Prompting instruction-tuned LLMs for Chinese semantic similarity

Yunchong Huang

Abstract

Evaluating lexical semantic similarity is essential for assessing language models (LMs) grounded in distributional semantics. While traditional benchmarks like SimLex-999 have been widely used for word embeddings, instruction-tuned LLMs require prompting for this evaluation as word embedding vectors are no longer directly accessible. Prior studies focus mainly on English and Germanic languages, leaving a gap for non-Western languages such as Mandarin Chinese. This report applies the prompting methods from previous research on English and Dutch to extract semantic similarity scores from GPT-4o and Doubao-pro-128k using the Multi-SimLex benchmark. By comparing their alignment with human judgments, we gain insights into their semantic representations and potential differences in training data and methodologies.

1 Introduction

Lexical-semantic similarity measurement has been a longstanding approach in evaluating Language Models (LMs). Unlike extrinsic tasks focusing on downstream performance (e.g. machine translation, question answering) or sentence similarity, measuring lexical-semantic similarity provides direct observation of how LMs encode and differentiate core lexical meanings without being influenced by higher-level factors like discoursal contexts and syntactic structures. Based on theories of distributional semantics (Harris, 1954; Boleda, 2020), LM word representations are in the form of vector embeddings reflecting subtle meaning differences quantitatively.

An important method to evaluate the general quality of lexical representations of an LM is by calculating the correlation between similarity scores of designated word pairs retrieved from an LM and the corresponding similarity judgments obtained from humans. This correlation is supposed to reveal how well the LM lexical representations align with human cognitive lexical representations. Classic benchmarks developed for this evaluation purpose include WordSim-353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015), which were widely applied to static word embeddings represented by Word2Vec (Mikolov et al., 2013) and Glove (Pennington et al., 2014). Their variants are also used to evaluate contextual word embeddings directly retrieved from Large Language Models (LLMs) with the Bidirectional Encoder Representations from Transformers (BERT) architecture (Brans and Bloem, 2024; Vulić et al., 2020).

When it comes to the contemporary instruction-tuned LLMs like Generative Pretrained Transformer (GPT) models which are closed-source (OpenAI, 2024b), as word embedding vectors are no longer directly accessible, prompting becomes the main approach to evaluate their lexical-semantic representations. Several works have explored this approach: Snelder (2024) extracted semantic similarity values of SimLex-999 word pairs utilising a set of prompting engineering strategies on GPT-3.5; De Deyne (2024) investigates GPT-4's capacity to produce similarity judgments similar to human's through lexical triad tasks; Trott (2024) explores the feasibility to use GPT-4 predictions on word-level semantic attributes as synthetic data for psycholinguistic dataset enhancement.

Viewing from the horizontal research landscape, current research efforts on evaluating lexical semantic representations of modern instruction-tuned LLMs are largely restricted to English and other Germanic languages such as Dutch. It is worth noting that for Mandarin Chinese, a non-Western language with abundant resources, less attention has been paid to this fundamental evaluation of the latest models. Although various valuable benchmarks on lexical semantic similarities have been proposed (Jin and Wu, 2012; Wu and Li, 2016; Huang et al., 2019; Vulić et al., 2020), an evaluation of Mandarin Chinese lexical-semantic abilities

of instruction-tuned LLMs is still absent. In this report, we will adopt the methodology introduced in Snelder (2024) to examine the effectiveness of prompting strategies on extracting lexical-semantic similarity scores for Mandarin Chinese from GPT-4o and Doubao-pro-128k[1]. Moreover, a correlational analysis will be carried out between model similarity scores and human similarity judgments from the Multi-SimLex benchmark (Vulić et al., 2020), and we will also compare the two models regarding this correlation measurement and investigate further their predicted similarity scores.

This evaluation task would serve as an insightful probe into models' implicit semantic representations, which is foundational for LLMs to perform NLP tasks related to natural language understanding and generation. Moreover, since Doubao is one of the leading Chinese instruction-tuned LLMs (8PixLabs, 2024), a comparative evaluation of the Mandarin Chinese lexical-semantic ability between Doubao-pro-128k and GPT-4o can potentially reveal similarities and differences of their technical details (e.g. the patterns of their training data).

2 Related Work

2.1 Instruction-tuned LLMs

GPT is an LLM based on the Transformer architecture (Vaswani et al., 2017) developed by OpenAI that consists of stacked decoders utilising causal self-attention mechanisms to ensure autoregressive generation, optimized by instruction tuning. The state-of-the-art GPT models include GPT-4 and GPT-4 Omni(o). They are both fine-tuned with Reinforcement Learning from Human Feedback (RLHF) (OpenAI, 2023, 2024a)[2] to achieve better interactions with users. They also both possess the multimodal ability to process text, audio and visual inputs, but GPT-4 depends on calling on other modal-specific OpenAI models (Dall-E and Whisper) while the multimodality of GPT-4o is achieved end-to-end within itself (OpenAI, 2024b; Craig, 2025). However, since GPT models after GPT-3 are closed-source, the detailed architecture of GPT-4 and GPT-4o are unavailable. In parallel, the Doubao LLM has rapidly gained popularity in the Chinese AI community since its initial release in May 2024. It is also based on the autoregressive transformer structure, and at the end of 2024, the latest version of its general mode (Doubao-pro-1215) has achieved comprehensive alignment with GPT-4o on several benchmarks, while several variants with multimodal processing capacity have been developed (Doubao Team, 2024). Similar to the latest GPT models, Doubao LLMs are also closed-source and their detailed structures remain unclear.

2.2 Lexical Semantics

Lexical semantics has always been an important aspect in the evaluation of LMs. The ideal lexical-semantic representations should converge with judgments based on human cognition, and this idea has motivated the curation of many benchmark datasets. Classic benchmark datasets for English include WordSimilarity 353 (WS-353) containing 353 word pairs annotated on semantic similarity and association (relatedness) by 13-16 human annotators (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015) which includes 999 word pairs rated on a scale from 0 to 10 on semantic similarity by about 50 native English speakers.

Similar datasets have also been developed in other languages or multilingually, effectively broadening the research landscape. More specifically, there are also many datasets involving Mandarin Chinese. To name a few, Multi-SimLex (Vulić et al., 2020) is a large-scale dataset covering 12 typologically diverse languages to evaluate both monolingual and crosslingual lexical semantic similarity. For Chinese, it contains 1,888 word semantically aligned word pairs across four different parts of speech (PoS). WordSim-297 (Jin and Wu, 2012) is a Chinese dataset for the evaluation of lexical semantic similarity derived and translated from the English WS-353. PKU500 (Wu and Li, 2016) is a Chinese word similarity dataset containing 500 pairs of Chinese words from various domains, including common and specialized vocabulary. COS960 (Huang et al., 2019) is a dataset consisting of 960 pairs of two-character Chinese words, developed to evaluate lexical similarity. The dataset includes words from different PoS selected from HowNet (Dong and Dong, 2003), a well-known linguistic knowledge base of Chinese.

With benchmark datasets of this sort and the continuous development of LLMs, research efforts have been made to evaluate their judgments on

[1] a representative instruction-tuned LLM developed by the Chinese company *ByteDance*.

[2] The official introduction to GPT-4o doesn't directly mention the RLHF technique, but the GPT-4o mini introduction explicitly mentions it and claims that the safety mitigations implemented in GPT-4o mini are identical to those of GPT-4o.

lexical semantic relationships and compare them with human judgments using correlational measurements. Snelder (2024) extracted semantic similarity values of English and Dutch SimLex-999 word pairs with GPT-3.5, utilising a set of prompting engineering strategies such as zero-shot in-batch prompting, few-shot in-batch prompting, single word-pair prompting, prompting with alternative scales, etc., concluding that single word-pair prompting extracts the best correlation results (ρ = 0.82 for English, ρ = 0.72 for Dutch). De Deyne (2024) investigates GPT-4's ability to replicate human similarity judgments through remote and basic-level triad tasks, showing correlations with human judgments up to ρ = 0.84, outperforming traditional models like SWOW and word2vec. It also effectively distinguishes between similarity and relatedness, adapting to tasks with concrete and abstract word diads/pairs respectively. Trott (2024) examines GPT-4's potential to enhance psycholinguistic datasets by using its predictions on word-level semantic attributes as synthetic data, where results show significant correlations with human annotations (with ρ between 0.39–0.86).

Zooming onto the specific language of Mandarin Chinese, Vulić et al. (2020) implemented the lexical semantic similarity (LSIM) task in several languages including Mandarin Chinese on a family of BERT LMs utilising the Multi-SimLex dataset, where the results show that monolingual contextual-encoded models with CLS tokens performed the best for Chinese, reaching a correlation value with human judgments (ρ) above 0.55. Within the Chinese research community, much attention has been paid to evaluating instruction-tuned LLMs' capacities in sentential understanding, logical reasoning and comprehensive knowledge (Xu et al., 2020; Wang et al., 2022; Huang et al., 2023; Sun et al., 2024). From the perspective of computational lexical semantics, most research has been focused on training more delicate word embeddings (Cao et al., 2018; Yin et al., 2016) and proposing innovative approaches to measure semantic relationships (Zhang et al., 2016; Huang et al., 2018; Pei et al., 2016). Little research has explored the extraction of lexical semantic representations from closed-source instruction-tuned LLMs and comparative studies on its basis.

2.3 Prompt Engineering

Prompting is the process of using natural language text to describe the task that a generative AI should perform (Radford et al., 2019). Prompt engineering, especially in interactions with text-to-text LMs, refers to measures such as query phrasing, style adjustment and word and grammar choice to elicit the best possible output (Wahle et al., 2024). In this report, we selected five prompt categories from Snelder (2024) to extract semantic similarity scores and verify the potential effects of differences between zero-shot or few-shot formats, inputting word pairs in batch or individually, different correlation measuring scales, and languages used in prompts. Detailed illustrations of prompt categories involved will be provided in Section 3.

3 Method

3.1 The Multi-SimLex benchmark

We use Mandarin Chinese word pairs from the Multi-SimLex dataset (Vulić et al., 2020) as the human judgment benchmark of the lexical-semantic similarity. Due to the financial constraints, we randomly sampled 25% of word pairs from each PoS category for the experiments, leading to 472 word pairs selected from the total 1888 pairs. The dataset adopts a scale from 0 to 6 (inclusive) for the pure *semantic similarity* discerned from *relatedness / association*, where 0 represents "no similarity" and 6 represents "very high similarity", and 11 valid annotations are collected for each Mandarin Chinese word pair.

3.2 Models

The language models evaluated in this report are *gpt-4o-2024-11-20* (OpenAI, 2024b) and *Doubao-pro-128k* (Doubao Team, 2024). Prompting interactions with the two models are achieved through API calls, with the former implemented with the OpenAI developer platform [3] and the latter with Volcano Engine [4].

3.3 Prompting Setup

Table 1 lists the five prompt categories selected from Snelder (2024). Prompting instructions are loyally translated into Mandarin Chinese (except for F-6) from the original English scripts to maintain the accurate representation of human language

[3] https://platform.openai.com/docs/overview

[4] https://console.volcengine.com/ark/region:ark+cn-beijing/model/detail?Id=doubao-pro-128k

ID	Category
F-1	Zero-shot, word pairs in batches
F-3	Few-shot, word pairs in batches
F-5	Zero-shot, categorical scale, word pairs in batches
F-6	Zero-shot, cross-linguistic, word pairs in batches
F-9	Zero-shot, single-word pair

Table 1: List of evaluated prompt categories.

understanding guaranteed in the previous research, adjusting only the numerical scale to align with the Multi-SimLex benchmark. The complete instructions for the prompts involved are provided in Appendix A.

The F-1 and F-3 categories both pack word pairs in batches with prompting instructions, but the difference is that the F-1 presents no examples while F-3 uses the first three word pairs from the selected 472 word pairs as examples. Few-shot learning is considered beneficial in prompting, and results in Snelder (2024) also reveal that F-3 obtain results with stronger correlations than F-1 for both English and Dutch. However, its high sensitivity towards specific instruction wordings and example quality also brings fluctuations in performance (Jiang et al., 2020; Ramlochan, 2024).

The F-5 category adopts a categorical scale inspired by works in psychology showing that verbal (categorical) rating is more natural and reliable while causing less cognitive costs (Krosnick and Fabrigar, 1997; Menold, 2020). We use labels from '非常不相似(very dissimilar)' to '非常相似(very similar)' that are mapped to [0, 1.5, 3, 4.5, 6] for computing correlations.

The cross-linguistic F-6 prompt adopts the original English instructions for Mandarin Chinese word pairs, with the scale altered to align with Multi-SimLex. The F-9 prompt, in contrast with all the other categories, is in the "instruction + a single word pair" format. Previous results from Trott (2024) and Snelder (2024) show that this prompting strategy can be beneficial for better results. It should be mentioned that due to financial constraints, a further 25% random sampling from each PoS category is carried out on the 472 selected word pairs, leading to 118 word pairs involved in the F-9 experiment.

Taking the potential consistency issue into consideration (Brown et al., 2020; Ouyang et al., 2024), each prompt is executed 15 times, which is higher than 11 human annotations for each word pair in Multi-SimLex. They are treated in a way equivalent to ratings from different human annotators. In the evaluation, Spearman's rank correlation (ρ) is calculated between the mean semantic similarity scores estimated by the models and the mean human similarity scores for each involved word pair. We will also present the prompt-wise distribution of mean similarity scores and the inter-sample dispersion within each word pair across prompts of both models and compare them with human annotations. Additionally, a comparison between model-model correlation and model-human correlation will also be carried out.

4 Results

4.1 The correlation with human annotations

	ρ-gpt-4o	ρ-Doubao-pro
F-1	**0.865**	0.802
F-3	0.863	0.810
F-5	0.859	0.775
F-6	0.857	0.801
F-9	0.859	**0.841**
Variance	0.00001	0.00056
SD	0.00329	0.02368
Mean	0.8606	0.8058

Table 2: The correlation coefficients between LLM similarity predictions and human-annotated similarity scores measured by Spearman's rank correlation (ρ) across different prompts.

Table 2 shows that similarity scores from GPT-4o generally have a stronger correlation with human-annotated scores compared to Doubao-pro-128k results (ρ-gpt-4o mean 0.8606 > ρ-Doubao-pro mean 0.8058). On the other hand, the effect of different prompting strategies is more significant in Doubao-pro-128k than in GPT-4o as observed from the variance (Doubao pro: 0.0005-0.002; GPT-4o: 0.00001-0.00007) and standard deviation (Doubao pro: 0.02-0.05; GPT-4o: 0.003-0.007). Similarity scores obtained with F-9 from Doubao-pro bear the strongest correlation with the human-annotated scores compared to other prompt categories applied to the model, which reveals a similar pattern observed in GPT-3.5-turbo (Snelder, 2024). For GPT-4o, despite the low variance and standard deviation of cross-prompt results, we can conclude that the F-1 prompt obtains the strongest correlation with human-annotated scores in two different measurements, while results from the F-9 prompt are only around the average level.

4.2 Prompt-wise distribution of similarity scores from GPT-4o and Doubao-pro

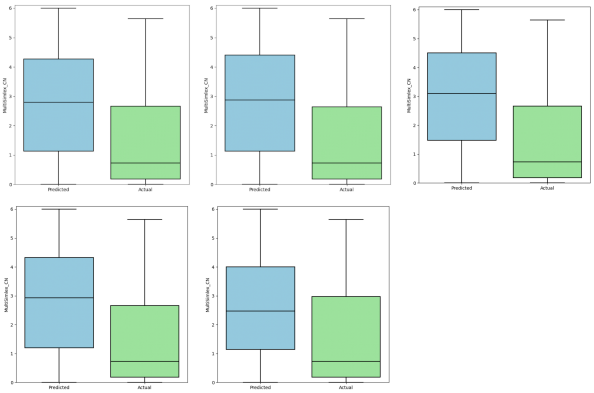

Figure 1: The comparison of mean similarity scores predicted by GPT-4o and human-annotated mean similarity scores based on F-1, F-3, F-5, F-6, and F-9 prompts respectively (from up left to bottom right).

As observed from Figure 1, across all prompts for GPT-4o, the medians (≈3) and the Interquartile Ranges (IQR) of predicted similarity scores are persistently higher than those of human-annotated similar scores (median=0.73). The highest prediction score reaches 6 (the largest value on the defined scale) in all prompts, which is higher than the maximal score of human annotations (5.64).

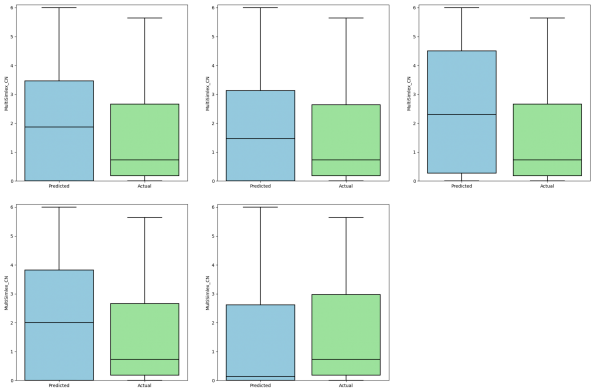

Figure 2: The comparison of mean similarity scores predicted by Doubao-pro-128k and human-annotated mean similarity scores based on F-1, F-3, F-5, F-6, and F-9 prompts respectively (from up left to bottom right).

Figure 2 shows that across all prompts other than F-9 with Doubao-pro, the medians (≈2) are also persistently higher than the median of human-annotated similar scores (0.73), but compared to the GPT-4o, their differences from the median of human-annotated results are smaller. It can also be observed that the IQR ranges of similarity scores predicted by Doubao-pro from prompts other than F-9 are significantly wider, revealing that the overall distribution of Doubao-pro estimated scores across word pairs is more dispersed than that of human-annotated scores. As for F-9, the distribution of predicted scores is comparatively peculiar: the median (0.17) is significantly lower than the human-annotated median (0.73), and the overall IQR range of predicted scores is lower than that of human-annotated scores. Meanwhile, similar to GPT-4o, the highest prediction score reaches 6 (the largest value on the defined scale) across all prompts, which is higher than the maximal score of human annotations (5.64).

4.3 Inter-sample dispersion within each word pair across prompts from GPT-4o and Doubao-pro

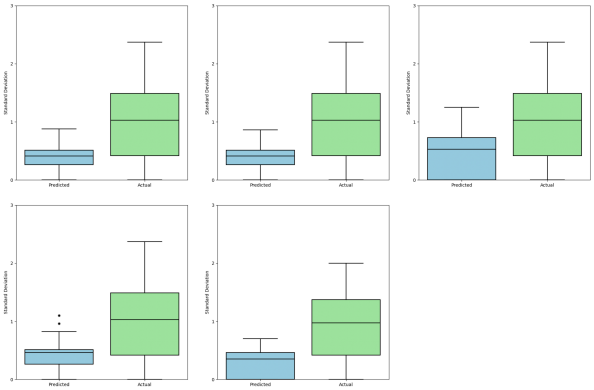

Figure 3: The comparison between standard deviations (SDs) of 15 samples for each word pair based on F1, F3, F5, F6 and F9 prompts respectively (from up left to bottom right) predicted by GPT-4o and the corresponding SD values of human-annotated scores.

As 15 samples of prediction scores are taken for each word pair, we also measure the standard deviation (SD) within samples of each word pair and present the overall distribution. Figure 3 and 4 show that across all prompts for both GPT-4o and Doubao-pro, the SD values of samples for word pairs are generally lower than those of human-annotated scores. More specifically, for GPT-4o, only F-5, F-6 prompts get samples with SD higher than 1; Doubao-pro, on the other hand, have more word pairs with a sample SD value greater than 1 and further outliers.

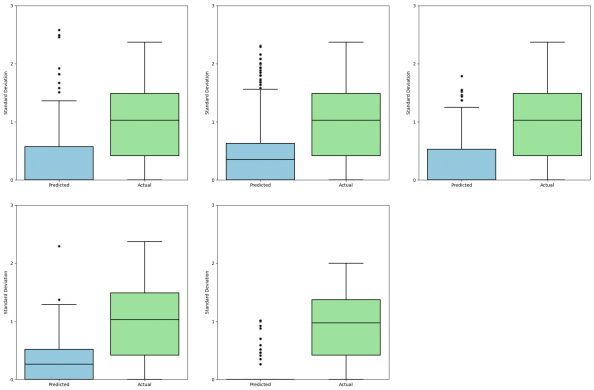

Figure 4: The comparison between standard deviations (SDs) of 15 samples for each word pair based on F1, F3, F5, F6 and F9 prompts respectively (from up left to bottom right) predicted by Doubao-pro and the corresponding SD values of human-annotated scores.

	Spearman ρ
gpt-4o & human	0.869
Doubao-pro & human	0.822
gpt-4o & Doubao-pro	**0.883**

Table 3: Correlation between cross-prompt mean similarity scores of models vs. Correlation between cross-prompt mean similarity scores of models and mean similarity scores of human annotations

4.4 Model-model correlation vs. model-human correlation

Table 3 shows another interesting result that the correlation between cross-prompt mean similarity scores of models is stronger than the correlation between cross-prompt mean similarity scores of both models and the mean similarity scores of human annotations.

5 Discussion

5.1 General Discussions

It is mentioned in Section 4.1 that the effect of different prompting categories is not obvious in results obtained from GPT-4o. The correlation between similarity scores obtained from the F-9 prompt (zero-shot, single word pair) and human annotations is weaker than that of the F-1 prompt (zero-shot, word pairs in batches). Both results are in contrast with the findings by Snelder (2024) on GPT-3.5-turbo. This may reveal improved output stability from GPT-3.5 to GPT-4o and a decreased sensitivity to prompt formats on similar tasks. The Doubao-pro model still presents the advantage of the F-9 prompting format and a more significant difference between the correlation results across different prompts, but the overall correlation strength is stronger than GPT-3.5-turbo.

Section 4.2 illustrates the prompt-wise distribution of mean similarity scores from both models. The GPT-4o distribution potentially reveals the model's overall tendency to rate higher semantic similarities between Chinese words than native-speaker human annotators. Interestingly, when introducing the Multi-SimLex dataset, Vulić et al. (2020) showed that Mandarin Chinese has one of the lowest average similarity scores, with over 55% of word pairs in the score interval [0,1). Many factors can be considered as potential reasons behind this tendency. For instance, similarity scores provided by multilingual LLMs may be more inclined to cross-linguistic general values instead of revealing the specific trend in one language. Additionally, human annotators recruited for Multi-SimLex curation received extended guidelines (e.g. the extra emphasis on differentiating semantic similarity and relatedness/association, to be discussed in Section 5.2), it is possible that they provided scores more cautiously than LLMs under more limited prompting instructions. Meanwhile, the Doubao-pro distribution shows a tendency similar to GPT-4o in rating higher semantic similarities between Chinese words than native-speaker human annotators, but this tendency is weaker with the observation as the numeric difference between medians is smaller. Potentially, this can be related to that although Doubao-pro is also a multilingual model, the Mandarin Chinese data may share a larger portion during its training and its lexical-semantic representations can better reveal the specific traits of Chinese lexical semantics. The comparatively peculiar distribution of predictions based on F-9 resonates with its outstanding correlation metrics with the human-annotated scores, while a similar pattern is not observed in GPT-4o. However, it remains uncertain if the smaller sample of word pairs for F-9 brings this effect.

From the inter-sample dispersion within each word pair across prompts from both models presented in Section 4.3, it can be concluded that both GPT-4o and Doubao-pro are more robust in the semantic similarity judgment over word pairs than different human annotators. This is not a surprising result, since an LLM can be regarded as a single agent despite the multiple sampling, while human annotators are truly multiple agents with potentially different linguistic backgrounds. Regarding

the comparison between the two models, we can conclude that GPT-4o is comparatively more robust than Doubao-pro from this aspect.

Lastly, Section 4.4 provides an insight such that instruction-tuned LLMs are more similar in lexical-semantic judgments with each other than with human judgments. This potentially reveals that instruction-tuned LLMs are still not completely aligning with the patterns of human cognition concerning lexical-semantic understanding.

5.2 Error Analysis

For a detailed error analysis, we calculated the difference between the mean prediction scores across prompts and the mean human-annotated scores for each word pair. Since we observe in 4.2 that LLMs in this research tend to predict similarity scores higher than those from human annotators, we carry out a detailed analysis on this and retrieve the 10 word pairs with the largest positive differences between LLM prediction scores and human annotation scores for both GPT-4o and Doubao-pro. Table 4 shows the 10 word pairs retrieved from GPT-4o predictions. It can be observed that one potential reason causing the positive difference between gpt-4o predictions and human annotations is the confusion between cognitive relatedness/association and semantic similarity. Retrieved word pairs potentially related to this include [荣誉(honour), 敬重(esteem)] (diff=4.07), [小说(novel), 作家(writer)] (diff=4.05), [病人(patient), 复诊(session)] (diff=3.88), [蜜蜂(bee), 蚂蚁(ant)] (diff=3.77), [牛(cow), 山羊(goat)] (diff=3.73) and [血液(blood), 骨髓(marrow)] (diff=3.64). Words in these pairs are highly related concepts that tend to co-occur in texts while they don't share essential functional features, which is the main contributor to semantic similarity. Since Vulić et al. (2020) addressed that discerning relatedness/association and semantic similarity is an important factor in the Multi-SimLex dataset curation, the low mean human annotations scores (all below 1) for these word pairs reveal this discerning effort. The much higher prediction scores from GPT-4o may reveal that this issue has not been addressed in the model's training, but it's also possible that the limited instructions in prompting categories included in our experiment didn't emphasize the issue as much as in the curation of Multi-SimLex. The latter speculation can be tested by including another prompting category with detailed instructions adapted from those provided for human annotators of Multi-SimLex.

R	Word 1	Word 2	SimL	GPT-4o
1	收到 *receive*	接受 *accept*	0.91	5.53
2	荣誉 *honor*	敬重 *esteem*	0.64	4.71
3	小说 *novel*	作家 *writer*	0.09	4.14
4	病人 *patient*	复诊 *session*	0.36	4.24
5	记得 *remember*	想 *think*	0.36	4.22
6	恶意软件 *malware*	蠕虫 *worm*	0.91	4.77
7	蜜蜂 *bee*	蚂蚁 *ant*	0.45	4.22
8	牛 *cow*	山羊 *goat*	0.45	4.18
9	血液 *blood*	骨髓 *marrow*	0.36	4.00
10	装作 *pretend*	似乎 *seem*	0.27	3.89

Table 4: The words with the highest absolute difference in human similarity score and model predicted similarity for GPT-4o, on a scale of 0 to 6.

Another potential reason for the observations above is the contextual limitations for the semantic similarity perception in some word pairs which is clear to humans but may not necessarily be obvious for LLMs. This is represented by word pairs such as [收到(receive), 接受(accept)] (diff=4.62), [恶意软件(malware), 蠕虫(worm)] (diff=3.86) and [装作(pretend), 似乎(seem)] (diff=3.62). For [收到(receive), 接受(accept)], their semantic similarity is easier to perceive when it comes to the delivery of physical entities; for [恶意软件(malware), 蠕虫(worm)], their similarity only becomes obvious when it comes to the domain of computers and the latter word is interpreted as a form of computer virus; while for [装作(pretend), 似乎(seem)], their similarity is only revealed in scenarios of subjective deception. Therefore, the large differences observed for these word pairs are potentially the results of the LLMs' over-generalisation of semantic similarities under certain contexts to general semantic similarity judgments.

Table 5 shows the 10 word pairs retrieved from Doubao-pro predictions. Comparatively, the confusion between relatedness/association and semantic similarity is not as visible as in GPT-4o results. Among the 10 retrieved word pairs, only [混合(mix), 溶解(dissolve)] (diff=4.08) and [蜜蜂(bee), 蚂蚁(ant)] (diff=4.08) can be strictly regarded as related instead of similar words. The large differences in other word pairs are more likely to be caused by the LLMs' over-generalisation of semantic similarities under certain contexts to

general semantic similarity judgments. For instance, the word pair with the largest difference [安排(arrange), 组织(organize)] (diff=5.14) reveals that the lack of part-of-speech information for words may be a reason for this over-generalisation, as the Chinese lexical form of 组织 can denote both the nominal meaning of "organization" and the verbal meaning of "organize".

R	Word 1	Word 2	SimL	Doubao
1	安排 *arrange*	组织 *organize*	0.82	5.14
2	收到 *receive*	接受 *accept*	0.91	5.03
3	看来 *apparently*	显然 *obviously*	1.91	6.00
4	钱包 *purse*	袋 *bag*	0.91	4.89
5	成长 *grow*	增加 *increase*	0.91	4.79
6	装作 *pretend*	似乎 *seem*	0.27	4.12
7	蜜蜂 *bee*	蚂蚁 *ant*	0.45	4.08
8	电线 *wire*	软线 *cord*	1.18	4.81
9	混合 *mix*	溶解 *dissolve*	0.55	4.08
10	享受 *enjoy*	娱乐 *entertain*	1.00	4.47

Table 5: The words with the highest absolute difference in human similarity score and model predicted similarity for Doubao-pro-128k, on a scale of 0 to 6.

Additionally, it's worth mentioning that 3 word pairs occur in both tables above: [收到(receive), 接受(accept)], [蜜蜂(bee), 蚂蚁(ant)] and [装作(pretend), 似乎(seem)]. Both LLMs involved in this report seem to overestimate their semantic similarities from the standard of human judgments, in which both speculations of similarity-relatedness confusion and over-generalisation of contextual similarity are revealed.

6 Conclusion and limitations

This report investigates the effectiveness of prompting instruction-tuned LLMs for evaluating Chinese lexical semantic similarity. By applying various prompting strategies to GPT-4o and Doubao-pro-128k and comparing their outputs with human-annotated similarity scores from the Multi-SimLex benchmark, we gain insights into the models' semantic representations. The results indicate that GPT-4o generally exhibits a stronger correlation with human judgments, whereas Doubao-pro-128k demonstrates greater variability across different prompting strategies. Notably, both models tend to overestimate semantic similarity compared to human annotators, suggesting that instruction-tuned LLMs still struggle to fully align with human lexical-semantic perception. In addition, this also brings questions about the ability of multilingual LLMs to present more accurate lexical-semantic representations specific to a designated target language.

Furthermore, our findings reveal that prompting strategies can significantly impact the performance of Doubao-pro, while GPT-4o remains relatively stable across different prompt formats. The observation of inter-sample dispersion within each word pair of LLMs being smaller than the inter-annotator variance of human judgments reflects that LLMs perform more like a single agent despite the multiple sampling. Additionally, the higher correlation between the two models' predictions compared to their respective correlation with human annotations suggests that these LLMs share similar underlying biases in their semantic representations.

Last but not least, our error analysis highlights two potential challenges for LLMs' lexical-semantic ability: the confusion between *semantic similarity* and *relatedness/association*; and the over-generalisation of *contextual* semantic similarity. These two challenges are present in samples we drew from both models, although the former challenge is more explicit for GPT-4o while the latter is more obvious for Doubao-pro.

Due to financial and time constraints, this report is with several limitations: (1) Experiments were implemented with a limited number of prompt categories. More prompt categories (e.g. the prompt in the fashion of guidelines provided for human annotators) to test open speculations proposed in this report. (2) The error analysis focuses only on word pairs with large positive differences between prediction scores from LLMs and human annotation scores. Future research can include more word pairs with large absolute differences to present a complete taxonomy of specific misalignment types between semantic similarity judgments of LLMs and humans.

Overall, this report provides valuable insights into the lexical-semantic capabilities of contemporary instruction-tuned LLMs for Mandarin Chinese. It also emphasizes the need for further refinements to improve their alignment with human lexical-semantic judgments, reflect language-specific lexical characteristics and include training for more fine-grained understanding of *semantic similarity*.

References

8PixLabs. 2024. Top 10 AI Model Created by Chinese Company. Retrieved January 15, 2025 from https://8pixlabs.com/top-10-ai-model-created-by-chinese-company/#Top_10_AI_Model_Created_by_Chinese_Company.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(Volume 6, 2020):213–234.

Lizzy Brans and Jelke Bloem. 2024. SimLex-999 for Dutch. *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14832–14845.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners.

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. 2018. cw2vec: Learning chinese word embeddings with stroke n-gram information. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Lev Craig. 2025. GPT-4o vs. GPT-4: How do they compare? Retrieved February 22, 2025 from https://www.techtarget.com/searchenterpriseai/feature/GPT-4o-vs-GPT-4-How-do-they-compare.

Simon De Deyne. 2024. Evaluating human-like similarity biases at every scale in large language models: Evidence from remote and basic-level triads. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Zhendong Dong and Qiang Dong. 2003. Hownet - a hybrid language and knowledge resource. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 820–824.

Doubao Team. 2024. 8 key moments of doubao big model in 2024. 2025-01-15.

Lev Finkelstein, Evgeniy Gabrilovich1, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20:116–131.

Zellig S. Harris. 1954. Distributional structure.

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

Degen Huang, Jiahuan Pei, Cong Zhang, Kaiyu Huang, and Jianjun Ma. 2018. Incorporating prior knowledge into word embedding for chinese word similarity measurement. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 17(3).

Junjie Huang, Fanchao Qi, Chenghao Yang, Zhiyuan Liu, and Maosong Sun. 2019. Cos960: A chinese word similarity dataset of 960 word pairs. *ArXiv*, abs/1906.00247.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. 2023. C-eval: a multi-level multi-discipline chinese evaluation suite for foundation models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Peng Jin and Yunfang Wu. 2012. SemEval-2012 task 4: Evaluating Chinese word similarity. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 374–377, Montréal, Canada. Association for Computational Linguistics.

Jon A Krosnick and Leandre R Fabrigar. 1997. Designing rating scales for effective measurement in surveys. *Survey measurement and process quality*, pages 141–164.

Natalja Menold. 2020. Rating-scale labeling in online surveys: An experimental comparison of verbal and numeric rating scales with respect to measurement quality and respondents' cognitive processes. *Sociological Methods & Research*, 49(1):79–107.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

OpenAI. 2023. GPT-4 Technical Report. Retrieved April 14, 2024 from https://doi.org/10.48550/arXiv.2303.08774.

OpenAI. 2024a. GPT-4o mini: advancing cost-efficient intelligence. Retrieved February 22, 2025 from https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

OpenAI. 2024b. Hello GPT-4o — OpenAI. Retrieved June 20, 2024 from https://openai.com/index/hello-gpt-4o/.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback.

Jiahuan Pei, Cong Zhang, Degen Huang, and Jianjun Ma. 2016. Combining word embedding and semantic lexicon for chinese word similarity computation. In *Natural Language Understanding and Intelligent Applications*, pages 766–777, Cham. Springer International Publishing.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Sunil Ramlochan. 2024. 0-shot vs few-shot vs partial-shot examples in language model learning. Retrieved February 23, 2025.

Xander Snelder. 2024. A new methodology to extract semantic similarities from gpts: Evaluating the effects of prompt engineering using simlex-999 benchmarks. Master's thesis, University of Amsterdam. Unpublished.

Jiaxing Sun, Weiquan Huang, Jiang Wu, Chenya Gu, Wei Li, Songyang Zhang, Hang Yan, and Conghui He. 2024. Benchmarking Chinese commonsense reasoning of LLMs: From Chinese-specifics to reasoning-memorization correlations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11205–11228, Bangkok, Thailand. Association for Computational Linguistics.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Ivan Vulić, Simon Baker, Edoardo Maria Ponti, Ulla Petti, Ira Leviant, Kelly Wing, Olga Majewska, Eden Bar, Matt Malone, Thierry Poibeau, Roi Reichart, and Anna Korhonen. 2020. Multi-SimLex: A Large-Scale Evaluation of Multilingual and Crosslingual Lexical Semantic Similarity. *Computational Linguistics*, 46(4):847–897.

Jan Philip Wahle, Terry Ruas, Yang Xu, and Bela Gipp. 2024. Paraphrase types elicit prompt engineering capabilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11004–11033, Miami, Florida, USA. Association for Computational Linguistics.

Lijie Wang, Yaozong Shen, Shuyuan Peng, Shuai Zhang, Xinyan Xiao, Hao Liu, Hongxuan Tang, Ying Chen, Hua Wu, and Haifeng Wang. 2022. A fine-grained interpretability evaluation benchmark for neural NLP. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–84, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Yunfang Wu and Wei Li. 2016. Overview of the nlpcc-iccpol 2016 shared task: Chinese word similarity measurement. In *Natural Language Understanding and Intelligent Applications*, pages 828–839, Cham. Springer International Publishing.

Liang Xu, Hai Hu, Xuanwei Zhang, Lu Li, Chenjie Cao, Yudong Li, Yechen Xu, Kai Sun, Dian Yu, Cong Yu, Yin Tian, Qianqian Dong, Weitang Liu, Bo Shi, Yiming Cui, Junyi Li, Jun Zeng, Rongzhao Wang, Weijian Xie, Yanting Li, Yina Patterson, Zuoyu Tian, Yiwen Zhang, He Zhou, Shaoweihua Liu, Zhe Zhao, Qipeng Zhao, Cong Yue, Xinrui Zhang, Zhengliang Yang, Kyle Richardson, and Zhenzhong Lan. 2020. CLUE: A Chinese language understanding evaluation benchmark. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4762–4772, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rongchao Yin, Quan Wang, Peng Li, Rui Li, and Bin Wang. 2016. Multi-granularity Chinese word embedding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 981–986, Austin, Texas. Association for Computational Linguistics.

Wen Zhang, Heng Wang, Kaijun Ren, and Junqiang Song. 2016. Chinese sentence based lexical similarity measure for artificial intelligence chatbot. In *2016 8th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–4.

A Appendix

ID	Category	Prompt
F-1	Zero-shot, word pairs in batches	请给每组词语的语义相似性按照从0到6的刻度打分，其中0代表语义完全不相似，6代表语义完全一致。所给分数仅保留整数。回答应该严格按照如下格式："[('词语1'，'词语2'，<分数>)，('词语3'，'词语4'，<分数>)，...]。" 请勿提供任何多余的解释或上下文。(Rate the semantic similarity of each word pair on a scale from 0 to 6, where 0 represents no semantic similarity, and 10 represents perfect semantic similarity. Use only integers. The response should strictly adhere to the structure: [('word1', 'word2', <score>), ('word3', 'word4', <score>), ...]. Do not provide additional explanations or context.)
F-3	Few-shot, word pairs in batches	请给每组词语的语义相似性按照从0到6的刻度打分，其中0代表语义完全不相似，6代表语义完全一致。所给分数仅保留整数。回答应该严格按照如下格式："[('词语1'，'词语2'，<分数>)，('词语3'，'词语4'，<分数>)，...]。" 请勿提供任何多余的解释或上下文。" "包含词语组和语义相似度分数的示例如下：[('长','狭窄', 1), ('内疚','羞愧', 6), ('暴力','被动', 0)]. ("Rate the semantic similarity of each word pair on a scale from 0 to 6, where 0 represents no semantic similarity, and 6 represents perfect semantic similarity. Use only integers. The response should strictly adhere to the structure: "[('word1', 'word2', <score>), ('word3', 'word4', <score>), ...]. " Do not provide additional explanations or context. Examples of word pairs and their semantic similarity scores are: [('长','狭窄', 1), ('内疚','羞愧', 6), ('暴力','被动', 0)].).
F-5	Zero-shot, categorical scale	请给每组词语的语义相似性按照如下等级分类："'非常不相似', '不相似', '中立'，'相似'，'非常相似'。"回答应该严格按照如下格式："[('词语1'，'词语2'，<分类>)，('词语3'，'词语4'，<分类>)，...]。"请勿提供任何多余的解释或上下文。(Classify the semantic similarity of each word pair in the hierarchical categories: 'very dissimilar', 'dissimilar', 'neutral', 'similar', and 'very similar'. The response should strictly adhere to the structure: [('word1', 'word2', <classification>), ('word3', 'word4', <classification>), ...]. Do not provide additional explanations or context.)
F-6	Zero-shot, cross-linguistic	Rate the semantic similarity of each English word pair on a scale from 0 to 6, where 0 represents no semantic similarity, and 6 represents perfect semantic similarity. Use two decimals. The response should strictly adhere to the structure: [('word1', 'word2', <score>), ('word3', 'word4', <score>), ...]. Do not provide additional explanations or context.
F-9	Zero-shot, single-word pair	请给该组词语的语义相似性按照从0到6的刻度打分：[('word1'), ('word2')]，其中0代表语义完全不相似，6代表语义完全一致。所给分数仅保留整数。回答应该严格按照如下格式："[('词语1'，'词语2'，<分数>)]。"请勿提供任何多余的解释或上下文。(Rate the semantic similarity of the word pair: [(<word1>), (<word2>)] on a scale from 0 to 6, where 0 represents no semantic similarity, and 6 represents perfect semantic similarity. Use two decimals. The response should strictly adhere to the structure: [('word1', 'word2', <score>)]. Do not provide additional explanations or context.)

Table 6: Description of prompt categories for Mandarin Chinese word pairs with English translations.