Context and POS in Action: A Comparative Study of Chinese Homonym Disambiguation in Human and Language Models

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

Abstract

Ambiguity is pervasive in language, yet we resolve it effortlessly and unconsciously, often aided by context and part-of-speech (POS) cues. This study investigates how context similarity and POS influence homonym disambiguation in humans and large language models (LLMs). To enable comparable analyses between humans and LLMs, we first built an expert-curated sentence-pair dataset, manipulating context similarity and homonym POS categories (nouns vs. verbs). Participants (n = 55) and LLMs (via prompting) were asked to rate the sense similarity of target homonyms embedded within each sentence on a 7-point Likert scale. We found that context similarity influenced both groups similarly, but only humans utilized POS information, likely contributing to their superior performance. Model-derived metrics (surprisal, entropy) predicted human reaction times, and angular similarity between homonym representations accounted for additional variance, highlighting the roles of both expectation-based and semantic processes. Psycholinguistic factors like age of acquisition affected only human responses, underscoring distinct language acquisition mechanisms. Together, our findings illustrate how context and POS information interactively shape homonym resolution in humans, while exposing the limitations of current language models in capturing these nuanced processes. Dataset and codes are publicly available at https://github.com/neurothew/context-and-pos-in-action.

1 Introduction

Language ambiguity is common in daily communication, in part because languages tend to maximize the use of individual words by allowing them to take on multiple meanings (Piantadosi et al., 2012; Wang, 2011). This process creates homonyms and helps reduce the overall vocabulary size and memory demands for speakers. Although homonyms could potentially lead to confusion, true ambiguity is rare, as speakers and listeners rely on mutual understanding and employ strategies, such as using sentential context and part of speech (POS), to quickly and automatically resolve ambiguous words (Zempleni et al., 2007).

Behavioral, electrophysiological, and fMRI studies show that context is crucial for resolving ambiguity in words with multiple meanings (Titone, 1998; Swaab et al., 2003; Zempleni et al., 2007). Even individuals with cognitive deficits, such as older adults, attempt to use context to resolve homonymous ambiguity (Dagerman et al., 2006). Additionally, the POS characteristics of homonyms affect processing; for example, greater brain activity occurs when senses share the same POS category (Grindrod et al., 2014). Furthermore, psycholinguistic factors such as word frequency and age of acquisition (AoA) also influence word processing (Elsherif et al., 2023; Brysbaert et al., 2017).

Distributed semantic models have been developed to generate dynamic word representations based on contextual information (Vaswani et al., 2017; Lenci et al., 2022). These models partially address the meaning conflation problem for homonyms that was previously criticized in static vector models (Mikolov et al., 2013; Navigli and Martelli, 2019). However, it is unclear whether these contextualized embeddings genuinely reflect how humans conceptualize word meaning across contexts, or if they depend on non-generalizable shortcuts, such as overreliance on previously encountered data (Lake and Murphy, 2023; Haber and Poesio, 2024). For example, it remains unknown whether such models, like humans, incorporate the influence of POS and AoA when representing senses of homonyms in context.

The main contributions of this study are twofold. First, we systematically investigate how context similarity and POS modulate both human and lan-

[*]These authors contributed equally.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27608–27625
November 4-9, 2025

guage model performance in homonym disambiguation, using these variables as the primary independent factors in our experimental design. Second, by leveraging the metrics generated by language models under varying conditions of context similarity and POS, we evaluate the extent to which these computational outputs can account for and predict human behavioral responses in homonym processing tasks. Additionally, while word frequency and AoA are not manipulated as independent variables, we incorporate these psycholinguistic factors in our analyses to further contextualize our findings and explore their potential influence on homonym processing.

2 Related work

2.1 Human processing of homonym

Homonyms are seldom interpreted in an isolated manner in real-world communication. Instead, surrounding context plays a crucial role in guiding comprehension and facilitating the integration of incoming linguistic information (Swinney, 1979; Rodd, 2018; Vu et al., 2000). Context provides listeners with prior expectations that help anticipate and resolve ambiguity, enabling efficient communication even when words have multiple meanings.

The influence of context on homonym processing has been widely studied, with major theories proposing that meaning selection involves either exhaustive access followed by contextual selection, direct context-driven access, or dynamic reordering based on context and meaning dominance (Swinney, 1979; Vu et al., 1998; Duffy et al., 1988). While these accounts have illuminated how context biases interpretation toward dominant or subordinate senses of a homonym, they do not fully explain the role of context similarity, such as in zeugmatic expressions, on meaning selection (DeLong et al., 2023); for example, it remains unclear how homonyms are processed in highly similar contexts, as in "they *found* the money and *found* the company." This gap highlights the need for further research into how contextual similarity shapes meaning selection during real-time comprehension.

Beyond contextual cues, POS is a key factor in homonym processing. Behavioral studies show that homonyms whose meanings share the same POS (e.g., noun-noun homonyms like “match”) are recognized more slowly than those with meanings from different classes (e.g., noun-verb homonyms like “bark”), likely due to increased competition among similar representations (Mirman et al., 2010). Yet, neuroimaging findings are mixed: faster reaction times for noun-noun homonyms coincide with less left inferior frontal gyrus (LIFG) activation, while slower reaction times for noun-verb homonyms are associated with greater LIFG activation, suggesting that noun-verb homonyms require more neural effort for lexical-syntactic retrieval processes (Grindrod et al., 2014). These differences may partly reflect limitations of the lexical decision task, as such studies present homonyms in isolation, potentially encouraging reliance on perceptual familiarity with the word form rather than full linguistic processing (Rogers et al., 2004).

Additionally, variables such as word frequency and age of acquisition (AoA) play significant roles. Frequently encountered meanings are accessed more rapidly and are more likely to be selected in ambiguous contexts, while meanings acquired earlier in life may have processing advantages due to their entrenchment in the mental lexicon (Jastrzembski, 1981; Brysbaert et al., 2000; McClelland and Rogers, 2003). Together, these factors, in conjunction with contextual cues, jointly shape the cognitive processes underlying homonym resolution and highlight the multifaceted nature of lexical ambiguity processing.

2.2 Homonym in contextualized word embeddings

Word sense disambiguation (WSD) is also a fundamental problem in natural language processing (NLP) and computational linguistics (Navigli, 2009; Vandenbussche et al., 2021). With the development of word vector approaches, significant progress has been made, especially with the introduction of contextualized word embeddings, which are capable of capturing context-sensitive semantic nuances (Lenci, 2018).

While LLMs have advanced our understanding of how models encode and recover word senses from diverse contexts (Loureiro et al., 2021), important limitations remain, particularly when homonyms appear in highly similar sentential environments, such as in “We’re going to the airport by *coach/bus*” (Garcia, 2021; Brivio and Coltekin, 2022). The effect of context similarity on sense embeddings is especially underexplored for Chinese, a language lacking explicit POS marking found in morphologically rich languages like English (Wang, 1973). For example, in English, verbs like “*founded*” in “they *found* the money and *founded*

the company" are readily identified by morphological cues, while Chinese relies more heavily on context, potentially reducing the distinction between same-POS and different-POS homonyms in contextualized embeddings (Ma et al., 2025). Thus, how context similarity and POS interact to shape homonym representations in Chinese remains unclear and warrants further investigation.

Moreover, recent studies show that computational models, particularly transformer-based language models, can capture aspects of human lexical processing, as contextualized embeddings reflect distinctions such as the polyseme advantage and homonym disadvantage observed in behavioral tasks (Wilson and Marantz, 2022; Rodd et al., 2002). However, while embedding-based similarity metrics like cosine similarity often correlate with human semantic judgments (Nair et al., 2020), they can misestimate similarities in highly similar contexts, for example, underestimating same-sense and overestimating different-sense similarity in cases like "He saw the furry/wooden *bat*" (Trott and Bergen, 2021).

Beyond embeddings, other model-derived metrics such as next-token probability (surprisal) and entropy are increasingly used to probe language processing, offering measures of prediction confidence and uncertainty that align with human neural activity during comprehension (Ryskin and Nieuwland, 2023; Willems et al., 2016; Goldstein et al., 2022; Frank et al., 2015; Heilbron et al., 2022). Together, these computational metrics may provide complementary insights into how language models approximate human lexical processing.

3 Methods

3.1 Sentence-pair dataset construction

Given the absence of datasets that systematically account for both POS and context similarity in Chinese homonyms, we constructed a dedicated sentence-pair dataset for this study. First, 64 noun or verb homonyms (32 same-POS and 32 different-POS) were selected from the seventh edition of the *Modern Chinese Dictionary* (*Xiandai Hanyu Cidian*), a widely used standardized Mandarin reference. Homonyms from other POS categories were too few to be included. For each homonym, relevant psycholinguistic properties, such as AoA and word frequency, were extracted from the Chinese Lexical Dataset (Sun et al., 2018). A full list of selected homonyms and their psycholinguistic properties can be found in Appendix A.2.

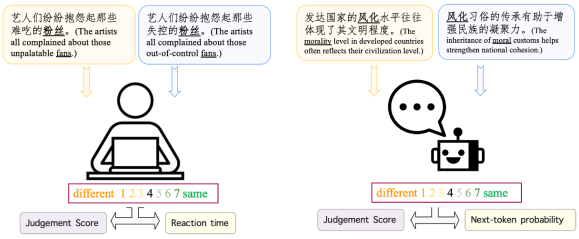

Figure 1: Schematic illustration of experimental procedures for both human and LLMs experiments.

For each target homonym, eight pairs of sentences were constructed, taking into account the two critical factors: POS and Context. Additionally, the senses of the two homonyms in the sentence pairs were balanced to avoid response bias, ensuring that neither "same" nor "different" judgments dominate. The length of the sentences ranged from 10 to 20 characters. Each sentence underwent thorough scrutiny for grammatical correctness and social acceptability. The Table 1 presents illustrative examples for "粉丝" and "风化", each with eight sentence pairs. Hereafter, homonyms will be described as same/diff-POS corresponding to same or different POS, and sentence-pair will be described as same/diff-sense and same/diff-context.

3.2 Experimental procedures

3.2.1 Human experiment

Sixty-one participants (31M) were recruited, although the data from one subject was excluded due to technical issues. All participants were right-handed Chinese native speakers and college students (24.1 $\pm$ 2.5 years old) whose majors were not related to linguistics or psychology. They visited the laboratory to complete the homonym judgment task. All participants gave informed consent in accordance with the requirements of The Hong Kong Polytechnic University's Human Subject Ethics Subcommittee (Reference Number: HSEARS20240515001). An honorarium (100 HKD) was paid to each participant.

Stimuli were presented electronically on a laptop using E-Prime 3.0 software (Schneider et al., 2016). After a fixation cross was displayed at the center of the screen for 600 ms, the target homonym appeared centrally for 500 ms. Subsequently, two sentences were presented simultaneously on either side of the screen, each occupying a single line. Participants were required to use the mouse to click the numbers below the two sentences to indicate

Word	POS1	POS2	POSWord	Context	Sense	Sentence1	Sentence2
粉丝	Noun	Noun	same	same	same	这种透明的粉丝是用绿豆淀粉制成的。 *This transparent vermicelli is made from mung bean starch.*	这种细长的粉丝是用绿豆淀粉制成的。 *This thin vermicelli is made from mung bean starch.*
粉丝	Noun	Noun	same	same	same	她的忠实粉丝常常为她的作品刷屏支持。 *Her loyal fan often supports her works online.*	她的铁杆粉丝常常为她的作品刷屏支持。 *Her die-hard fan often supports her works online.*
粉丝	Noun	Noun	same	same	different	艺人们纷纷抱怨起那些难吃的粉丝。 *The artists all complained about those unpalatable fans.*	艺人们纷纷抱怨那些失控的粉丝。 *The artists all complained about those out-of-control fans.*
粉丝	Noun	Noun	same	same	different	那些激动的粉丝让专家们感到非常吃惊。 *Those excited fans surprised the experts very much.*	那些过期的粉丝让专家们感到非常吃惊。 *Those expired vermicelli surprised the experts very much.*
粉丝	Noun	Noun	same	different	same	我妈妈做的粉丝汤味道真鲜美。 *The vermicelli soup my mom made is really delicious.*	粉丝是广东菜中常见的食材之一。 *Vermicelli is one of the common ingredients in Cantonese cuisine.*
粉丝	Noun	Noun	same	different	same	周杰伦的粉丝会唱他的每一首歌。 *Jay Chou's fans can sing every one of his songs.*	粉丝们在演唱会上的热情非常高涨。 *The fans were very enthusiastic at the concert.*
粉丝	Noun	Noun	same	different	different	饥肠辘辘的他只想来一碗热腾腾的酸辣粉丝。 *Starving, he just wanted a bowl of hot and sour vermicelli.*	前来接机的粉丝把整个机场大厅都挤满了。 *The fans who came to pick up at the airport filled the entire hall.*
粉丝	Noun	Noun	same	different	different	这场签售会吸引了上万的粉丝。 *This signing event attracted tens of thousands of fans.*	妈妈正在厨房里准备凉拌粉丝。 *Mom is preparing cold vermicelli salad in the kitchen.*
风化	Noun	Verb	different	same	same	他的善举对年轻人起到了良好的风化示范作用。 *His good deeds set a good example of moral influence for young people.*	他的善举对年轻人起到了良好的风化促进作用。 *His good deeds had a positive moralizing effect on young people.*
风化	Noun	Verb	different	same	same	某些矿物在潮湿环境中容易发生风化反应。 *Certain minerals are prone to weathering reactions in humid environments.*	某些矿物在潮湿环境中容易发生风化侵蚀。 *Certain minerals are prone to weathering erosion in humid environments.*
风化	Noun	Verb	different	same	different	一个地区的风化观念往往是在长年累月的结果。 *A region's concept of public morals is often the result of many years.*	一个地区的风化作用往往是在长年累月的结果。 *A region's weathering process is often the result of many years.*
风化	Noun	Verb	different	same	different	专家们正在调查社会风化对人民生活的影响。 *Experts are investigating the effects of social morality on people's lives.*	专家们正在调查岩石风化对人民生活的影响。 *Experts are investigating the effects of rock weathering on people's lives.*
风化	Noun	Verb	different	different	same	发达国家的风化水平往往在体现了其文明程度。 *The level of morality in developed countries often reflects their level of civilization.*	风化习俗的传承有助于增强民族的凝聚力。 *The inheritance of moral customs helps strengthen national cohesion.*
风化	Noun	Verb	different	different	same	风化后的岩石表面会出现许多微小的裂缝。 *The rock surface after weathering will have many tiny cracks.*	地貌的形成与频繁的风化密切相关。 *The formation of landforms is closely related to frequent weathering.*
风化	Noun	Verb	different	different	different	村里新修建的祠堂成了传播风化理念的中心。 *The newly built shrine in the village became a center for promoting morality ideas.*	石碑上的文字因受风化影响而变得模糊不清了。 *The text on the stele became blurred due to the influence of weathering.*
风化	Noun	Verb	different	different	different	那块花岗岩的表面出现了明显的风化裂纹。 *Obvious weathering cracks appeared on the surface of that granite.*	这次文化活动的目的是为了促进地方风化建设。 *The purpose of this cultural activity is to promote the local morality construction.*

Table 1: Example sentence pairs for the homonyms "粉丝" and "风化" under different POS, contexts, and senses.

their response on a 7-point Likert scale, as illustrated in Figure 1. A rating of 1 indicated the most different, while 7 indicated the most similar meaning between the senses of the homonym in the two sentences. All 512 sentence pair stimuli were presented in a pseudo-random sequence and divided into 8 blocks. This pseudo-randomization ensured that no three consecutive stimuli were from the same homonym, nor were there three consecutive sentence pairs in which the senses of the target homonym were the same or different. After each block, a rest interval was provided. Eight practice trials were administered to familiarize participants with the experiment. Participants' rating scores and reaction times were recorded. The entire procedure lasted approximately 90 minutes.

3.2.2 LLM experiment

Experiments were conducted on three different model families: Llama3 (Dubey and Zhao, 2024), Qwen2.5 (Yang and Fan, 2024) and Qwen3 (Qwen, 2025). The prompts used to elicit the responses closely resemble the experimental instructions given to human participants. Further details of the prompt can be found in Appendix A.1.

4 Analysis I: Comparison between human and LLMs' responses

All data preprocessing and statistical analyses reported were conducted via custom *R* (R Core Team, 2021) scripts. Linear mixed effect models (LMEMs) and post-hoc comparisons were conducted with the *lme4* (Bates et al., 2014) and *emmeans* (Lenth, 2025).

4.1 Data preprocessing

We preprocessed the human participants' data based on their rating scores and reaction times. First of all, we computed the subject-specific median and the median absolute deviation (MAD) of the reaction time. Then, trials with reaction time deviated more than 2.5 times the subject-specific MAD were removed (Leys et al., 2013). Then, we proceeded to compute the accuracy of the homonym judgment task for each participant. For same-sense sentence pairs, ratings of $\{1, 2, 3\}$ were defined as correct responses. For diff-sense sentence pairs, ratings of $\{5, 6, 7\}$ were defined as correct responses. The answer $\{4\}$ which corresponds to uncertain, was excluded. Participants whose accuracy was below 70% were excluded, leaving 55 participants in the following analyses.

To evaluate model performance, we extracted the probabilities that the next token belonged to a specific set of answers (here, 1 to 7), following previous studies on the evaluation of multiple-choice questions (Wang et al., 2024a; Dominguez-Olmedo et al., 2024; Santurkar et al., 2023; Hendrycks et al., 2021). We considered the model answered correctly when the sum of the probabilities of the cor-

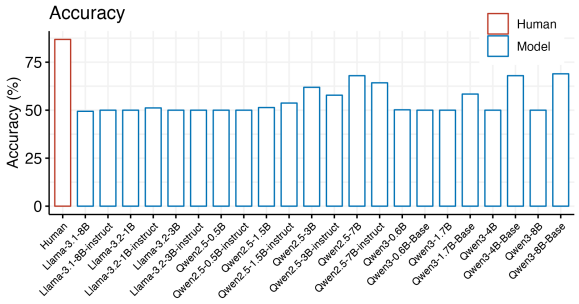

Figure 2: Accuracy of the homonym judgment task.

rect answers was greater than that of the incorrect answers. For instance, in the same-sense condition, the model was determined to answer correctly if the sum of probabilities of responses $\{5, 6, 7\}$ was greater than that of $\{1, 2, 3\}$. The answer $\{4\}$, which corresponds to uncertain, was excluded.

In addition to grouping the 7-point ratings into correct and incorrect responses, we also considered another common way to evaluate LLMs' responses, which is to extracted a single answer for each question from the LLMs (see also Appendix A.3). This was done by selecting the rating with the highest next-token probability. The distributions of these extracted ratings, shown in Appendix Figure 5, reveals that several LLMs (e.g., *Llama-3.1-8B*, *Qwen3-8B*) exhibit a strong bias, with their responses skewed towards only one or two ratings. The correlation results between human ratings and LLM ratings are shown in Appendix Table 4 for readers' reference. As we can see, the correlation cannot be computed for certain LLMs due to a complete lack of diversity in their ratings. Consequently, we focused on grouped responses rather than the raw numeric ratings to effectively evaluate LLMs' performances and compare them with humans.

4.2 Accuracy of both human and LLMs on homonym judgment task

The performance of both humans and LLMs on the homonym judgment task is shown in Figure 2. As depicted, human participants significantly outperformed all included LLMs, achieving an accuracy of 86.9%. Notably, only a few models performed above chance level, with *Qwen3-4B-Base* and *Qwen3-8B-Base* being the top performers, achieving accuracies of 68.0% and 68.9%, respectively. Although these two models showed similar accuracy, they exhibited an unexpected dichotomous response pattern (see details in Appendix A.6). While *Qwen2.5-7B* and *Qwen2.5-3B* demonstrated comparable performance, we focused subsequent analyses on the two Qwen3 models, as they represent the latest developments in open-weight language models (Qwen, 2025).

4.3 Analysis of human reaction time

To examine how human judgments were modulated by POS and Context, we fitted a linear mixed effect model (LMEM) with log reaction time as the dependent variable, Context and POS being the independent variables, with Sense, Trial and other psycholinguistic variables being the covariates. A full list of the included psycholinguistic variables, with their formal definitions, can be found in Appendix Table 3. Trial indicates the presentation order of trials corresponding to the stimuli. The LMEM was fitted as in Equation 1:

$$\begin{aligned}\log(\text{RT}) \sim\ & \text{Context} * \text{POS} * \text{Sense} + \text{Trial} \\ & + \ldots \text{psycholinguistic variables} \ldots \\ & + (1|\text{Subject}) + (1|\text{Word})\end{aligned} \quad (1)$$

After fitting Equation 2, backward elimination was conducted using likelihood ratio tests. The final LMEM is shown in Table 2. The Type-III ANOVA result of the final model is presented in Appendix Table 5.

As shown in Table 2, there is a significant three-way interaction among Sense, Context, and POS; follow-up analyses revealed a significant POS × Context interaction only in the diff-sense condition ($t(22554.68) = -3.63$, $p < .001$), but not in the same-sense condition (see Section 4.5 for further discussion). Additionally, Trial and PSPMI were negatively correlated with reaction time, while AoA was positively correlated: reaction times decreased across trials (likely reflecting practice or fatigue effects (Lanthier et al., 2013)); higher PSPMI (reflecting greater co-occurrence frequency of constituent characters) predicted faster lexical processing (Gertel et al., 2020; Brysbaert et al., 2017); and higher AoA led to slower responses, a pattern not observed in language models and potentially reflecting different underlying mechanisms (see Section 4.4 for further discussion).

4.4 Analysis of LLM surprisal

Similar to the analysis of human responses in Section 4.3, we fitted two LMEMs (Equation 2), one for each of *Qwen3-4B-Base* and *Qwen3-8B-Base*.

	Final model
Human	log(RT) ∼ Sense + Context + POS + Trial + AoA + PSPMI + (1|Subject) + (1|Word) + Sense:Context + Sense:POS + Context:POS + Sense:Context:POS
Qwen3-4B-Base	$\text{Surprisal}_{\text{sum}}$ ∼ Sense + Context + Word_logW_CD + PMI + PSPMI + Sense:Context
Qwen3-8B-Base	$\text{Surprisal}_{\text{sum}}$ ∼ Sense + Context + POSWord + EntropyCharacterFrequencies + Sense:Context + Sense:POS + Context:POS + Sense:Context:POS

Table 2: The final models fitted on human reaction time or LLM $\text{Surprisal}_{\text{sum}}$, obtained via backward elimination. The format of the model follows the convention in *lme4* (Bates et al., 2014). Detailed procedures can be found in Section 4.3 and 4.4.

$\text{Surprisal}_{\text{sum}}$ is the negative logarithm of the sum of probabilities of the correct answers. The lower the $\text{Surprisal}_{\text{sum}}$ value, the better the model performance.

$$\begin{aligned}\text{Surprisal}_{\text{sum}} \sim\ & \text{Context} * \text{POS} * \text{Sense} \\ & + \ldots \text{psycholinguistic variables} \ldots \\ & + (1|\text{Word})\end{aligned} \tag{2}$$

As shown in Table 2, POS significantly predicted $\text{Surprisal}_{\text{sum}}$ in the 8B model but not in the 4B model (see Appendix Tables 6 and 7 for Type III ANOVA results). Further analysis of the 8B model revealed a significant POS × Context interaction in the same-sense condition ($t(344)$ = -2.49, p = 0.013), but not in the diff-sense condition, in contrast to human results (see Section 4.5 for discussion). In both models, additional factors such as contextual similarity, word frequency, PMI, and PSPMI also significantly predicted $\text{Surprisal}_{\text{sum}}$. Moreover, entropy of character frequencies was negatively associated with surprisal, indicating that words with more similar character frequencies elicited lower surprisal. Finally, the word-specific random effect was excluded from both final models, suggesting that the LLMs may respond homogeneously to the 64 homonyms.

Unlike the human LMEM, AoA is not a significant predictor of language model responses. In humans, homonyms learned earlier are processed more quickly, as shown in Appendix Figure 6. Early-acquired words are more robustly represented due to frequent exposure during critical developmental periods, leading to more efficient access and deeper integration within semantic networks (Juhasz, 2005; Perret et al., 2014; Ellis and Lambon Ralph, 2000; Steyvers and Tenenbaum, 2005). In contrast, LLMs are trained on large corpora without a curriculum that prioritizes foundational vocabulary (Kirkpatrick et al., 2017; Houlsby et al., 2019; Wang et al., 2022; Lopez-Paz and Ranzato, 2017). As a result, sense-specific representations in LLMs are formed homogeneously, and ratings are influenced mainly by token frequency or contextual distinctiveness rather than human-like AoA effects. These findings highlight a fundamental divergence between biological and transformer-based learning.

4.5 Similarity and differences between human and models

To compare how humans and language models use POS and contextual cues during homonym processing, we examined their interaction effects on human reaction times and *Qwen3-8B-Base* $\text{Surprisal}_{\text{sum}}$ (Figure 3), the detailed statistics of the post-hoc tests can be found in Appendix Table 8 and 9. Both systems benefit from same context in same-sense conditions, with humans responding faster (Figure 3E) and LLMs showing lower surprisal (Figure 3F). This facilitation reverses in different-sense trials, where different contexts aid performance (Figure 3G-H), indicating that contextual similarity is only helpful when the underlying sense matches.

POS effects, however, diverge. The language model shows only marginal POS × context interactions for same-sense pairs (Figure 3B; same-context: $t(344)$ = -1.78, p = 0.076; diff-context: $t(344)$ = 1.82, p = 0.070), while human reaction times do not significantly differ by POS. In contrast, for different-sense, same-context trials, humans (Figure 3C) display a significant POS effect, with slower responses for different-POS pairs ($t(90.68)$ = -3.02, p = 0.003), a pattern absent in the model (Figure 3D). This POS effect disappears when context is maximally different, suggesting both systems rely less on grammatical cues when context

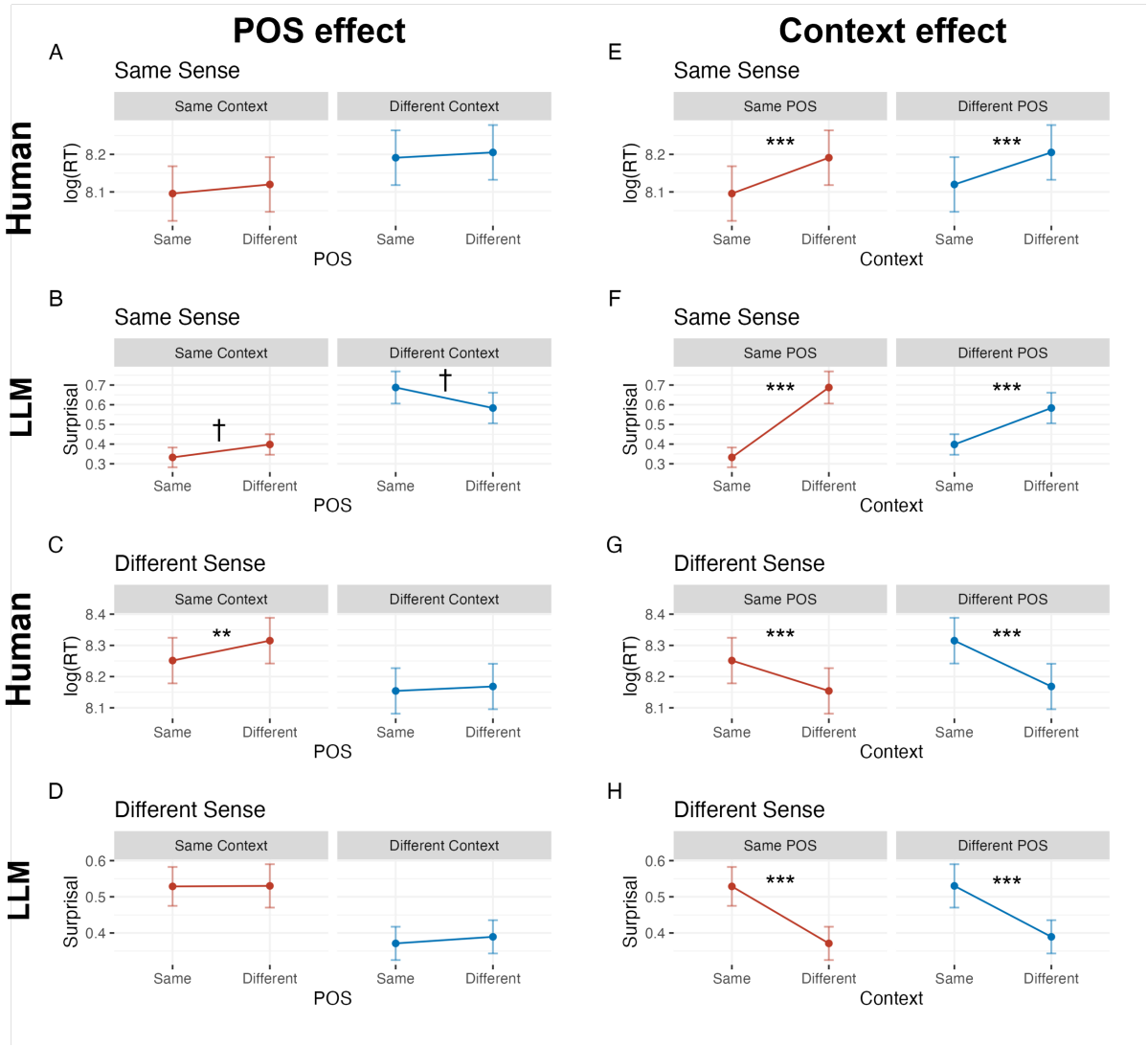

Figure 3: Interaction plots between POS and Context at same-sense and diff-sense conditions. Left: Context effect by POS. Right: POS effect by Context. First and third rows: effects from human data. Second and fourth rows: effects from *Qwen3-8B-Base*. Significance level: †: $p < .1$, *: $p < .05$, **: $p < .01$, ***: $p < .001$.

alone is informative.

In summary, while both humans and LLMs benefit from context in sense judgments, only humans leverage POS information to resolve ambiguity in challenging cases. LLMs do not significantly use POS for homonym processing, which may underlie their lower task accuracy.

5 Analysis II: Associating human responses with LLMs

In Section 4, we fitted LMEMs with the same set of independent variables separately to human reaction time and LLM-derived $\text{Surprisal}_{\text{sum}}$ values to examine the effects of POS and Context. In this section, we further investigate whether three model-derived metrics, including (1) surprisal, (2) entropy, and (3) angular similarity, can improve the modeling of human reaction time. Surprisal is defined as the negative logarithm of the probability assigned to a given response (i.e., cross-entropy) (Goldstein et al., 2022).

Entropy is defined as follows in Equation 3 (Goldstein et al., 2022):

$$H(X) = \sum_{i=1}^{7} P(i) \times \log P(i) \tag{3}$$

where $P(i)$ denotes the probability of the model's next token being i, with i ranging from 1 to 7. A higher entropy corresponds to a lower confidence in the model's next token prediction.

The angular similarity is defined as follows in

Equation 4 (Ma et al., 2025):

$$AngSim = 90 - \arccos(CosSim) \times \frac{180}{\pi} \quad (4)$$

where $CosSim$ is the cosine similarity between two homonym representations in a sentence pair.

5.1 Contribution of surprisal and entropy

To examine the contribution of surprisal and entropy, we added these two variables to Equation 1 for *Qwen3-4B-Base* and *Qwen3-8B-Base* (see Type-III ANOVA results in Appendix Table 10 and 11), and conducted backward elimination, respectively. For metrics derived from *Qwen3-4B-Base*, the surprisal, instead of entropy, contributed significantly to the prediction of human reaction time with a positive association (F(1, 21816.065)=87.2, p<.001). On the other hand, for metrics derived from *Qwen3-8B-Base*, the entropy, instead of surprisal, contributed significantly with a positive association (F(1, 22081.381)=4.813, p=0.028).

Given that autoregressive language models are trained to predict the next token from prior context, model-derived surprisal serves as a post-hoc measure of how unexpected a token is in context (Slaats and Martin, 2025). A positive correlation between surprisal and human reaction time suggests that less likely tokens are associated with longer reaction times and greater cognitive effort, possibly because language models can reduce surprisal for easier prompts by extracting relevant information during inference. In contrast, entropy measures uncertainty about potential outcomes of a future event; its positive correlation with reaction time implies that higher uncertainty is likewise linked to increased cognitive effort (Heilbron et al., 2022).

An interesting observation emerged when comparing metrics derived from two language models of different sizes: surprisal from the 4B model predicted reaction times, while entropy from the 8B model did. However, when these two metrics were entered into the same LMEM (i.e., Equation 1), only surprisal from the 4B model remained significant after backward elimination. This suggests that surprisal from the smaller model better captures human-like processing in this task, subsuming the predictive value of entropy from the larger model.

5.2 Contribution of angular similarity

After obtaining the final LMEMs from each of the Qwen models in Section 5.1, we tested whether adding angular similarity between homonym representations can significantly improve model fit. For each layer, we conducted likelihood ratio tests by comparing nested models with and without the angular similarity. The chi-square statistic quantifies the improvement in model fit based on the difference in log-likelihoods between the two models.

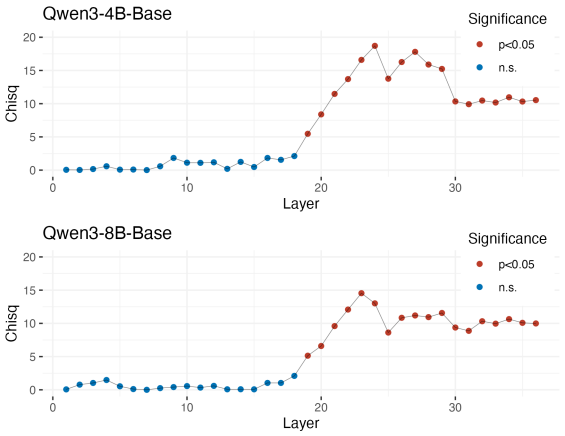

Figure 4: Layer-wise results of likelihood ratio tests comparing nested models with and without angular similarity.

As shown in Figure 4, angular similarity from both *Qwen3-4B-Base* and *Qwen3-8B-Base* significantly improved reaction time prediction beginning after layer 19, with contributions peaking in the late middle layers. It revealed that the angular similarity between contextual representations of the homonyms in different sentences contributed unique variance to reaction time prediction, with this effect emerging in middle layers. This suggests that these layers encode critical semantic information relevant to human processing.

Importantly, angular similarity contributed unique variance complementing surprisal (4B) and entropy (8B), indicating that reaction times reflect both expectation-based processing and semantic integration. Whereas surprisal and entropy may capture the cost of updating predictions when encountering unexpected input, angular similarity may reflect the cognitive effort required for resolving lexical ambiguity. Incorporating both types of metrics thus provides a more complete account of the cognitive processes underlying homonym disambiguation.

6 Conclusion

We presented a comparative study of Chinese homonym disambiguation in humans and language models, collecting their responses toward

a homonym judgment task. Our expert-curated sentence pairs allowed us to systematically examine how context similarity and POS information modulate responses. We found that context similarity had similar effects on both humans and models. However, only humans leveraged POS information during homonym disambiguation, which may account for the models' relatively poorer performance. Model-derived metrics such as surprisal and entropy had significant predictive power while modeling human behavioural responses (reaction time). On top of these expectation-based metrics, incorporating angular similarity between homonyms in sentence pairs contributes unique variance in predicting reaction time, highlighting that human responses are predicted by both expectation-based and semantic information. Furthermore, psycholinguistic properties like AoA influenced human, but not model, response, underscoring fundamental differences in language acquisition mechanisms. Together, these findings put context and POS in action, highlighting how the interplay of contextual and syntactic cues shapes human homonym resolution and revealing the current limitations of language models in capturing these nuanced processes.

7 Limitations

This study has several limitations that should be acknowledged. First, although we manually assessed the naturalness of our curated sentences, some may still be unnatural, particularly those involving rarely used homonyms. This ecological issue, to some extent, could affect both human and model performance. Also, we only adopted the next-token probabilities to evaluate models' performances, which maybe suboptimal for instruction-tuned models as reported in existing studies (Wang et al., 2024b).

Second, we found that the LLMs show significant response skew when we selected the highest next-token probability rating as their responses. Some of them elicited a single rating across all 512 sentence pairs. As such, we made an empirical decision to focus on grouped responses (correct/incorrect) to tackle our research questions. We caution that grouping ratings together could mask the fine-grained differences in semantic similarity. Since scalar scoring is only one response modality, future study could consider others like uncertainty handling (Kadavath et al., 2022), and choice-set or label-space presentation (Zheng et al., 2024). Moreover, exploration of prompt format and instruction framing (Chiang and yi Lee, 2023) can be conducted to elicit more nuanced responses from LLMs. Lastly, there could also be a possibility that certain LLMs do not possess the ability to make fine-grained judgments.

Third, although we observed AoA effects in humans but not in models, AoA and other potential confounds were not explicitly controlled in the experimental design. Moreover, our AoA measures apply to the homonym as a whole rather than to its individual senses, which may not capture sense-specific acquisition. In addition, our dataset intentionally targets noun/verb homonyms to enable balanced manipulations, limiting generalization to other POS categories (e.g., adjectives, adverbs, function words); extending to additional POS combinations will be necessary to test whether the observed human–LLM divergences in POS utilization persist beyond noun/verb homonymy. Future work should better control these factors and, where possible, collect sense-level AoA data.

Last, we found evidence of dichotomous patterns in both human participants and models, see in Appendix A.6. However, due to the lack of individual difference data, such as cognitive ability and language history, in our human sample, it remains unclear what underlies these dichotomies and how they relate across humans and LLMs. Collecting more detailed participant profiles will be important for understanding these patterns in future research.

References

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

Matteo Brivio and Cagri Coltekin. 2022. [re] exploring the representation of word meanings in context. In *ML Reproducibility Challenge*.

Marc Brysbaert, Paweł Mandera, and Emmanuel Keuleers. 2017. The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, 27(1):45–50. Doi: 10.1177/0963721417727521.

Marc Brysbaert, Ilse Van Wijnendaele, and Simon De Deyne. 2000. Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica*, 104(2):215–226.

Cheng-Han Chiang and Hung yi Lee. 2023. A closer look into automatic evaluation using large language models. *Preprint*, arXiv:2310.05657.

Karen Stevens Dagerman, Maryellen C. MacDonald, and Michael W. Harm. 2006. Aging and the use of context in ambiguity resolution: Complex changes from simple slowing. *Cognitive Science*, 30(2):311–345.

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

R. L. Dobrushin. 1970. Prescribing a System of Random Variables by Conditional Distributions. *Theory of Probability & Its Applications*, 15(3):458–486.

Ricardo Dominguez-Olmedo, Moritz Hardt, and Celestine Mendler-Dünner. 2024. Questioning the survey responses of large language models. *Advances in Neural Information Processing Systems*, 37:45850–45878.

Abhimanyu et al. Dubey and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

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

Andrew W. Ellis and Matthew A. Lambon Ralph. 2000. Age of acquisition effects in adult lexical processing reflect loss of plasticity in maturing systems: Insights from connectionist networks. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(5):1103–1123.

M. M. Elsherif, E. Preece, and J. C. Catling. 2023. Age-of-acquisition effects: A literature review. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 49(5):812–847.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The erp response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. *arXiv preprint arXiv:2106.13553*.

Victoria H. Gertel, Hossein Karimi, Nancy A. Dennis, Kristina A. Neely, and Michele T. Diaz. 2020. Lexical frequency affects functional activation and accuracy in picture naming among older and younger adults. *Psychology and Aging*, 35(4):536–552.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380.

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. 2024. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P. de Lange. 2022. A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, 119(32):e2201968119. Doi: 10.1073/pnas.2201968119.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. *Preprint*, arXiv:2009.03300.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

James E. Jastrzembski. 1981. Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology*, 13(2):278–305.

Barbara J. Juhasz. 2005. Age-of-acquisition effects in word and picture identification. *Psychological Bulletin*, 131(5):684–712.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National*

Academy of Sciences, 114(13):3521–3526. Doi: 10.1073/pnas.1611835114.

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

Sophie Lanthier, Evan Risko, Daniel Smilek, and Alan Kingstone. 2013. Measuring the separate effects of practice and fatigue on eye movements during visual search. In *Proceedings of the annual meeting of the cognitive science society*, volume 35.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual Review of Linguistics*, 4(Volume 4, 2018):151–171.

Alessandro Lenci, Magnus Sahlgren, Patrick Jeuniaux, Amaru Cuba Gyllensten, and Martina Miliani. 2022. A comparative evaluation and analysis of three generations of distributional semantic models. *Language Resources and Evaluation*, 56(4):1269–1313.

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

Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

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.

Matthew King-Hang Ma, Chenwei Xie, Wenbo Wang, and William Shiyuan Wang. 2025. Exploring layer-wise representations of english and chinese homonymy in pre-trained language models (accepted).

James L. McClelland and Timothy T. Rogers. 2003. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4):310–322.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

Daniel Mirman, Ted J. Strauss, James A. Dixon, and James S. Magnuson. 2010. Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science*, 34(1):161–173.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. *arXiv preprint arXiv:2010.13057*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2).

Roberto Navigli and Federico Martelli. 2019. An overview of word and sense similarity. *Natural Language Engineering*, 25(6):693–714.

Cyril Perret, Patrick Bonin, and Marina Laganaro. 2014. Exploring the multiple-level hypothesis of aoa effects in spoken and written object naming using a topographic erp analysis. *Brain and Language*, 135:20–31.

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

Team Qwen. 2025. Qwen3 Technical Report.

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

Jennifer Rodd, Gareth Gaskell, and William Marslen-Wilson. 2002. Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, 46(2):245–266.

Jennifer M. Rodd. 2018. Lexical ambiguity.

Timothy T. Rogers, Lambon Ralph Matthew A., Hodges John R., , and Karalyn Patterson. 2004. Natural selection: The impact of semantic impairment on lexical and object decision. *Cognitive Neuropsychology*, 21(2-4):331–352. Doi: 10.1080/02643290342000366.

Rachel Ryskin and Mante S. Nieuwland. 2023. Prediction during language comprehension: what is next? *Trends in Cognitive Sciences*, 27(11):1032–1052. Doi: 10.1016/j.tics.2023.08.003.

Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose opinions do language models reflect? In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 29971–30004. PMLR.

Walter Schneider, Amy Eschman, and Anthony Zuccolotto. 2016. E-prime (version 3.0). *Computer software and manual]. Pittsburgh, PA: Psychology Software Tools Inc*.

Sophie Slaats and Andrea E. Martin. 2025. What's Surprising About Surprisal. *Computational Brain & Behavior*.

Mark Steyvers and Joshua B. Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Ching Chu Sun, Peter Hendrix, Jianqiang Ma, and Rolf Harald Baayen. 2018. Chinese lexical database (cld). *Behavior research methods*, 50(6):2606–2629.

Tamara Swaab, Colin Brown, and Peter Hagoort. 2003. Understanding words in sentence contexts: The time course of ambiguity resolution. *Brain and Language*, 86(2):326–343.

David A. Swinney. 1979. Lexical access during sentence comprehension: (re)consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, 18(6):645–659.

Debra Titone. 1998. Hemispheric differences in context sensitivity during lexical ambiguity resolution. *Brain and Language*, 65(3):361–394.

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

Pierre-Yves Vandenbussche, Tony Scerri, and Ron Daniel Jr. 2021. Word sense disambiguation with transformer models. In *Proceedings of the 6th Workshop on Semantic Deep Learning (SemDeep-6)*, pages 7–12.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Hoang Vu, George Kellas, Ktmberly Metcalf, and Ruth Herman. 2000. The influence of global discourse on lexical ambiguity resolution. *Memory Cognition*, 28(2):236–252.

Hoang Vu, George Kellas, and Stephen T. Paul. 1998. Sources of sentence constraint on lexical ambiguity resolution. *Memory Cognition*, 26(5):979–1001.

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

William Shi Yuan Wang. 2011. Ambiguity in language. *Korea Journal of Chinese Language and Literature*, 1:3–20.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Xinpeng Wang, Chengzhi Hu, Bolei Ma, Paul Röttger, and Barbara Plank. 2024a. Look at the text: Instruction-tuned language models are more robust multiple choice selectors than you think. *arXiv preprint arXiv:2404.08382*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "My Answer is C": First-Token Probabilities Do Not Match Text Answers in Instruction-Tuned Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416, Bangkok, Thailand. Association for Computational Linguistics.

Roel M. Willems, Stefan L. Frank, Annabel D. Nijhof, Peter Hagoort, and Antal van den Bosch. 2016. Prediction during natural language comprehension. *Cerebral Cortex*, 26(6):2506–2516.

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.

An et al. Yang and Zhihao Fan. 2024. Qwen2 Technical Report. *Preprint*, arXiv:2407.10671.

Monika-Zita Zempleni, Remco Renken, John C. J. Hoeks, Johannes M. Hoogduin, and Laurie A. Stowe. 2007. Semantic ambiguity processing in sentence context: Evidence from event-related fmri. *NeuroImage*, 34(3):1270–1279.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

A Appendix

A.1 Prompt example for eliciting LLMs' judgments on homonyms

Aligning with the experimental requirements and instructions for human participants, we created the following prompt template, using "风化" as an example, to elicit judgments about the two senses of target homonyms.

"You are a university student whose major is not related to linguistics or psychology, and your native language is Mandarin Chinese. You are now participating in a language experiment. In this experiment, you are asked to use intuition to judge whether a word has the same meaning in two sentences. If the meanings are exactly the same, please choose 7; if they are completely different, please choose 1; if they are somewhere in between, please choose 2, 3, 4, 5, or 6.

Please judge whether the meaning of "风化(fenghua)" is the same in the following two sentences:

(1) "发达国家的风化水平往往体现了其文明程度。(In developed countries, the level of fenghua/public morality often reflects their degree of civilization.)"

(2) "风化习俗的传承有助于增强民族的凝聚力。(The transmission of fenghua/social-mores-related customs helps to enhance national cohesion.)"

Please answer directly with a number. Your choice is:"

A.2 Homonym stimuli list

Table 3 shows the 64 homonyms included in the rating experiment, along with their Part of Speech (POS) categories and ten other Age of Acquisition (AOA) and word frequency-related psychological properties.

A.3 Distributions of Human and LLM responses

Figure 5 show the distributions of the ratings from human and all models examined in the present study. To obtain the a single response for each question from LLMs, we first extracted the probabilities that the next token belonged to a specific set of answers (1 to 7), then selected the answer with the highest probability. It should be noted that we were plotting the distributions for exploration purpose and all our other results reported in the main text were based on the summation of probabilities method described in Section 4.1.

Table 4 shows the correlation between between human and LLM responses. We computed the correlation between all of our 55 participants and LLMs, then averaged the correlation to give an overview.

A.4 Statistical results

Table 5, 6, 7 correspond the Type-III ANOVA results of the models fitted in Section 4.3 and 4.4. Table 8, 9 show the post-hoc tests examining the interaction between POS and Context, corresponding to Section 4.5 and Figure 3. Table 10, 11 correspond to the Type-III ANOVA results of the models fitted in Section 5.1.

A.5 AoA effect on human reaction time

Figure 6 shows the relationship between Age of Acquisition (AoA) and log-transformed reaction time (log RT).

A.6 Divergent response patterns in both human and models

We observed that the 4B and 8B models yielded different results, including different sets of independent variables for modeling model responses (Section 4) and differential contributions of surprisal and entropy (Section 5). Given that their overall accuracies were very similar, it is intriguing to see such divergent outcomes in more detailed analyses. We then visualized the response patterns of both models by condition, as shown in Figures 7A and 7C. The most notable difference between the two language models is their responses in the diff-context condition: the 4B model performed much better in the same-sense than the diff-sense condition, while the 8B model performed much better in the diff-sense than the same-sense condition. Moreover, in the same-context condition, the 8B model showed comparable performance between same-sense and diff-sense conditions, whereas the 4B model showed drastically better performance only in the same-sense, but not the diff-sense, condition. These two figures indicate that both models are biased toward certain response options.

Motivated by these dichotomous patterns, we computed the accuracy of human participants by condition and compared its distribution to those of the 4B and 8B models using Earth mover's distance (also known as Wasserstein distance; Dobrushin, 1970), computed via the *R* package *emdist*. Our analyses revealed substantial individual differences in response patterns, mirroring those observed in the LLMs: some participants were closer to the 4B model (Figure 7B), while others resembled the 8B model (Figure 7D). However, due to the absence of cognitive ability and language history measurements from participants, it is difficult to draw definitive conclusions regarding the source of these differences. Future research should consider including such measures to better interpret individual variability.

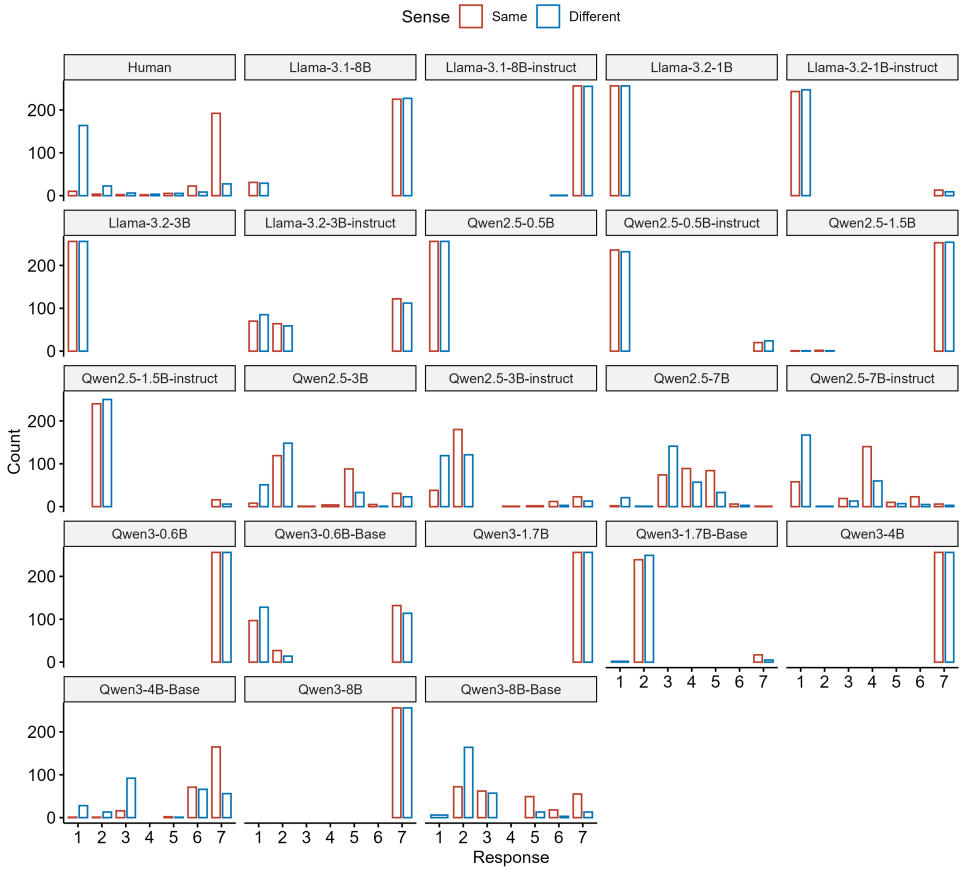

Figure 5: Distributions of responses generated by Human (averaged across all participants) and all LLMs.

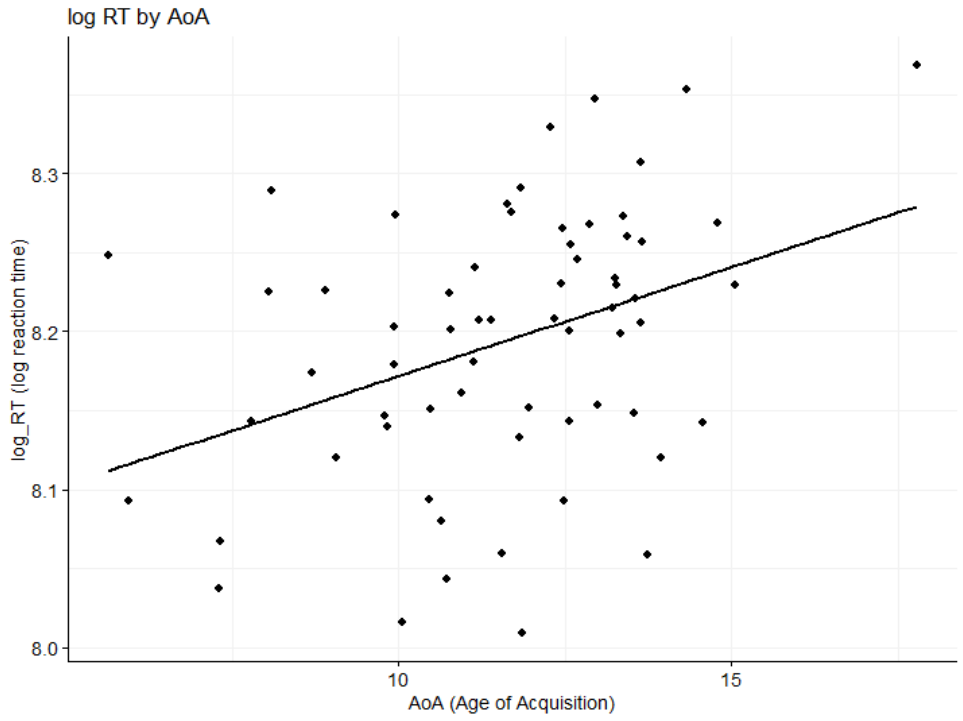

Figure 6: Scatter plot showing the relationship between Age of Acquisition (AoA) and log-transformed reaction time (log RT). Each point represents an individual homonym. The positive slope of the regression line indicates that words acquired later in life tend to elicit longer reaction times, suggesting slower processing for later-acquired words.

Word	POSWord	AoA	Word_logW	Word_logW_CD	Word_W_million	CLD_Frequency	PMI	PSPMI	TScore	PSTScore	EntropyCharacterFrequencies
内心	Same	11.125	3.1596	3.0082	43.04	106.338	4.2085	5.2797	41.9416	62.6172	0.5404
开播	Same	11.9565	1.6532	1.6021	1.34	3.6101	0.8778	1.695	1.174	2.3629	0.3432
自重	Different	12.5789	1.4314	1.3802	0.8	2.4759	-2.3084	-0.2295	-2.795	-0.2506	0.7709
口服	Same	7.7895	0.9542	0.9031	0.27	1.7346	0.742	2.7612	0.6849	2.9235	0.9993
下手	Different	11.8333	2.8319	2.7143	20.24	13.699	-0.9587	1.4823	-2.5051	3.9723	0.8739
站台	Different	10.7826	2.1847	1.9031	4.56	4.2995	5.0736	8.5102	11.6752	39.485	0.9933
开口	Same	8.05	3.0245	2.9165	31.54	20.9118	2.0228	2.9363	6.9501	10.9991	0.6722
下线	Different	14.3333	1	0.9031	0.3	2.5278	-0.3804	1.1514	-0.4205	1.3028	0.3552
偏食	Different	10.65	0	0	0.03	0.7339	3.7022	4.7977	2.8531	4.3554	0.5578
赛会	Same	13.9474	1	0.8451	0.3	0.2669	-1.9853	0.8279	-0.7683	0.3005	0.3138
跑车	Different	9.8235	2.1959	2.0969	4.68	5.1223	4.9418	6.6793	12.139	22.6889	0.6511
名家	Same	12.5714	1.0414	1	0.33	1.5715	-2.9003	-1.2307	-2.9665	-1.1022	0.77
上任	Different	12.2857	1.9494	1.8751	2.65	1.4455	-2.3323	0.967	-2.1624	0.8211	0.4984
风化	Different	13.3333	1.2553	1.1761	0.54	0.6597	-0.3528	0.6328	-0.1991	0.3591	0.9602
抽水	Same	11.6316	1.2041	1.0414	0.48	0.4003	0.9925	1.9126	0.4439	0.9015	0.8034
人流	Same	13.7368	1.5051	1.3802	0.95	3.9807	-0.852	2.3293	-1.1954	3.5827	0.2832
积分	Different	12.4375	1.7076	1.5441	1.52	60.5411	5.5038	8.2162	51.2572	133.7279	0.2527
作为	Different	12.9474	3.7682	3.4891	174.8	107.9911	3.27	5.0709	28.9298	58.4561	0.7671
抽风	Same	13.5455	1.1139	1.1139	0.39	4.255	4.0924	5.7494	8.0201	14.848	0.8672
光盘	Different	11.8182	1.5682	1.4472	1.1	2.854	3.5315	5.866	5.248	12.6805	0.8248
生气	Different	5.95	3.583	3.3456	114.11	88.5322	2.5563	3.9784	18.9402	34.9853	0.8092
空门	Same	15.0625	1.1139	1.0792	0.39	0.5634	-0.9768	0.012	-0.518	0.0062	0.9822
大气	Same	11.2105	2.0934	1.8573	3.7	10.5486	-0.446	0.0528	-1.0079	0.1188	0.6475
会意	Different	12.6818	0.4771	0.4771	0.09	0.2372	-4.8948	-1.1785	-2.5672	-0.409	0.8305
应变	Different	12.4737	1.1761	1.1461	0.45	1.4974	-0.2429	1.5205	-0.2063	1.3502	0.9989
粉丝	Same	11.85	2.9974	2.6693	29.63	105.0482	10.2615	10.8692	358.8025	443.044	0.9768
上手	Different	13.6667	1.7709	1.7076	1.76	4.9741	-1.889	0.464	-3.1333	0.7204	0.8583
枪手	Same	13	2.4771	2.233	8.94	4.6924	3.2399	5.842	5.9534	16.1198	0.3917
开展	Same	9.9545	2.0043	1.959	3.01	6.8718	1.6045	2.8801	3.068	6.1464	0.342
火花	Same	8.0952	2.4378	2.2695	8.17	5.1223	2.6881	4.1054	4.8539	8.8443	0.963
转机	Different	12.5652	2.1173	2.0334	3.91	4.3069	0.3845	1.3245	0.5547	1.9729	0.8579
效力	Different	13.2778	2.356	2.2718	6.77	2.5797	1.64	2.7533	1.9257	3.552	0.4846
配方	Same	10.7647	2.3945	2.0645	7.39	4.8184	2.9721	4.6081	5.3653	10.3959	0.6476
关门	Different	5.6316	2.934	2.8014	25.61	19.0808	2.9539	3.448	10.59	13.1082	0.9751
金星	Same	14.5714	1.5563	1.4472	1.07	2.5204	1.1606	2.8665	1.3119	3.6994	0.9999
上报	Same	13.2105	2.1703	2.0682	4.41	1.7569	-1.3754	1.0725	-1.312	1.0082	0.3869
下场	Different	12.875	2.4928	2.4548	9.27	8.8807	-0.1733	1.1935	-0.3582	2.5363	0.621
蘑菇	Different	7.3	2.3502	2.1239	6.68	11.8681	13.9613	13.9658	435.053	435.7319	0.8858
抄袭	Same	10.95	1.5185	1.4314	0.98	5.2261	12.1359	13.4931	153.3326	245.4549	0.8768
点播	Same	12.35	1.4472	1.3802	0.83	1.5641	0.6964	3.8494	0.6096	4.4188	0.3573
上身	Different	9.9375	2.2175	2.1271	4.92	7.361	-0.9112	1.8247	-1.7423	3.6647	0.7226
满月	Different	7.3182	1.8451	1.6902	2.09	3.788	3.4833	4.7223	5.9267	9.6209	0.9462
花红	Same	10.4783	0.699	0.6021	0.15	0.4522	-1.0758	2.2583	-0.5131	1.1634	0.9931
印花	Different	14.8	1.4314	1.3222	0.8	5.2557	4.9225	5.9346	12.2086	17.636	0.5195
地头	Same	13.65	1.301	1.2553	0.6	0.4893	-3.7659	-2.9014	-2.3902	-1.656	0.921
参见	Same	11.7037	1.7243	1.6335	1.58	0.8821	-1.4986	-0.9256	-1.0201	-0.613	0.9212
相机	Different	8.7	2.7388	2.5079	16.34	32.8095	2.196	3.1572	9.5851	15.1906	0.9672
制服	Different	12.4762	2.9425	2.7745	26.11	10.5337	3.8721	5.7844	11.571	23.6581	0.9693
改编	Same	13.4286	2.2788	2.1931	5.66	4.7517	5.9949	8.1197	17.135	36.2245	0.6643
大作	Different	13.6316	1.8195	1.7709	1.97	2.7428	-2.4409	-1.6329	-3.1484	-1.9762	0.6384
家居	Different	13.25	1.7324	1.6902	1.61	12.209	1.3624	6.1013	3.4236	28.5306	0.3979
点子	Same	9.8	2.7789	2.5922	17.92	6.4047	-1.5556	1.1637	-2.863	2.0971	0.9888
编制	Different	17.8	1.5315	1.4624	1.01	1.401	4.2677	5.2657	4.9252	7.1509	0.6321
分数	Same	8.8947	2.3945	2.2148	7.39	9.103	1.1784	2.2864	2.5335	5.298	0.5721
命题	Different	11.15	1.301	1	0.6	1.5789	-0.1102	2.2289	-0.096	2.1402	0.9698
做工	Different	11.4	1.6628	1.5051	1.37	2.8688	1.203	3.4898	1.4537	5.1717	0.8813
火星	Same	10.45	2.5302	2.2095	10.11	21.9051	4.4799	6.0872	21.1176	38.023	0.9829
虎口	Same	11.5556	1.5051	1.4624	0.95	1.6605	4.8548	7.789	6.692	19.0768	0.6947
调剂	Same	13.5556	1.2041	1.1461	0.48	1.1119	7.1228	8.3983	12.3595	19.3117	0.4765
上天	Different	10.7222	2.1173	2.0969	3.91	17.2128	-1.3964	0.1337	-4.1744	0.3846	0.9952
台风	Same	10.05	2.0043	1.6812	3.01	3.4322	2.4836	5.2424	3.5979	11.0974	0.9562
弹奏	Same	9.0526	2.0531	1.8865	3.37	2.1349	7.1317	8.8493	17.1797	31.3116	0.7435
农家	Same	9.9333	1.2553	1.2041	0.54	3.7361	1.8838	2.2737	2.707	3.3715	0.1537
燃点	Different	13.3704	0.7782	0.699	0.18	0.1408	-0.5096	0.0848	-0.1332	0.0221	0.1071

Table 3: Sixty-four homonyms were included in the rating experiment, along with their part-of-speech (*POS*) categories and ten other age of acquisition (*AoA*) and word frequency-related psychological properties. *CD* in *Word-log-WCD* refers to the number of film titles in which the word or character appears. *CLD-Frequency* indicates the frequency from the Chinese Lexical Database, which is based on a large-scale corpus of simplified Chinese (the Simplified Chinese Corpus of Webpages). *PMI* (Pointwise Mutual Information) measures how much more likely the joint occurrence of two variables is compared to what would be expected if the variables were independent; it is calculated as the logarithm of the ratio between the observed and expected frequencies. *PSPMI* refers to position-specific PMI, meaning that character frequencies are calculated according to their specific positions. *TScore* (t-score) provides a measure of the association strength between two characters in a homonym, giving higher scores to pairs with high co-occurrence frequencies and reflecting the non-randomness of their co-occurrence. *PSTScore* refers to a position-specific t-score. *EntropyCharacterFrequencies* denotes the entropy over the probability distribution of both characters in a two-character word; values are higher when the frequencies of the two characters are more similar. For detailed data and explanations of these psychological properties, please see Sun et al. (2018).

Model	Correlation (averaged)	Standard deviation
Llama-3.1-8B	-0.006	0.024
Llama-3.1-8B-instruct	0.053	0.009
Llama-3.2-1B	NA	NA
Llama-3.2-1B-instruct	0.068	0.026
Llama-3.2-3B	NA	NA
Llama-3.2-3B-instruct	0.050	0.028
Qwen2.5-0.5B	NA	NA
Qwen2.5-0.5B-instruct	-0.023	0.024
Qwen2.5-1.5B	0.016	0.023
Qwen2.5-1.5B-instruct	0.095	0.020
Qwen2.5-3B	0.307	0.027
Qwen2.5-3B-instruct	0.246	0.028
Qwen2.5-7B	0.383	0.031
Qwen2.5-7B-instruct	0.455	0.031
Qwen3-0.6B	NA	NA
Qwen3-0.6B-Base	0.053	0.028
Qwen3-1.7B	NA	NA
Qwen3-1.7B-Base	0.112	0.026
Qwen3-4B	NA	NA
Qwen3-4B-Base	0.512	0.037
Qwen3-8B	NA	NA
Qwen3-8B-Base	0.444	0.032

Table 4: Correlation between human and LLM responses, averaged across 55 participants. NA indicates that correlation cannot be computed as those model only produced one of the seven answers in all questions.

Factor	Sum Sq	Mean Sq	NumDF	DenDF	F-value	p-value
Sense	26.174	26.174	1.000	22,594.020	216.254	$< 0.001^{***}$
Context	1.398	1.398	1.000	22,550.642	11.552	0.001^{***}
POSWord	0.282	0.282	1.000	60.116	2.328	0.132
Trial	176.215	176.215	1.000	22,546.701	1,455.893	$< 0.001^{***}$
AoA	0.828	0.828	1.000	60.128	6.839	0.011^{*}
PSPMI	0.939	0.939	1.000	59.748	7.761	0.007^{**}
Sense:Context	62.869	62.869	1.000	22,552.566	519.424	$< 0.001^{***}$
Sense:POSWord	0.538	0.538	1.000	22,591.785	4.443	0.035^{*}
Context:POSWord	1.242	1.242	1.000	22,550.085	10.261	0.001^{**}
Sense:Context:POSWord	0.544	0.544	1.000	22,551.996	4.497	0.034^{*}

Table 5: Type III ANOVA results of the final model fitted on human reaction time.

Factor	Sum Sq	Df	F-value	p-value
(Intercept)	0.816	1.000	31.701	$< 0.001^{***}$
Sense	2.313	1.000	89.876	$< 0.001^{***}$
Context	1.337	1.000	51.931	$< 0.001^{***}$
Word_logW_CD	0.141	1.000	5.484	0.020^{*}
PMI	0.142	1.000	5.535	0.019^{*}
PSPMI	0.115	1.000	4.478	0.035^{*}
Sense:Context	0.333	1.000	12.943	$< 0.001^{***}$

Table 6: Type III ANOVA results of the final model fitted on surprisal computed from *Qwen3-4B-Base*.

Factor	Sum Sq	Df	F-value	p-value
(Intercept)	5.033	1.000	146.362	$< 0.001^{***}$
Sense	0.947	1.000	27.550	$< 0.001^{***}$
Context	1.822	1.000	52.975	$< 0.001^{***}$
POSWord	0.109	1.000	3.167	0.076
EntropyCharacterFrequencies	0.562	1.000	16.340	$< 0.001^{***}$
Sense:Context	2.456	1.000	71.436	$< 0.001^{***}$
Sense:POSWord	0.047	1.000	1.355	0.245
Context:POSWord	0.214	1.000	6.217	0.013^{*}
Sense:Context:POSWord	0.161	1.000	4.691	0.031^{*}

Table 7: Type III ANOVA results of the final model fitted on surprisal computed from *Qwen3-8B-Base*.

Agent	Context effect	POSWord	Sense	estimate	SE	df	t.ratio	p.value
Human	Same - Different	Same	Same	0.095	0.009	22,547.185	10.662	$< 0.001^{***}$
Human	Same - Different	Different	Same	0.085	0.009	22,546.864	9.498	$< 0.001^{***}$
Human	Same - Different	Same	Different	0.097	0.010	22,550.048	10.088	$< 0.001^{***}$
Human	Same - Different	Different	Different	0.147	0.010	22,559.439	15.195	$< 0.001^{***}$
Qwen3-8B-Base	Same - Different	Same	Same	0.355	0.049	344.000	7.278	$< 0.001^{***}$
Qwen3-8B-Base	Same - Different	Different	Same	0.185	0.048	344.000	3.892	$< 0.001^{***}$
Qwen3-8B-Base	Same - Different	Same	Different	0.158	0.036	344.000	4.380	$< 0.001^{***}$
Qwen3-8B-Base	Same - Different	Different	Different	0.141	0.038	344.000	3.664	$< 0.001^{***}$

Table 8: Post-hoc comparison between same-context and diff-context conditions. Correspond to Figure 3.

Agent	POS effect	Context	Sense	estimate	SE	df	t.ratio	p.value
Human	Same - Different	Same	Same	0.024	0.020	79.647	1.184	0.240
Human	Same - Different	Different	Same	0.014	0.021	82.264	0.686	0.495
Human	Same - Different	Same	Different	0.064	0.021	90.678	3.018	0.003^{**}
Human	Same - Different	Different	Different	0.014	0.021	82.347	0.688	0.493
Qwen3-8B-Base	Same - Different	Same	Same	0.066	0.037	344.000	1.779	0.076
Qwen3-8B-Base	Same - Different	Different	Same	0.104	0.057	344.000	1.819	0.070
Qwen3-8B-Base	Same - Different	Same	Different	0.001	0.041	344.000	0.035	0.972
Qwen3-8B-Base	Same - Different	Different	Different	0.018	0.033	344.000	0.549	0.583

Table 9: Post-hoc comparison between same-POS and diff-POS conditions. Correspond to Figure 3.

Factor	Sum Sq	Mean Sq	NumDF	DenDF	F-value	p-value
Sense	1.796	1.796	1.000	22,369.455	14.891	$< 0.001^{***}$
Context	0.378	0.378	1.000	22,565.927	3.132	0.077
POSWord	0.251	0.251	1.000	59.877	2.077	0.155
Trial	177.773	177.773	1.000	22,545.570	1,473.910	$< 0.001^{***}$
AoA	0.946	0.946	1.000	59.853	7.843	0.007^{**}
PSPMI	0.950	0.950	1.000	59.457	7.876	0.007^{**}
surprisal	10.518	10.518	1.000	21,816.065	87.200	$< 0.001^{***}$
Sense:Context	40.025	40.025	1.000	22,599.685	331.847	$< 0.001^{***}$
Sense:POSWord	0.434	0.434	1.000	22,594.006	3.599	0.058
Context:POSWord	0.905	0.905	1.000	22,551.405	7.500	0.006^{**}
Sense:Context:POSWord	0.604	0.604	1.000	22,551.776	5.007	0.025^{*}

Table 10: Type III ANOVA results of the final model fitted on human reaction time with model-derived metrics from *Qwen3-4B-Base*.

Factor	Sum Sq	Mean Sq	NumDF	DenDF	F-value	p-value
Sense	26.222	26.222	1.000	22,593.151	216.677	$< 0.001^{***}$
Context	1.352	1.352	1.000	22,549.860	11.174	0.001^{***}
POSWord	0.282	0.282	1.000	60.075	2.330	0.132
Trial	176.340	176.340	1.000	22,545.598	1,457.125	$< 0.001^{***}$
AoA	0.843	0.843	1.000	60.090	6.969	0.011^{*}
PSPMI	0.931	0.931	1.000	59.719	7.694	0.007^{**}
entropy_z	0.582	0.582	1.000	22,081.381	4.813	0.028^{*}
Sense:Context	62.260	62.260	1.000	22,552.424	514.468	$< 0.001^{***}$
Sense:POSWord	0.512	0.512	1.000	22,590.680	4.233	0.040^{*}
Context:POSWord	1.256	1.256	1.000	22,549.355	10.378	0.001^{**}
Sense:Context:POSWord	0.544	0.544	1.000	22,551.065	4.497	0.034^{*}

Table 11: Type III ANOVA results of the final model fitted on human reaction time with model-derived metrics from *Qwen3-8B-Base*.

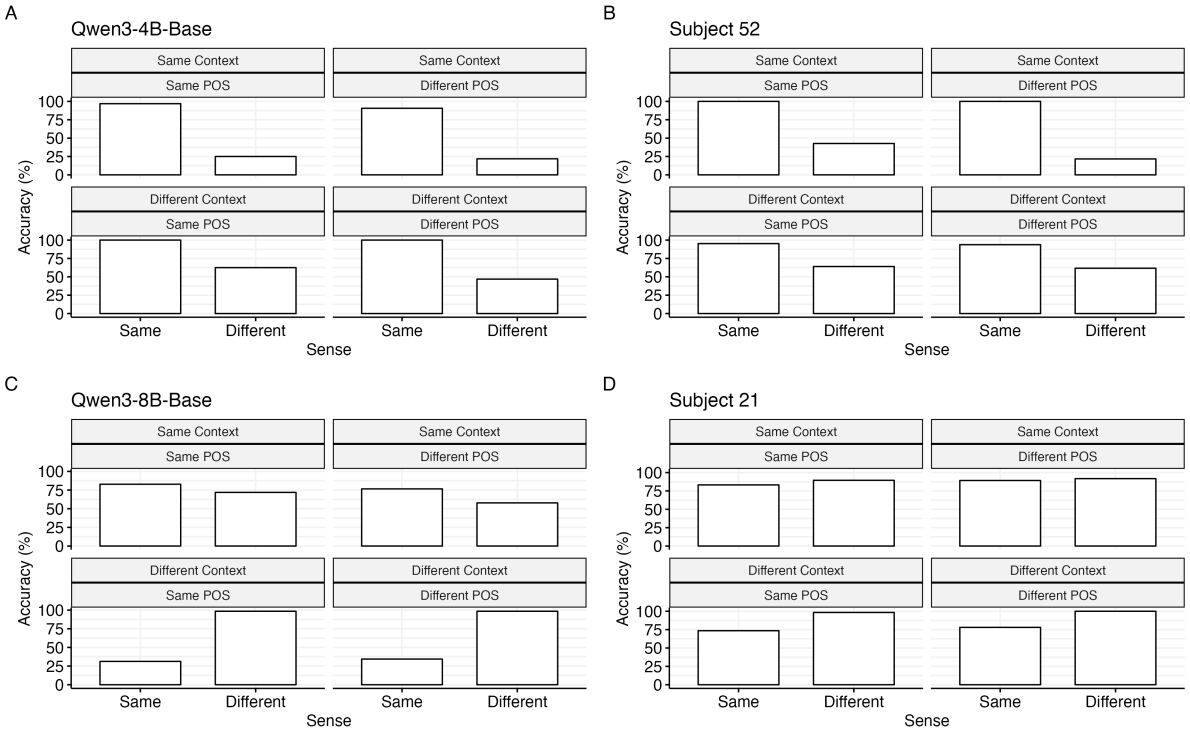

Figure 7: The accuracy of the homonym judgment tasks by humans and models. A: *Qwen3-4B-Base*. B: Subject 52, whose response pattern is closer to 4B than 8B model. C: *Qwen3-8B-Base*. D: Subject 21, whose response pattern is closer to 4B than 8B model.