Exploring Layer-wise Representations of English and Chinese Homonymy in Pre-trained Language Models

Matthew King-Hang Ma*♠, Chenwei Xie*♠, Wenbo Wang♣, William Shiyuan Wang♠
Research Centre for Language, Cognition, and Neuroscience
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
♠{khmma,cwxie,wsywang}@polyu.edu.hk
♣wenbo99.wang@connect.polyu.hk

Abstract

Homonymy can easily raise lexical ambiguity due to the misunderstanding of its multiple senses. Correct recognition of homonym sense greatly relies on its surrounding context. This ambiguous nature makes homonyms an appropriate testbed for examining the contextualization capability of pre-trained (PLM) and large language models (LLMs). Considering the impact of part of speech (POS) on homonym disambiguation and the prevalence of English-focused studies in word embedding research, this study extends to Chinese and provides a comprehensive layer-wise analysis of homonym representations in both languages, spanning same and different POS categories, across four families of PLMs/LLMs (BERT, GPT-2, Llama 3, Qwen 2.5). Through the creation of a synthetic dataset and computation of disambiguation score (*D-Score*), we found that: (1) no universal layer depth excels in differentiating homonym representations; (2) bidirectional models produce better contextualized homonym representations compared to much larger autoregressive models; (3) most importantly, POS affects homonym representations in models in ways that differ from human research findings. The individual differences between LLMs uncovered in our study challenge the simplistic understanding of their inner workings. This reveals a compelling research frontier: conducting controlled experiments with purposefully manipulated inputs to enhance the interpretability of LLMs. We have made our dataset and codes available publicly at `https://github.com/neurothew/exploring-homonym-rep-in-llm`.

1 Introduction

The efficient and economic use of lexical inventory results in multiple word senses converging into a single lexical item, leading to lexical ambiguity (Wang, 2011; Piantadosi et al., 2012). Among these lexical items, homonyms represent a common type. They denote two (or more) semantically and etymologically unrelated meanings. For instance, "bank" can refer to a financial institution or the side of a river.

Resolution of lexical ambiguity, while rarely conscious in everyday language use, can pose specific challenges to human readers or listeners. Psycholinguistic and neuroimaging research suggest that homonyms, with their unrelated meanings, often make comprehension more difficult as indicated by enhanced neural activation and longer response time (Frazier and Rayner, 1990; Rodd, 2018; Huang and Lee, 2018).

Lexical ambiguity also poses challenges to distributional semantic models, though the nature of these difficulties differs from those encountered in human (Lake and Murphy, 2023). Early static word representation models such as LSA (Deerwester et al., 1990) and Word2Vec (Mikolov et al., 2013) cannot be modulated by surrounding contexts after the training process. Different senses of a word must share the same representation, hindering its ability to differentiate word senses. Contextualized word representations and language models such as ELMo (Peters et al., 2018) and BERT (Devlin et al., 2019) were proposed to address the problem. Followed by this development, modern PLMs and LLMs are all contextualized language models.

Contextualized word representations from these models are influenced by its surrounding contexts, allowing the representations to vary instead of remain static. The context can be preceding tokens in autoregressive models like GPT (Radford et al., 2019), or both preceding and following tokens in bidirectional, autoencoding models like BERT (Devlin et al., 2019). Homonyms, which greatly rely on their surrounding context to be accurately understood, serve as appropriate candidates to experiment on the contexutalization capability of PLMs and LLMs.

*These authors contributed equally.

Findings of the Association for Computational Linguistics: ACL 2025, pages 19705–19724

Besides, the difference between homonyms whose senses belong to the same or different syntactic categories or parts of speech (POS) also warrants attention (MacDonald et al., 1994). For example, 'fly' can serve as both a noun (a small insect) and a verb (to move through air). Electrophysiological responses and blood-oxygen-level-dependent (BOLD) signals, as observed during lexical ambiguity processing, are significantly influenced by POS variability, according to previous research (Federmeier et al., 2000; Gennari et al., 2007). Homonyms with different POS senses elicit greater neural activation compared to those with the same POS meanings, suggesting a greater cognitive resource demand for syntactic category processing (Grindrod et al., 2014).

Given these findings in human language processing, it becomes interesting to explore whether these two types of homonyms would be represented differently in language models, as they are in the human brain.

Our main contributions are:

- Constructed a new Chinese dataset to facilitate homonym representation analysis.
- Presented a cross-linguistic analysis on the layer-wise, contextualized representations of Chinese and English homonyms across various families of PLMs and LLMs.
- Contributed to a deeper understanding on how homonyms are represented in language models, specifically on how POS modulates model representations.

2 Related Works

Our present work is largely related to previous studies on the investigation of the contextualization capabilities of PLMs. Contextualized embeddings can be easily translated into their corresponding definitions in an English dictionary, capturing sense-specific information (Chang and Chen, 2019). They can also be used to predict human behaviours and explain variances of human judgements on meanings (Wilson and Marantz, 2022; Rivière et al., 2024). A study that focuses on polysemes and homonyms revealed that while word embeddings can differentiate ambiguous words in terms of cosine similarity, the distinctive power was much less compared to human ratings (Haber and Poesio, 2021). It reveals the limitations of word embeddings. The distance between ambiguous word embeddings was found to show different trajectories across model layers, which in turn depended on the architectural factors, including but not limited to model size (Rivière et al., 2024).

Ethayarajh (2019) analyzed the contextual word representations in ELMo, BERT and GPT-2. It was found that the representations were more context-specific in the higher layers, consistent with previous studies (Peters et al., 2018; Liu et al., 2019a; Clark et al., 2019). Additionally, Ethayarajh (2019) observed that the word embeddings in language models suffered from anisotropy, referring to the non-uniform distribution of the embeddings in the embedding space. They proposed subtracting the cosine similarity from the baseline computed from the text materials to create an adjusted measure that enhances interpretability.

While there is existing research on contextuality and lexical ambiguity, our work makes significant contributions in several ways. For instance, while Sevastjanova et al. (2021) examined the contextualization of words along the functionality continuum ranging from homonyms, modals, to articles, their study was not positioned to explore the nuances, such as the contrast between contextualized homonym representations computed from homonyms of the same or different meanings. In contrast, we specifically targeted homonyms with two different meanings by developing our own dataset. We compared the performance of 21 models across four model families, varying in size (from 110M to 8B) and architecture, which is more extensive than the few models used in Ethayarajh (2019) and Rivière et al. (2024). We proposed an angle-based disambiguation score to account for the nonlinearity of cosine similarity (Section 3.2.1).

Furthermore, we investigated the impact of POS on contextualized embeddings by controlling for the selection of homonyms. While POS tagging has attracted substantial research attention in NLP (Chiche and Yitagesu, 2022), the inner workings of POS processing in LLMs remain unclear. This issue is particularly intriguing, as POS is known to affect human comprehension of homonyms (Grindrod et al., 2014).

Last but not least, our work extends the analysis of contextualized embeddings to Chinese, in contrast to previous studies that have predominantly focused on English (Haber and Poesio, 2021) and other Western languages such as Spanish (Rivière et al., 2024). Due to its logographic writing system, a single sinogram in Chinese can often represent multiple meanings depending on context (Wang, 1973; Huang and Lee, 2018; Wang et al., 2023). In many cases, the meaning of a bisyllabic compound word, which constitutes a major portion of the vocabulary, can be inferred from its monosyllabic sinogram components. This ambiguity is especially pronounced in Mandarin, whose writing system underwent a revolutionary simplification that merged different traditional sinograms into a single simplified form (e.g., both “台风” meaning “stage manners” and “颱風” meaning “typhoon” are merged as “台风”). This merging complicates the homonym system in Chinese compared to English. The inherent ambiguity of Chinese sinograms thus makes them an ideal testbed for exploring the capabilities and limitations of contextualized representations.

3 Methods

3.1 Synthetic data construction

Existing datasets did not adequately address how language models represent homonyms in context, especially for the Chinese language. Therefore, we created custom datasets containing English and Chinese sentences composed with homonyms via LLMs and validated with experts, following the practical instructions of prompt design from previous studies (Schick and Schütze, 2021; Yu et al., 2023). Prompts for generating English sentences included additional instruction to guarantee that the written form of the target homonym remains unchanged (e.g., no inflections, capitalization, etc.). It is noted that all homonyms we used in the current paper are also homographs and homophones, as they do not differ in both orthographical and phonological representations. Details can be found in Appendix A.1.

To select appropriate homonyms, we referred to existing and established resources. For English, we referred to the *British eDom Norms* database (Maciejewski and Klepousniotou, 2016). The database includes 100 homonyms that have two unrelated meanings, with the relative frequency of each meaning rated by 100 monolingual British-English native speakers aged from 19 - 39 (mean 28.1 $\pm$ 5.3). For Chinese, due to the lack of a suitable existing dataset, we curated one on our own by collecting possible homonyms from a comprehensive Chinese dictionary *XianDai HanYu CiDian (7th Edition)*. After identifying the homonym candidates, we designed a prompt protocol to instruct LLMs to construct sentences. An example is shown in Figure 1, with the full prompt in Appendices A.2 and A.3.

Figure 1: An example prompt for generating sentence pairs that illustrate the distinct meanings of homonyms. Details are provided in the Appendix A.

For sentence generation, we employed GPT-4o (OpenAI, 2024) for English and ChatGLM4 (GLM, 2024) for Chinese. For each homonym, we generated 20 sentences, with 10 sentences corresponding to each of their two senses. Three linguistic experts (including two of the authors) manually examined all sentences, especially for Chinese. 100 homonyms were selected for this study, with half of them having the same POS for both of their senses and the other half having different POS.

3.2 Metrics

3.2.1 Angular similarity between homonyms

To obtain layer-wise homonym representations, the prepared sentences were fed into pre-trained language models to extract token-level representations. Word-level representations were derived by mean pooling on token-level representations, as it had been shown to achieve satisfactory performance to determine word similarity (Bommasani et al., 2020). While cosine similarity between word representations is commonly used to assess the closeness of semantic meanings, we opted to compute the angular similarity. This choice was made because the cosine similarity varies nonlinearly as higher values represent progressively smaller angular differences (see Figure 4 for visualization). Angle-based measures have also been shown to improve embedding performance (Cer et al., 2018). The angular similarity is defined as in Equation 1:

$$AngSim = 90 - \arccos(CosSim) \times \frac{180}{\pi} \quad (1)$$

where $CosSim$ is the cosine similarity between any two word representations.

To properly assess contextual disambiguation of homonyms, it is essential to consider both same-sense and cross-sense similarities. While considering the cross-sense similarity alone, if one observes that it is close to the baseline, the intuition might be that the model shows satisfactory contextualization capability because the model treats senses as distinct as random words. However, this intuition cannot be confirmed without examining same-sense similarity. If a model successfully encodes the word sense information, it is expected that the similarity between same-sense representations to be higher than that of the cross-sense. Therefore, it is essential to consider both same-sense and cross-sense similarity, as relying on either one of them may lead to incomplete or biased interpretations. By examining both metrics alongside their difference, we can quantify how well a model uses context to differentiate between various word senses while maintaining consistent representations for the same sense. The same-sense and cross-sense angular similarity are defined as in Equation 2 and 3 respectively.

$$AngSim_{same}(l, w) = \mathbb{E}\left[\sum_{\substack{i,j=1 \\ i \neq j}}^{n} \angle(f_l(w_i^s), f_l(w_j^s))\right] \tag{2}$$

$$AngSim_{cross}(l, w) = \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} \angle(f_l(w_i^1), f_l(w_j^2))\right] \tag{3}$$

where both of them are the averaged angular similarity between hidden representations of the same or different senses of the homonym w at layer l. $\angle$ is the angular similarity function (Equation 1); f_l is the function to map from a homonym in a sentence to its hidden representation at layer l; n is 10; w_i^1 and w_j^2 correspond to the homonym of meaning 1 and meaning 2 in the i-th and j-th sentence. The $\mathbb{E}$ denotes that we are averaging all angular similarity values computed. Note that we denoted $i \neq j$ in Equation 2 as we do not want to compute the similarity of the two homonyms that come from the same sentence.

3.2.2 Adjusting for anisotropy

There is an ongoing debate regarding the implications of anisotropy. Some studies propose that it is beneficial (Biś et al., 2021; Hämmerl et al., 2023), while others argue it could be potentially detrimental and impair task performance (Ding et al., 2022; Rudman and Eickhoff, 2024; Mickus et al., 2024). Although we recognize the significance of anisotropy and various anisotropy measures, our study does not aim to resolve this debate. Instead, we follow a similar approach as in Ethayarajh (2019), which involves adjusting the similarity measure for anisotropy through a baseline computed as the similarity between randomly sampled words. We computed the same-sense and cross-sense baseline as the angular similarity between the randomly sampled words within each set of sentences and between the two sets, respectively (Details can be found in Appendix C). Since different sentences vary in linguistic features/information (syntax, vocabulary, etc.), both the baseline and angular similarity would be affected. Correcting the angular similarity with their respective baselines can therefore isolate the observed measure from other linguistic factors.

3.2.3 Disambiguation score (*D-score*)

After computing both same-sense and cross-sense adjusted angular similarity ($AngSimAdj_{same}$ and $AngSimAdj_{cross}$), we computed their difference and termed it as the disambiguation score (*D-score*):

$$\textit{D-score} = AngSimAdj_{same} - AngSimAdj_{cross} \tag{4}$$

The *D-score* is our primary measure in this study. It measures how much the model relies on context to modulate representations (contextual disambiguation). A larger *D-score* indicates robust disambiguation, while a smaller *D-score* suggests poor differentiation, the model either conflates senses or is insensitive to the context. The *D-score* ranges from 0 to a theoretical maximum of 90.

3.3 Models

Experiments were conducted on 21 models from four different model families: BERT (Devlin et al., 2019; Liu et al., 2019b; He et al., 2023), GPT-2 (Radford et al., 2019), Llama3 (Llama, 2024), and Qwen (Yang et al., 2024). These four models were chosen because they are (1) open-weights, (2) representing both PLM and LLMs, (3) various model sizes are available and (4) Qwen is particularly trained with a significant amount of Chinese data. *bert-base-chinese* was not included in the English analyses since its tokenizer could not effectively tokenize English input. *bert-base-uncased* and *bert-large-uncased* were not included in the Chinese analyses since their tokenizers do not include most of the Chinese characters in their vocabularies. Unrecognized tokens would be represented by UNK tokens by these tokenizers, such that the computation of meaningful word representations is not possible. More details are listed in Table 1 and 3.

4 Analyses

The results on layer-wise representation of English and Chinese homonyms are shown in Figure 2, showing the averaged *D-score* over all homonyms. Table 1 shows layers at which the *D-scores* are highest. Hereafter we will describe the layer depth as: lower (Layer Depth (%) $\leq$ 33%); middle (33% $<$ Layer Depth (%) $\leq$ 67%) and higher (67% $<$ Layer Depth (%)). The details of all statistical analyses conducted in this Section can be found in Appendix D, all multiple comparisons were corrected via false discovery rate (FDR; Benjamini and Hochberg, 1995). All results, tables and figures can be found in our online repository `https://github.com/neurothew/exploring-homonym-rep-in-llm`.

4.1 Contextualization of English homonyms

For each of both Sections 4.1 and 4.2, we conducted two major statistical tests. For each language, we first fitted a linear mixed effect model, with *D-score* as the dependent variable, layer as the independent variable, and a word-specific random intercept. Through the fitted models, it can be observed that the main effect of layer was significant for all models and language (Appendix Table 4). Then, to examine the differences of the best *D-score* (*D-score*$_{best}$ hereafter) between models, we fitted a linear mixed effect model with *D-score*$_{best}$ being the dependent variable and model being the independent variable, and a word-specific random intercept (details in Appendix D.1). The results on layer-wise representation of English and Chinese homonyms are shown in Table 1, and the upper and lower rows of Figure 2 respectively.

For BERT model family (Devlin et al., 2019; Liu et al., 2019b; He et al., 2023), the layer-wise *D-score* trajectories differ drastically among models. In general, BERT-based models perform best in the middle and higher layers (Table 1). *deberta-v3-large* has the highest *D-score*$_{best}$ (25.32, $ps <$ .001) among BERT model family at the middle layer.

For GPT2 family (Radford et al., 2019), the layer-wise *D-score* trajectories of all four models increase from lower layers and peak in higher layers. Numerically, *gpt2-xl* has the highest *D-score*$_{best}$ at 15.83 at the final layer, although not significantly different from *gpt2-large* ($p = .865$).

For Llama3 family (Llama, 2024), the *D-score* trajectories of the three models almost overlap with each other. All three lines show a rapid increase in the lower layers, reaching a global peak and then decreasing in the middle layers, following a small peak in the later layers. The highest *D-score*$_{best}$ is observed in *Llama-3.1-8B* (16.87) at the lower layer.

For Qwen2.5 family (Yang et al., 2024), similar to Llama3 family, the layer-wise *D-score* trajectories are similar across all three models, which increase in lower layers, peak in lower to middle layers and retrace in higher layers. While *Qwen2.5-1.5B* has the highest *D-score*$_{best}$ (15.11), the *D-score*$_{best}$ from all three models are not significantly different ($ps > .05$).

4.2 Contextualization of Chinese homonyms

For BERT model family, it can be observed that the layer-wise *D-score* trajectories differ significantly among models. *mdeberta-v3-base* has the highest *D-score*$_{best}$ (10.56) among BERT model family at higher layer. *bert-base-chinese*, the original BERT model variant pre-trained on Chinese Wikipedia (Devlin et al., 2019), also has a comparable performance. The two models exhibit significantly better performance than others ($ps < .001$).

Within the GPT2 family, the *D-score* trajectories are almost entirely flat, with their maximum in a wide range spanning middle layers. While the two larger models have relatively higher *D-score*$_{best}$, the *D-score*$_{best}$ are not significantly different within the family ($ps > .05$).

For Llama3 family, the layer-wise *D-score* trajectories are similar across all models. They increase rapidly in the lower layers, peak in the lower to middle layers, and decline until reaching a plateau in the middle and higher layers. *Llama3.2-3B* has the highest *D-score*$_{best}$ at 6.66 at a higher layer.

For Qwen2.5 family, the layer-wise *D-score* trajectories are similar across all three models, which increase rapidly in lower layers, reach a plateau in lower and middle layers, and then decline in higher layers. The three models exhibit similar *D-score*$_{best}$ ($ps > .05$). The highest *D-score*$_{best}$ is observed in *Qwen2.5-3B* at 6.99 at a middle layer.

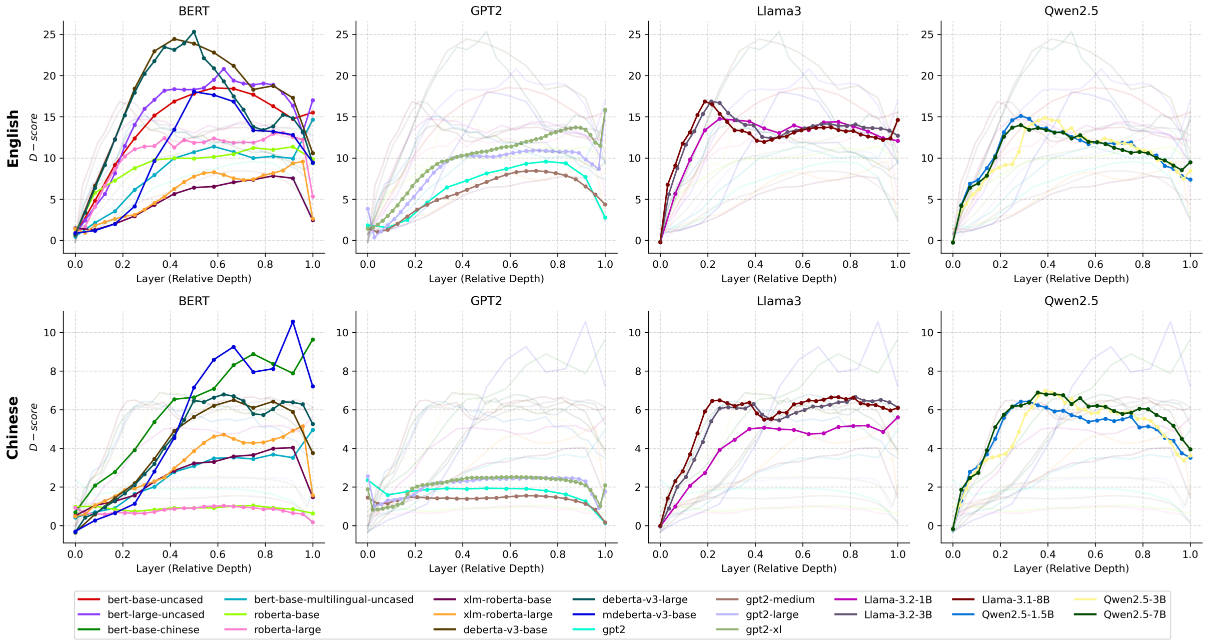

Figure 2: Average *D-score* between the contextualized representation pairs of English (upper) and Chinese (lower) homonyms across different language models. The x-axis indicates the relative layer depth, computed as the layer number divided by the total number of layers for each model. Note that between English and Chinese, the scale of the y-axis is different.

			English				Chinese			
Model Family	Model	Parameters	Layer	Layer Depth (%)	Layer Depth	*D-score*	Layer	Layer Depth (%)	Layer Depth	*D-score*
BERT	bert-base-uncased	110M	7.00	58.30	middle	18.50	—	—	—	—
	bert-large-uncased	340M	15.00	62.50	middle	20.81	—	—	—	—
	bert-base-chinese	102M	—	—	—	—	12.00	100.00	higher	9.63
	bert-base-multilingual-uncased	167M	12.00	100.00	higher	14.65	12.00	100.00	higher	4.96
	roberta-base	125M	11.00	91.70	higher	11.36	9.00	75.00	higher	1.04
	roberta-large	355M	21.00	87.50	higher	13.04	15.00	62.50	middle	1.03
	xlm-roberta-base	278M	10.00	83.30	higher	7.82	11.00	91.70	higher	4.04
	xlm-roberta-large	560M	23.00	95.80	higher	9.57	23.00	95.80	higher	5.15
	deberta-v3-base	183M	5.00	41.70	middle	24.44	8.00	66.70	middle	6.50
	deberta-v3-large	434M	12.00	50.00	middle	**25.32**	15.00	62.50	middle	6.79
	mdeberta-v3-base	278M	6.00	50.00	middle	18.04	11.00	91.70	higher	**10.56**
GPT2	gpt2	124M	9.00	75.00	higher	9.57	6.00	50.00	middle	1.93
	gpt2-medium	355M	17.00	70.80	higher	8.44	16.00	66.70	middle	1.55
	gpt2-large	774M	36.00	100.00	higher	15.72	26.00	72.20	higher	2.49
	gpt2-xl	1.5B	48.00	100.00	higher	**15.83**	29.00	60.40	middle	**2.53**
Llama3	Llama-3.2-1B	1B	4.00	25.00	lower	14.77	16.00	100.00	higher	5.61
	Llama-3.2-3B	3B	6.00	21.40	lower	16.86	23.00	82.10	higher	**6.66**
	Llama-3.1-8B	8B	6.00	18.80	lower	**16.87**	24.00	75.00	higher	6.65
Qwen2.5	**Qwen2.5-1.5B**	1.5B	8.00	28.60	lower	**15.11**	9.00	32.10	lower	6.43
	Qwen2.5-3B	3B	14.00	38.90	middle	14.89	14.00	38.90	middle	**6.99**
	Qwen2.5-7B	7B	8.00	28.60	lower	13.99	10.00	35.70	middle	6.89

Table 1: The best layer of representing English and Chinese homonyms of each model based on *D-score*. The model with the highest *D-score* within each model family is in boldface for two languages. Layer Depth is categorized as: lower (Layer Depth (%) $\leq 33\%$); middle ($33\% <$ Layer Depth (%) $\leq 67\%$) and higher ($67\% <$ Layer Depth (%)).

4.3 Layer-wise comparisons of English and Chinese homonym representations

To examine the differences between English and Chinese homonym representations, we fitted a linear mixed effect model per PLM/LLM with *D-score* being the dependent variable, language, layer and their interactions being the independent variables, and a word-specific random intercept (details in Appendix D.2).

First of all, the main effect of language was found to be significant in which the *D-scores* computed from English homonym representations were significantly higher than from Chinese (Table 5). This suggests that the included LLMs are better at contextualizing English homonyms. It is of no surprise since many of the models were all pre-

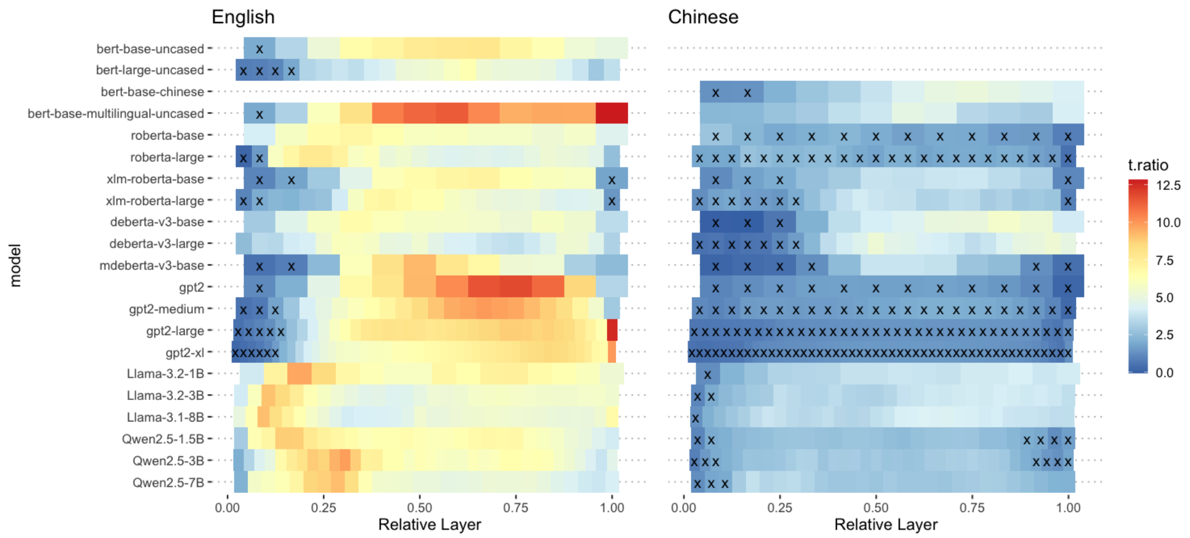

Figure 3: A heatmap showing the t-ratio computed from the post-hoc comparisons between same-POS *D-score* and different-POS *D-score*. Multiple comparisons were corrected via FDR (Benjamini and Hochberg, 1995). Nonsignificant results are marked with a cross.

trained with English-dominant data. For Chinese homonym representations, models involving multilingual and Chinese training data and achieve better *D-Scores*, such as *mdeberta-v3-base* and *bert-base-Chinese*. We suggest that this can be attributed to the fact that these models included a significant portion of Chinese data in the pre-training phase. On the other hand, as discussed in Section 2, unlike alphabets, each sinogram in Chinese can represent multiple meanings which can lead to the two-sinogram word becoming a homonym (Huang and Lee, 2018). For instance, "一/*yi*" can be both "one" and "first". And the homonym "一线/*yi xian*" can be interpreted into "one piece of" or "battlefront" based on its context. It is possible this inherent sinogram-level ambiguity is not captured by the models, causing the lower performance of word-sense disambiguation in Chinese.

Across languages, the layer-wise *D-score* trajectories from the same model families show both similarities and differences. For instance, BERT and DeBERTa based models tend to excel in higher layers for both English and Chinese. In contrast, the trajectories observed from GPT-2 models show significant differences between languages. Trajectories from English homonyms show an increasing trend from lower to higher layers, while those from Chinese homonyms show an inverted U-shaped trend with a broad peak at middle layers.

Similar trajectories may suggest that models employ comparable strategies for homonym contextualization in both English and Chinese, possibly indicating the use of language-universal features or processing mechanisms. Conversely, divergent trajectories imply that models adapt their approach based on language-specific characteristics, recognizing that different linguistic cues or structures may be more relevant for homonym disambiguation in one language versus another. Further research with linguistically well-designed sentences would be needed to confirm these hypotheses about the models' internal representations.

4.4 Does increasing model size help with differentiating homonym representations?

While larger models generally perform better than smaller models in language tasks (Kaplan et al., 2020), this is not always observed in our results. As we observed from Table 1, in terms of English homonym disambiguation, larger models in BERT and GPT2 families indeed performed better. *Llama-3.2-3B*, however, exhibited a similar performance as *Llama-3.1-8B*; the smallest model of Qwen family even performed best. For Chinese homonym disambiguation, in the GPT2 family the largest model performed best, though *gpt2* is still better than *gpt2-medium*. As such, the homonym disambiguation capability did not consistently scale with the size of the language models in general.

4.5 How does model architecture affect homonym representations?

PLMs and LLMs can roughly be divided into bidirectional or autoregressive models (Yang, 2019). This division is based on which part of the context these models can attend to. Of the four model families, only the BERT model family is bidirectional, while the others are autoregressive.

Regarding English homonym representations, the bidirectional *deberta-v3-large* model has the highest *D-score* among all models tested, surpassing autoregressive models that are around 16 times larger (*Llama-3.1-8B* and *Qwen2.5-7B*). This result is even more significant given that *deberta-v3-base* was pre-trained on a much smaller dataset (same as RoBERTa (He et al., 2023)) than those of LLama3 and Qwen2.5 model families.

Regarding Chinese homonym disambiguation, the multilingual *mdeberta-v3-base* performed the best, while both *deberta-v3-base* and *deberta-v3-large* also performed at a similar level as the best model from Qwen and Llama family. This result highlights the need of multilingual data in homonym disambiguation other than English, as well as the strength of bidirectional architecture.

4.6 Do same or different parts of speech affect homonym representations?

In our main analyses, the *D-score* were compared across homonyms regardless of their POS. To further investigate how POS might modulate the *D-score*, we fitted a linear mixed effect model per language and PLM/LLM with *D-score* being the dependent variable, layer and POS as the independent variables, and a word-specific random intercept (details can be found in Appendix D.3). The interaction effects between POS and layer were significant except for *gpt2*, *gpt2-large*, *gpt2-xl* and *roberta-large* on Chinese homonyms (Table 6). Post-hoc pairwise comparisons were conducted to examine at which layer the differences (different-POS *D-score* − same-POS *D-score*) were significant. A heatmap showing the t-ratio resulted from the comparisons is shown in Figure 3, and the layers where the t-ratios were largest are shown in Table 7. A higher t-ratio suggests that the difference observed is more robust, and less likely to be caused by random noises. In other words, the POS information contributes more to the prediction of the *D-score* when the t-ratio is higher. First of all, it can be observed that almost all t-ratios were positive, indicating that the different-POS *D-scores* were always larger than the same-POS *D-scores*. These results results suggest that the two meanings of same-POS homonyms are more challenging for models to differentiate than those of different-POS homonyms. Intriguingly, this finding contrasts with previous studies on humans, which showed that different-POS homonyms elicited greater neural activations and required higher cognitive efforts (Grindrod et al., 2014). Our analysis indicates that PLMs and LLMs are actually more adept at representing different-POS homonyms.

This contrast between human and machine could potentially be attributed to the underlying processing mechanism. While humans need to actively switch between grammatical frameworks to interpret different-POS homonyms (Federmeier et al., 2000), the models already possess all relevant POS information in the embedding space during inference without a switching process. However, this benefit in resolving different-POS homonyms might be an obstacle for resolving same-POS homonyms in models.

Second, it can be observed from Figure 3 that the t-ratios resulted from English homonym comparisons were much larger than that of Chinese. This difference may lie in the fact that the markedness of POS information are asymmetric between English and Chinese (Greenberg, 1966; Wang, 1973). For instance, the plural form or past tense are marked by inflectional suffixes in English, while it is unmarked in Chinese. For instance, "关门 (*guan men*)" can function both as a verb (to close the door) and as a noun (the gate of a pass). No additional suffix will be added when it is used to express plural meaning in its noun function, such as "他们要在所有的关门设防 (*ta men yao zai suo you de guan men she fang*) / they will set up defences at all gates of the pass)". It is possible that the marked POS information in its orthographic representation of English is encoded and stored in its hidden representations during training, which lacks in Chinese. It is likely that this additional embedded POS information in English helps improving models' performance in contextualizing homonym representations.

Third, it can be observed that the best POS layers (where the POS contributes the most, Figure 3) and the best *D-score* layers (where the *D-score* is the highest, Table 1) are not always equal (more details can be found in Table 7). For English, the best POS layers are almost always the earlier layer (except for *bert-base-uncased*). For Chinese, this pattern can still be observed (e.g., *mdeberta-v3-base*), albeit less consistent for other models. Language, as a complex adaptive system (Mufwene et al., 2017), employs various strategies, such as POS, word order, and contextual cues to address lexical ambiguity that arises from the reuse of existing forms (Ogura and Wang, 2022). Our results indicate that, beyond relying on POS information at lower layers, the models also incorporate other linguistic information at higher layers to enhance their disambiguation performance.

5 Conclusion

We presented a comprehensive analysis on English and Chinese homonym representations, spanning same and different POS categories, across 21 PLMs and LLMs in four model families. Patterns of layer-wise trajectories of *D-Score* were found to differ across models and languages, suggesting that these different models might excel differently in encoding distinct levels of linguistic information (e.g., meanings, POS, etc..) towards encoding distinct homonym representations. Model architecture and pretraining data portion appear to be important factors as bidirectional and multilingual models excel in homonym disambiguation. We also highlighted the functional role of POS in word-sense disambiguation as models disambiguate same-POS homonyms better than different-POS homonyms. Intriguingly, this is opposite to what have been observed in human studies. To conclude, the individual differences between LLMs complicate our understanding of their inner workings, there is a need to conduct rigorous, controlled experiments using purposefully manipulated input, in order to enhance interpretability in future LLM research.

Limitations

Several limitations of this study need to be addressed. Firstly, LLM-generated sentences may be biased and unnatural. Although most English sentences and all Chinese sentences were manually examined, this examination was limited to semantic acceptability and syntactic correctness. This unnatural issue is particularly pronounced for low-frequency homonyms.

Secondly, to facilitate the computation of similarity for target words, all English sentences were restricted to using the base form of homonyms. This constraint creates an artificial situation where English homonyms must rely solely on context for POS differentiation, without the benefit of morphological strategies such as suffixation. As a result, this limitation may reduce the observed differences between English and Chinese in representing homonyms with the same or different POS, given that Chinese inherently has limited morphological variation. Future studies should compare these two types of homonyms using sentences with more diverse morphological transformations.

Thirdly, while we observed that LLMs performed differently in representing same-POS and different-POS homonyms compared to findings from human research, our study lacks direct human-related data for comparison. A recent dataset on English with human judgments on meaning relatedness has been developed, finding that humans and models perform similarly in word-sense disambiguation (Trott and Bergen, 2021). Future research should examine whether similar phenomena can be observed in the Chinese context and directly compare model performance with human judgments across languages.

Reproducibility

We have made our dataset and codes available publicly at `https://github.com/neurothew/exploring-homonym-rep-in-llm`.

Acknowledgement

This research was funded by the Departmental General Research Fund (P0051135) provided by the Hong Kong Polytechnic University awarded to W.S.W. (PI) and the HKRGC Postdoctoral Fellowship Scheme awarded to M.K-H.M.. We thank the University Research Facility for Big Data Analytics (UBDA), HKPolyU, for making the GPU virtual machines available for the computations involved.

References

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2014. Fitting Linear Mixed-Effects Models using lme4.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. 2021. Too Much in Common: Shifting of Embeddings in Transformer Language Models and its Implications.

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5117–5130, Online. Association for Computational Linguistics.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal Sentence Encoder. *arXiv preprint*. ArXiv:1803.11175 [cs].

Ting-Yun Chang and Yun-Nung Chen. 2019. What Does This Word Mean? Explaining Contextualized Embeddings with Natural Language Definition. 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 6064–6070, Hong Kong, China. Association for Computational Linguistics.

Alebachew Chiche and Betselot Yitagesu. 2022. Part of speech tagging: a systematic review of deep learning and machine learning approaches. *Journal of Big Data*, 9(1):10.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

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

Yue Ding, Karolis Martinkus, Damian Pascual, Simon Clematide, and Roger Wattenhofer. 2022. On Isotropy Calibration of Transformer Models. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

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.

Kara D. Federmeier, Jessica B. Segal, Tania Lombrozo, and Marta Kutas. 2000. Brain responses to nouns, verbs and class-ambiguous words in context. *Brain*, 123(12):2552–2566.

Lyn Frazier and Keith Rayner. 1990. Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, 29(2):181–200.

Silvia P. Gennari, Maryellen C. MacDonald, Bradley R. Postle, and Mark S. Seidenberg. 2007. Context-dependent interpretation of words: Evidence for interactive neural processes. *NeuroImage*, 35(3):1278–1286.

Team GLM. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *Preprint*, arXiv:2406.12793.

Joseph H Greenberg. 1966. *Language universals*. Mouton The Hague.

Christopher M. Grindrod, Emily O. Garnett, Svetlana Malyutina, and Dirk B. den Ouden. 2014. Effects of representational distance between meanings on the neural correlates of semantic ambiguity. *Brain and Language*, 139:23–35.

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.

Katharina Hämmerl, Alina Fastowski, Jindřich Libovický, and Alexander Fraser. 2023. Exploring Anisotropy and Outliers in Multilingual Language Models for Cross-Lingual Semantic Sentence Similarity. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7023–7037, Toronto, Canada. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. *Preprint*, arXiv:2111.09543.

Hsu-Wen Huang and Chia-Ying Lee. 2018. Number of meanings and number of senses: An erp study of sublexical ambiguities in reading chinese disyllabic compounds. *Frontiers in Psychology*, 9:324.

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.

Brenden M. Lake and Gregory L. Murphy. 2023. Word meaning in minds and machines. *Psychological Review*, 130(2):401–431.

Russell V. Lenth. 2025. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.10.7-100001, https://rvlenth.github.io/emmeans/.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic Knowledge and Transferability of Contextual Representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *Preprint*, arXiv:1907.11692.

Team Llama. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Maryellen C. MacDonald, Neal J. Pearlmutter, and Mark S. Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676–703.

Greg Maciejewski and Ekaterini Klepousniotou. 2016. Relative Meaning Frequencies for 100 Homonyms: British eDom Norms. *Journal of Open Psychology Data*, 4(1):e6.

Timothee Mickus, Stig-Arne Grönroos, and Joseph Attieh. 2024. Isotropy, Clusters, and Classifiers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 75–84, Bangkok, Thailand. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *Preprint*, arXiv:1301.3781.

Salikoko S Mufwene, Christophe Coupé, and François Pellegrino. 2017. *Complexity in language: Developmental and evolutionary perspectives*. Cambridge University Press.

Mieko Ogura and William S-Y Wang. 2022. Ambiguity resolution and the evolution of homophones in english. *Amsterdam Studies in the Theory and History of Linguistic Science. Series 4: Current Issues in Linguistic Theory*, 358:61–90.

OpenAI. 2024. Hello GPT-4o. https://openai.com/index/hello-gpt-4o/.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

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

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Pamela D. Rivière, Anne L. Beatty-Martínez, and Sean Trott. 2024. Bidirectional Transformer Representations of (Spanish) Ambiguous Words in Context: A New Lexical Resource and Empirical Analysis. *Preprint*, arXiv:2406.14678.

Jennifer Rodd. 2018. *Lexical Ambiguity*, pages 95–117. Oxford University Press.

William Rudman and Carsten Eickhoff. 2024. Stable Anisotropic Regularization. *Preprint*, arXiv:2305.19358.

Timo Schick and Hinrich Schütze. 2021. Generating Datasets with Pretrained Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6943–6951, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rita Sevastjanova, Aikaterini-Lida Kalouli, Christin Beck, Hanna Schäfer, and Mennatallah El-Assady. 2021. Explaining Contextualization in Language Models using Visual Analytics. 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 464–476, Online. Association for Computational Linguistics.

Sean Trott and Benjamin Bergen. 2021. Raw-c: Relatedness of ambiguous words–in context (a new lexical resource for english). *arXiv preprint arXiv:2105.13266*.

Tianqi Wang, Xu Xu, Xurong Xie, and Manwa Lawrence Ng. 2023. Probing lexical ambiguity in chinese characters via their word formations: Convergence of perceived and computed metrics. *Cognitive Science*, 47(11):e13379.

William S-Y Wang. 1973. The chinese language. *Scientific American*, 228(2):50–63.

William S-Y Wang. 2011. Ambiguity in Language. *Korea Journal of Chinese Language and Literature*, 1:3–20.

Kyra Wilson and Alec Marantz. 2022. Contextual Embeddings Can Distinguish Homonymy from Polysemy in a Human-Like Way. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 144–155, Trento, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *Preprint*, arXiv:1910.03771.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Technical Report. *Preprint*, arXiv:2407.10671.

Zhilin Yang. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias. *Preprint*, arXiv:2306.15895.

A Appendix

A.1 Procedures for constructing the synthetic dataset

To select appropriate English homonyms, we referred to an existing dataset, the *British eDom Norms* database (Maciejewski and Klepousniotou, 2016). The database includes 100 homonyms that have two unrelated meanings, with the relative frequency of each meaning rated by 100 monolingual British-English native speakers aged from 19 - 39 (mean 28.1 $\pm$ 5.3). This dataset also provided psycholinguistic properties such as semantic distance and relatedness, which is valuable for comparing the similarities and differences in language processing between humans and machines. For Chinese, due to the lack of a suitable existing dataset, we curated one on our own. We first collected all possible homonyms from a comprehensive Chinese dictionary *XianDai HanYu CiDian (7th Edition)*, whose two or three meanings are semantically and etymologically unrelated. We then applied strict criteria, excluding words used in early vernacular Chinese, rare Buddhist terms, dialects, and those that could be variant sinograms, polyphonic sinograms, or elements of synonyms. This process yielded 329 Chinese homonyms, including 130 with different POS and 193 with the same POS.

After identifying the homonym candidates, we designed a prompt protocol for each homonym to instruct LLMs to construct sentences demonstrating the specified meanings of the target homonym. Both English and Chinese prompts follow the same general structure. Note that the exact communication flow between the human expert and LLMs differed for Chinese and English; details are provided in Appendix A.2 and A.3. Besides here is one key difference: English prompts require that the written form of the target homonym remains unchanged across different meanings or contexts (e.g., no inflections, capitalization, etc.), whereas Chinese prompts do not include this requirement. This distinction arises because Chinese homonyms consistently use the same characters regardless of their different meanings or contexts.

To build the dataset, we generated 20 sentences for each homonym, with 10 sentences corresponding to each of its two meanings. The Chinese sentences required additional scrutiny due to occasional inconsistencies in adhering to the specified meanings. Consequently, three linguistic experts (including two of the authors) manually examined all Chinese sentences. Any ungrammatical and incorrect sentences that did not include the target homonyms or failed to represent the intended meanings were manually modified or replaced with newly generated sentences to ensure dataset quality. Additionally, we confirmed that each sentence includes the targeted homonym only once. The average accuracy rates of generated sentences are 94.97% for Chinese homonyms whose meanings belong to the same POS and 90.47% for those belonging to different POS.

A.2 English prompt example for sentence construction

This section illustrates prompts used for generating English sentence pairs containing target homonyms. Note that only the human expert instruction is manually input by the authors. To ensure independence between generations, the chat history is cleared before generating each new 10 sentences for each meaning of the homonym.

Human expert: Please generate ten sentences using word "lean" and the semanteme of “lean” here is "to bend or incline ". The length of each generated sentence is around 15 words. Please do not change the morphological structure of the word "fan". Do not capitalize the letters of the word "fan". Do not add a clitic to the word “fan”.

GPT-4o: 1. She likes to lean against the old oak tree while reading her favorite book. 2. The tall grass would often lean towards the ground during the strong summer winds. 3. He would lean forward in his chair to get a better view of the board. 4. You could see the tower slightly lean to one side after the earthquake. 5. As the sun set, the flowers began to lean towards the fading light. 6. During the concert, fans would lean over the barricades to get closer to the stage. 7. The tired hiker decided to lean on his walking stick for a brief rest. 8. The cat loves to lean its body against the warm window on cold days. 9. Watching the game, he would often lean closer to the screen in excitement. 10. The old bookshelf started to lean due to the weight of the heavy books.

Human expert: Please generate ten sentences using word "lean" and the semanteme of “lean” here is " having little flesh ". The length of each generated sentence is around 15 words. Please do not change the morphological structure of the word "fan". Do not capitalize the letters of the word "fan". Do not add a clitic to the word “fan”.

GPT-4o: 1. The greyhound, known for being lean, darted across the field with incredible speed. 2. Her lean figure, honed by years of ballet, moved gracefully across the stage. 3. After months of training, his once bulky frame transformed into a lean, muscular build. 4. The stray cat was lean, surviving on scraps it found in the alleyways. 5. Despite his lean body, he had incredible strength and endurance from years of rock climbing. 6. The lean wolves hunted in packs, their slender bodies moving silently through the forest. 7. The fashion model’s lean silhouette was highlighted by the tailored dress she wore. 8. His diet consisted mainly of lean meats and vegetables to maintain his slim physique. 9. The marathon runner’s lean legs carried her swiftly to the finish line. 10. The lean boxer danced around the ring; his agility unmatched by his heavier opponents.

A.3 Chinese prompt example for sentence construction

This section illustrates prompts used for generating Chinese sentence pairs containing target homonyms. The prompt structure follows a predefined system protocol developed by the authors, which consists of three initial instructions: two provided by human experts and one input from GLMchat. These instructions guide the subsequent sentence generation process (Table 2).

B Visualization of comparison between cosine similarity and angular similarity

The cosine similarity varies nonlinearly as higher values represent progressively smaller angular differences, as shown in Figure 4.

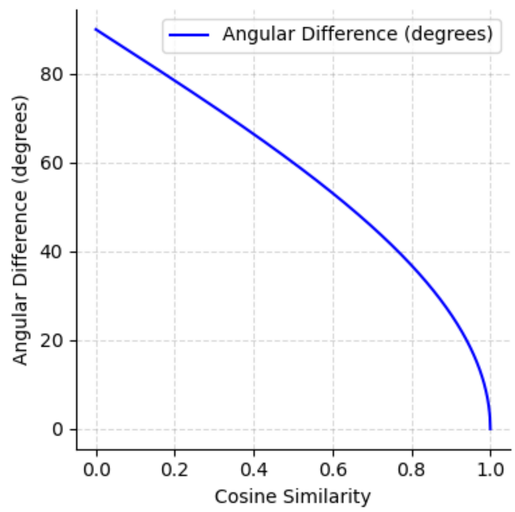

Figure 4: Cosine similarity vs. angular similarity. Cosine similarity varies nonlinearly: higher values represent progressively smaller angular differences.

C Computation of the angular similarity baseline

Following a similar approach as in Ethayarajh (2019), we calculated a baseline as the averaged angular similarity between randomly sampled words from two distinct sentences. Conceptually, the baseline represents the intrinsic similarity between random words. It serves to create an adjusted measure for better interpretation of results. Similar to Section 3.2.1, we computed both same-sense and cross-sense baselines. For brevity, we will only introduce the computation for the cross-sense baseline as follows. Define s_i^1 and s_j^2 as the i-th and j-th sentences composed of the two meanings of a homonym w: w^1 and w^2. We then randomly sampled words r_i and r_j from s_i^1 and s_j^2 respectively, and computed their angular similarity. The sampling was done 30 times for each combination of s_i^1 and s_j^2. Finally, the baseline for homonym w was computed as the average of all angular similarity values across all samples and combinations of sentences. This process is described in Equation 5:

$$\begin{gathered} Baseline(l, w) = \mathbb{E}\left[\sum_{i,j} \angle(f_l(r_i), f_l(r_j))\right] \\ s_i^1 = [r_1^1, r_2^1, ..., r_k^1], s_j^2 = [r_1^2, r_2^2, ..., r_k^2] \\ r_i^1 \neq w_i^1, r_j^2 \neq w_j^2 \end{gathered} \tag{5}$$

where $\angle$ and f_l are defined as in Equation 3, k is the number of words in the corresponding sentence.

D Statistical analyses

We have conducted various statistical analyses to support our claims in the main text. All statistical tests were conducted via custom *R* (R Core Team, 2021) scripts. Linear mixed effect models and post-hoc comparisons were conducted with the *lme4* (Bates et al., 2014) and *emmeans* (Lenth, 2025). All multiple comparisons were corrected via false discovery rate (FDR; Benjamini and Hochberg, 1995). The dependent variable appears on the left-hand side of the tilde (~), and the independent variables (fixed and random effects) are listed on the right. All our linear mixed-effects model notations (Equations 6–9) follow the conventions used in the *lme4* package in *R*.

D.1 Contextualization of English and Chinese homonyms

In Section 4.1 and 4.2, we conducted two statistical analyses to backup our claims and descriptions over the *D-score* trajectories. Firstly, we examined whether the main effect of layer was significant for all models and languages. For each model and language, a linear mixed effect model was constructed as in Equation 6:

$$\textit{D-score} \sim \text{layer} + (1|\text{word}) \tag{6}$$

An omnibus F-test was then conducted for each fitted model, the results are shown in Table 4, where the main effects of layer were significant for all models and languages. Second, to examine the differences between the best *D-score* (notated as *D-score*$_{best}$ in the main text) across models, we fitted one linear mixed effect model per language, with *D-score*$_{best}$ being the dependent variable, model as the independent variable and word-specific random intercept. The model is shown in Equation 7.

$$\textit{D-score}_{best} \sim \text{model} + (1|\text{word}) \tag{7}$$

We then conducted post-hoc pairwise comparisons between every two models. Considering the substantial number of comparisons involved, the result table has been uploaded in our publicly available repository (`https://github.com/neurothew/exploring-homonym-rep-in-llm`).

D.2 Layer-wise comparisons of English and Chinese homonym representations

In Section 4.3, we conducted one statistical analysis, aiming to illustrate the overall differences between the computed *D-scores* from the English and Chinese homonym representations. A linear mixed effect model was fitted per language model, with *D-score* as the dependent variable, language and layer as independent variables, with a word-specific random intercept, as in Equation 8:

$$\textit{D-score} \sim \text{language} * \text{layer} + (1|\text{word}) \tag{8}$$

After that, post-hoc comparisons were conducted to examine the averaged differences between the two languages across all layers. The results are shown in Table 5.

D.3 Do same or different parts of speech affect homonym representations?

In Section 4.6, we conducted one statistical analysis, aiming to examine how POS information might modulate the *D-scores*. For each model and language, we fitted a linear mixed effect model, with *D-score* as the dependent variable, layer and POS as independent variables, with a word-specific random intercept. The model is shown in Equation 9:

$$\textit{D-score} \sim \text{layer} * \text{POS} + (1|\text{word}) \tag{9}$$

Type III ANOVA results of each model are shown in Table 6. Post-hoc pairwise comparisons were conducted to examine at which layer the differences (different-POS *D-score* − same-POS *D-score*) were significant. The results are visualized in Figure 3 as a heatmap in the main text. Considering the substantial number of comparisons involved, the result table has been uploaded in our publicly available repository (`https://github.com/neurothew/exploring-homonym-rep-in-llm`).

	prompt	translation
Human expert	你是一名语言学家，非常擅长回答与词汇歧义相关的问题， 你的任务解释同形词的意思并用同形词造句。 请问“上任”是个同形词吗?	You are a linguist who is very good at answering questions related to lexical ambiguity, and you have been tasked with explaining the meaning of homonyms and making sentences with them. Is ‘shàngrèn’ a homonym?
GLMchat	是的，“上任”是一个同形词。 它有两个意思，一个是“动词，指官吏就职”， 一个是指“名词，称前一任的官吏”。	Yes, ‘shàngrèn’ is a homonym that has two meanings. One means ‘verb, to bend or incline’ and the other means ‘noun, the predecessor.
Human expert	请分别用“上任”造二十个句子， 其中十个句子表达“动词，指官吏就职”的意思， 另外十个句子表达“名词，称前一任的官吏”的意思。	Please use ‘lean’ to make twenty sentences, ten of which express the meaning of ‘verb, to take a post’ and the other ten express the meaning of ‘noun, the predecessor’.
GLMchat	以下是使用“上任”这个词来表示“动词，指官吏就职”的10个句子: 1. 他刚上任不久，还在熟悉新的工作环境。 2. 市长上任后，立即开始着手解决交通拥堵问题。 3. 新上任的校长对教育改革充满热情。 4. 经过严格的选拔，她终于在上任成为了部门的负责人。 5. 他上任的第一天，就给全体员工开了一个鼓舞人心的会议。 6. 这位年轻的法官上任后，受到了社会各界的关注。 7. 上任仅一周，他已经展现出了出色的领导能力。 8. 她上任后，推动了公司的技术创新。 9. 他刚上任就被派去处理一个棘手的案子。 10. 上任以来，他一直致力于改善市民的生活质量。 以下是使用“上任”这个词来表示“名词，称前一任的官吏”的10个句子: 1. 我们需要分析上任领导留下的政策，看看有哪些需要改进。 2. 上任市长在位期间，城市的基础设施得到了很大的改善。 3. 新市长表示，他将继承上任市长的良好传统。 4. 上任经理的离职，给公司带来了不少挑战。 5. 他试图了解上任留下的工作计划，以便更好地继续推进项目。 6. 上任部长在任时，推出了一系列受欢迎的教育政策。 7. 员工们对上任领导的离开感到非常惋惜。 8. 这座大桥是上任政府投资建设的。 9. 她在评价上任领导的工作时，给予了高度的评价。 10. 上任领导在离任前，留下了一套完善的管理体系。	Here are 10 sentences that use the word ‘shàngrèn’ to mean ‘verb, to bend or incline’: 1. He has just taken office and is still familiarizing himself with his new working environment. 2. The mayor began to tackle the problem of traffic congestion as soon as he took office. 3. 3. The new headmaster is enthusiastic about education reform. 4. 4. After a rigorous selection process, she finally became the head of the department on her first day in office. 5. On his first day in office, he gave a pep talk to all the staff. 6. The young judge received a lot of attention from the community when he took office. 7. After only one week in office, he has already demonstrated excellent leadership skills. 8. She has promoted technological innovation in the company since she took office. 9. He was assigned to a difficult case just after he took office. 10. Since taking office, he has been working to improve the quality of life of the citizens. Here are 10 sentences that use the word ‘shàngrèn’ to mean ‘noun, the predecessor’: 1. We need to analyse the policies left behind by the previous leader to see what needs to be improved. 2. During the previous mayor’s tenure, the city’s infrastructure was greatly improved. 3. 3. The new mayor said that he would carry on the good traditions of the previous mayor. 4. The departure of the previous manager has created a number of challenges for the company. 5. He tried to understand the work programme left behind by his predecessor so that he could better continue with the project. 6. When the previous minister was in office, he introduced a number of popular education policies. 7. The employees were very sorry to see the previous leader leave. 8. The bridge was invested in by the previous government. 9. She spoke highly of the work of the previous leader. 10. The previous leader left behind a sound management system before he left office.

Table 2: Prompts used for generating Chinese sentence pairs containing target homonyms.

Model family	Languages		Number of layer	Parameters
	English	Chinese		
BERT	bert-base-uncased		12	110M
	bert-large-uncased		24	340M
		bert-base-chinese	12	102M
	bert-base-multilingual-uncased	bert-base-multilingual-uncased	12	167M
	roberta-base	roberta-base	12	125M
	roberta-large	roberta-large	12	355M
	xlm-roberta-base	xlm-roberta-base	12	278M
	xlm-roberta-large	xlm-roberta-large	12	560M
	deberta-v3-based	deberta-v3-based	12	183M
	deberta-v3-large	deberta-v3-large	12	434M
	mdeberta-v3-based	mdeberta-v3-based	12	278M
GPT	gpt2	gpt2	12	124M
	gpt2-medium	gpt2-medium	24	355M
	gpt2-large	gpt2-large	36	774M
	gpt2-xl	gpt2-xl	48	1.5B
Llama	Llama-3.2-1B	Llama-3.2-1B	16	1B
	Llama-3.2-3B	Llama-3.2-3B	28	3B
	Llama-3.1-8B	Llama-3.1-8B	32	8B
Qwen	Qwen-2.5-1.5B	Qwen-2.5-1.5B	28	1.5B
	Qwen-2.5-3B	Qwen-2.5-3B	36	3B
	Qwen-2.5-7B	Qwen-2.5-7B	28	7B

Table 3: List of models included in our analyses. All models are available on Huggingface via the *transformers* library (Wolf et al., 2020).

Model	Language	Factor	df1	df2	F-ratio	p-value
bert-base-uncased	en	layer	11.000	1,089.000	275.838	$< 0.001^{***}$
bert-large-uncased	en	layer	23.000	2,277.000	347.023	$< 0.001^{***}$
bert-base-chinese	zh	layer	11.000	1,089.000	176.867	$< 0.001^{***}$
bert-base-multilingual-uncased	en	layer	11.000	1,089.000	241.625	$< 0.001^{***}$
bert-base-multilingual-uncased	zh	layer	11.000	1,089.000	112.165	$< 0.001^{***}$
roberta-base	en	layer	11.000	1,089.000	105.724	$< 0.001^{***}$
roberta-base	zh	layer	11.000	1,089.000	9.527	$< 0.001^{***}$
roberta-large	en	layer	23.000	2,277.000	266.928	$< 0.001^{***}$
roberta-large	zh	layer	23.000	2,277.000	28.143	$< 0.001^{***}$
xlm-roberta-base	en	layer	11.000	1,089.000	209.648	$< 0.001^{***}$
xlm-roberta-base	zh	layer	11.000	1,089.000	126.957	$< 0.001^{***}$
xlm-roberta-large	en	layer	23.000	2,277.000	268.824	$< 0.001^{***}$
xlm-roberta-large	zh	layer	23.000	2,277.000	148.892	$< 0.001^{***}$
deberta-v3-base	en	layer	11.000	1,089.000	230.244	$< 0.001^{***}$
deberta-v3-base	zh	layer	11.000	1,089.000	103.987	$< 0.001^{***}$
deberta-v3-large	en	layer	23.000	2,277.000	338.160	$< 0.001^{***}$
deberta-v3-large	zh	layer	23.000	2,277.000	143.651	$< 0.001^{***}$
mdeberta-v3-base	en	layer	11.000	1,089.000	139.872	$< 0.001^{***}$
mdeberta-v3-base	zh	layer	11.000	1,089.000	91.247	$< 0.001^{***}$
gpt2	en	layer	11.000	1,089.000	158.189	$< 0.001^{***}$
gpt2	zh	layer	11.000	1,089.000	11.645	$< 0.001^{***}$
gpt2-medium	en	layer	23.000	2,277.000	133.047	$< 0.001^{***}$
gpt2-medium	zh	layer	23.000	2,277.000	9.674	$< 0.001^{***}$
gpt2-large	en	layer	35.000	3,465.000	206.389	$< 0.001^{***}$
gpt2-large	zh	layer	35.000	3,465.000	4.265	$< 0.001^{***}$
gpt2-xl	en	layer	47.000	4,653.000	128.006	$< 0.001^{***}$
gpt2-xl	zh	layer	47.000	4,653.000	6.147	$< 0.001^{***}$
Llama-3.2-1B	en	layer	15.000	1,485.000	97.628	$< 0.001^{***}$
Llama-3.2-1B	zh	layer	15.000	1,485.000	82.293	$< 0.001^{***}$
Llama-3.2-3B	en	layer	27.000	2,673.000	85.032	$< 0.001^{***}$
Llama-3.2-3B	zh	layer	27.000	2,673.000	83.865	$< 0.001^{***}$
Llama-3.1-8B	en	layer	31.000	3,069.000	70.306	$< 0.001^{***}$
Llama-3.1-8B	zh	layer	31.000	3,069.000	82.185	$< 0.001^{***}$
Qwen2.5-1.5B	en	layer	27.000	2,673.000	111.454	$< 0.001^{***}$
Qwen2.5-1.5B	zh	layer	27.000	2,673.000	56.554	$< 0.001^{***}$
Qwen2.5-3B	en	layer	35.000	3,465.000	130.752	$< 0.001^{***}$
Qwen2.5-3B	zh	layer	35.000	3,465.000	56.042	$< 0.001^{***}$
Qwen2.5-7B	en	layer	27.000	2,673.000	99.746	$< 0.001^{***}$
Qwen2.5-7B	zh	layer	27.000	2,673.000	56.780	$< 0.001^{***}$

Table 4: Testing the main effect of layer per language and PLM/LLM (***: $\leq .001$, **: $\leq .01$, *: $\leq .05$). Multiple comparisons corrected via FDR (Benjamini and Hochberg, 1995). Details of model fitting can be found in Appendix D.1.

Model	contrast	estimate	SE	df	t-ratio	p-value
bert-base-multilingual-uncased	en - zh	6.130	0.530	198.000	11.570	$< 0.001^{***}$
roberta-base	en - zh	8.741	0.395	198.000	22.117	$< 0.001^{***}$
roberta-large	en - zh	9.776	0.369	198.000	26.464	$< 0.001^{***}$
xlm-roberta-base	en - zh	2.421	0.309	198.000	7.830	$< 0.001^{***}$
xlm-roberta-large	en - zh	2.667	0.302	198.000	8.818	$< 0.001^{***}$
deberta-v3-base	en - zh	13.706	0.702	198.000	19.512	$< 0.001^{***}$
deberta-v3-large	en - zh	11.792	0.567	198.000	20.802	$< 0.001^{***}$
mdeberta-v3-base	en - zh	5.295	0.574	198.000	9.232	$< 0.001^{***}$
gpt2	en - zh	4.849	0.583	198.000	8.313	$< 0.001^{***}$
gpt2-medium	en - zh	4.504	0.454	198.000	9.915	$< 0.001^{***}$
gpt2-large	en - zh	6.623	0.751	198.000	8.820	$< 0.001^{***}$
gpt2-xl	en - zh	7.896	0.796	198.000	9.923	$< 0.001^{***}$
Llama-3.2-1B	en - zh	8.575	0.795	198.000	10.788	$< 0.001^{***}$
Llama-3.2-3B	en - zh	7.900	0.805	198.000	9.817	$< 0.001^{***}$
Llama-3.1-8B	en - zh	7.350	0.780	198.000	9.426	$< 0.001^{***}$
Qwen2.5-1.5B	en - zh	6.040	0.834	198.000	7.242	$< 0.001^{***}$
Qwen2.5-3B	en - zh	5.617	0.783	198.000	7.177	$< 0.001^{***}$
Qwen2.5-7B	en - zh	5.371	0.784	198.000	6.852	$< 0.001^{***}$

Table 5: Testing the overall effect of language averaged over all layers per model by the pairwise comparisons between the *D-scores* of the two languages (***: $\leq .001$, **: $\leq .01$, *: $\leq .05$). Multiple comparisons corrected via FDR (Benjamini and Hochberg, 1995). Details of model fitting can be found in Appendix D.2.

Model	Language	Factor	df1	df2	F-ratio	p-value
bert-base-uncased	en	POS:layer	11.000	1,078.000	20.732	$< 0.001^{***}$
bert-large-uncased	en	POS:layer	23.000	2,254.000	11.595	$< 0.001^{***}$
bert-base-chinese	zh	POS:layer	11.000	1,078.000	7.248	$< 0.001^{***}$
bert-base-multilingual-uncased	en	POS:layer	11.000	1,078.000	48.921	$< 0.001^{***}$
bert-base-multilingual-uncased	zh	POS:layer	11.000	1,078.000	2.635	0.003^{**}
roberta-base	en	POS:layer	11.000	1,078.000	3.167	$< 0.001^{***}$
roberta-base	zh	POS:layer	11.000	1,078.000	3.019	0.001^{***}
roberta-large	en	POS:layer	23.000	2,254.000	12.184	$< 0.001^{***}$
roberta-large	zh	POS:layer	23.000	2,254.000	1.442	0.084
xlm-roberta-base	en	POS:layer	11.000	1,078.000	16.885	$< 0.001^{***}$
xlm-roberta-base	zh	POS:layer	11.000	1,078.000	3.693	$< 0.001^{***}$
xlm-roberta-large	en	POS:layer	23.000	2,254.000	9.831	$< 0.001^{***}$
xlm-roberta-large	zh	POS:layer	23.000	2,254.000	2.531	$< 0.001^{***}$
deberta-v3-base	en	POS:layer	11.000	1,078.000	3.627	$< 0.001^{***}$
deberta-v3-base	zh	POS:layer	11.000	1,078.000	11.075	$< 0.001^{***}$
deberta-v3-large	en	POS:layer	23.000	2,254.000	2.594	$< 0.001^{***}$
deberta-v3-large	zh	POS:layer	23.000	2,254.000	8.391	$< 0.001^{***}$
mdeberta-v3-base	en	POS:layer	11.000	1,078.000	14.018	$< 0.001^{***}$
mdeberta-v3-base	zh	POS:layer	11.000	1,078.000	3.280	$< 0.001^{***}$
gpt2	en	POS:layer	11.000	1,078.000	54.885	$< 0.001^{***}$
gpt2	zh	POS:layer	11.000	1,078.000	1.765	0.059
gpt2-medium	en	POS:layer	23.000	2,254.000	28.487	$< 0.001^{***}$
gpt2-medium	zh	POS:layer	23.000	2,254.000	2.109	0.002^{**}
gpt2-large	en	POS:layer	35.000	3,430.000	38.495	$< 0.001^{***}$
gpt2-large	zh	POS:layer	35.000	3,430.000	0.643	0.949
gpt2-xl	en	POS:layer	47.000	4,606.000	21.148	$< 0.001^{***}$
gpt2-xl	zh	POS:layer	47.000	4,606.000	0.973	0.531
Llama-3.2-1B	en	POS:layer	15.000	1,470.000	11.654	$< 0.001^{***}$
Llama-3.2-1B	zh	POS:layer	15.000	1,470.000	5.489	$< 0.001^{***}$
Llama-3.2-3B	en	POS:layer	27.000	2,646.000	8.432	$< 0.001^{***}$
Llama-3.2-3B	zh	POS:layer	27.000	2,646.000	4.869	$< 0.001^{***}$
Llama-3.1-8B	en	POS:layer	31.000	3,038.000	8.704	$< 0.001^{***}$
Llama-3.1-8B	zh	POS:layer	31.000	3,038.000	5.111	$< 0.001^{***}$
Qwen2.5-1.5B	en	POS:layer	27.000	2,646.000	6.972	$< 0.001^{***}$
Qwen2.5-1.5B	zh	POS:layer	27.000	2,646.000	3.108	$< 0.001^{***}$
Qwen2.5-3B	en	POS:layer	35.000	3,430.000	12.456	$< 0.001^{***}$
Qwen2.5-3B	zh	POS:layer	35.000	3,430.000	2.996	$< 0.001^{***}$
Qwen2.5-7B	en	POS:layer	27.000	2,646.000	12.519	$< 0.001^{***}$
Qwen2.5-7B	zh	POS:layer	27.000	2,646.000	2.970	$< 0.001^{***}$

Table 6: Interaction effect between POS and layer (***: $\leq .001$, **: $\leq .01$, *: $\leq .05$). Multiple comparisons corrected via FDR (Benjamini and Hochberg, 1995). Details of model fitting can be found in Appendix D.3.

		English			Chinese		
Model Family	Model	Layer (Best POS)	Layer (Best *D-score*)	*D-score*	Layer (Best POS)	Layer (Best *D-score*)	*D-score*
BERT	bert-base-uncased	8	7	18.50	—	—	—
	bert-large-uncased	15	15	20.81	—	—	—
	bert-base-chinese	—	—	—	9	12	9.63
	bert-base-multilingual-uncased	12	12	14.65	7	12	4.96
	roberta-base	4	11	11.36	1	9	1.04
	roberta-large	5	21	13.04	4	15	1.03
	xlm-roberta-base	8	10	7.82	9	11	4.04
	xlm-roberta-large	10	23	9.57	11	23	5.15
	deberta-v3-base	4	5	24.44	10	8	6.50
	deberta-v3-large	7	12	25.32	23	15	6.79
	mdeberta-v3-base	6	6	18.04	6	11	10.56
GPT2	gpt2	9	9	9.57	8	6	1.93
	gpt2-medium	16	17	8.44	15	16	1.55
	gpt2-large	36	36	15.72	28	26	2.49
	gpt2-xl	48	48	15.83	29	29	2.53
Llama3	Llama-3.2-1B	3	4	14.77	14	16	5.61
	Llama-3.2-3B	3	6	16.86	23	23	6.66
	Llama-3.1-8B	3	6	16.87	22	24	6.65
Qwen2.5	Qwen2.5-1.5B	4	8	15.11	20	9	6.43
	Qwen2.5-3B	11	14	14.89	11	14	6.99
	Qwen2.5-7B	8	8	13.99	5	10	6.89

Table 7: The layer where the t-ratio computed from the comparison between same-POS *D-score* and different-POS *D-score* is the largest. The columns *D-score* show the *D-score* from the best *D-score* layer, same as Table 1.