Bidirectional Transformer Representations of (Spanish) Ambiguous Words in Context: A New Lexical Resource and Empirical Analysis

Anonymous ACL submission

Abstract

Lexical ambiguity—where a single wordform takes on distinct, context-dependent meanings—serves as a useful tool to compare across different large language models' (LLMs') ability to form distinct, contextualized representations of the same stimulus. Few studies have systematically compared LLMs' contextualized word embeddings for languages beyond English. Here, we evaluate multiple bidirectional transformers' (BERTs') semantic representations of Spanish ambiguous nouns in context. We develop a novel dataset of minimal-pair sentences evoking the same or different sense for a target ambiguous noun. In a pre-registered study, we collect contextualized human relatedness judgments for each sentence pair. We find that various BERT-based LLMs' contextualized semantic representations capture some variance in human judgments but fall short of the human benchmark, and for Spanish—unlike English—model scale is uncorrelated with performance. We also identify *stereotyped trajectories* of target noun disambiguation as a proportion of traversal through a given LLM family's architecture, which we partially replicate in English. We contribute (1) a dataset of controlled, Spanish sentence stimuli with human relatedness norms, and (2) to our evolving understanding of the impact that LLM specification (architectures, training protocols) exerts on contextualized embeddings.

1 Introduction

Large language models (LLMs) display a remarkable level of formal linguistic competence (Mahowald et al., 2024). To date, however, we currently lack a precise accounting of the mechanisms underlying LLMs' fundamental linguistic capabilities. The opacity of model internals has motivated work probing the transformations that inputs undergo as they are processed through various model components (Tenney et al., 2019; Hu et al., 2020; Wang et al., 2022; Zou et al., 2023). This work has focused on LLMs trained with English-language corpora, with a smaller subset of studies investigating cross-linguistic representations in multilingual models (Chang et al., 2022; Wendler et al., 2024; Michaelov et al., 2023) or comparing representations across multiple monolingual models (Edmiston, 2020). As others have noted (Blasi et al., 2022b; Bender, 2009), an over-reliance on English as a "model language" limits the generalizability of findings, as well as potential applications (Blasi et al., 2022a). Here, we extend interpretability work to Spanish, a language spoken by almost 600M people (with almost 500M native speakers)[1]. Specifically, we: (1) evaluate the trajectory of ambiguous Spanish words' semantic representations within mono- and multilingual LLMs and (2) identify layers along a model's architecture whose semantic representations best capture human judgments of semantic relatedness.

Lexical ambiguity—where a given wordform evokes multiple related or unrelated meanings—offers a unique opportunity to dissociate a word's form from the contextualized, semantic representations that it can take on as it interacts with a given model's architecture. Specifically, we can evaluate whether and how LLMs integrate the surrounding lexical items in a sentence to produce flexible, context-dependent representations. The representation and processing of ambiguous words is also well-studied in humans (Rodd et al., 2004; Martin et al., 1999; Duffy et al., 1988), offering a convenient comparison group. Finally, ambiguity appears to pervade language, with some estimates in English positing that more than 80% of words have multiple meanings (Rodd et al., 2004); more frequent words are also more likely to evoke multiple senses (Zipf, 1945; Piantadosi et al., 2012). Ambiguity, then, is an important *phenomenon* to

[1] https://www.exteriores.gob.es/en/PoliticaExterior/Paginas/ElEspanolEnElMundo.aspx

contend with in LLMs and a useful *tool* to understand LLM representations.

We first provide a survey of related work on ambiguity and interpretability in LLMs (Section 2), then present a novel dataset of human relatedness judgments about Spanish ambiguous words—in context (SAW-C) (Section 3). Section 4 documents our use of this dataset to empirically probe the representation of ambiguous words in pre-trained spanish LLMs, focusing first on pre-registered analyses[2] of the monolingual BETO (Cañete et al., 2020). We next systematically compare LLM architectures (multiple monolingual Spanish and multilingual BERT-based models) (Section 5). We conclude with a discussion of the implications of the current results (Section 6), as well as limitations and directions for future work (Section 7).

2 Related work

To date, lexical ambiguity has been largely explored within English monolingual models (Haber and Poesio, 2020; Trott and Bergen, 2021; DeLong et al., 2023). Less work is available with models trained on other languages, such as Spanish (Garcia, 2021; Garí Soler and Apidianaki, 2021), and crucial distinctions emerge across studies within this literature: (1) the operationalization of SAME and DIFFERENT SENSE conditions, (2) the degree to which sentential context is controlled around the target word, and (3) the degree to which human semantic judgments are leveraged to set usage-based expectations for the *context-dependence* and *graded quality* of ambiguous word meanings (Erk et al., 2013; Trott and Bergen, 2023).

In Spanish, mono- and multilingual BERT-based models can capture information about semantic relationships between homonyms and their synonyms (Garcia, 2021) and can approximate words' degree of polysemy (Garí Soler and Apidianaki, 2021). However, this work tends to leverage naturalistic sentence stimuli from sense-annotated corpora. While valuable, the variability in token sequence length and target word position within naturalistic sentential contexts may make it challenging to isolate the precise effect of the context's semantic *content* from the uncontrolled effects of sentence frame (Haber and Poesio, 2020). We thus follow as closely as possible the experimental structure leveraged (for English) in Trott and Bergen (2021), creating Spanish-language sentence pairs that vary along a single context cue evoking either the same or different sense of the target ambiguous noun (see Section 3.1). For our dataset, we document the extent to which context cues presented true minimal pairs (e.g. zero-token-differences across sentence pairs) for the LLMs tested (see **Appendix A.1; Table 3**).

Using more controlled sentence stimulus design, coupled with empirically collected human benchmarks, prior studies in English have shown that BERT-based models' contextualized embeddings capture some—though not all—variance in human similarity (Haber and Poesio, 2020) and relatedness (Trott and Bergen, 2021; DeLong et al., 2023) judgments for ambiguous English words. Some of this work has also argued that the *continuous* nature of LLM contextualized representations makes them well-suited as models of human word meanings, which are likely graded to some extent (Elman, 2009; Li and Armstrong, 2024; Li and Joanisse, 2021; Trott and Bergen, 2023; Rodd et al., 2004; Nair et al., 2020)—though importantly, may also exhibit marked *categoriality* compared to LLM representational spaces (Trott and Bergen, 2023).

Finally, another important line of work has used techniques like classifier probes (Tenney et al., 2019) and activation patching (Meng et al., 2022; Wang et al., 2022; Merullo et al., 2024) to decode the putative functional role of different model components (e.g., layers, attention heads, etc.) in producing observed behavior.

To our knowledge, there is little to no work combining these strands of research: i.e., making use of graded human judgments about ambiguous Spanish words to trace the dynamics of contextualized representations in pre-trained Spanish language models. This is the gap we aim to address.

3 Human Annotation Study

We created a dataset containing graded human judgments about Spanish ambiguous words—in context (SAW-C). This process involved first producing and curating materials (i.e., target words and sentences), collecting judgments from native Spanish speakers, and validating those judgments for quality and reliability. As noted previously, this study was pre-registered on OSF.

[2] A link to the pre-registration, code, and data will be provided after the anonymity period has passed.

3.1 Materials

All sentences were developed by two native Spanish speakers. 102 target words were drawn from noun lists collected in earlier studies of Spanish lexical ambiguity (Estévez Monzó, 1991; Fraga et al., 2017), as well as spontaneously generated and then verified using the online Real Academia Española[3] dictionary. We excluded wordform meanings that corresponded to distinct parts of speech, but we accepted a small fraction (8/102) of nouns whose grammatical gender changed across meanings. Each target ambiguous noun was embedded within sentence pairs that differed by a single modifier[4], termed context cue[5]. The average number of words in sentences was 4.72. The context cues across the sentence pair could evoke either the SAME or DIFFERENT sense of the word across the contexts.

1a. Compró el **aceite** de oliva (*[S/he] bought the olive oil*)
1b. Compró el **aceite** de cocina (*[S/he] bought the cooking oil*)
2a. Compró el **aceite** de motor (*[S/he] bought the motor oil*)
2b. Compró el **aceite** de carro (*[S/he] bought the car oil*)

The minimum number of *sentence pairs* per target noun was 6, with a maximum of 28 (M = 7.96). A total of 812 sentence pairs and 102 target nouns were included in the dataset.

Finally, for the purpose of human norming, the 812 sentence pairs were assigned to 10 experimental lists using a Latin Square design, where each list had approximately 81 or 82 sentence pairs.

3.2 Participants

Our goal was to collect a minimum of 10 judgments per sentence pair (i.e., a minimum of 100 participants). Because we anticipated a non-zero exclusion rate, our pre-registration specified: 1) an initial goal of 120 participants; and 2) a plan to sample more participants as needed, if any sentence pairs had fewer than 10 observations after applying the pre-registered exclusion criteria.

Using the two-step process described above, we recruited an initial pool of 139 participants through Prolific. Participants received $2.40 for participating and the median completion time was 12 minutes and 23 seconds, for an average rate of $11.64 per hour. On Prolific, we screened for participants who reported that their primary language was Spanish; we specifically recruited participants from the United States, as well as countries in which the dominant language was Spanish (including Chile, México, and Spain).

We excluded participants (1) who failed "catch" trials (where the sentences in the pair were identical), (2) whose task completion times exceeded the sample mean by 3 standard deviations, (3) whose inter-annotator agreement for the items they rated was very low (Pearson's $r^2 < 0.1$), and (4) who self-identified as non-native Spanish speakers. After all exclusions, we considered data from a total of 131 participants.

Participants' self-identified nationalities corresponded heavily to México (69), Spain (39), and Chile (20); 1 participant was from the United States and 2 participants were from Venezuela.[6] 54 participants self-identified as female (74 male, 3 non-binary). The average age was 30.97 (median = 28), and ranged from 20 to 59.

3.3 Procedure

After providing consent, participants were given instructions explaining that some words can have different meanings in different contexts (using an example that was *not* included in the experimental materials), and that the goal of the experiment was to collect ratings about the relatedness of the meanings expressed by a given word across two sentence contexts.

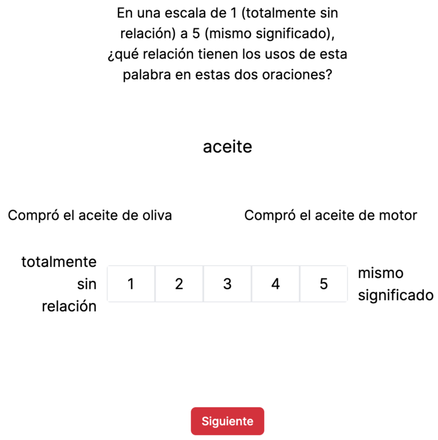
Each sentence pair was presented on a separate page. Participants were instructed to rate each word on a scale from 1 (*totalmente sin relación*, "totally unrelated") to 5 (*mismo significado*, "same meaning"). The target word (e.g., *aceite*, "oil") was centered in larger font, and the target sentences were presented side-by-side (the side was randomized across trials). Participants indicated their response via button-press; see **Figure 1** for an example. We also included one "catch" trial in the

[3] https://www.rae.es/

[4] For 187/812 sentence pairs, the modifiers varied along more than just a single word, as was the case when the modifiers required different prepositions, or the grammatical gender of the modifier differed across sentences and required distinct determiners and contractions. Of these, 173 pairs differed by 1 word; 14 differed by 2 words.

[5] Context cues were adjectival modifiers for 100/102 target nouns; verbs for 2/102.

[6] As described in Section 3.4, we found relatively high correlations between mean relatedness judgments produced by participants across Chile, México, and Spain.



En una escala de 1 (totalmente sin relación) a 5 (mismo significado), ¿qué relación tienen los usos de esta palabra en estas dos oraciones?

aceite

Compró el aceite de oliva

Compró el aceite de motor

Figure 1: Sample item from task. Participants emitted graded relatedness judgments from a scale of 1 (totally unrelated) to 5 (same sense) for a given target word (here: **aceite** – *oil*), using information provided by the context cue across the sentence pair (here: **de oliva / de motor** – *olive / motor*).

experiment, which simply contained the same sentence, repeated (i.e., the correct answer was 5).

The entire experiment (including consent form, instructions, and debrief page) was conducted in Spanish. Participants were randomly assigned to lists, and the order of items within each list was randomized. After completing the primary experiments, participants read a debrief form explaining once more that the goal of the experiment was to collect judgments about ambiguous Spanish words, and that their data would be anonymized before analysis and publication.

3.4 Validation of Final Dataset

We validated the final dataset using several approaches. First, we applied multiple exclusion criteria (Section 3.2) and collected a minimum of 10 ratings per sentence pair. The average number of ratings per pair was 13.1; the maximum was 17.

After applying exclusion criteria, we recalculated inter-annotator agreement to estimate the reliability of the ratings in the final dataset. Following past work (Hill et al., 2015; Trott and Bergen, 2021) we calculated inter-annotator agreement using a *leave-one-annotator-out* scheme. For each of the final 131 participants, we calculated Spearman's ρ between the judgments produced by that participant and the mean judgment for those same sentence pairs, leaving out that participant's data. The resulting distribution of correlation coefficients ranged from 0.39 to 0.88, with an average correlation of 0.77. This number is comparable to past work using similar methods (Hill et al., 2015; Trott and Bergen, 2021).

Finally, we compared average relatedness judgments across the three main demographic groups reported by participants (Chile, México, and Spain). Judgments across each group were all strongly correlated ($r > 0.82$ in all cases).

3.5 Relatedness of Same vs. Different Sense Contexts

We then asked whether and how human relatedness judgments varied as a function of whether two uses of a word (e.g., "aceite") corresponded to the SAME SENSE or DIFFERENT SENSE. These will hereafter be considered the levels of the binary variable we call Sense Relationship.

Using the entire dataset of trial-level judgments (10639 observations), we fit a linear mixed effects model in R using the *lme4* package (De Boeck et al., 2011), which had Relatedness as a dependent variable and a fixed effect of Sense Relationship. The model also contained by-participant and by-list random slopes for the effect of SAME SENSE, and random intercepts for participants, lists, and words. (The specification of random effects was determined by beginning with the maximal model, then reducing as needed for model convergence (Barr et al., 2013).) The full model explained significantly more variance than a reduced model omitting only the effect of Sense Relationship [$\chi^2(1) = 39.59, p < .001$]. As expected, SAME Sense contexts were rated as more related on average ($M = 4.35, SD = 1.14$) than DIFFERENT Sense contexts ($M = 2.11, SD = 1.41$; **Figure 2**).

4 Analysis of BETO, a Pre-Trained Spanish LLM

Using SAW-C as a probe, we conducted multiple pre-registered analyses on how BETO—a pre-trained monolingual Spanish LLM (Cañete et al., 2020)—represents ambiguous words in context, whether and to what extent these representations are predictive of human semantic representations, and *which layers* of the model contained the most information (Tenney et al., 2019). **Table 1** summarizes the research questions and their results.

4.1 Model Details

Our pre-registered analyses used the cased version of a Spanish monolingual BERT-based model:

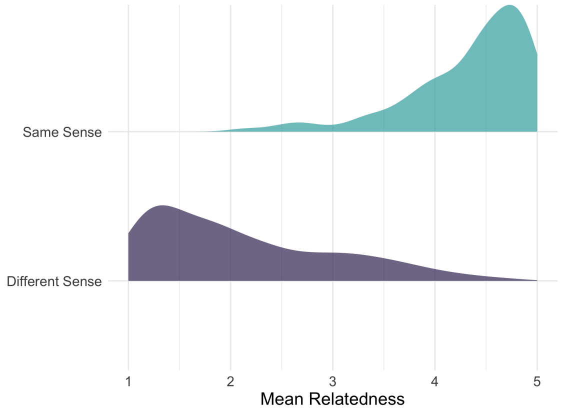


Figure 2: Density plot representing the distribution of mean relatedness judgments for sentence pairs. As expected, word meanings were rated as more related when used in SAME Sense than DIFFERENT Sense contexts.

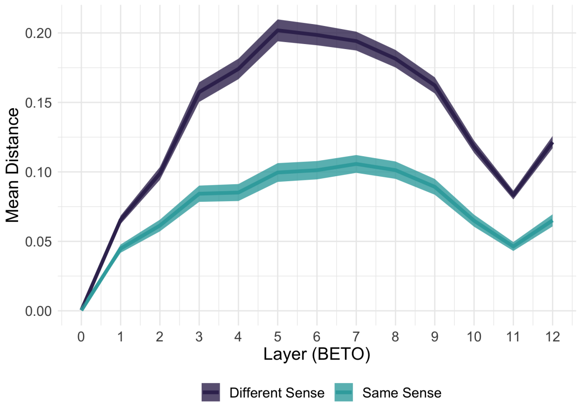


Figure 3: Average Cosine Distance between the contextualized representations of the target ambiguous word across each layer of BETO, depicted as a function of whether the contexts cued the SAME SENSE or DIFFERENT SENSE.

BETO, comprised of 12 layers, each made up of 12 self-attention heads, and hidden size of 768, trained on a corpus of approximately 3B words (Cañete et al., 2020)[7]. Exploratory analyses leveraged the models summarized in the **Appendix A.1** (**Table 2**). Each sentence in the dataset (bracketed by special tokens [CLS] and [SEP]) was tokenized separately according to each model's tokenizer. Sentences did not contain periods, to match the version of the sentences viewed by human participants. We report sentence pair tokenization differences across models in **Table 3** (**Appendix A.1**). When the target noun was represented by multiple subword tokens, we took the average embedding across tokens. We extracted embeddings from each model layer.

4.2 Which layer of BETO best captures sense boundaries?

We first assessed which layers of BETO produced representations that best distinguished between SAME SENSE and DIFFERENT SENSE uses of word. To address this question, we calculated the Cosine Distance between these contextualized representations of the target word from each sentence pair for each layer. Concretely, this yielded 812 Cosine Distance values for each layer of BETO.

Then, we asked how Cosine Distance evolved through the network's layers with respect to the SAME/DIFFERENT SENSE distinction. We built a series of logistic regression models in R with Sense Relationship as a dependent variable, and Cosine Distance from a given layer ℓ_i of BETO as an independent variable. We then measured the Akaike Information Criterion (or AIC) (Akaike, 2011; Burnham and Anderson, 2004) of the resulting model as a measure of model fit. The best-fitting model used Cosine Distance from *layer 5*. **Figure 3** highlights the change in Cosine Distance across layers of BETO as a function of Sense Relationship, suggesting that the *difference* between conditions was largest at this layer.

4.3 Which layer of BETO best predicts relatedness?

Our second question was whether certain layers of BETO produced representations that better predicted *human relatedness judgments* than others. For each layer, we calculated the correlation coefficient (both Pearson's r and Spearman's ρ) between Cosine Distance values obtained from BETO and the distribution of Mean Relatedness judgments obtained for each sentence pair. We also calculated R^2 as an estimate of the amount of *variance explained* in human relatedness judgments as a function of Cosine Distance from that layer alone.

As depicted in **Figure 7**, the layer of BETO with the highest R^2 was *layer 7* ($R^2 = 0.33, r = -.57, \rho = -.59$). However, performance did not meaningfully improve beyond *layer 5* ($R^2 = 0.328$). This suggests that the operations performed by later layers were less useful in terms of producing contextualized representations that captured relevant variance in relatedness judgments (see also **Appendix A.2**).

[7] Accessed via https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased

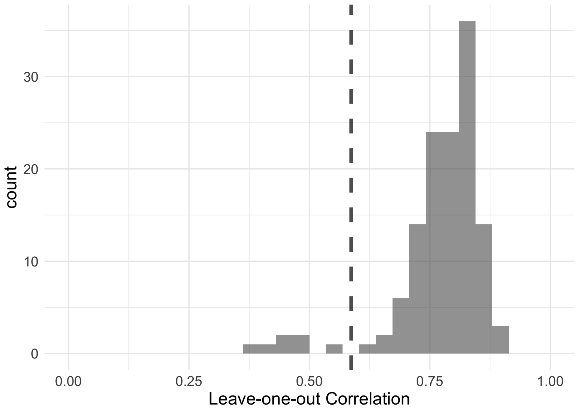


Figure 4: Distribution of human inter-annotator agreement scores, calculated using a *leave-one-annotator out* scheme. The vertical dashed line represents the correlation between human judgments and Cosine Distance values extracted from BETO.

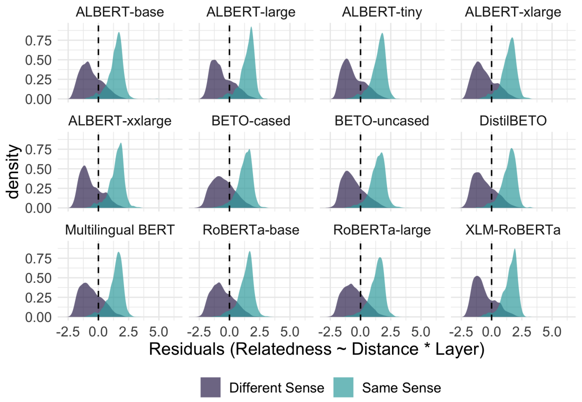


Figure 5: Residuals of linear regression models fit for each LLM, predicting relatedness from the interaction between cosine distance and layer position; residual distributions are separable as a function of Sense Relationship.

4.4 BETO under-performs inter-annotator agreement

We then compared the best-performing layer ($\ell = 7$) to human inter-annotator agreement (calculated in Section 3.4). The correlation of the best-performing layer ($\rho = 0.58$) was considerably lower than the average inter-annotator agreement ($\bar{X}_\rho = 0.77$), with BETO's performance lying in the bottom 5% of the distribution of agreement values (**Figure 4**).

4.5 BETO is Less Sensitive to Sense Boundaries

Past work conducted in English (Trott and Bergen, 2021, 2023) suggests that LLMs are less sensitive to sense boundaries—the distinction between SAME and DIFFERENT SENSE—than humans. However, it is unclear whether this effect generalizes to Spanish speakers and Spanish LLMs.

We constructed a linear mixed effects model in R with Relatedness as a dependent variable, fixed effects of Cosine Distance and Sense Relationship, by-participant random slopes for both fixed effects, and random intercepts for participants, words, and lists. The full model explained significantly more variance than a model omitting only the effect of Sense Relationship [$\chi^2(1) = 331.55, p < .001$]. The full model also explained more variance than a model omitting only Cosine Distance [$\chi^2(1) = 208.27, p < .001$]. Further, the R^2 of a linear regression model predicting Mean Relatedness using Sense Relationship alone ($R^2 = .61$) explained almost twice as much variance as Cosine Distance alone ($R^2 = 0.33$); adding both predictors resulted in a modest improvement over the Sense Relationship model ($R^2 = 0.66$).

Finally, we extracted the *residuals* of a linear regression model predicting Mean Relatedness from Cosine Distance alone. We then plotted the distribution of these residuals according to Sense Relationship. As illustrated in **Figure 5**, BETO (as well as all the Spanish language models tested in Section 5) consistently *underestimated* the relatedness of SAME sense pairs, and consistently *overestimated* the relatedness of DIFFERENT sense pairs.

5 Comparing Pre-Trained Spanish Language Models

Section 4 tested several pre-registered hypotheses (**Table 1**) with respect to a single pre-trained Spanish LLM. Here, we extend this work in exploratory analyses of additional pre-trained Spanish LLMs. Testing multiple models is an important step towards establishing the *external validity* of a finding; additionally, it is useful for testing hypotheses about model scale (Kaplan et al., 2020) or other model specifications (e.g., architecture, multilingual status).

5.1 Models

We considered 10 monolingual Spanish language models (including BETO) and 2 multilingual models. LLMs varied in their training procedures (e.g., BERT vs. RoBERTa; Liu et al. (2019)), tokenization scheme, number of layers, training corpus size, total number of parameters, and whether or not

Research Question	Result	Section
Do humans judge SAME SENSE uses to be more related than DIFFERENT SENSE uses?	Yes	3.5
Which layer of BETO is most sensitive to the SAME/DIFFERENT SENSE distinction?	$\ell = 5$	4.2
Which layer of BETO is most correlated with human relatedness judgments?	$\ell = 7$	4.3
Does BETO match human inter-annotator agreement?	No	4.4
Does BETO "explain away" the effect of categorical Sense Relationship in humans (e.g. "sense boundaries")?	No	4.5

Table 1: Summary of pre-registered research questions and their results.

they were multilingual (**Table 2**).

5.2 Impact of model scale

Past work (Kaplan et al., 2020) suggests that increases in a model's number of parameters may correlate with metrics of model performance (e.g., perplexity). At the same time, there is some evidence that increasing scale does not always produce more human-like representations, i.e., large models with lower perplexity do not always better predict human reading times (Kuribayashi et al., 2021).

To assess this question, we compared the maximum R^2 achieved by each of the 12 language models[8] and asked whether an LLM's best R^2 was related to its size. We found virtually no evidence that a model's size was correlated with its ability to predict human relatedness judgments (**Figure 6**). The best-performing model (BETO-cased) was not the largest tested, and larger models (e.g., ALBERT-xxlarge or XLM-RoBERTa) performed just as poorly as models with many fewer parameters. None of the models approached the variance explained in average judgments by individual human annotators (i.e., inter-annotator R^2).

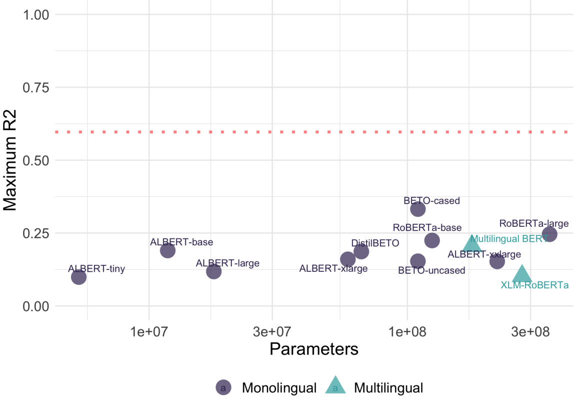


Figure 6: Maximum R^2 achieved by each model by number of parameters and multilingual status. Horizontal dashed line depicts the average variance explained a leave-one-annotator-out scheme.

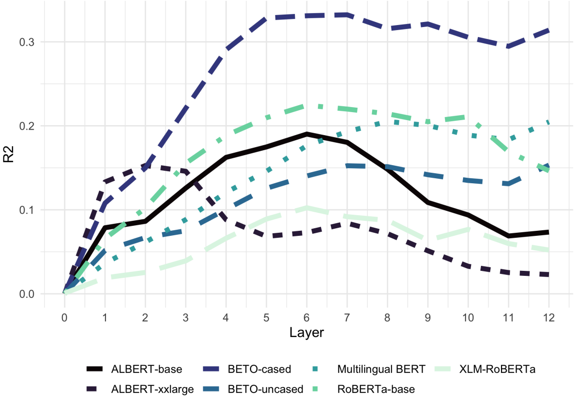


Figure 7: Depiction of seven pre-trained Spanish LLMs' ability to predict human relatedness judgments across layers (measured as R^2). For ease of illustration, this plot shows only LLMs with 12 layers.

Of course, this was not an ideal test of the scaling hypothesis, given that many features of these models (e.g., amount of training data, specific training corpus, model architecture) were not controlled. It is worth noting, however, than even among the Spanish ALBERT family of models (Cañete et al., 2022), there was not a clear, consistent relationship between size and performance.

5.3 Performance across layers

In Section 4, we found that cosine distance measures extracted from the *middle layers* of BETO were most useful for predicting whether two contexts belonged to the same sense, and for predicting human relatedness judgments. Do other models or model families show the same trajectory of performance across layers?

Models varied in their number of layers. Thus,

[8] I.e., the R^2 from the best-performing layer of each model.

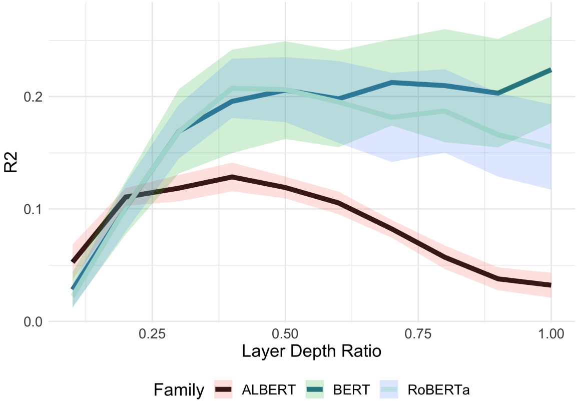


Figure 8: Depiction of LLM ability to predict mean relatedness judgments (measured as R^2), broken down by *Model Family* and *Layer Depth Ratio*, i.e., with each layer divided by the the total number of layers in a given model. 4 BERT models (1 multilingual), 3 RoBERTa models (1 multilingual), and 5 ALBERT models.

we first compared the trajectory of R^2 across layers for the subset of models with the same number of layers, i.e., 12 layers (**Figure 7**) or 24 layers (**Figure 9**). In each case, we identified two qualitatively distinct "classes" of trajectory: a *rise and plateau* trajectory, in which performance improves up until a point (e.g., layer 6) and then stays relatively stable; and a *rise and fall* trajectory, in which performance improves and then decays substantively in the final layers (**Figure 8**).

In order to compare all models on the same axis, we calculated the *layer depth ratio*, which divides each layer position in a given network by the total number of layers in that network. We then visualized the average R^2 by *layer depth ratio* across the three model families tested: ALBERT, BERT, and RoBERTa. As **Figure 8** suggests, the two putative trajectories appear to covary with model family: the ALBERT family of models shows a *rise and fall* trajectory, while the BERT and RoBERTa family of models shows a *rise and plateau* trajectory.

5.4 Sensitivity to sense categoriality across models

Finally, we asked whether any of the models tested *explained away* the sense boundary effect in humans. Each model showed the same pattern as BETO: Cosine Distance *underestimated* the relatedness of SAME SENSE meanings and *overestimated* the relatedness of DIFFERENT SENSE meanings (**Figure 5**).

6 Discussion

We introduced a novel dataset (SAW-C) containing human relatedness judgments about ambiguous Spanish words in controlled, minimal pair contexts. Using this dataset, we probed pre-trained LLMs' representations of ambiguous words, finding that: 1) LLM representations correlate with human judgments but do not match inter-annotator agreement, and exhibit *systematic errors*; 2) performance varies across layers, with model families showing distinct *trajectories* of performance; and 3) there was no discernible effect of model *size*.

The systematic *underestimation* and *overestimation* errors observed with respect to SAME vs. DIFFERENT sense contexts (**Figure 5**) is consistent with past work conducted in English (Trott and Bergen, 2021, 2023). One explanation for this is that the initial (static) embedding for an ambiguous wordform might entangle all of its multiple meanings (Grindrod, 2024), which must then be "teased apart" in context—but which might nevertheless persist as "attractors" in subsequent layers. Disambiguation could be made even more difficult by the presence of minimal pair contexts (Garcia, 2021).

Psycholinguistic research suggests that humans also activate uncued, dominant meanings in certain tasks (Duffy et al., 1988; Martin et al., 1999); here, however, humans appeared to distinguish target meanings with relative ease. It is possible that humans represent distinct homonymous (though not necessarily polysemous) meanings along clearly differentiable regions of meaning space (Rodd et al., 2004; Trott and Bergen, 2023, 2021; Haber and Poesio, 2020), which would be consistent with the fact that categorical sense boundaries (SAME vs. DIFFERENT SENSE conditions) explained an overwhelming share of human relatedness judgments (Section 4.5, Trott and Bergen (2023)).

Lastly, this work contributes to expanding the linguistic diversity of both human-annotated benchmarks and interpretability research. Although Spanish ranks among the most widely spoken languages, it suffers from a surprising dearth of resources, pre-trained models, and interpretability research—particularly when compared to English. In one study, a sample of 550 corpora (spanning 22 languages) contained >50% English-language corpora, while <10% represented Spanish-language corpora (Anand et al., 2020). A wider research perspective—considering varied languages—is critical for ensuring the generalizability of findings.

7 Limitations

The current work is limited both in terms of the novel dataset we created, as well as the analyses we conducted. The sections below address each issue in turn.

7.1 Limitations of the Dataset

SAW-C is limited in *scope*, containing only 812 sentence pairs. This is considerably smaller than many English benchmarks, such as BLiMP (Warstadt et al., 2020), which contain tens of thousands of examples. However, it is larger than or comparable in size to other, more targeted datasets involving crowd-sourced human annotations (Erk et al., 2013; Haber and Poesio, 2020; Trott and Bergen, 2021). Relative to specifically Spanish-language datasets, ours is the only one we are aware of that collects human judgments for target ambiguous words embedded *within* minimal pair stimuli[9]. Importantly, SAW-C includes not only the sentence pairs but also over 10000 validated *human judgments* from 131 participants about those sentence pairs.

Another limitation concerns the generation process for the materials. The ambiguous nouns included in our dataset were spontaneously produced by native Spanish speakers, or selected from previously published lists (Estévez Monzó, 1991; Fraga et al., 2017), rather than via automated searches. Spontaneous production of ambiguous words may (1) overrepresent homonymous words (Estévez Monzó, 1991), and (2) underrepresent words whose multiple meanings have large dominance asymmetries (Duffy et al., 1988). In future work, we intend to collect dominance norms for these items.

Finally, our sentence pairs are not *naturalistic*. The creation of controlled minimal pairs was an intentional design feature of the dataset, which enabled us to identify key differences between LLM and human representations, e.g., LLMs display less sensitivity than humans to the manipulation of word meaning across minimal pair contexts (**Figure 5**). At the same time, it is important to know whether and to what extent the current results replicate with naturalistic stimuli. Thus, we aim to augment SAW-C with naturalistic examples of the target ambiguous words, which would also increase its scope.

7.2 Limitations of the Analysis

All the analyses presented here are essentially *correlational*. As others have noted (Hewitt and Liang, 2019; Niu et al., 2022; Zhou and Srikumar, 2021), supervised methods for probing LLM representations are more informative about the ability of the *probe* to learn specific features than the question of whether the LLM "naturally" encodes that feature and deploys it for token prediction. Future work would benefit from the selective application of "knock-out" methods or "activation patching" (Meng et al., 2022), both of which have proven more successful in characterizing the causal, mechanistic role of model components. We view the current work as a useful starting point, which can motivate future work isolating the mechanistic role of specific model circuits within each layer.

Similarly, the finding that distinct model families exhibit distinct *trajectories* in performance (**Figure 8**) is intriguing, but due to its exploratory nature, it is unclear to what extent this finding is reliable and robust to different datasets or probing methods. In a supplementary analysis (see **Appendix 11**), we found qualitatively similar clusters of trajectories in pre-trained English models, though these differences appeared considerably weaker than in the Spanish models. Thus, future work could build on the question of whether—and more importantly, *why*— distinct model *architectures* and training schemes lead to different *processing mechanisms*.

8 Ethical Considerations

This research was conducted with IRB approval. All data from human participants has been fully anonymized before analysis and publication.

Acknowledgments

References

Hirotugu Akaike. 2011. Akaike's information criterion. *International Encyclopedia of Statistical Science*, page 25.

Pranav Anand, Sandra Chung, and Matthew Wagers. 2020. Widening the net: Challenges for gathering linguistic data in the digital age. *Response to NSF SBE*.

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for

[9] For examples of other Spanish-language datasets that collect human judgments under different experimental conditions, see Estévez Monzó (1991); Gómez-Veiga et al. (2010); Domínguez et al. (2001); Haro et al. (2017); Fraga et al. (2017).

confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3):255–278.

Emily M. Bender. 2009. Linguistically naïve != language independent: Why NLP needs linguistic typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, pages 26–32, Athens, Greece. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022a. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022b. Over-reliance on English Hinders Cognitive Science. *Trends in cognitive sciences*, 26(12):1153–1170.

Kenneth P Burnham and David R Anderson. 2004. Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304.

José Cañete, Sebastian Donoso, Felipe Bravo-Marquez, Andrés Carvallo, and Vladimir Araujo. 2022. ALBETO and DistilBETO: Lightweight Spanish language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4291–4298, Marseille, France. European Language Resources Association.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained BERT model and evaluation data. In *PML4DC at ICLR 2020*.

Tyler Chang, Zhuowen Tu, and Benjamin 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.

Paul De Boeck, Marjan Bakker, Robert Zwitser, Michel Nivard, Abe Hofman, Francis Tuerlinckx, and Ivailo Partchev. 2011. The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, 39:1–28.

Katherine A DeLong, Sean Trott, and Marta Kutas. 2023. Offline dominance and zeugmatic similarity normings of variably ambiguous words assessed against a neural language model BERT. *Behavior Research Methods*, 55(4):1537–1557.

Danica Dillion, Niket Tandon, Yuling Gu, and Kurt Gray. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences*, 27(7):597–600.

Alberto Domínguez, Fernando Cuetos, and Manuel de Vega. 2001. 100 palabras polisémicas con sus acepciones. *Revista Electrónica de Metodología Aplicada*, 6(2):63–84.

Susan A Duffy, Robin K Morris, and Keith Rayner. 1988. Lexical ambiguity and fixation times in reading. *Journal of memory and language*, 27(4):429–446.

Daniel Edmiston. 2020. A systematic analysis of morphological content in BERT models for multiple languages. *arXiv preprint arXiv:2004.03032*.

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

Katrin Erk, Diana McCarthy, and Nicholas Gaylord. 2013. Measuring word meaning in context. *Computational Linguistics*, 39(3):511–554.

Adelina Estévez Monzó. 1991. Estudio normativo sobre ambigüedad en castellano. *Cognitiva*, 3(2):237–246.

Isabel Fraga, Isabel Padrón, Manuel Perea, and Montserrat Comesaña. 2017. I saw this somewhere else: The Spanish ambiguous words (SAW) database. *Lingua*, 185:1–10.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. 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 3625–3640, Online. Association for Computational Linguistics.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Isabel Gómez-Veiga, Nuria Carriedo López, M Rucián Gallego, and José O Vila Cháves. 2010. Estudio normativo de ambigüedad léxica en castellano, en niños y en adultos. *Psicológica*, 31(1):25–47.

Jumbly Grindrod. 2024. Transformers, contextualism, and polysemy. *Preprint*, arXiv:2404.09577.

Janosch Haber and Massimo Poesio. 2020. Word sense distance in human similarity judgements and contextualised word embeddings. In *Proceedings of the Probability and Meaning Conference (PaM 2020)*, pages 128–145, Gothenburg. Association for Computational Linguistics.

Juan Haro, Pilar Ferré, Roger Boada, and Josep Demestre. 2017. Semantic ambiguity norms for 530 Spanish words. *Applied Psycholinguistics*, 38(2):457–475.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. 2021. Lower perplexity is not always human-like. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 5203–5217. Association for Computational Linguistics.

Jiangtian Li and Blair C Armstrong. 2024. Probing the representational structure of regular polysemy via sense analogy questions: Insights from contextual word vectors. *Cognitive Science*, 48(3):e13416.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2020. RoBERTa: A robustly optimized BERT pretraining approach.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, pages S1364–6613.

Charles Martin, Hoang Vu, George Kellas, and Kimberly Metcalf. 1999. Strength of discourse context as a determinant of the subordinate bias effect. *The Quarterly Journal of Experimental Psychology Section A*, 52(4):813–839.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*.

James Michaelov, Catherine Arnett, Tyler Chang, and Ben Bergen. 2023. Structural priming demonstrates abstract grammatical representations in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3703–3720, Singapore. Association for Computational Linguistics.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

Jingcheng Niu, Wenjie Lu, and Gerald Penn. 2022. Does bert rediscover a classical nlp pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3143–3153.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Jennifer M Rodd, M Gareth Gaskell, and William D Marslen-Wilson. 2004. Modelling the effects of semantic ambiguity in word recognition. *Cognitive science*, 28(1):89–104.

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)*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Sean Trott. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, pages 1–19.

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 and Benjamin Bergen. 2023. Word meaning is both categorical and continuous. *Psychological Review*, 130(5):1239–1261.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *NeurIPS ML Safety Workshop*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do Llamas Work in English? on the Latent Language of Multilingual Transformers. *Preprint*, arXiv:2402.10588.

Yichu Zhou and Vivek Srikumar. 2021. DirectProbe: Studying representations without classifiers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5070–5083.

George Kingsley Zipf. 1945. The meaning-frequency relationship of words. *The Journal of General Psychology*, 33(2):251–256.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*.

A Appendix

A.1 LLM Specifications and Dataset Tokenization

See **Table 2** for a summary of the LLMs considered in Section 5, including architecture, multilingual status, corpus size, tokenization scheme, training objective, number of layers, and number of parameters. Because models used different tokenizers, we also calculated summary statistics about the number of tokens in each sentence in each sentence pair, as well as the average number of *token differences* (i.e., 5 vs. 4 tokens across the members of a given pair) for each tokenizer (see **Table 3**).

A.2 Analysis of Expected Layer

In the primary manuscript, we identified which layers of BETO provided the *best* fit, i.e., which were most effective for predicting Sense Relationship (layer 5) and which were most effective at predicting Mean Relatedness judgments (layer 7). However, in some cases, the improvements across layers are fairly marginal. Thus, in this supplementary analysis, we implemented a version of the *Expected Layer* analysis described by Tenney et al. (2019). This analysis considers the size of the improvement across layers and estimates the layer at which particular kinds of information is expected to resolve in the network.

A.2.1 Methods

The *Expected Layer* statistic considers the *improvement in performance* (measured here as AIC or R^2, depending on the analysis in question) across progressively more complex regression models fit with cosine distance information from each layer. This improvement in performance measure was defined as:

$$\Delta^{\ell} = Score(P_T^{\ell}) - Score(P_T^{\ell-1}) \qquad (1)$$

Where $Score(P_T^{\ell})$ is defined as the performance (AIC or R^2) of a regression model equipped with cosine distance information from a given layer ℓ and each previous layer, i.e., such that the number of parameters in the regression model was equal to ℓ. (Note that this was distinct from the approach taken in the primary manuscript, in which distinct univariate regression models were fit for each layer.) The Expected Layer statistic itself was defined as follows:

$$\bar{E}_{\Delta}[\ell] = \frac{\sum_{\ell=1}^{L} \ell \cdot \Delta_T^{(\ell)}}{\sum_{\ell=1}^{L} \Delta_T^{(\ell)}} \qquad (2)$$

A.2.2 Results

Using this approach, we obtained Expected Layer statistics for both predicting Sense Relationship (3.2) and for predicting Mean Relatedness (2.98).

[11] **10:** See https://github.com/google-research/bert/blob/master/multilingual.md for available details on language-specific corpus size selection.

[11] See https://github.com/google-research/bert for notes on the English BERT pre-training update using whole-word masking.

Model	# Lang	Corpus	Tokenization	Trn Obj	# Layers	# Params
BETO	1	$\sim$ 3B	SentencePiece	DM, WWM	12	$\sim$ 110M
BETO-uncased	1	$\sim$ 3B	SentencePiece	DM, WWM	12	$\sim$ 110M
mBERT	104	?	WordPiece	MLM, WWM$^?$, NSP	12	$\sim$ 178M
DistilBETO	1	$\sim$ 3B	SentencePiece	DistilLoss, MLM	6	$\sim$ 66M
ALBETO-tiny	1	$\sim$ 3B	SentencePiece	MLM	4	$\sim$ 5M
ALBETO-base	1	$\sim$ 3B	SentencePiece	MLM	12	$\sim$ 12M
ALBETO-large	1	$\sim$ 3B	SentencePiece	MLM	24	$\sim$ 18M
ALBETO-xlarge	1	$\sim$ 3B	SentencePiece	MLM	24	$\sim$ 59M
ALBETO-xxlarge	1	$\sim$ 3B	SentencePiece	MLM	12	$\sim$ 223M
RoBERTa-BNE-base	1	$\sim$ 135B	byte-BPE	MLM (DM$^?$)	12	$\sim$ 125M
RoBERTa-BNE-large	1	$\sim$ 135B	byte-BPE	MLM (DM$^?$)	24	$\sim$ 355M
XLM-RoBERTa	100	?	SentencePiece	MLM	12	$\sim$ 278M

Table 2: (Spanish) Language model properties and training procedures. Models are cased (distinguish between upper and lowercase characters) unless otherwise specified. All monolingual models are trained on Spanish-language corpora; multilingual models include Spanish-language corpora. ***Model Notes:*** For mBERT, (a) the corpus size per language varied and we are unsure of the total corpus size[10], (b) it was unclear to us whether the current version of the model on HuggingFace is updated with the whole-word masking (WWM) technique[11]during pre-training. For RoBERTa-BNE models, it was unclear to us whether authors used dynamic masked (DM) modeling, as in the English RoBERTa. For XLM-RoBERTa, the corpus size per language may have varied, but we were uncertain to what extent pretraining text was sampled proportionally to its representation in the corpus (Conneau et al., 2020). ***Acronyms:*** **BNE:** Biblioteca Nacional de España (National Library of Spain); **byte-BPE:** byte-level Byte-Pair Encoding; **DistilLoss:** Distillation loss (Sahn et al., 2019; Cañete et al., 2022). **MLM:** Masked Language Modeling; **DM:** Dynamic Masking (Liu et al., 2020); **WWM:** Whole-Word Masking; **NSP:** Next Sentence Prediction. This summary represents our best attempt at gathering and reconstructing some model specifications—they are necessarily incomplete, and may contain inaccuracies borne from either a lack of knowledge regarding more recent updates or an imperfect understanding of the training protocols as described in the relevant primary literature and repositories.

.

Model	Avg ; Modal ; Max Token Differences	Target Noun # Tokens
BETO	∼ 0.682 ; 1 ; 3	∼ 1.07
BETO-uncased	∼ 0.670 ; 1 ; 3	∼ 1.03
mBERT	∼ 0.997 ; 1 ; 4	∼ 1.27
DistilBETO	∼ 0.670 ; 1 ; 3	∼ 1.03
ALBETO-tiny	∼ 0.619 ; 1 ; 3	∼ 1.04
ALBETO-base	∼ 0.619 ; 1 ; 3	∼ 1.04
ALBETO-large	∼ 0.619 ; 1 ; 3	∼ 1.04
ALBETO-xlarge	∼ 0.619 ; 1 ; 3	∼ 1.04
ALBETO-xxlarge	∼ 0.619 ; 1 ; 3	∼ 1.04
RoBERTa-BNE-base	∼ 0.643 ; 0 ; 3	∼ 1.05
RoBERTa-BNE-large	∼ 0.643 ; 0 ; 3	∼ 1.05
XLM-RoBERTa	∼ 0.929 ; 1 ; 4	∼ 1.28

Table 3: Average, modal, and maximum token differences across sentence pairs per LLM. Tokenization schemes for all Spanish monolingual LLMs were heavily represented by either non-zero or single-token differences across sentence pair stimuli, whereas the multilingual models tested here tended to more frequently generate non-zero token differences across the sentence pairs.

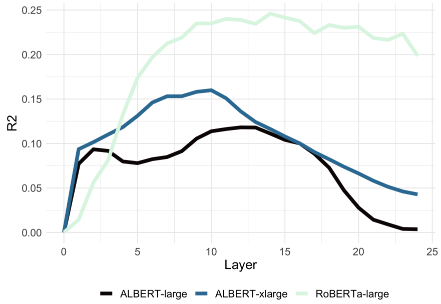


Figure 9: Depiction of three pre-trained Spanish LLMs' ability to predict human relatedness judgments across layers (measured as R^2). For ease of illustration, this plot shows only LLMs with 24 layers.

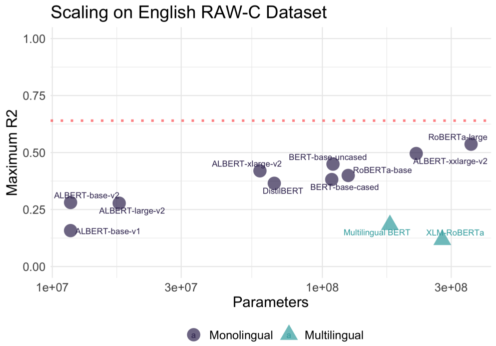


Figure 10: Pre-trained monolingual English language models show evidence of *scaling*, i.e., models with more parameters achieve a higher R^2 in predicting human relatedness judgments about ambiguous English words.

Note that in both cases, the Expected Layer was smaller than the layer at which optimal performance was achieved; this is consistent with the observation that past a certain point, additional LLM layers resulted in only marginal gains in prediction.

A.3 Additional analyses of pre-trained English LLMs

In the primary manuscript, we reported the results of work using ambiguity as a probe for understanding and interpreting how pre-trained Spanish LLMs process word meanings. We found two intriguing results: first, that larger models did not exhibit consistently better performance; and second, that different model *families* exhibited different trajectories of performance across layers. We asked whether these results replicated in pre-trained English models, using an openly available dataset of relatedness judgments about ambiguous English words, in context (RAW-C) (Trott and Bergen, 2021).

A.3.1 A correlation between model scale and performance

We tested English versions of the Spanish language models tested in the primary manuscript (with the exception of *ALBERT-tiny*, and with the addition of two different versions of *ALBERT-base*). We then asked whether there was a relationship between the

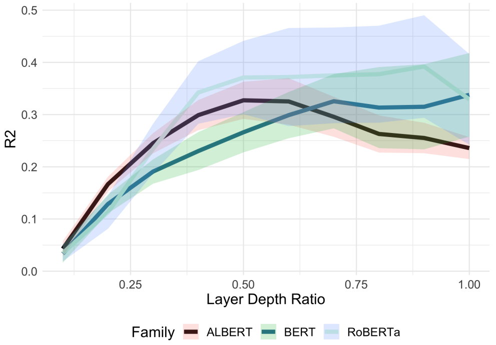


Figure 11: Depiction of English LLMs' ability to predict mean relatedness judgments (measured as R^2), broken down by *Model Family* and *Layer Depth Ratio*, i.e., with each layer divided by the the total number of layers in a given model.

number of parameters in each model and the maximum R^2 achieved in predicting human relatedness judgments about English words, in context. Unlike Spanish, we found a clear, positive relationship between the logarithm of the number of parameters and the maximum R^2 (see Figure 10): a linear model estimating maximum R^2 from number of parameters and a model's multilingual status estimated that for every order of magnitude increase in a model's number of parameters, R^2 increased by approximately 0.2 $[\beta = 0.198, SE = 0.03, p < .001]$. On average, multilingual models also performed worse (adjusting for number of parameters), though the small number of multilingual models tested makes it difficult to determine whether this is a reliable finding.

A.3.2 Layer-wise trajectories by model family

We also asked whether different model *families* displayed different performance trajectories across *layers*. In the primary manuscript, we found that the ALBERT family models displayed a *rise and fall* trajectory, while both BERT and RoBERTa displayed *rise and plateau* trajectories (see Figure 8). Surprisingly, we found qualitatively similar (albeit weaker) classes of trajectories using the pre-trained English models on the RAW-C dataset (see Figure 11).

A.4 Analysis of GPT-4 Turbo

In the primary manuscript, we found that the Large Language Models tested produced representations that were *correlated* with human judgments, but nonetheless systematically underestimated how related people judged SAME SENSE meanings to be—and overestimated how related people judged DIFFERENT SENSE meanings to be (see Figure 5). However, recent work (Trott, 2024; Dillion et al., 2023) suggests that state-of-the-art LLMs like GPT-4 are capable of producing "norms" that accurately predict human judgments across various domains, including the relatedness of English words.

Because these models are "closed source", this work typically relies on *prompting* the models with instructions and directly eliciting a judgment (e.g., a relatedness rating). Thus, a key limitation is that even if these judgments are highly correlated with human judgments, it is very difficult (in some cases impossible) to know *why*, i.e., which representations or mechanisms give rise to the behavior in question—making them less well-suited to questions about model interpretability.

Nonetheless, the question of the empirical fit between these LLM judgments and human judgments is still an interesting one—particularly because past work has primarily focused on judgments in English, and it is unclear whether these LLMs would excel at other languages. Thus, in this supplementary analysis, we asked whether GPT-4 Turbo, a state-of-the-art LLM, produced judgments that were more predictive of human judgments than the models tested in the primary manuscript.

A.4.1 Methods

Following past work (Trott, 2024; Dillion et al., 2023), we prompted GPT-4 Turbo (*gpt-4-1106-preview*) using the OpenAI Python API. GPT-4 Turbo was presented with a system prompt containing the same instructions (in Spanish) that were presented to human participants, explaining the purpose of the task. Then, for each sentence pair, Turbo was presented with the same instructions given to human participants (again in Spanish) asking them to rate the relatedness of the target word across the two contexts. The two sentences were presented on separate lines, as was the target word (e.g., "Word: aceite"). Finally, we included an additional instruction requesting a single number in response. Turbo was prompted using a temperature of 0 and its responses were limited to a maximum of 3 tokens.

A.4.2 Results

The ratings produced by Turbo were highly correlated with human judgments, approaching or even exceeding average human inter-annotator agreement ($\rho = 0.79$).

We then asked whether Turbo's ratings explained

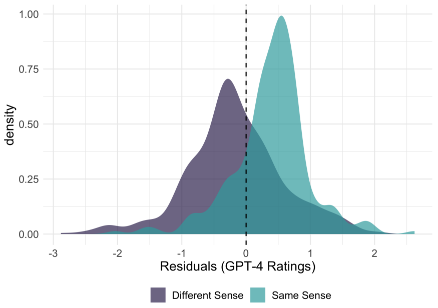


Figure 12: Residuals from a linear regression predicting Mean Relatedness from ratings elicited from GPT-4 Turbo. Although Turbo's ratings are highly correlated with human judgments, they still systematically underestimate the relatedness of Same Sense pairs and overestimate the relatedness of Different Sense pairs.

away the sense boundary effect observed in the primary manuscript. First, we fit a linear model with Mean Relatedness as a dependent variable and two predictors: Rating (from GPT-4 Turbo) and Sense Relationship. The coefficients assigned to each predictor were significant, suggesting that they explained some amount of independent variance: both Rating $[\beta = 0.45, SE = 0.02, p < .001]$ and SAME SENSE $[\beta = 1.12, SE = 0.05, p < .001]$ exhibited a positive relationship.

As in Figure 5, we also visualized the residuals of a linear model containing only Rating as a predictor (Figure 12). Notably, even though Rating was highly correlated with Mean Relatedness, the residuals suggest that Turbo's ratings follow a similar pattern with respect to sense boundaries as was observed with BETO and the other models tested: GPT-4 consistently *underestimates* the relatedness of SAME SENSE pairs, and consistently *overestimates* the relatedness of DIFFERENT SENSE pairs.

A.4.3 Discussion

In this supplemental analysis, we found that a larger model trained with Reinforcement Learning from Human Feedback (RLHF) produced behavior that was much more correlated with human relatedness judgments than the other Spanish language models tested.

On the one hand, this analysis has considerable limitations: as noted above, a major challenge with relying on LLMs such as GPT-4 is that, despite their impressive performance, much is still unclear about how exactly they were trained—either in terms of the original training data or the procedure for implementing RLHF. Thus, as a tool for testing scientific hypotheses, they may be less useful than open-source models, i.e., the ones tested in the primary manuscript, especially if the scientific question under investigation concerns the *representations* or *mechanisms* used by the LLM. These are, of course, the kinds of questions that research on interpretability is generally interested in.

On the other hand, the fact that an LLM produces behavior that rivals inter-annotator agreement suggests that the representations required to produce this behavior can be learned provided sufficient training data and fine-tuning; in this case, direct data contamination is an unlikely concern given that the materials were entirely novel and the ratings had never been published before. This analysis is also notable in that it reveals that even a very large-scale model trained with RLHF appears to be less sensitive to sense boundaries than human judgments, as depicted in Figure 12.

However, because Turbo is a closed-source model, it is difficult to draw firm conclusions from these results precisely because we cannot investigate where or how this behavior emerges. For example, we cannot investigate which *layer* of GPT-4 Turbo appears to be most helpful for producing high-quality relatedness judgments. Future work would thus benefit from investigating which architectural features, training objectives, or fine-tuning procedures are likely candidates for producing this improvement in performance, ideally in open-source Spanish language models.