

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

Anonymous ACL submission

Abstract

001 Homonymy can easily raise lexical ambiguity
002 due to the misunderstanding of its multiple
003 senses. Correct recognition of homonym sense
004 greatly relies on its surrounding context. This
005 ambiguous nature makes homonyms an appropriate
006 testbed for examining the contextualization
007 capability of pre-trained (PLM) and large
008 language models (LLMs). Considering the
009 impact of part-of-speech (POS) on homonym
010 disambiguation and the dominance of English
011 in word embedding research, this study provides
012 a comprehensive layer-wise analysis of
013 homonym representations in both English and
014 Chinese, spanning same and different POS categories,
015 across four families of PLMs/LLMs (BERT, GPT-2,
016 Llama 3, Qwen 2.5). Through the creation of a
017 synthetic dataset and computation of disambiguation
018 score (*D-Score*), we found that: (1) no universal
019 layer depth excels in differentiating homonym
020 representations; (2) bidirectional models produce
021 better contextualized homonym representations
022 compared to much larger autoregressive models;
023 (3) most importantly, POS affects homonym
024 representations in models in ways that differ
025 from human research findings. The individual
026 differences between LLMs uncovered in our
027 study challenge the simplistic understanding
028 of their inner workings. This reveals a compelling
029 research frontier: conducting controlled
030 experiments with purposefully manipulated
031 inputs to enhance the interpretability of LLMs.
032 We have made our dataset and codes available
033 publicly at <https://anonymous.4open.science/r/ehril/>.
034
035

036 1 Introduction

037 The efficient and economic use of lexical inventory
038 results in multiple word senses converging into
039 a single lexical item, leading to lexical ambiguity
040 (Wang, 2011; Piantadosi et al., 2012). Among these
041 lexical items, homonyms represent a common type.

*These authors contributed equally.

042 They denote two (or more) semantically and etymo-
043 logically unrelated meanings. For instance, "bank"
044 can refer to a financial institution or the side of a
045 river.

046 Resolution of lexical ambiguity, while rarely
047 conscious in everyday language use, can pose specific
048 challenges to human readers or listeners. Psycholinguistic
049 and neuroimaging research suggest that homonyms,
050 with their unrelated meanings, often make comprehension
051 more difficult (Frazier and Rayner, 1990; Rodd, 2018;
052 Huang and Lee, 2018).
053

054 Lexical ambiguity also poses challenges to distributional
055 semantic models, though the nature of these difficulties
056 differs from those encountered in human language
057 processing (Lake and Murphy, 2023). Early static
058 word representation models such as LSA (Deerwester
059 et al., 1990) and Word2Vec (Mikolov et al., 2013)
060 cannot be modulated by surrounding contexts after
061 the training process. Different senses of a word
062 must share the same representation, hindering its
063 ability to differentiate word senses. Contextualized
064 word representations and language models such as
065 ELMo (Peters et al., 2018) and BERT (Devlin et al.,
066 2019) were proposed to address the problem. Followed
067 by this development, modern PLMs and LLMs are
068 all contextualized language models.
069

070 Contextualized word representations from these
071 models are influenced by its surrounding contexts,
072 allowing the representations to vary instead of
073 remain static. The context can be preceding tokens
074 in autoregressive models like GPT (Radford et al.,
075 2019), or both preceding and following tokens in
076 bidirectional, autoencoding models like BERT (Devlin
077 et al., 2019). Homonyms, which greatly rely on
078 their surrounding context to be accurately understood,
079 serve as appropriate candidates to experiment on the
080 contextualization capability of PLMs and LLMs.
081

082 Besides, the difference between homonyms

083 whose senses belong to the same or different syn- 132
084 tactic categories or parts of speech (POS) also war- 133
085 rants attention (MacDonald et al., 1994). For exam- 134
086 ple, 'fly' can serve as both a noun (a small insect) 135
087 and a verb (to move through air). Electrophysiolog- 136
088 ical responses and blood-oxygen-level-dependent 137
089 (BOLD) signals, as observed during lexical ambi- 138
090 guity processing, are significantly influenced by 139
091 POS variability, according to previous research 140
092 (Federmeier et al., 2000; Gennari et al., 2007). 141
093 Homonyms with different POS senses elicit greater 142
094 neural activation compared to those with the same 143
095 POS meanings, suggesting a greater cognitive re- 144
096 source demand for syntactic category processing 145
097 (Grindrod et al., 2014). Given these findings in 146
098 human language processing, it becomes interesting 147
099 to explore whether these two types of homonyms 148
100 would be represented differently in language mod- 149
101 els, as they are in the human brain. 150

102 Our main contributions are:

- 103 • Constructed a new Chinese dataset to facilitate 152
104 homonym representation analysis. 153
- 105 • Presented a cross-linguistic analysis on the 154
106 layer-wise, contextualized representations of 155
107 Chinese and English homonyms across vari- 156
108 ous families of PLMs and LLMs. 157
- 109 • Contributed to a deeper understanding on how 158
110 homonyms are represented in language mod- 159
111 els, specifically on how POS modulates model 160
112 representations. 161

113 2 Related Works 162

114 Our present work is largely related to previous stud- 163
115 ies on the investigation of the contextualization 164
116 capabilities of PLMs. Contextualized embeddings 165
117 were found to be able to capture the nuanced seman- 166
118 tic information encoded from English dictionary 167
119 (Chang and Chen, 2019). It can also be used to 168
120 predict human behaviours and explain variances 169
121 of human judgements on meanings (Wilson and 170
122 Marantz, 2022; Rivière et al., 2024). A study that 171
123 focuses on polysemes and homonyms revealed that 172
124 while word embeddings can differentiate ambigu- 173
125 ous words in terms of cosine similarity, the dis- 174
126 tinctive power was much less compared to human 175
127 ratings (Haber and Poesio, 2021). It suggests the 176
128 limitations of word embeddings. The distance be- 177
129 tween ambiguous word embeddings were found 178
130 to show different trajectories across model layers, 179
131 which in turn depended on the architectural factors 180

including but not limited to model size (Rivière 132
et al., 2024). 133

Ethayarajh (2019) analyzed the contextual word 134
representations in ELMO, BERT and GPT-2. It was 135
found that the representations were more context- 136
specific in the higher layers, consistent with previ- 137
ous studies (Peters et al., 2018; Liu et al., 2019a; 138
Clark et al., 2019). Additionally, Ethayarajh (2019) 139
observed that the word embeddings in language 140
models suffered from anisotropy, referring to as the 141
non-uniform distribution of the embeddings in the 142
embedding space. They proposed subtracting the 143
cosine similarity from the baseline computed from 144
the text materials to create an adjusted measure that 145
enhances interpretability. 146

While there is existing research on contextuality 147
and lexical ambiguity, our work makes significant 148
contributions in several ways. For instance, unlike 149
Sevastjanova et al. (2021), who did not explicitly 150
focus on identifying homonymous words within a 151
large corpus, we specifically targeted homonyms 152
with different meanings by developing our own 153
dataset. We compared the performance of 21 mod- 154
els across four model families, varying in size 155
(from 110M to 8B) and architecture, which is more 156
extensive than the few models used in Ethayarajh 157
(2019) and Rivière et al. (2024). We proposed an 158
angle-based disambiguation score to account for 159
the nonlinearity of cosine similarity. Furthermore, 160
we investigated the impact of POS on the contex- 161
tualized embeddings by controlling the selected 162
homonyms. 163

Last but not least, our work extends contextual- 164
ized embeddings analysis to Chinese, unlike previ- 165
ous studies which have predominantly focused on 166
English (Haber and Poesio, 2021) and other West- 167
ern languages like Spanish (Rivière et al., 2024). 168
With a logographic writing system, a single sino- 169
gram can often represent multiple meanings de- 170
pending on its surroundings (Wang, 1973; Huang 171
and Lee, 2018; Wang et al., 2023). This inherent 172
ambiguity in Chinese sinograms makes it an ideal 173
testbed for exploring the capabilities and limita- 174
tions of contextualized representations. 175

176 3 Methods 176

177 3.1 Synthetic data construction 177

Existing datasets did not adequately address how 178
language models represent homonyms in context, 179
especially for the Chinese language. Therefore, 180
we created custom datasets containing English 181



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

and Chinese sentences composed with homonyms via LLMs and validated with experts, following the procedures from previous studies (Schick and Schütze, 2021; Yu et al., 2023). 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). 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 Appendix A.2 and A.3.

For sentence generation, we employed GPT-4o (OpenAI, 2024) for English and ChatGLM4 (GLM et al., 2024) for Chinese. For each homonym, we generated 20 sentences, with 10 sentences corresponding to each of its 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 lan-

guage 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 the difference between the same-sense similarity and cross-sense similarity, but not either one of them. 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.

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

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

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 . g 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.

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 the 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):

$$D\text{-score} = AngSimAdj_{same} - AngSimAdj_{cross} \quad (4)$$

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.

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 (Dubey et al., 2024),

and Qwen (Yang et al., 2024). More details are listed in Table 1 and 3. Due to tokenization issue, *bert-base-chinese* was not included in the English analyses, *bert-base-uncased* and *bert-large-uncased* were not included in the Chinese analyses.

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 ($33 < \text{Layer Depth} (\%) \leq 67\%$) and higher ($67\% < \text{Layer Depth} (\%)$). The details of all statistical analyses conducted in this Section can be found in Appendix D, all multiple comparisons were corrected via FDR (Benjamini and Hochberg, 1995). All results, tables and figures can be found in our online repository <https://anonymous.4open.science/r/ehril/>.

4.1 Contextualization of English homonyms

Linear mixed effect models were fitted to examine the difference of the best D -score ($D\text{-score}_{best}$ hereafter) between models (details can be found in Appendix D.1). 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. RoBERTa based models show relatively poor D -scores at around 10 with comparatively flat trajectories across layers. The three BERT based models (Devlin et al., 2019), in contrast, show increasing D -score in middle layers and plateaued in higher layers. In general, BERT-based models perform best in the middle and higher layers (Table 1). *deberta-v3-large* has the highest $D\text{-score}_{best}$ (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. The two larger models exhibit different patterns where they show sudden spikes of D -score at the final layer, leading to their significantly better performance compared to smaller models ($ps < .001$). This behaviour is not observed in the two smaller models. Numerically, *gpt2-xl* has the highest $D\text{-score}_{best}$ at 15.83 at the final layer, although not significantly different from *gpt2-large* ($p = .865$).

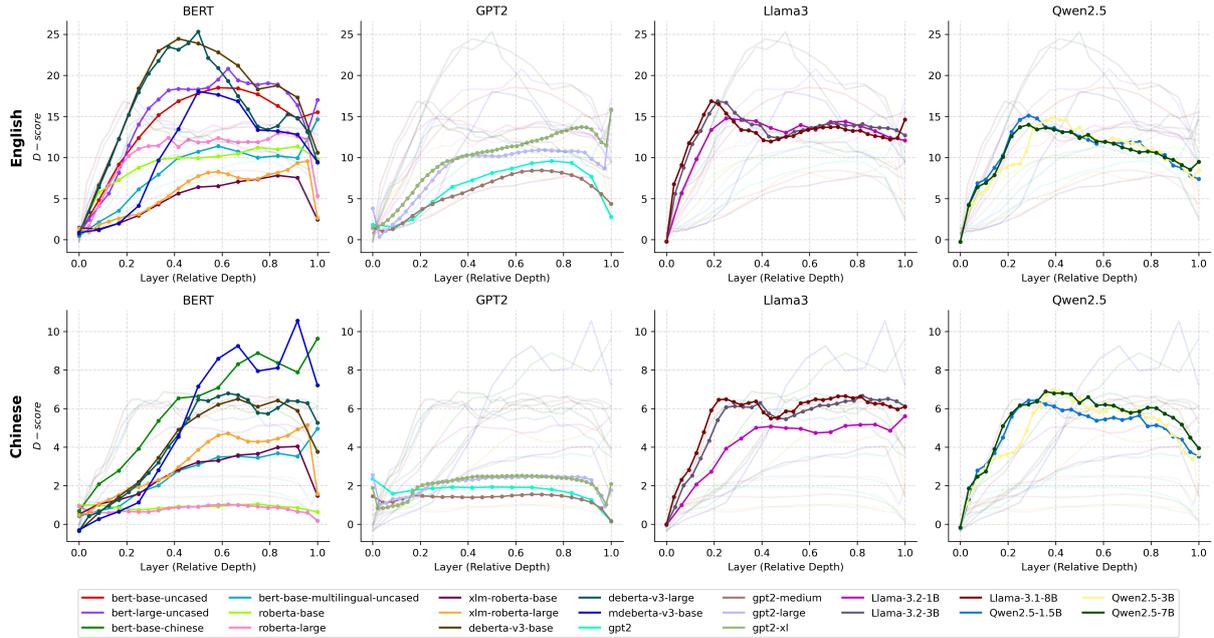


Figure 2: Average D -score between the contextualized representation pairs of English (upper) and Chinese (lower) homonyms across different language models. The x-axis indicates the relative layer depth, computed as the layer number divided by the total number of layers for each model. Note that between 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 ($33\% < \text{Layer Depth} (\%) \leq 67\%$) and higher ($67\% < \text{Layer Depth} (\%)$).

For Llama3 family (Dubey et al., 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. Compared to the two larger models, the 1B model exhibits a significantly smaller peak at the lower layers ($ps < .001$).

The 8B model also shows a sharp increase at the final layer, similar to the two larger models of GPT2. 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

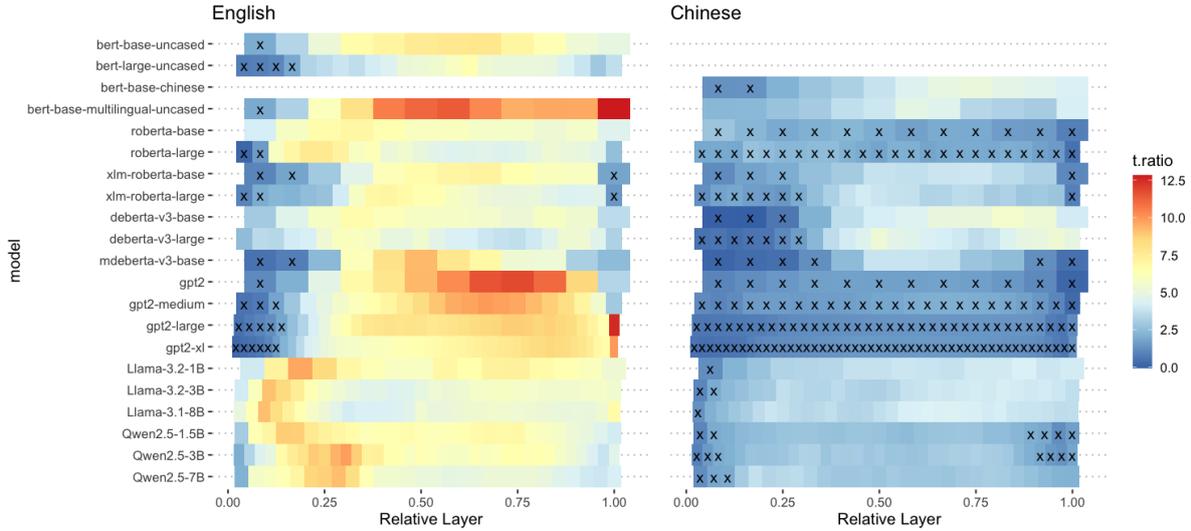


Figure 3: A heatmap showing the t -ratio computed from the post-hoc comparisons between same-POS D -score and different-POS D -score. Multiple comparisons were corrected via FDR (Benjamini and Hochberg, 1995). Nonsignificant results are marked with a cross.

layers and retrace in higher layers. In contrast to 1.5B and 7B, the 3B model exhibits a two-stage ascent at the lower layers and reaches its peak at a somewhat middle layer. 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

The results on layer-wise representation of Chinese homonyms are shown in the lower row of Figure 2 and Table 1. For BERT model family, it can be observed that the layer-wise D -score trajectories differ among BERT, RoBERTa and DeBERTa based models. BERT based models show a monotonically increasing trend, achieving the best score in the final layer. RoBERTa based models show relatively poor performance with the highest D -score_{best} only reaching around 5. DeBERTa based models exhibit two minor peaks in the middle layer and higher layer, with the optimal score observed in the middle layer. *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 high D -score_{best} at 9.63 at a higher layer (Table 1). The two models exhibit significantly better performance than others ($ps < .001$), although they are not significantly different from each other.

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 3B and 8B models show similar D -score_{best} ($p = .989$) values, which are significantly better than that of the 1B version ($ps < .001$). The layer-wise D -score trajectories are similar across all three models. The D -scores increase rapidly in the lower layers and peak in the lower to middle layers. After that, they decline until reaching a plateau and maintain similar scores 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 a middle layer.

4.3 Layer-wise comparisons of English and Chinese homonym representations

First of all, the main effect of language was found to be significant via linear mixed effect models fitted per language model, in which the D -scores computed from English homonym representations were significantly higher than from Chinese (Table 5, Appendix D.2). This suggests that the in-

cluded LLMs are better at contextualizing English homonyms. It is of no surprise since many of the models were all pre-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 that these models can attend on. Of the four model families, only the BERT model family is bidirectional, while 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-*

538 *score* – same-POS *D-score*) were significant. A
539 heatmap showing the *t*-ratio resulted from the compar-
540 isons was shown in Figure 3, and the layers
541 where the *t*-ratios were largest were shown in Table
542 7. A higher *t*-ratio suggests that the difference ob-
543 served is more robust, and less likely to be caused
544 by random noises. In other words, the POS infor-
545 mation contributes more to the prediction of the
546 *D-score* when the *t*-ratio is higher. First of all, it
547 can be observed that almost all *t*-ratios were posi-
548 tive, indicating that the different-POS *D-scores*
549 were always larger than the same-POS *D-scores*.
550 These results suggest that the two meanings
551 of same-POS homonyms are more challenging for
552 models to differentiate than those of different-POS
553 homonyms. Intriguingly, this finding contrasts with
554 previous studies on humans, which showed that
555 different-POS homonyms elicited greater neural
556 activations and required higher cognitive efforts
557 (Grindrod et al., 2014). Our analysis indicates that
558 PLMs and LLMs are actually more adept at repre-
559 senting different-POS homonyms.

560 This contrast between human and machine could
561 potentially be attributed to the underlying process-
562 ing mechanism. While humans need to actively
563 switch between grammatical frameworks to inter-
564 pret different-POS homonyms (Federmeier et al.,
565 2000), the models already possess all relevant POS
566 information in the embedding space during infer-
567 ence without a switching process. However,
568 this benefit in resolving different-POS homonyms
569 might be an obstacle for resolving same-POS
570 homonyms in models.

571 Second, it can be observed from Figure 3 that the
572 *t*-ratios resulted from English homonym compar-
573 isons were much larger than that of Chinese. This
574 difference may lie in the fact that the markedness of
575 POS information are asymmetric between English
576 and Chinese (Greenberg, 1966; Wang, 1973). For
577 instance, the plural form or past tense are marked
578 by inflectional suffixes in English, while it is un-
579 marked in Chinese. For instance, "关门 (*guan*
580 *men*)" can function both as a verb (to close the
581 door) and as a noun (the gate of a pass). No addi-
582 tional suffix will be added when it is used to express
583 plural meaning in its noun function, such as "他
584 们要在所有的关门设防 (*ta men yao zai suo you*
585 *de guan men she fang*) / they will set up defences
586 at all gates of the pass)". It is possible that the
587 marked POS information in its orthographic rep-
588 resentation of English is encoded and stored in its
589 hidden representations during training, which lacks

590 in Chinese. It is likely that this additional embed-
591 ded POS information in English helps improving
592 models' performance in contextualizing homonym
593 representations.

594 Third, it can be observed that the best POS layers
595 (where the POS contributes the most, Figure 3) and
596 the best *D-score* layers (where the *D-score* is the
597 highest, Table 1) are not always equal (more details
598 can be found in Table 7). For English, the best POS
599 layers almost always are the earlier layer (except
600 for *bert-base-uncased*). For Chinese, this pattern
601 can still be observed (e.g., *mdeberta-v3-base*), al-
602 beit less consistent for other models. Language,
603 as a complex adaptive system, employs various
604 strategies, such as POS, word order and contex-
605 tual cues to address lexical ambiguity (Mufwene
606 et al., 2017) that arises from the reuse of existing
607 forms (Ogura and Wang, 2022). Our results indi-
608 cate that, beyond relying on POS information at
609 lower layers, the models also incorporate other lin-
610 guistic information at higher layers to enhance their
611 disambiguation performance.

5 Conclusion 612

613 We presented a comprehensive analysis on En-
614 glish and Chinese homonym representations, span-
615 ning same and different POS categories, across 21
616 PLMs and LLMs in four model families. Patterns
617 of layer-wise trajectories of *D-Score* were found
618 to differ across models and languages, suggesting
619 that these different models might excel differently
620 in encoding distinct levels of linguistic informa-
621 tion (e.g., meanings, POS, etc..) towards encoding
622 distinct homonym representations. Model archi-
623 tecture and pretraining data portion are important
624 factors as bidirectional and multilingual models
625 excel in homonym disambiguation. We also high-
626 lighted the functional role of POS in word-sense
627 disambiguation as models disambiguate same-POS
628 homonyms better than different-POS homonyms.
629 Intriguingly, this is opposite to what have been
630 observed in human studies. To conclude, the indi-
631 vidual differences between LLMs complicate our
632 understanding of their inner workings, there is a
633 need to conduct rigorous, controlled experiments
634 using purposefully manipulated input, in order to
635 enhance interpretability in future LLM research.

6 Limitations 636

637 Several limitations of this study need to be ad-
638 dressed. Firstly, LLM-generated sentences may

639 be biased and unnatural. Although most English
640 sentences and all Chinese sentences were manually
641 examined, this examination was limited to seman-
642 tic acceptability and syntactic correctness. This
643 unnatural issue is particularly pronounced for low-
644 frequency homonyms.

645 Secondly, to facilitate the computation of simi-
646 larity for target words, all English sentences were
647 restricted to using the base form of homonyms.
648 This constraint creates an artificial situation where
649 English homonyms must rely solely on context for
650 POS differentiation, without the benefit of mor-
651 phological strategies such as suffixation. As a re-
652 sult, this limitation may reduce the observed differ-
653 ences between English and Chinese in representing
654 homonyms with the same or different POS, given
655 that Chinese inherently has limited morphologi-
656 cal variation. Future studies should compare these
657 two types of homonyms using sentences with more
658 diverse morphological transformations.

659 Thirdly, while we observed that LLMs per-
660 formed differently in representing same-POS and
661 different-POS homonyms compared to findings
662 from human research, our study lacks direct human-
663 related data for comparison. A recent dataset on
664 English with human judgments on meaning relat-
665 edness has been developed, finding that humans
666 and models perform similarly in word-sense dis-
667 ambiguation (Trott and Bergen, 2021). Future re-
668 search should examine whether similar phenom-
669 ena can be observed in the Chinese context and
670 directly compare model performance with human
671 judgments across languages.

672 References

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

676 Yoav Benjamini and Yosef Hochberg. 1995. [Control-](#)
677 [ling the False Discovery Rate: A Practical and Pow-](#)
678 [erful Approach to Multiple Testing](#). *Journal of the*
679 *Royal Statistical Society. Series B (Methodological)*,
680 57(1):289–300.

681 Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. 2021.
682 [Too Much in Common: Shifting of Embeddings in](#)
683 [Transformer Language Models and its Implications](#).
684 In *Proceedings of the 2021 Conference of the North*
685 *American Chapter of the Association for Computa-*
686 *tional Linguistics: Human Language Technologies*,
687 pages 5117–5130, Online. Association for Computa-
688 tional Linguistics.

689 Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020.

[Interpreting Pretrained Contextualized Representa-](#)
690 [tions via Reductions to Static Embeddings](#). In *Pro-*
691 *ceedings of the 58th Annual Meeting of the Asso-*
692 *ciation for Computational Linguistics*, pages 4758–
693 4781, Online. Association for Computational Lin-
694 guistics. 695

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]. 701

Ting-Yun Chang and Yun-Nung Chen. 2019. [What](#)
702 [Does This Word Mean? Explaining Contextualized](#)
703 [Embeddings with Natural Language Definition](#). In
704 *Proceedings of the 2019 Conference on Empirical*
705 *Methods in Natural Language Processing and the*
706 *9th International Joint Conference on Natural Lan-*
707 *guage Processing (EMNLP-IJCNLP)*, pages 6064–
708 6070, Hong Kong, China. Association for Computa-
709 tional Linguistics. 710

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. 711

Scott Deerwester, Susan T. Dumais, George W. Furnas,
Thomas K. Landauer, and Richard Harshman. 1990.
[Indexing by latent semantic analysis](#). *Journal of the*
716 *American Society for Information Science*, 41(6):391–
717 407. 718 719

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and
Kristina Toutanova. 2019. [BERT: Pre-training of](#)
720 [Deep Bidirectional Transformers for Language Un-](#)
721 [derstanding](#). In *Proceedings of the 2019 Conference*
722 *of the North American Chapter of the Association for*
723 *Computational Linguistics: Human Language Tech-*
724 *nologies, Volume 1 (Long and Short Papers)*, pages
725 4171–4186, Minneapolis, Minnesota. Association for
726 Computational Linguistics. 727 728

Yue Ding, Karolis Martinkus, Damian Pascual, Si-
mon Clematide, and Roger Wattenhofer. 2022. [On](#)
729 [Isotropy Calibration of Transformer Models](#). In *Pro-*
730 *ceedings of the Third Workshop on Insights from Neg-*
731 *ative Results in NLP*, pages 1–9, Dublin, Ireland.
732 Association for Computational Linguistics. 733 734

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey,
Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman,
Akhil Mathur, Alan Schelten, Amy Yang, Angela
Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang,
Archi Mitra, Archie Sravankumar, Artem Korenev,
Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien
Rodriguez, Austen Gregerson, Ava Spataru, Bap-
tiste Roziere, Bethany Biron, Binh Tang, Bobbie
Chern, Charlotte Caucheteux, Chaya Nayak, Chloe
Bi, Chris Marra, Chris McConnell, Christian Keller,
Christophe Touret, Chunyang Wu, Corinne Wong,
Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-
lonsius, Daniel Song, Danielle Pintz, Danny Livshits,
735 736 737 738 739 740 741 742 743 744 745 746 747

748	David Esiobu, Dhruv Choudhary, Dhruv Mahajan,	Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,	812
749	Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes,	Ashwin Bharambe, Assaf Eisenman, Azadeh Yaz-	813
750	Egor Lakomkin, Ehab AlBadawy, Elina Lobanova,	dan, Beau James, Ben Maurer, Benjamin Leonhardi,	814
751	Emily Dinan, Eric Michael Smith, Filip Radenovic,	Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi	815
752	Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Geor-	Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Han-	816
753	gia Lewis Anderson, Graeme Nail, Gregoire Mi-	cock, Bram Wasti, Brandon Spence, Brani Stojkovic,	817
754	alon, Guan Pang, Guillem Cucurell, Hailey Nguyen,	Brian Gamido, Britt Montalvo, Carl Parker, Carly	818
755	Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan	Burton, Catalina Mejia, Changhan Wang, Changkyu	819
756	Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan	Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu,	820
757	Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan	Chris Cai, Chris Tindal, Christoph Feichtenhofer, Da-	821
758	Geffert, Jana Vranes, Jason Park, Jay Mahadeokar,	mon Civin, Dana Beaty, Daniel Kreymer, Daniel Li,	822
759	Jeet Shah, Jelmer van der Linde, Jennifer Billock,	Danny Wyatt, David Adkins, David Xu, Davide Tes-	823
760	Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi,	tuggine, Delia David, Devi Parikh, Diana Liskovich,	824
761	Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu,	Didem Foss, Dingkan Wang, Duc Le, Dustin Hol-	825
762	Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph	land, Edward Dowling, Eissa Jamil, Elaine Mont-	826
763	Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia,	gomery, Eleonora Presani, Emily Hahn, Emily Wood,	827
764	Kalyan Vasuden Alwala, Kartikeya Upasani, Kate	Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan	828
765	Plawiak, Ke Li, Kenneth Heafield, Kevin Stone,	Frank Kanayet, Frank Seide, Gabriela Medina Flo-	829
766	Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuen-	rez, Gabriella Schwarz, Gada Badeer, Georgia Swee,	830
767	ley Chiu, Kunal Bhalla, Lauren Rantala-Yearly, Lau-	Gil Halpern, Govind Thattai, Grant Herman, Grigory	831
768	rens van der Maaten, Lawrence Chen, Liang Tan, Liz	Sizov, Guangyi, Zhang, Guna Lakshminarayanan,	832
769	Jenkins, Louis Martin, Lovish Madaan, Lubo Malo,	Hamid Shojanazeri, Han Zou, Hannah Wang, Han-	833
770	Lukas Blecher, Lukas Landzaat, Luke de Oliveira,	wen Zha, Haroun Habeeb, Harrison Rudolph, He-	834
771	Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh,	len Suk, Henry Aspegren, Hunter Goldman, Ibrahim	835
772	Manohar Paluri, Marcin Kardas, Mathew Oldham,	Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena	836
773	Mathieu Rita, Maya Pavlova, Melanie Kambadur,	Veliche, Itai Gat, Jake Weissman, James Geboski,	837
774	Mike Lewis, Min Si, Mitesh Kumar Singh, Mona	James Kohli, Japhet Asher, Jean-Baptiste Gaya,	838
775	Hassan, Naman Goyal, Narjes Torabi, Nikolay Bash-	Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen,	839
776	lykov, Nikolay Bogoychev, Niladri Chatterji, Olivier	Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong,	840
777	Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan	Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill,	841
778	Zhang, Pengwei Li, Petar Vasic, Peter Weng, Praj-	Jon Shepard, Jonathan McPhie, Jonathan Torres,	842
779	jjwal Bhargava, Pratik Dubal, Praveen Krishnan,	Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou	843
780	Punit Singh Koura, Puxin Xu, Qing He, Qingxiao	U, Karan Saxena, Karthik Prasad, Kartikay Khan-	844
781	Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon	delwal, Katayoun Zand, Kathy Matosich, Kaushik	845
782	Calderer, Ricardo Silveira Cabral, Robert Stojnic,	Veeraraghavan, Kelly Michelena, Keqian Li, Kun	846
783	Roberta Raileanu, Rohit Girdhar, Rohit Patel, Ro-	Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang,	847
784	main Sauvestre, Ronnie Polidoro, Roshan Sumbaly,	Lailin Chen, Lakshya Garg, Lavender A, Leandro	848
785	Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar	Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng	849
786	Hosseini, Sahana Chennabasappa, Sanjay Singh,	Yu, Liron Moshkovich, Luca Wehrstedt, Madian	850
787	Sean Bell, Seohyun Sonia Kim, Sergey Edunov,	Khabsa, Manav Avalani, Manish Bhatt, Maria Tsim-	851
788	Shaoliang Nie, Sharan Narang, Sharath Rapparthi,	poukelli, Martynas Mankus, Matan Hasson, Matthew	852
789	Sheng Shen, Shengye Wan, Shruti Bhosale, Shun	Lennie, Matthias Reso, Maxim Groshev, Maxim	853
790	Zhang, Simon Vandenhende, Soumya Batra, Spencer	Naumov, Maya Lathi, Meghan Keneally, Michael L.	854
791	Whitman, Sten Sootla, Stephane Collot, Suchin Gu-	Seltzer, Michal Valko, Michelle Restrepo, Mihir	855
792	rurangan, Sydney Borodinsky, Tamar Herman, Tara	Patel, Mik Vyatskov, Mikayel Samvelyan, Mike	856
793	Fowler, Tarek Sheasha, Thomas Georgiou, Thomas	Clark, Mike Macey, Mike Wang, Miquel Jubert Her-	857
794	Scialom, Tobias Speckbacher, Todor Mihaylov, Tong	moso, Mo Metanat, Mohammad Rastegari, Mun-	858
795	Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor	ish Bansal, Nandhini Santhanam, Natascha Parks,	859
796	Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent	Natasha White, Navyata Bawa, Nayan Singhal, Nick	860
797	Gonguet, Virginie Do, Vish Vogeti, Vladan Petro-	Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev,	861
798	vic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whit-	Ning Dong, Ning Zhang, Norman Cheng, Oleg	862
799	ney Meers, Xavier Martinet, Xiaodong Wang, Xiao-	Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem	863
800	qing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei	Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pa-	864
801	Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine	van Balaji, Pedro Rittner, Philip Bontrager, Pierre	865
802	Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue	Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratan-	866
803	Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng	chandani, Pritish Yuvraj, Qian Liang, Rachad Alao,	867
804	Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh,	Rachel Rodriguez, Rafi Ayub, Raghotham Murthy,	868
805	Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam	Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah	869
806	Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva	Hogan, Robin Battey, Rocky Wang, Rohan Mah-	870
807	Goldstand, Ajay Menon, Ajay Sharma, Alex Boesen-	eswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu,	871
808	berg, Alex Vaughan, Alexei Baevski, Allie Feinstein,	Samyak Datta, Sara Chugh, Sara Hunt, Sargun	872
809	Amanda Kallet, Amit Sangani, Anam Yunus, An-		873
810	drei Lupu, Andres Alvarado, Andrew Caples, An-		874
811	drew Gu, Andrew Ho, Andrew Poulton, Andrew		875

876	Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma,	Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu	936
877	Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lind-	Hou, and Zihan Wang. 2024. ChatGLM: A Family of	937
878	say, Shaun Lindsay, Sheng Feng, Shenghao Lin,	Large Language Models from GLM-130B to GLM-4	938
879	Shengxin Cindy Zha, Shiva Shankar, Shuqiang	All Tools . <i>Preprint</i> , arXiv:2406.12793.	939
880	Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agar-		
881	wal, Soji Sajuyigbe, Soumith Chintala, Stephanie	Joseph H Greenberg. 1966. <i>Language universals</i> . Mou-	940
882	Max, Stephen Chen, Steve Kehoe, Steve Satterfield,	ton The Hague.	941
883	Sudarshan Govindaprasad, Sumit Gupta, Sungmin		
884	Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury,	Christopher M. Grindrod, Emily O. Garnett, Svetlana	942
885	Sydney Goldman, Tal Remez, Tamar Glaser, Tamara	Malyutina, and Dirk B. den Ouden. 2014. Effects of	943
886	Best, Thilo Kohler, Thomas Robinson, Tianhe Li,	representational distance between meanings on the	944
887	Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook	neural correlates of semantic ambiguity . <i>Brain and</i>	945
888	Shaked, Varun Vontimitta, Victoria Ajayi, Victoria	<i>Language</i> , 139:23–35.	946
889	Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal		
890	Mangla, Vitor Albiero, Vlad Ionescu, Vlad Poenaru,	Janosch Haber and Massimo Poesio. 2021. Patterns	947
891	Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li,	of Polysemy and Homonymy in Contextualised Lan-	948
892	Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will	guage Models . In <i>Findings of the Association for</i>	949
893	Constable, Xiaocheng Tang, Xiaofang Wang, Xiao-	<i>Computational Linguistics: EMNLP 2021</i> , pages	950
894	jian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo	2663–2676, Punta Cana, Dominican Republic. Asso-	951
895	Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li,	ciation for Computational Linguistics.	952
896	Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam,		
897	Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach	Katharina Hämmerl, Alina Fastowski, Jindřich Li-	953
898	Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen,	bovický, and Alexander Fraser. 2023. Exploring	954
899	Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3	Anisotropy and Outliers in Multilingual Language	955
900	Herd of Models . <i>Preprint</i> , arXiv:2407.21783.	Models for Cross-Lingual Semantic Sentence Sim-	956
		ilarity . In <i>Findings of the Association for Compu-</i>	957
901	Kawin Ethayarajh. 2019. How Contextual are Con-	<i>tational Linguistics: ACL 2023</i> , pages 7023–7037,	958
902	textualized Word Representations? Comparing the	Toronto, Canada. Association for Computational Lin-	959
903	Geometry of BERT, ELMo, and GPT-2 Embeddings .	guistics.	960
904	In <i>Proceedings of the 2019 Conference on Empirical</i>		
905	<i>Methods in Natural Language Processing and the</i>	Pengcheng He, Jianfeng Gao, and Weizhu Chen.	961
906	<i>9th International Joint Conference on Natural Lan-</i>	2023. DeBERTaV3: Improving DeBERTa us-	962
907	<i>guage Processing (EMNLP-IJCNLP)</i> , pages 55–65,	ing ELECTRA-Style Pre-Training with Gradient-	963
908	Hong Kong, China. Association for Computational	Disentangled Embedding Sharing . <i>Preprint</i> ,	964
909	Linguistics.	arXiv:2111.09543.	965
910	Kara D. Federmeier, Jessica B. Segal, Tania Lombrozo,	Hsu-Wen Huang and Chia-Ying Lee. 2018. Number	966
911	and Marta Kutas. 2000. Brain responses to nouns,	of meanings and number of senses: An erp study of	967
912	verbs and class-ambiguous words in context . <i>Brain</i> ,	of sublexical ambiguities in reading chinese disyllabic	968
913	123(12):2552–2566.	compounds. <i>Frontiers in Psychology</i> , 9:324.	969
914	Lyn Frazier and Keith Rayner. 1990. Taking on seman-	Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B.	970
915	tic commitments: Processing multiple meanings vs.	Brown, Benjamin Chess, Rewon Child, Scott Gray,	971
916	multiple senses . <i>Journal of Memory and Language</i> ,	Alec Radford, Jeffrey Wu, and Dario Amodei. 2020.	972
917	29(2):181–200.	Scaling Laws for Neural Language Models . <i>Preprint</i> ,	973
		arXiv:2001.08361.	974
918	Silvia P. Gennari, Maryellen C. MacDonald, Bradley R.	Brenden M. Lake and Gregory L. Murphy. 2023. Word	975
919	Postle, and Mark S. Seidenberg. 2007. Context-	meaning in minds and machines . <i>Psychological Re-</i>	976
920	dependent interpretation of words: Evidence for in-	<i>view</i> , 130(2):401–431.	977
921	teractive neural processes . <i>NeuroImage</i> , 35(3):1278–		
922	1286.		
923	Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chen-	Russell V. Lenth. 2025. emmeans: Estimated Marginal	978
924	hui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu	Means, aka Least-Squares Means . R package version	979
925	Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning	1.10.7-100001, https://rvlenth.github.io/emmeans/ .	980
926	Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi		
927	Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li,	Nelson F. Liu, Matt Gardner, Yonatan Belinkov,	981
928	Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu,	Matthew E. Peters, and Noah A. Smith. 2019a. Lin-	982
929	Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu,	guistic Knowledge and Transferability of Context-	983
930	Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun	tual Representations . In <i>Proceedings of the 2019</i>	984
931	Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao	<i>Conference of the North American Chapter of the</i>	985
932	Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan	<i>Association for Computational Linguistics: Human</i>	986
933	Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai	<i>Language Technologies, Volume 1 (Long and Short</i>	987
934	Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao	<i>Papers)</i> , pages 1073–1094, Minneapolis, Minnesota.	988
935	Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi,	Association for Computational Linguistics.	989

1097	An Yang, Baosong Yang, Binyuan Hui, Bo Zheng,	instruct LLMs to construct sentences demonstrat-	1151
1098	Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan	ing the specified meanings of the target homonym.	1152
1099	Li, Dayiheng Liu, Fei Huang, Guanting Dong, Hao-	Both English and Chinese prompts follow the same	1153
1100	ran Wei, Huan Lin, Jialong Tang, Jialin Wang,	general structure. Note that the exact communi-	1154
1101	Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin	cation flow between the human expert and LLMs	1155
1102	Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai,	differed for Chinese and English; details are pro-	1156
1103	Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Ke-	vided in Appendix A.2 and A.3. Besides here is	1157
1104	qin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni,	one key difference: English prompts require that	1158
1105	Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize	the written form of the target homonym remains	1159
1106	Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan,	unchanged across different meanings or contexts	1160
1107	Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge,	(e.g., no inflections, capitalization, etc.), whereas	1161
1108	Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren,	Chinese prompts do not include this requirement.	1162
1109	Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing	This distinction arises because Chinese homonyms	1163
1110	Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan,	consistently use the same characters regardless of	1164
1111	Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang,	their different meanings or contexts.	1165
1112	Zhifang Guo, and Zhihao Fan. 2024. Qwen2 Techni-	To build the dataset, we generated 20 sentences	1166
1113	cal Report . <i>Preprint</i> , arXiv:2407.10671.	for each homonym, with 10 sentences correspond-	1167
1114	Zhilin Yang. 2019. Xlnet: Generalized autoregres-	ing to each of its two meanings. The Chinese	1168
1115	sive pretraining for language understanding. <i>arXiv</i>	sentences required additional scrutiny due to occa-	1169
1116	<i>preprint arXiv:1906.08237</i> .	sional inconsistencies in adhering to the specified	1170
1117	Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng,	meanings. Consequently, three linguistic experts	1171
1118	Alexander Ratner, Ranjay Krishna, Jiaming Shen,	(including two of the authors) manually examined	1172
1119	and Chao Zhang. 2023. Large Language Model as	all Chinese sentences. Any ungrammatical and	1173
1120	Attributed Training Data Generator: A Tale of Diver-	incorrect sentences that did not include the tar-	1174
1121	sity and Bias . <i>Preprint</i> , arXiv:2306.15895.	get homonyms or failed to represent the intended	1175
1122	A Appendix	meanings were manually modified or replaced with	1176
1123	A.1 Procedures for constructing the synthetic	newly generated sentences to ensure dataset qual-	1177
1124	dataset	ity. Additionally, we confirmed that each sentence	1178
1125	To select appropriate English homonyms, we refer-	includes the targeted homonym only once. The	1179
1126	red to an existing dataset, the <i>British eDom</i>	average accuracy rates of generated sentences are	1180
1127	<i>Norms</i> database (Maciejewski and Klepousniotou,	94.97% for Chinese homonyms whose meanings	1181
1128	2016). The database includes 100 homonyms that	belong to the same POS and 90.47% for those be-	1182
1129	have two unrelated meanings, with the relative fre-	longing to different POS.	1183
1130	quency of each meaning rated by 100 monolingual	A.2 English prompt example for sentence	1184
1131	British-English native speakers aged from 19 - 39	construction	1185
1132	(mean 28.1 ± 5.3). This dataset also provided psy-	This section illustrates prompts used for generating	1186
1133	cholinguiistic properties such as semantic distance	English sentence pairs containing target homonyms.	1187
1134	and relatedness, which is valuable for comparing	Note that only the human expert instruction is man-	1188
1135	the similarities and differences in language process-	ually input by the authors. To ensure independence	1189
1136	ing between humans and machines. For Chinese,	between generations, the chat history is cleared	1190
1137	due to the lack of a suitable existing dataset, we cu-	before generating each new 10 sentences for each	1191
1138	rated one on our own. We first collected all possible	meaning of the homonym.	1192
1139	homonyms from a comprehensive Chinese dictio-	Human expert: Please generate ten sentences	1193
1140	nary <i>XianDai HanYu CiDian (7th Edition)</i> , whose	using word "lean" and the semanteme of "lean"	1194
1141	two or three meanings are semantically and etymo-	here is "to bend or incline ". The length of each	1195
1142	logically unrelated. We then applied strict criteria,	generated sentence is around 15 words. Please do	1196
1143	excluding words used in early vernacular Chinese,	not change the morphological structure of the word	1197
1144	rare Buddhist terms, dialects, and those that could	"fan". Do not capitalize the letters of the word	1198
1145	be variant sinograms, polyphonic sinograms, or	"fan". Do not add a clitic to the word "fan".	1199
1146	elements of synonyms. This process yielded 329	GPT-4o: 1. She likes to lean against the old oak	1200
1147	Chinese homonyms, including 130 with different	tree while reading her favorite book. 2. The tall	1201
1148	POS and 193 with the same POS.		
1149	After identifying the homonym candidates, we		
1150	designed a prompt protocol for each homonym to		

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 pre-defined 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.

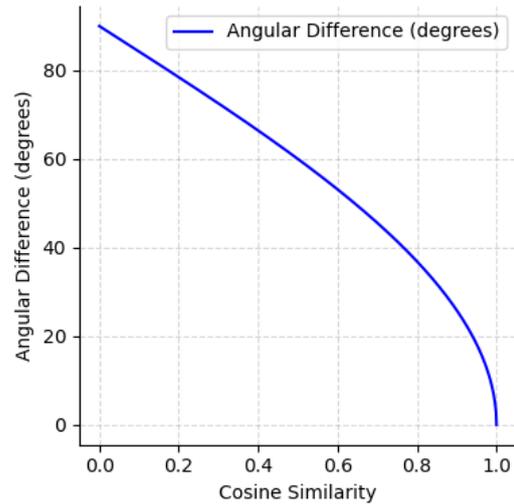


Figure 4: Cosine similarity vs. angular similarity. Cosine similarity varies nonlinearly: higher values represent progressively smaller angular differences.

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. This baseline 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 baseline. 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:

$$\begin{aligned} \text{Baseline}(l, w) &= \mathbb{E} \left[\sum_{i,j} g(f_l(r_i), f_l(r_j)) \right] \\ s_i^1 &= [r_1^1, r_2^1, \dots, r_k^1], s_j^2 = [r_1^2, r_2^2, \dots, r_k^2] \\ & r_i^1 \neq w_i^1, r_j^2 \neq w_j^2 \end{aligned} \quad (5)$$

where g 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) script. 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 FDR (Benjamini and Hochberg, 1995).

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:

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

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 $D\text{-score}_{best}$ in the main text) across models, we fitted one linear mixed effect model per language, with $D\text{-score}_{best}$ being the dependent variable, model as the independent variable and word-specific random intercept. The model is shown in Equation 7.

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

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://anonymous.4open.science/r/ehril/>)

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:

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

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:

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

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://anonymous.4open.science/r/ehril/>).

	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 'shàngren' 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, 'shàngren' 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 'shàngren' 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. The new headmaster is enthusiastic about education reform. 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 'shàngren' 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. 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	12	167M
		roberta-base	12	125M
		roberta-large	12	355M
		xlm-roberta-base	12	278M
		xlm-roberta-large	12	560M
		deberta-v3-based	12	183M
		deberta-v3-large	12	434M
	mdeberta-v3-based	12	278M	
GPT		gpt2	12	124M
		gpt2-medium	24	355M
		gpt2-large	36	774M
		gpt2-xl	48	1.5B
Llama		Llama-3.2-1B	16	1B
		Llama-3.2-3B	28	3B
		Llama-3.1-8B	32	8B
Qwen		Qwen-2.5-1.5B	28	1.5B
		Qwen-2.5-3B	36	3B
		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	<i>F</i> -ratio	<i>p</i> -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	<i>t</i> -ratio	<i>p</i> -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	<i>F</i> -ratio	<i>p</i> -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.

Model Family	Model	English			Chinese		
		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.