status was associated with a 0.22 *decrease* in R^2 relative to models of an equivalent size tested on the same dataset.

C Additional Isotropy Metrics

As noted in the primary manuscript, researchers use different metrics for evaluating embedding isotropy, which suffer from different advantages and disadvantages (Rudman et al., 2022). In addition to Mean Centered Isotropy, we evaluated Mean Cosine Distance (the average cosine distance between all token embeddings for a given sentence from a given layer) and Intra-Sentence Similarity (the cosine distance between each individual token embedding and the average embedding for the sentence from a given layer). In general, higher Mean Cosine Distance is interpreted as *more* isotropic (i.e., embeddings are more widely dispersed), and higher Intra-Sentence Similarity is interpreted as *less* isotropic (i.e., embeddings are all similar to the sentence average). As with Mean Centered Isotropy, we observed evidence of decreased isotropy in multilingual models compared to their monolingual counterparts, particularly at later layers, in both metrics (see Figure 4).



(a) Embeddings from multilingual models were closer on average than embeddings from layers of an equivalent depth of monolingual models.



(b) Individual token embeddings from multilingual models were more similar on average to the mean sentence embedding than embeddings from layers of an equivalent depth of monolingual models.

Figure 4: Multilingual models showed evidence of reduced isotropy (lower average cosine distance between token embeddings; and higher cosine similarity between individual token embeddings and the sentence average) relative to monolingual models.

D Increased Rate of Multi-Token Words

As described in the primary manuscript, tokenizers for multilingual models were more likely to segment target words (and disambiguating words) into multiple tokens. A visual comparison of this difference is illustrated in Figure 5 below.

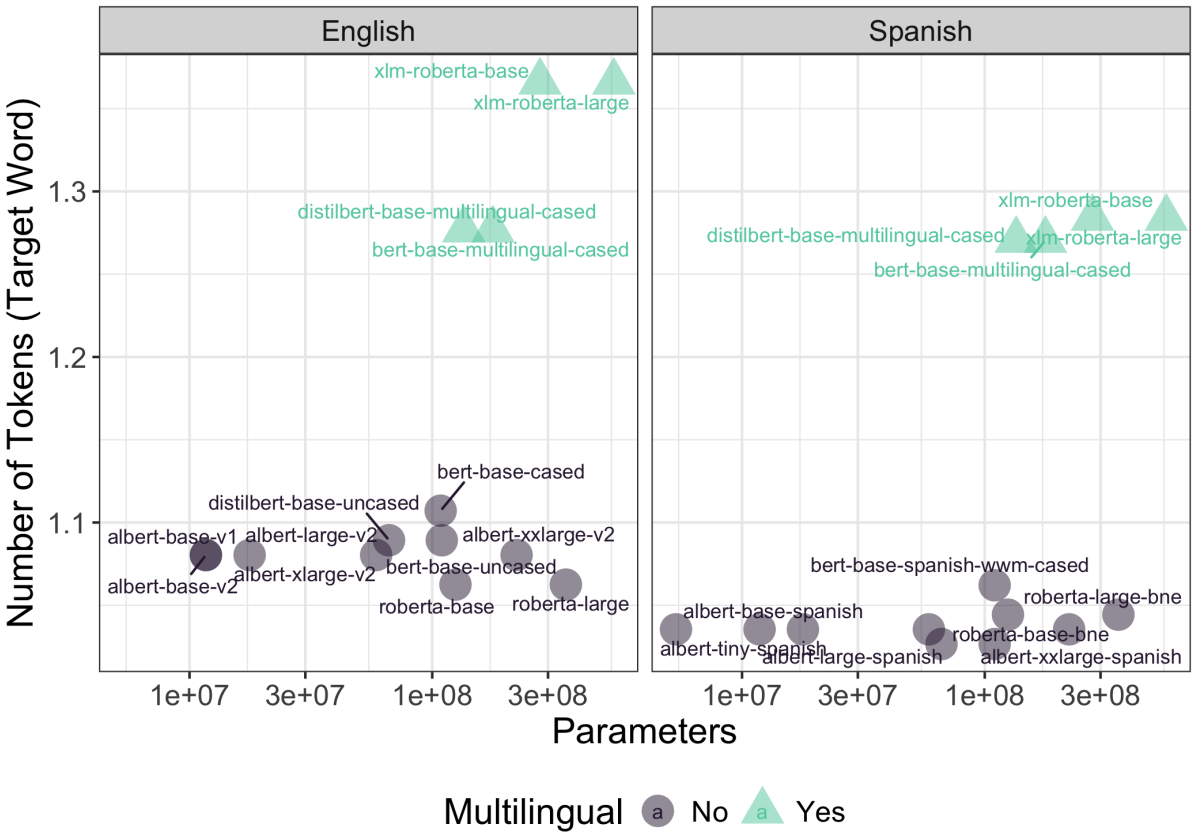


Figure 5: Multilingual models consistently segmented target words into more tokens than monolingual models.