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

Anonymous ACL submission

Abstract

Homonymy can easily raise lexical ambiguity due to the misunderstanding of its multiple meanings. Correct recognition of homonym meaning greatly relies on its surrounding context. The ambiguous nature poses significant challenges to the contextualization capability of pre-trained (PLM) and large language models (LLMs), and therefore homonyms can serve as an appropriate testbed. Considering the dominance of English in contextualized word embeddings research, the present work provides a comprehensive layer-wise analysis of both English and Chinese homonym representations across four families of PLMs/LLMs (BERT, GPT2, Llama3, Qwen2.5). Through the creation of a synthetic dataset and a novel anisotropy-adjusted measure, we found that: (1) no universal layer depth was found which excels in differentiating homonym representations; (2) bidirectional models produce better contextualized homonym representations compared to much larger autoregressive models; (3) part-of-speech (POS) affects homonym representations in models in ways that slightly 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.

1 Introduction

Lexical ambiguity is a type of ambiguity that concerns about the multiplicity of meaning of word forms (Wang, 2011; Haber and Poesio, 2024). It can be categorized into polysemy and homonymy, determined by the relationship between word senses. Polysemes refer to word forms that have multiple semantically related senses (Rodd, 2018) that can be connected to one another. In contrast, homonyms have two (or more) semantically and etymologically unrelated meanings associated with a single word form. 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 (Frazier and Rayner, 1990; Rodd, 2018; Huang and Lee, 2018). Homonyms also elicit greater neural activation and require higher cognitive effort to process compared to polysemes (Liang et al., 2024).

Lexical ambiguity, especially homonyms, also poses challenges to distributional semantic models, though the nature of these difficulties differs from those encountered in human language processing (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 contextualization capability of PLMs and LLMs.

Our main contributions are:

- Create a new Chinese dataset constructed to facilitate homonym representation analysis.
- Offer a novel cross-linguistic analysis on the layer-wise, contextualized representations of Chinese and English homonyms across various families of PLMs and LLMs.
- Contribute to a deeper understanding on models' inner workings, and a more nuanced view of homonym representations across typologically distinct languages.

2 Related Works

Our present work is largely related to previous studies on the investigation of the contextualization capabilities of PLMs. Contextualized embeddings were found to be able to capture the nuanced semantic information encoded from English dictionary (Chang and Chen, 2019). It 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 suggests the limitations of word embeddings. The distance between ambiguous word embeddings were 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). The contextuality of word embeddings was found to be affected by the functionality of the word. Homonymous words, compared to function words, have high sense variation and can appear in semantically very different contexts, thus are highly context-specific (Sevastjanova et al., 2021).

Ethayarajh (2019) conducted a comprehensive analysis of the contextuality with ELMo, BERT and GPT-2. The study revealed that the contextualized word representations became more context-specific in the higher layers, matching previous studies (Peters et al., 2018; Liu et al., 2019a; Clark et al., 2019). Importantly, the study mentioned how anisotropy might affect our interpretations on word embeddings. Anisotropy refers to the non-uniform distribution of word embeddings in the embedding space of language models. It has been found that embeddings tend to cluster together in a narrow cone (Cai et al., 2020; Fuster Baggetto and Fresno, 2022). It was believed to negatively impact models' performance in semantic task and viewed as a representation degradation problem (Gao et al., 2019). Anisotropy might be caused by underlying bias in the model such as frequency bias (Fuster Baggetto and Fresno, 2022), or by architectural factors.

While there is existing research on contextuality and lexical ambiguity, our work makes significant contributions in several ways. Unlike Sevastjanova et al. (2021), who did not explicitly focused on identifying homonymous words within a large corpus, we specifically targeted homonyms with different meanings by developing our own dataset of homonyms from dictionaries and creating an equal number of sentences for each one. 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). Furthermore, we proposed an improved method for correcting anisotropy, accounting for fluctuations in the baseline.

Last but not least, our work extends contextualized embeddings analysis to Chinese, unlike previous studies which have predominantly focused on English (Haber and Poesio, 2021) and other Western languages like Spanish (Rivière et al., 2024). With a logographic writing system, a single sinogram can often represent multiple meanings depending on context (Wang, 1973; Huang and Lee, 2018; Wang et al., 2023). This inherent ambiguity in Chinese sinograms makes it 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 procedures from previous studies (Schick and Schütze, 2021; Yu et al., 2023). Details can be found in Appendix Section 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 Chi-

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

nese, 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 Section 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 meanings. 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 meanings and the other half having different POS. This balanced selection allowed us to account for the varying difficulties in representing different types of homonyms, as this meaning granularity of POS has been reported to impact both human language processing (Liang et al., 2024) and language model processing (Loureiro et al., 2021).

3.2 Metrics

3.2.1 Cosine similarity between homonyms

To obtain layer-wise homonym representations, the prepared sentences were fed into pre-trained language models to obtain token-level representations. Word-level representations were obtained by doing mean pooling on token-level representations, as it had been shown to achieve satisfactory performance to determine word similarity (Bommasani et al., 2020). Then, the cosine similarity between the homonym representations that encode either of the two unrelated meanings was computed as in Equation 1, which is conceptually similar to the $IntraSim$ measurement from Ethayarajh (2019):

$$CosSim(l, w) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g(f_l(w_i^1), f_l(w_j^2)) \tag{1}$$

where $CosSim(l, w)$ is the averaged cosine similarity between embedding of the homonym w that encode two different meanings w^1 and w^2 at layer l. g is the cosine similarity function, f_l is the function to map from a homonym in a sentence to its hidden representation at layer l. 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. n is equal to 10 in the present experiments. $CosSim_l(w)$ close to 1 indicates poor contextualization and differentiation of w^1 and w^2. $CosSim_l(w)$ close to 0 suggests the model effectively differentiates meanings based on context.

3.2.2 Adjusting for anisotropy

We followed a similar approach as in Ethayarajh (2019) to adjust for anisotropy, where a baseline was computed by computing averaged cosine similarity between randomly sampled words from two different sentences. 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 cosine 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 cosine similarity values across all samples and combinations of sentences. This process is described in Equation 2:

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

where g and f_l are defined as in Equation 1, k is the number of words in the corresponding sentence and $\mathbb{E}$ denotes the arithmetic average.

After computing the baseline, we followed Ethayarajh (2019) to compute the adjusted $CosSim$:

$$\begin{gathered} CosSimAdj(l, w) = \\ CosSim(l, w) - Baseline(l, w) \end{gathered} \tag{3}$$

while Equation 3 accounts for anisotropy, it might not be suitable for comparisons across layers and models since it does not consider the variations of the baseline level. We argue that if the representation space is already highly anisotropic, it is questionable whether the contextualization of the homonyms is robust even if the adjusted value is close to zero. We proposed another improved correction method (Equation 4):

$$CosScore(l, w) = (1 - Baseline(l, w)) \times (1 - |CosSim(l, w) - Baseline(l, w)|) \quad (4)$$

Our proposed adjustment accounts for anisotropy while penalizing a high baseline level. A value of $CosScore$ close to zero indicates poor contextualization of homonyms, which can be attributed to either the inadequacy to differentiate between the two homonyms or an excessively high baseline level. We believe this approach allows us to more effectively compare models and quantify their homonym representation capability.

3.3 Models

Identical computations and analyses were conducted for via custom Python script for four different model architectures: 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). A total of 21 models were included: *bert-base-uncased*, *bert-large-uncased*, *bert-base-chinese*, *bert-base-multilingual-uncased*, *roberta-base*, *roberta-large*, *xlm-roberta-base*, *xlm-roberta-large*, *deberta-v3-base*, *deberta-v3-large*, *mdeberta-v3-base*, *gpt2*, *gpt2-medium*, *gpt2-large*, *gpt2-xl*, *Llama-3.2-1B*, *Llama-3.2-3B*, *Llama-3.1-8B*, *Qwen2.5-1.5B*, *Qwen2.5-3B*, *Qwen2.5-7B*. More details are listed in Table 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. All models are available on Huggingface via the *transformers* library (Wolf et al., 2020).

4 Analyses

4.1 Anisotropy correction

The layer-wise trajectories of $CosSim$, $Baseline$, $CosSimAdj$, and $CosScore$ were computed from English homonym representations across all 21 models examined in this study.

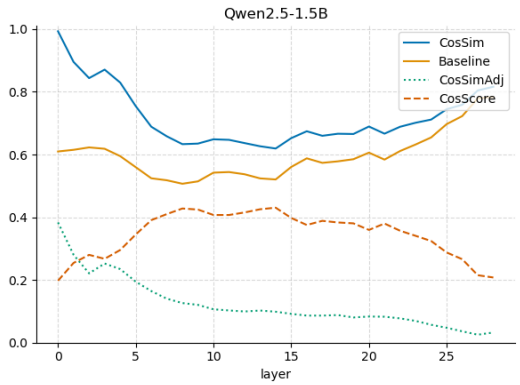

Figure 2: Cosine similarity-derived measurements computed from English homonym representations using *Qwen2.5-1.5B*.

As shown in Figure 2, using *Qwen2.5-1.5B* as a typical example, a clear inconsistency between $CosSimAdj$ and $CosScore$ is observed. $CosSimAdj$ decreases almost monotonically, reaching a minimum at higher layers, while $CosScore$ increases in lower layers, reaches its peak in middle layers, and decreases afterward. This discrepancy is driven by both the decreased difference between $CosSim$ and $Baseline$, and the increase in $Baseline$ in higher layers. $CosSimAdj$ does not account for the increase in $Baseline$, which could lead to misleading conclusion that higher layers perform better. Our proposed $CosScore$, on the other hand, attempted to strike a balance between the two factors. We provided the same figure for all models in the Appendix (English homonyms: Figure 6; Chinese homonyms: Figure 7).

4.2 Contextualization of English homonyms

The results on layer-wise representation of English homonyms are shown in Figure 4, with the upper and lower rows showing layer-wise $CosSim$ and $CosScore$ trajectories respectively. All measurements were computed by averaging over all homonyms at each layer. Table 1 shows the highest $CosScore$ and its corresponding layers in each included model, a more detailed table with $Baseline$ and $CosSimAdj$ can be found in Table 4. Hereafter we will describe the layer depth as: lower (Layer Depth (%) $\leq$ 33%); middle (%33 $<$ Layer Depth (%) $\leq$ 67%) and higher (67% $<$ Layer Depth (%)).

For BERT model family (Devlin et al., 2019; Liu et al., 2019b; He et al., 2023), the layer-wise $CosScore$ trajectories differ drastically among

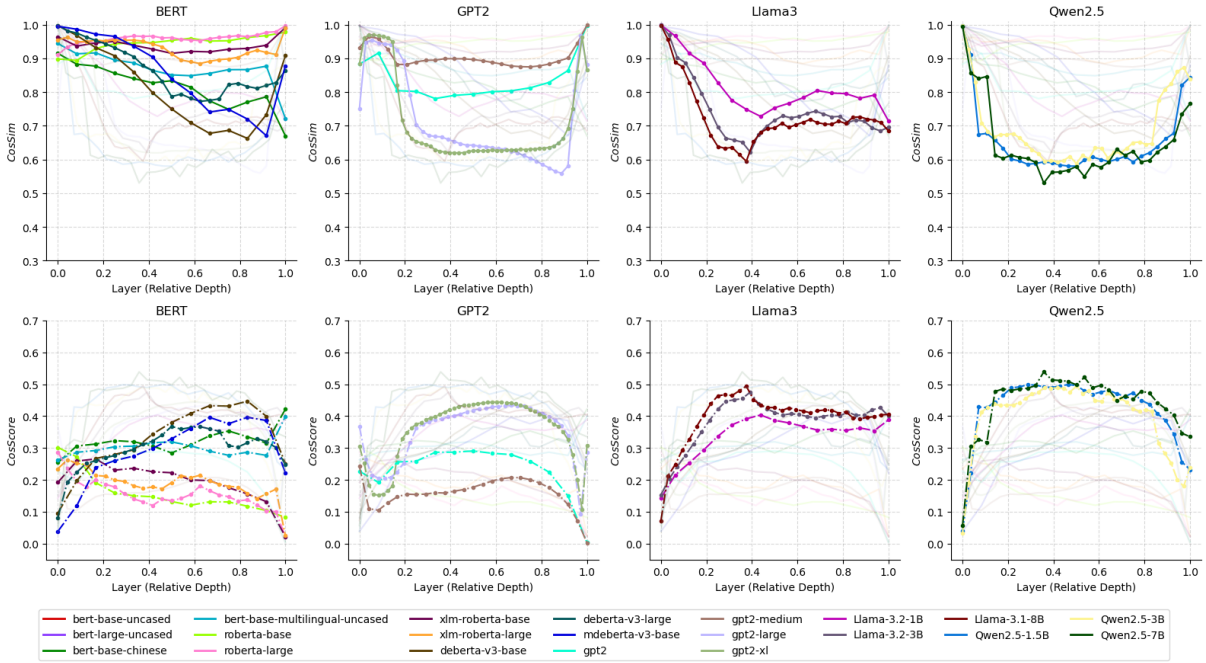

Figure 3: Average $CosSim$ and $CosScore$ between the contextualized representation pairs of Chinese homonyms. The x-axis indicates the relative layer depth, computed as the layer number divided by the total number of layers for each model.

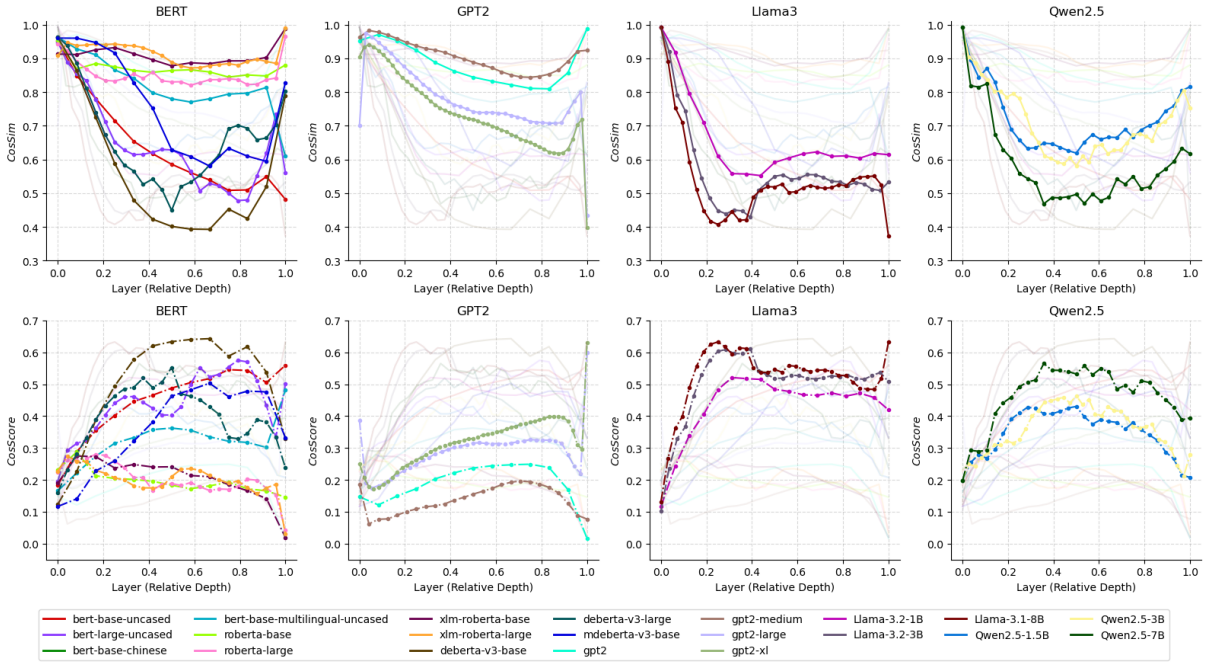

Figure 4: Average $CosSim$ and $CosScore$ between the contextualized representation pairs of English homonyms across different language models. The x-axis represents the relative layer depth, calculated as the layer number divided by the total number of layers for each model.

all models. RoBERTa based models show poor $CosScore$ values at below 0.3 with relatively flat trajectories across layers. The three BERT based models (Devlin et al., 2019), in contrast, show increasing $CosScore$ in higher layers, achieving comparable performance with the later DeBERTa-v3 based models (He et al., 2023), although their best layers were at different depth. From Table 1 it can be observed that BERT based models perform best in higher layers, with the *bert-large-uncased*

			English					Chinese				
Model Family	Model	Parameters	Layer	Layer Depth (%)	Layer Depth	CosSim	CosScore	Layer	Layer Depth (%)	Layer Depth	CosSim	CosScore
BERT	bert-base-uncased	110M	12	100	higher	0.482	0.559	—	—	—	—	—
	bert-large-uncased	340M	19	79.2	higher	0.478	0.575	—	—	—	—	—
	bert-base-chinese	102M	—	—	—	—	—	12	100	higher	0.670	0.422
	bert-base-multilingual-uncased	167M	12	100	higher	0.610	0.482	12	100	higher	0.722	0.399
	roberta-base	125M	1	8.33	lower	0.869	0.288	1	8.33	lower	0.894	0.272
	roberta-large	355M	4	16.7	lower	0.847	0.280	1	4.17	lower	0.935	0.235
	xlm-roberta-base	278M	1	8.33	lower	0.912	0.275	2	16.7	lower	0.946	0.266
	xlm-roberta-large	560M	1	4.17	lower	0.947	0.274	1	4.17	lower	0.963	0.264
	deberta-v3-base	183M	8	66.7	middle	0.393	**0.643**	10	83.3	higher	0.663	**0.447**
	deberta-v3-large	434M	12	50.0	middle	0.449	0.553	12	50.0	middle	0.787	0.368
	mdeberta-v3-base	278M	8	66.7	middle	0.581	0.503	8	66.7	middle	0.741	0.396
GPT2	gpt2	124M	9	75.0	higher	0.811	0.249	6	50.0	middle	0.794	0.291
	gpt2-medium	355M	17	70.8	higher	0.845	0.197	16	66.7	middle	0.878	0.208
	gpt2-large	774M	36	100	higher	0.434	0.599	24	66.7	middle	0.625	0.434
	gpt2-xl	1.5B	48	100	higher	0.397	**0.630**	29	60.4	middle	0.627	**0.445**
Llama3	Llama-3.2-1B	1B	5	31.3	lower	0.558	0.521	7	43.8	middle	0.728	0.404
	Llama-3.2-3B	3B	11	39.3	middle	0.429	0.612	11	39.3	middle	0.621	0.475
	Llama-3.1-8B	8B	8	25.0	lower	0.407	**0.633**	12	37.5	middle	0.595	**0.494**
Qwen2.5	Qwen2.5-1.5B	1.5B	14	50.0	middle	0.619	0.430	8	28.6	lower	0.586	0.499
	Qwen2.5-3B	3B	18	50.0	middle	0.581	0.464	18	50.0	middle	0.588	0.494
	Qwen2.5-7B	7B	10	35.7	middle	0.468	**0.565**	10	35.7	middle	0.530	**0.539**

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

model achieving the highest $CosScore$ at 0.575. *deberta-v3-base* has the highest $CosScore$ among BERT model family at 0.643 in a middle layer.

For GPT2 family (Radford et al., 2019), the layer-wise $CosScore$ trajectories of all four models increase from lower layers and peak in higher layers. Performance improves as model size increases. The two larger models also exhibit different patterns where they show sudden spikes of $CosScore$ at the final layer, leading to their significantly better performance compared to smaller models. This behaviour is not observed in the two smaller models. GPT2-xl has the highest $CosScore$ at 0.630 at the final layer (Table 1).

For Llama3 family (Dubey et al., 2024), $CosScore$ from the 3B and 8B models are similar, surpassing that of the 1B model by a noticeable margin. Yet, the layer-wise $CosScore$ trajectories are similar across all three models, which show rapid increase in lower layers, reaching a local minimum then decreases for a small amount before reaching a plateau at middle layers. The 8B model further shows a sharp increase at the final layer, resembling the patterns observed in the two larger models of GPT2. *Llama-3.2-8B* has the highest $CosScore$ at 0.633 at the final layer (Table 1).

For Qwen2.5 family (Yang et al., 2024), the layer-wise $CosScore$ trajectories are similar across all three models, which increase in lower layers, peak in lower to middle layers and retrace in higher layers. While the 1.5B and 3B models performed similarly, it can be seen that the 7B model performs significantly better. *Qwen2.5-7B* has the highest $CosScore$ at 0.565 in a middle layer (Table 1).

4.3 Contextualization of Chinese homonyms

The results on layer-wise representation of Chinese homonyms are shown in Figure 3 and Table 1. A more detailed table with $Baseline$ and $CosSimAdj$ can be found in Appendix Table 5.

For BERT model family, it can be observed that the layer-wise $CosScore$ 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 poor overall performance with the highest $CosScore$ below 0.3. DeBERTa based models show increasing $CosScore$ from lower to middle layers, peaking in later layers and retracing in the final layer. *deberta-v3-base* has the highest $CosScore$ among BERT model family, at 0.447 in a higher layer. *bert-base-chinese*, the original BERT model variant pre-trained on Chinese Wikipedia (Devlin et al., 2019), also has a comparable $CosScore$ at 0.422 (Table 5).

For GPT2 family, the layer-wise $CosScore$ trajectories of all four models show inverted U-shaped trends, with their peaks in the middle layers. The two larger models have significantly higher $CosScore$ compared to the two smaller models. *GPT2-xl* has the highest score at 0.445 in a middle layer (Table 5).

For Llama3 family, the 3B and 8B models show similar $CosScore$ values, which are significantly better than that of the 1B version. The layer-wise $CosScore$ trajectories are similar across all three

models. The $CosScore$ values 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. *Llama3.1-8B* has the highest $CosScore$ at 0.494 in a middle layer.

For Qwen2.5 family, the layer-wise $CosScore$ 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. Notably, all three models have comparable $CosScore$ values. *Qwen2.5-7B* scores highest at 0.539 in a middle layer.

4.4 Layer-wise comparisons of English and Chinese homonym representations

First of all, the overall $CosScore$ is lower for Chinese compared to English homonym representations. This suggests that the included LLMs are better at contextualizing English homonyms. It is of no surprise since BERT, GPT2 and Llama3 models were all pre-trained with entirely English or English-dominant data. Their superior performance in representing English homonyms were expected. For Chinese homonym representations, the *Qwen2.5-7B* achieves the best $CosScore$. Different sizes of the Qwen2.5 models also performed similarly. We suggest that this can be attributed to the fact that Qwen models included a significant portion of Chinese data in the pre-training phase (Bai et al., 2023; Yang et al., 2024).

Across languages, the layer-wise $CosScore$ trajectories from the same model families show both similarities and differences. For instance, BERT and DeBERTa based models tend to excel in higher layers, while RoBERTa-based models perform better in lower layers for both English and Chinese. Similar trajectories across languages are also observed in Llama3 and Qwen2.5 models. Llama3 models show increasing $CosScore$ values in lower layers, peaking in middle layers. Qwen2.5 models also show increasing $CosScore$ in lower layers across languages, with steeper slopes for Chinese than English homonyms. 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 peak at middle layers.

Similar trajectories 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.

4.5 Does increasing model size help with differentiating homonym representations?

In general, larger models perform better than smaller models in language tasks (Kaplan et al., 2020). An immediate question is whether a larger model could encode more distinctive homonym representations. As we observed from Table 1, larger models in GPT2, Llama3 and Qwen2.5 families indeed perform better than their smaller counterparts. However, this trend does not hold universally. In BERT model family, we found that the *deberta-v3-base* model achieved higher $CosScore$ in both English and Chinese analyses, surpassing its larger counterpart. This is despite the larger model's better performance in various NLP tasks compared to the base model (He et al., 2023). Whether or not larger model can achieve better performance might be dependent on the model architecture and task.

4.6 How does model architecture affect homonym representations?

PLMs and LLMs can roughly be divided into autoencoding, bidirectional models, or autoregressive models. This division is based on which part of the context that these models can attend on. Of the four model families investigated, only the BERT model family can attend to bidirectional contexts, while the other three model families are autoregressive.

Regarding English homonym representations, the bidirectional *deberta-v3-base* model has the lowest $CosSim$ and highest $CosScore$ among all models tested, surpassing autoregressive models that are around 40 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 the same dataset as RoBERTa (He et al., 2023), which was much smaller than those of LLama3 and Qwen2.5 model families.

However, the drop in performance for Chinese homonyms highlights an important caveat: the ad-

vantages of bidirectionality are only fully realized when the model has been adequately exposed to the language during training. DeBERTa lacks Chinese data during pretraining, which likely explains why the model could not effectively use surrounding contexts to represent Chinese homonyms.

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

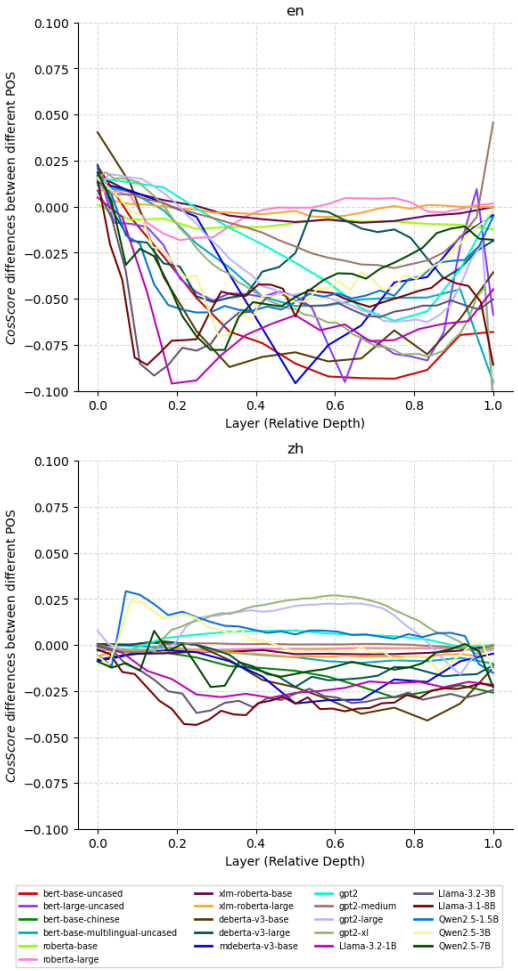

Figure 5: $CosScore$ differences between same-POS homonyms and different-POS homonyms. Top: English homonyms; Bottom: Chinese homonyms.

In our main analyses, the $CosSim$ and $CosScore$ were computed by averaging across all homonym pairs. To further investigate the role of POS information in homonym representation, we computed $CosScore$ differences by subtracting the $CosScore$ of same-POS homonym representations from that of different-POS homonyms. Figure 5 reveals that for English homonym representations, these $CosScore$ differences are predominantly negative across all layers and models. A similar trend is observed for Chinese homonym representations, albeit with less pronounced differences. These 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), LLMs might have encoded relevant information during the pre-training process where such information can be instantaneously retrieved from mathematical operations.

5 Conclusion

We presented a comprehensive analysis on English and Chinese homonym representations across 21 PLMs and LLMs in four model families. Patterns of layer-wise trajectories of cosine similarity and anisotropy-corrected measurement 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 are important factors that determine robust homonym representations. We showed that increasing model size did not directly translate to increasingly distinct representations. Bidirectional models might possess greater capability to produce distinct homonym representations, surpassing autoregressive models of much larger sizes. Finally, we highlighted the importance to have a language-balanced pre-training phase, where it might help capturing semantic information in the target language. 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.

6 Limitations

Several limitations of this study need to be addressed. Firstly, generating sentences with LLMs may introduce biases and produce unnatural language. 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 cosine 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 homonyms with and without the same POS 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.

References

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *Preprint*, arXiv:2309.16609.

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.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. 2020. Isotropy in the Contextual Embedding Space: Clusters and Manifolds. In *International Conference on Learning Representations*.

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.

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.

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, Baptiste 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 Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan

Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin

Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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.

Alejandro Fuster Baggetto and Victor Fresno. 2022. Is anisotropy really the cause of BERT embeddings not being semantic? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4271–4281, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jun Gao, Di He, Xu Tan, Tao Qin, Liwei Wang, and Tie-Yan Liu. 2019. Representation Degeneration Problem in Training Natural Language Generation Models. *Preprint*, arXiv:1907.12009.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. *Preprint*, arXiv:2406.12793.

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.

Janosch Haber and Massimo Poesio. 2024. Polysemy—Evidence from Linguistics, Behavioral Science, and Contextualized Language Models. *Computational Linguistics*, 50(1):351–417.

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.

Xinyuan Liang, Fuchun Huang, Danqing Liu, and Min Xu. 2024. Brain representations of lexical ambiguity: Disentangling homonymy, polysemy, and their meanings. *Brain and Language*, 253:105426.

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.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word

sense disambiguation. *Computational Linguistics*, 47(2):387–443.

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

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

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.

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.

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. 2011. Ambiguity in Language. *Korea Journal of Chinese Language and Literature*, 1:3–20.

William Shi Yuan Wang. 1973. The chinese language. *Scientific American*, 228(2):50–63.

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.

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 parts of speech 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 Section 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).

	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		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.

Figure 6: Average cosine similarity between the contextualized representation of English homonyms across each layer of selected model.

Model Family	Model	Layer	Layer Depth (%)	Layer Depth	CosSim	Baseline	CosSimAdj	CosScore
BERT	bert-base-uncased	12.000	100.000	higher	0.482	0.370	0.112	0.559
	bert-large-uncased	19.000	79.167	higher	0.478	0.306	0.172	0.575
	bert-base-multilingual-uncased	12.000	100.000	higher	0.610	0.325	0.285	0.482
	roberta-base	1.000	8.333	lower	0.869	0.596	0.273	0.288
	roberta-large	4.000	16.667	lower	0.847	0.643	0.204	0.280
	xlm-roberta-base	1.000	8.333	lower	0.912	0.316	0.596	0.275
	xlm-roberta-large	1.000	4.167	lower	0.947	0.512	0.435	0.274
	deberta-v3-base	8.000	66.667	middle	0.393	0.235	0.158	**0.643**
	deberta-v3-large	12.000	50.000	middle	0.449	0.403	0.046	0.553
	mdeberta-v3-base	8.000	66.667	middle	0.581	0.295	0.285	0.503
GPT2	gpt2	9.000	75.000	higher	0.811	0.726	0.085	0.249
	gpt2-medium	17.000	70.833	higher	0.845	0.792	0.053	0.197
	gpt2-large	36.000	100.000	higher	0.434	0.334	0.100	0.599
	gpt2-xl	48.000	100.000	higher	0.397	0.305	0.092	**0.630**
Llama3	Llama-3.2-1B	5.000	31.250	lower	0.558	0.296	0.262	0.521
	Llama-3.2-3B	11.000	39.286	middle	0.429	0.291	0.137	0.612
	Llama-3.1-8B	8.000	25.000	lower	0.407	0.252	0.156	**0.633**
Qwen2.5	Qwen2.5-1.5B	14.000	50.000	middle	0.619	0.520	0.099	0.430
	Qwen2.5-3B	18.000	50.000	middle	0.581	0.486	0.095	0.464
	Qwen2.5-7B	10.000	35.714	middle	0.468	0.375	0.094	**0.565**

Table 4: The best layer of representing English homonyms of each model based on $CosScore$. The model with the highest $CosScore$ within each model family is in boldface.

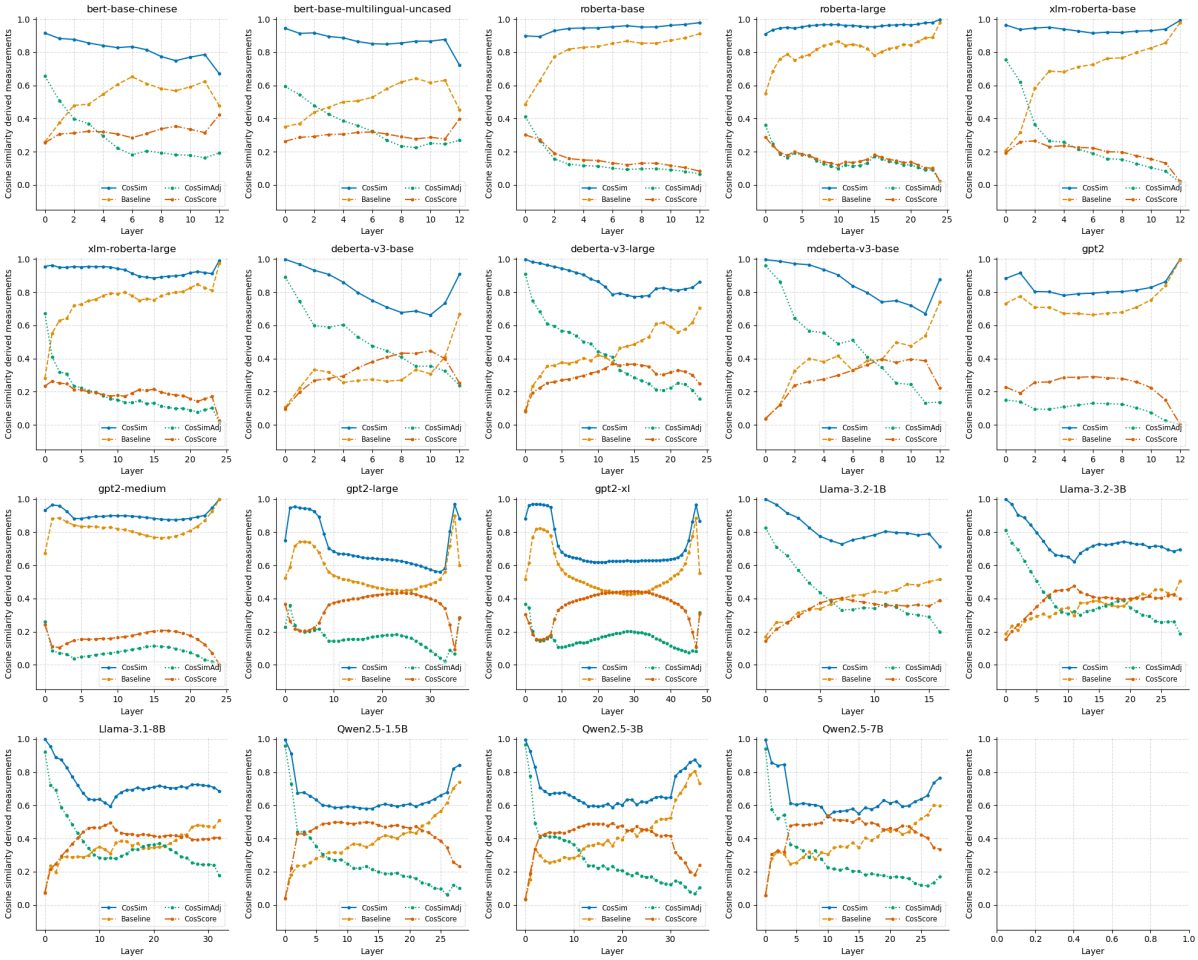

Figure 7: Average cosine similarity between the contextualized representation of Chinese homonyms across each layer of selected model.

Model Family	Model	Layer	Layer Depth (%)	Layer Depth	CosSim	Baseline	CosSimAdj	CosScore
BERT	bert-base-chinese	12.000	100.000	higher	0.670	0.477	0.193	0.422
	bert-base-multilingual-uncased	12.000	100.000	higher	0.722	0.453	0.268	0.399
	roberta-base	1.000	8.333	lower	0.894	0.629	0.265	0.272
	roberta-large	1.000	4.167	lower	0.935	0.687	0.248	0.235
	xlm-roberta-base	2.000	16.667	lower	0.946	0.582	0.364	0.266
	xlm-roberta-large	1.000	4.167	lower	0.963	0.553	0.409	0.264
	deberta-v3-base	10.000	83.333	higher	0.663	0.308	0.355	**0.447**
	deberta-v3-large	12.000	50.000	middle	0.787	0.377	0.410	0.368
	mdeberta-v3-base	8.000	66.667	middle	0.741	0.394	0.347	0.396
GPT2	gpt2	6.000	50.000	middle	0.794	0.664	0.131	0.291
	gpt2-medium	16.000	66.667	middle	0.878	0.766	0.112	0.208
	gpt2-large	24.000	66.667	middle	0.625	0.446	0.178	0.434
	gpt2-xl	29.000	60.417	middle	0.627	0.425	0.202	**0.445**
Llama3	Llama-3.2-1B	7.000	43.750	middle	0.728	0.397	0.331	0.404
	Llama-3.2-3B	11.000	39.286	middle	0.621	0.297	0.324	0.475
	Llama-3.1-8B	12.000	37.500	middle	0.595	0.310	0.285	**0.494**
Qwen	Qwen2.5-1.5B	8.000	28.571	lower	0.586	0.316	0.270	0.499
	Qwen2.5-3B	18.000	50.000	middle	0.588	0.357	0.231	0.494
	Qwen2.5-7B	10.000	35.714	middle	0.530	0.306	0.224	**0.539**

Table 5: The best layer of representing Chinese homonyms of each model based on $CosScore$. The model with the highest $CosScore$ within each model family is in boldface.