Prompting Instruction-Tuned LLMs for Semantic Similarity Values

Xander Snelder[♦], **Yunchong Huang**[♣], **Jelke Bloem**[♣♠]

♦ Data Science (Information Studies), University of Amsterdam
♣ Institute for Logic, Language and Computation, University of Amsterdam
♠ Data Science Centre, University of Amsterdam
j.bloem@uva.nl

Abstract

The impressive few-shot performance of generative decoder transformer language models at novel tasks has raised interest in using them to estimate lexical-semantic properties of words, word pairs or multi-word expressions. We explore the task of eliciting semantic similarity scores between word pairs through prompting, comparing these scores to human benchmarks. We investigate different prompting approaches, different model architectures and different languages using the Dutch, English and Mandarin Chinese SimLex-999 benchmarks. The results show that prompting each word pair individually yields better correlations, and that models struggle with the distinction between similarity and relatedness, just as static and contextual word embedding models did. The new, open-weight gpt-oss-20b model yields the highest correlation with human ratings out of the models we evaluated.

Keywords: semantic similarity, lexical semantics, embeddings, prompting, evaluation

1. Introduction

Various recent studies have advocated for the use of LLM-elicited lexical-semantic scoring or judgements for factors such as word concreteness (Martínez et al., 2024), word familiarity (Brysbaert et al., 2025), age of acquisition (Alzahrani et al., 2025), word association (Abramski et al., 2024), dispersion (Zhong et al., 2025), lexical aspect (Ma, 2024) lexical-semantic equivalence (Hayashi, 2025), perceptual strength (Lee et al., 2025), word iconicity, word similarity and others (Trott, 2024). Human ratings of lexical-semantic properties are costly to collect, and when in-context judgements or judgements over multi-word expressions (Martínez et al., 2025) are required, the combinatorial explosion of items to rate becomes difficult to manage without automation.

These types of ratings are often used in psycholinguistic and corpus linguistic research, but also in natural language processing. For example, Littlemore et al. (2018) used concreteness and valence ratings to evaluate whether these factors drive metaphor appreciation and understanding, which can benefit the task of automatic metaphor generation. So, it is worth investigating how to best elicit such ratings from autoregressive decoder LLMs (e.g. OpenAI, 2024a), whether such automated ratings are accurate and what types of mistakes the models tend to make. Trott (2024) propose that large-scale automated rating can beneficially augment psycholinguistic datasets.

The factor of word similarity is a particularly interesting one. Ever since the use of continuous vector representations of words in language models became widespread (Lenci, 2018), semantic similarity has been used as a benchmark of language model quality. The use of continuous vector representations is grounded in distributional theories of meaning, according to which "difference of meaning correlates with difference of distribution" (Harris, 1954, p. 156). These language models rely exclusively on distributional information to learn word representations. In this view, a model that has a higher correlation with word similarity ratings elicited from humans, has learned more accurate semantic representations. Benchmarks like WordSim-353 (Finkelstein et al., 2002) and SimLex-999 (Hill et al., 2015) were used for this purpose, becoming particularly popular when Word2Vec (Mikolov et al., 2013) and other static word embedding models were the state of the art. With more recent contextual embedding models, these benchmarks continue to be used to evaluate those models' lexical-semantic representations 'distilled' from contextual embeddings (Brans and Bloem, 2024) or for interpretability purposes (Rogers et al., 2020). The use of these benchmarks as a core intrinsic language model evaluation metric means that word pair similarity benchmarks are available for at least 27 languages (Brans and Bloem, 2024), and for many of them with parallel word pairs (Vulić et al., 2020).

Prompting for similarity scores differs from earlier benchmarking efforts as vector representations are not used directly, and models receive more context to solve the task. Indeed, better human correlations are achieved by prompting generative models: Trott (2024) finds that GPT-4's judgements have a correlation of 0.86 with SimLex-999, while static and contextual word embedding models typically achieve correlations of 0.4-0.6. De Deyne (2024) shows that GPT-4 also provides good word similar-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 11390–11403
11–16 May 2026.

ity judgements in a triad task, similar to comparative intrinsic evaluation tasks for word embedding evaluation (Schnabel et al., 2015). However, these studies do not investigate different prompting approaches, do not investigate error patterns, do not compare to modern embedding-based approaches, and they only pertain to English.

We explore this topic in more detail from a NLP perspective, comparing different model architectures, prompting techniques, and performing error analysis to evaluate weaknesses in the model's similarity judgements. We compare similarity scores elicited by prompting (from GPT-3.5) to similarity scores from contemporary embeddings (OpenAI's *text-embedding-3-large*, Neelakantan et al., 2022) to gain insight into the effectiveness of both approaches. We also compare the prompting of equivalent instruction-tuned and non-instruction-tuned models (Mistral-7b-v0.3 and Mistral-7b-v0.3-Instruct) and include a recent large open-weight model (gpt-oss-20b).

To move beyond English, we include Dutch and Mandarin Chinese in our study, comparing GPT-4o, Doubao-pro (Doubao Team, 2024) and Qwen2.5-7b-Instruct for Chinese. We confirm strong correlations with human ratings across models and languages, but nevertheless find that the generative models make similar mistakes as older word embedding models in mixing up the concepts of similarity and relatedness.[1]

2. Related Work

2.1. Lexical-semantic representation in language models

Static word embedding models such as Word2Vec (Mikolov et al., 2013) represent words as dense vectors in a continuous vector space, where similar words are positioned closely, and this reflects their semantic similarities (Bengio et al., 2003). Encoder transformer models such as BERT (Devlin et al., 2019) consist of stacked encoders that are bidirectionally trained on unlabeled data to encode language in context. As a side effect of the pre-training process, BERT can generate contextual word embeddings, which can be extracted from the model to evaluate the semantic relationships between word pairs. Neelakantan et al.'s (2022) *text-embedding-3* models have a transformer encoder architecture with a contrastive learning objective to differentiate between positive and negative pairs of text. This yields contextual word embeddings that are used for tasks like knowledge retrieval in ChatGPT (OpenAI, 2024c).

[1]Code and data available at https://github.com/XanderSnelder/semantic-similarity-prompting

Autoregressive transformer models consisting of stacked decoders, such as the GPT series of models, are pre-trained generatively and can be scaled to larger parameter sizes. GPT-3 is scaled to 175B parameters and 96 layers (Radford et al., 2019), applying a few-shot, one-shot, and zero-shot learning approach to tune for specific tasks (Brown et al., 2020). GPT-4 is fine-tuned using Reinforcement Learning from Human Feedback (RLHF) (OpenAI, 2023), while GPT-4 Omni (o) is an end-to-end multimodal model (OpenAI, 2024a). Explicit details about the architectures of GPT-3, GPT-4, and GPT-4o are not available (OpenAI, 2023), and the models are closed-source.

The recently released gpt-oss-20b (Agarwal et al., 2025) is a smaller open-weight model that has been pretrained on 'trillions of tokens' and post-trained with chain-of-thought reinforcement learning. In addition to OpenAI's GPTs, several state-of-the-art decoder LLM families have been developed, including the also closed-source Doubao-pro for Chinese (Doubao Team, 2024), Mistral models such as Mistral-7B-v0.3 and Mistral-7B-v0.3-Instruct (Jiang et al., 2023) and Qwen-2.5 models from Alibaba Cloud (Qwen et al., 2025).

Using word embedding models and encoder transformers, semantic similarity between word pairs is estimated by calculating the cosine similarity between word embeddings of the target words. While embeddings can be extracted from (open-weight) decoder LLMs as well, Arnold et al.'s (2024) experiment with a T5-based sequence-to-sequence model shows that embeddings from decoders do not yield strong correlations with English SimLex-999 ratings (up to $\rho = 0.179$ in a Mixture of Experts setup, while BERT achieves $\rho = 0.48$). Recently, Brans and Bloem (in press) showed that Dutch word embeddings extracted from the decoder models Falcon-7B, BLOOM-560m and Schaapje-2B (a Dutch model) correlate very weakly with human ratings (all correlations $\rho < 0.1$).

However, the instruction tuning and generalization capabilities of generative decoder models also facilitate the task of eliciting semantic similarity scores by prompting, even if the model has never been tuned explicitly for this task.

2.2. Semantic Similarity

Semantic similarity and association benchmarks are used to evaluate the capabilities of language models related to lexical semantics, specifically, to what extent they represent the meaning and relationships between words similarly to human judgements. In 2015, the English SimLex-999 benchmark (Hill et al., 2015) was introduced and became the gold standard for evaluating language models on their ability to estimate semantic similarities between word pairs. The authors differentiate be-

tween semantic similarity and association, emphasizing that these concepts are not mutually exclusive or independent.[2] The benchmark contains 999 word pairs rated on a scale from 0 to 10 for semantic similarity by approximately 50 native speaker annotators from the US.

Subsequently, many variations of this benchmark have been developed, including for verbs (Gerz et al., 2016), cross-lingual semantic similarity (Camacho-Collados et al., 2017; Barzegar et al., 2018), rare words (Pilehvar et al., 2018) and specific domains (Chiu et al., 2018). It has also been rerated and translated for a range of languages (see Brans and Bloem, 2024 for a recent overview), and MultiSimLex (Vulić et al., 2020) was developed for broader multilingual coverage of lexical-semantic ratings. While these benchmarks have been used extensively, these studies are clustered around contextual embeddings from encoder models (Brans and Bloem, 2024; Vulić et al., 2020) and on static embedding models (Leviant and Reichart, 2015; Chiu et al., 2018).

The difference between semantic association and similarity has shown to be a difficult distinction for language models to make in the application of these benchmarks, and even for humans to rate (Gladkova and Drozd, 2016). As autoregressive decoder models still store semantic information in dense vector spaces, this raises concerns about similar errors: do recent generative LLMs still make the same mistakes in providing semantic similarity ratings, or does the additional context and instruction tuning address this problem?

2.3. Prompting for lexical-semantic scores

Prompt engineering involves structuring clear and effective textual prompts to make a LLM generate the most accurate response in a downstream task. Models often benefit from being prompted with a few examples in a few-shot setting (Schick and Schütze, 2022), though they are quite sensitive to prompt specifics — alterations to prompts can lead to inconsistent performance in fact retrieval, question answering, and natural language inference (Leidinger et al., 2023; Ye et al., 2023; Zhou et al., 2022; Jiang et al., 2020; Zhao et al., 2021; Lu et al., 2022; Webson and Pavlick, 2022). Variability in performance using few-shot learning can be attributed to the prompt format, prompt training examples, and the order of training examples.

Leidinger et al.'s (2023) evaluation of 550 semantically equivalent prompts that varied in linguistic structure shows that the models are not affected by the frequency of synonyms, prompt length, ambiguity, or word frequency. High perplexity prompts often outperform simpler prompts, and the performance of prompts does not transfer to other LLMs or datasets (Lu et al., 2022; Gonen et al., 2023).

As for the LLM-elicited lexical-semantic scoring studies mentioned in the introduction, several prompting strategies were applied. Martínez et al. (2025) prompt for concreteness using a short prompt with three examples of words that should get the lowest rating and three examples that should get the highest rating, asking for a five-point scale. The prompt is repeated for each word. The authors state other strategies were explored but this is not discussed. Brysbaert et al. (2025) perform zero-shot prompting for familiarity, also repeating the prompt.

As for the aforementioned GPT-based semantic similarity studies, Trott's (2024) study only prompts with the human instructions. De Deyne (2024) uses a few-shot approach with step-by-step instructions, a 20-point rating scale, and separately presented prompts. We investigate these assumptions in our experiment, although our task is pairwise rather than triadic.

We extract word similarity scores from GPT models under 9 different prompting conditions. We also perform a cross-linguistic comparison of English and Dutch, a lesser-resourced language with possible interference from English, using the Dutch SimLex-999 benchmark (Brans and Bloem, 2024), including a cross-lingual prompting condition. Furthermore, we compare our prompt-based results to *text-embedding-3* (Neelakantan et al., 2022) cosine similarity scores, directly measuring similarity of representations used by OpenAI models. Lastly, we conduct a comparison of the state-of-the-art GPT-4o with a similarly performing Chinese model on Mandarin Chinese.

3. Method

3.1. Benchmarks

We use SimLex-999 (Hill et al., 2015), Dutch SimLex-999 (Brans and Bloem, 2024) and the Mandarin Chinese part of MultiSimLex (Vulić et al., 2020) as benchmarks of word similarity. These datasets have semantic similarity scores rated on a scale of 0 to 10 (SimLex) and 0 to 6 (MultiSimLex) by native speakers. The same (translated) instructions were used for all benchmarks, instructing annotators to distinguish synonymy (high simi-

[2] Semantic similarity refers to how much two words share common characteristics, which is sometimes more specifically called paradigmatic semantic similarity. Semantic association, also known as syntagmatic semantic similarity or semantic relatedness, indicates a broader relationship between words pairs, including how frequently two words interact or co-occur with each other (Barzegar et al., 2018; Vulić et al., 2020).

ID	Category
F-1	Zero-shot, default
F-2	One-shot, default
F-3	Few-shot, default
F-4	Zero-shot, small scale (0-5)
F-5	Zero-shot, categorical scale
F-6	Zero-shot, cross-linguistic
F-7	Few-shot, detailed questionnaire
F-8	Few-shot, conversational
F-9	Zero-shot, single-word pair

Table 1: List of evaluated prompt categories.

larity) from antonymy (low similarity, even though antonyms may be strongly associated with each other). The Dutch and English benchmarks contain 999 word pairs, and we randomly sampled 472 from the Chinese benchmark. English SimLex-999 pairs have been rated by 50 annotators, Dutch pairs by 15 and Mandarin pairs by 11 annotators.

3.2. Models

For English, we use several different types of model from OpenAI: *gpt-3.5-turbo-0125*, a probabilistic chat model based on InstructGPT and GPT-3.5 developed for conversational use (OpenAI, 2022, 2024b; Ouyang et al., 2024), *gpt-oss-20b*, a recent open-weight model tuned for instruction following (Agarwal et al., 2025), as well as *text-embedding-3-large* and *text-embedding-3-small*, deterministic models trained by OpenAI that are designed to evaluate semantic similarities and associations by generating word embeddings (large: 3072 dimensions, small: 1536 dimensions, OpenAI, 2024b,c; Neelakantan et al., 2022). Further important details, such as information on the training data size for these models or whether GPT-3.5 and text-embeddings-3 were trained on the same data, are not available, although they were developed around the same time. We also include Mistral-7b-v0.3 and Mistral-7b-v0.3-Instruct (Jiang et al., 2023) to quantify the effect of instruction tuning, and include Qwen2.5-7b (Qwen et al., 2025) for cross-linguistic comparison with Chinese in an open-weight model.

For the *text-embedding* models, we embed each word and compute the cosine similarities of the embeddings, as is done in semantic similarity benchmarking, without prompting or sampling.

For Chinese, we evaluate GPT-4o (OpenAI, 2024a, version of 2024-11-20), doubao-pro-128k (Doubao Team, 2024) and Qwen2.5-7b. All generative models were used with their default temperature setting.

> Rate the semantic similarity of the word pair: [$word_1$, $word_2$] on a scale from 0 to 10, where 0 represents no semantic similarity, and 10 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.

Figure 1: The F-9 prompt template (zero-shot, single word pair).

3.3. Prompting Experiments

We start with a prompt engineering experiment to explore optimal strategies for this task, evaluating on English and Dutch benchmarks. Table 1 lists nine distinct prompt categories, inspired by previous work on eliciting lexical-semantic ratings and prompting in general (Section 2.3). Prompts are based on the human instructions for the relevant (Multi)SimLex-999 questionnaires (Brans and Bloem, 2024; Hill et al., 2015; Vulić et al., 2020). Figure 1 shows the F-9 prompt template, which yielded the best results. The complete instructions for all English prompt templates are provided in Appendix A.

The F-1, F-2, and F-3 prompting conditions are relatively simple baseline prompts, lacking the explanation on synonymy, antonymy and relatedness from the human instructions, but with increasing numbers of examples. Few-shot learning is often considered beneficial in prompting, but in a context-less semantic task, it may bias the model towards specific word senses related to the few-shot examples through semantic priming (cf. Jumelet et al., 2024). Both zero-shot (Brysbaert et al., 2025) and few-shot (Martínez et al., 2025) prompting has been used in lexical-semantic elicitation tasks.

The other prompts are based on F-1. The F-4 and F-5 prompts include alternative scales, which are normalized after extraction. Prompt F-4 uses a scale of 0 to 5 as used by Martínez et al. (2025), rather than 0 to 10, while F-5 requests five categorial values. This was inspired by work in psychology, where it has been argued that verbal (categorial) rating scales are more natural (Krosnick and Fabrigar, 1997). Verbal rating scales were shown to have higher reliability and require fewer eye fixations to process (Menold, 2020). Instruction-tuned language models may have acquired some of these human preferences. We use labels from 'very dissimilar' to 'very similar' that are mapped to [0, 2.5, 5, 7.5, and 10] for computing correlations.

In the cross-linguistic F-6 prompt, the instructions are phrased in English for Dutch SimLex-999, and vice versa for English SimLex-999. The F-7 prompt has the highest token count, including de-

tailed instructions similar to the human SimLex-999 questionnaires as done by De Deyne (2024), testing the idea that higher-perplexity prompts might work better (Leidinger et al., 2023). The F-8 prompt uses a conversational approach that separates the instructions from the word pairs in a chat thread. Lastly, the F-9 prompt includes a single word pair, whereas the other prompts include batches of word pairs for efficiency. In a contextless judgement task, avoiding distraction from other word pairs may be beneficial as also done by Trott (2024), but this contrasts with the idea that examples may be beneficial, tested in conditions F1-F3. Each prompt includes the structured format of the desired response.

In subsequent experiment comparisons, we use the more effective prompting strategies from this experiment.

3.4. Evaluation

Similar to prior studies using word embedding models, Spearman's rank correlation is calculated between the estimated semantic similarity scores and observed human similarity scores. To account for potential consistency issues of the generative models (Brown et al., 2020; Ouyang et al., 2024), each prompt is executed 20 times (Experiment 1) or 15 times (Experiment 2). In the evaluation, these are treated as samples in the same way that the different ratings by different human annotators are.

4. Experiment 1: Prompt engineering, Dutch-English

Results are obtained through API calls to the OpenAI API. All word pairs are grouped into batches to reduce computational and financial costs, except for the F-9 prompt, which processes each word pair individually. For the F-2 and F-3 prompts that have examples, word pairs and their semantic similarity scores are extracted from the SimLex-999 dataset and embedded in the prompt. To prevent data leakage, these word pairs are excluded from the datasets used for these prompts. Regular expressions are used to extract the word pairs and their similarity scores from GPT's output, and this process is manually checked for inconsistencies and corrected.

Since the F-4 and F-5 prompts generate similarity scores on a different scale, normalization is applied to align these scores with the SimLex-999 benchmarks. For *text-embedding-3-large* and *text-embedding-3-small*, which are deterministic models that directly compute word embeddings, the cosine similarity is calculated between the word embeddings. Subsequently, the cosine similarity is scaled to a range of 0 to 10.

Model	Prompt	ρ-NL	ρ-EN
`gpt-3.5-turbo-0125`	F-1	0.60	0.53
`gpt-3.5-turbo-0125`	F-2	0.64	0.57
`gpt-3.5-turbo-0125`	F-3	0.68	0.67
`gpt-3.5-turbo-0125`	F-4	0.60	0.51
`gpt-3.5-turbo-0125`	F-5	0.68	0.65
`gpt-3.5-turbo-0125`	F-6	0.54	0.56
`gpt-3.5-turbo-0125`	F-7	0.56	0.57
`gpt-3.5-turbo-0125`	F-8	0.71	0.57
`gpt-3.5-turbo-0125`	F-9	**0.72**	**0.82**
`text-embedding-3-large`	—	0.41	0.57
`text-embedding-3-small`	—	0.34	0.50

Table 2: Correlations (ρ) between model-estimated and human-observed similarity scores for Dutch (NL) and English (EN) word pairs.

4.1. Results

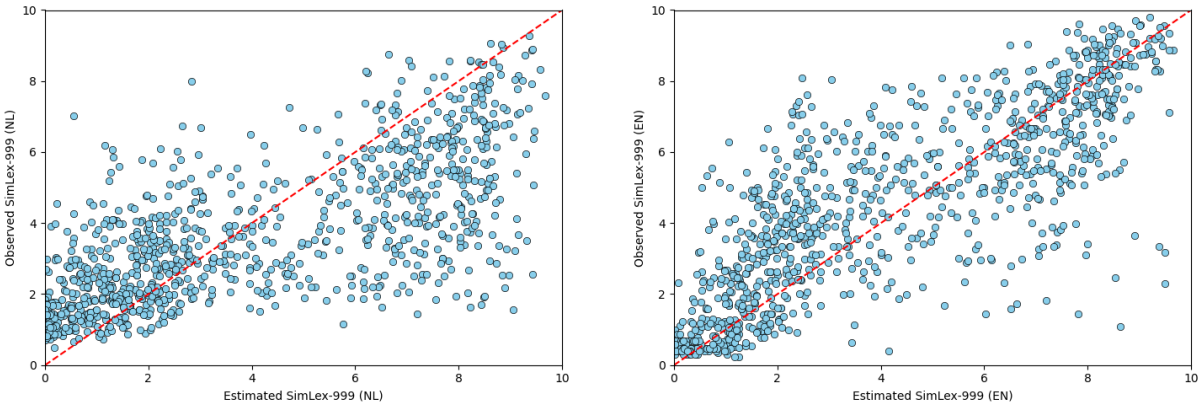
Table 2 shows that the F-9 prompt (zero-shot, single word pair per prompt) yields the highest correlations with human ratings across both benchmarks and prompt categories. These correlations are visualized in Figure 2. The F-9 prompt for English SimLex-999 has the highest correlation of 0.82 and the lowest mean standard deviation of 0.96, which are 0.10 higher and 0.06 lower compared to Dutch SimLex-999, respectively. Although the F-1 and F-9 prompts have similar instructions, the difference in their correlations is large. This shows that prompting in batches is detrimental, perhaps due to irrelevant information in the context.

The differences for the Dutch experiment are closer, with the F-9 prompt being 0.01 higher than the F-8 prompt (conversational). The lowest correlations are found in the F-6 (0.54, cross-linguistic) and F-7 (0.56, detailed instructions) prompts for Dutch SimLex-999, and in the F-1 (0.53, zero-shot) and F-4 (0.51, 0-5 scale) prompts for English SimLex-999.

When comparing the prompts between the two languages, there is no consistent difference across prompts. The largest difference between the languages of 0.14 is found in the F-8 (conversational) prompt. Perhaps this is due to more limited conversational instruction tuning for Dutch.

The model does not appear to suffer from semantic priming by the few-shot examples — the correlation consistently increases for both benchmarks when more word pairs and their respective similarity scores are added to the prompts, as illustrated by the F-1 (zero-shot), F-2 (one-shot), and F-3 (few-shot) prompts.

Providing more detailed instructions to the model, similar to the instructions that SimLex-999 annotators got, did not improve performance. We did not observe any benefit from prompting for a smaller scale (0-5 ratings), but the use of a categorial scale



(a) Dutch (NL) word pairs

(b) English (EN) word pairs

Figure 2: Scatter plots of model-estimated vs. human-observed similarity scores using the F-9 prompt template.

('very dissimilar', 'dissimilar', 'neutral' etc.) did lead to better human correlations for both languages. Verbally labeled categories are more natural to human participants in psychological studies (Menold, 2020), and this preference may have been conveyed through instruction tuning.

It appears that there is an overall preference for simple, low-context prompting conditions such as F-9 and F-8, though few-shot examples do help. This contrasts with findings that high-perplexity prompts are better (Leidinger et al., 2023). Interestingly, we did not observe a clear effect of cross-linguistic prompting. Recent studies have observed that models often perform better on under-resourced languages when prompted in English (Liu et al., 2025). Perhaps Dutch and English are too closely related to observe this effect here.

4.1.1. Embedding Models

Table 2 shows that the *text-embedding-3* models have clearly lower correlations than the best prompted scores. The numbers are comparable to those obtained in word embedding model benchmarking, such as for Dutch BERTje (0.42, Brans and Bloem, 2024), English BERT (0.48, Ehrmanntraut et al., 2021) and English fastText (0.55, Vulić et al., 2020).

5. Experiment 2: Model comparison

Next, we compare different types of English LLMs. For the instruction-tuned models, the F-9 prompt is used, and we prompted for JSON response formats. For the non-instruction-tuned Mistral-7b-v0.3, we use a simple sentence completion prompt: "*On a scale from 0 to 10, the semantic similarity of the word pair: word1 - word2 is* " and extracted the first generated number between 0 and 10. Besides the listed models, we also tried Llama-3.3-70B, Llama-4-Scout and Mixtral-8x7B-Instruct, but got high rates of invalid responses and did not include the results. Table 3 shows our results along with some SimLex-999 correlations from previous work.

Model	ρ
`Word2Vec` (Leviant and Reichart, 2015)	0.27
`BERT` (Ehrmanntraut et al., 2021)	0.48
`fastText` (Vulić et al., 2020)	0.55
`GPT-4` (Trott, 2024)	0.86
`gpt-3.5turbo`	0.821
`text-emb-3-l`	0.566
`Mistral-7b-v0.3`	0.581
`Mistral-7b-v0.3-Instruct`	0.829
`Qwen2.5-7b-Instruct`	0.834
`gpt-oss-20b`	**0.902**

Table 3: Correlations (ρ) between model-estimated and human-observed similarity scores for English pairs.

5.1. Error analysis

We perform an error analysis of the best-performing and newest model, gpt-oss-20b. Table 4 shows test items with the largest human-model differences for the English F-9 prompt. We see many instances of words that are related but not synonymous, which the model rates too highly. There are mostly co-hyponyms (husband-wife, dog-cat, bee-ant, woman-man, decade-century). A few pairs are antonyms (send-receive, sunset-sunrise) which receive and should receive low scores from humans.

However, the model is not incapable of rating antonyms — among the best predictions at the bottom of Table 4, we see the antonym pair necessary-unnecessary (0.63), as well as the co-hyponym pair man-child (4.13) and the similar word pair buddy-

R	Word 1	Word 2	SimL	F-9
1	wife	husband	2.3	8.38
2	dog	cat	1.75	7.07
3	bee	ant	2.78	7.32
4	woman	man	3.33	7.80
5	sunset	sunrise	2.47	6.82
6	decade	century	3.48	7.80
7	send	receive	1.08	5.32
...	...	...	...	
997	unnecessary	necessary	0.63	0.63
998	man	child	4.13	4.13
999	buddy	companion	8.65	8.65

Table 4: Highest and lowest absolute differences to human gold standard for *gpt-oss-20b* with the F-9 prompt, on a scale of 0 to 10.

Model	ρ
`fastText` (Vulić et al., 2020)	0.583
`M-BERT` (Vulić et al., 2020)	0.487
`Chinese BERT` (Vulić et al., 2020)	0.59
`GPT-4o`	**0.865**
`Doubao-pro`	0.841
`Qwen2.5-7b-Instruct`	0.794

Table 5: Correlations (ρ) between model-estimated and human-observed similarity scores for Mandarin word pairs.

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

companion (8.65).

Overall, more of the top scoring pairs are highly similar words, and more of the worst scoring pairs are antonyms and co-hyponyms. Distinguishing between relatedness and similarity is still a struggle, which was an issue that Word2Vec also had and a main motivation for developing the SimLex-999 benchmark originally (Hill et al., 2015).

6. Experiment 3: Mandarin Chinese

Experiment 1 prompted us to perform a follow-up experiment for a typologically unrelated language. We use the more recent GPT-4o model, shown to perform well on this task in English by Trott (2024). In the Chinese AI community, the Doubao LLM has rapidly gained popularity since its initial release in May 2024. *Doubao-pro-1215* has achieved comprehensive alignment with GPT-4o on several benchmarks (Doubao Team, 2024).

Similar to the latest GPT models, Doubao LLMs are also closed-source and their detailed structures remain unclear. Results are obtained through API calls to the OpenAI API and Doubao API provided by Volcano Engine. We experimented with the F-1 (zero-shot), F-3 (few-shot), F-5 (categorial scale), F-6 (cross-lingual) and F-9 (unbatched) prompt, but found near-identical performance with these prompts using GPT-4o. Doubao performs best with the F-9 prompt and otherwise showed different prompt-specific results than GPT-3.5, showing that prompt engineering findings are model-specific also for this task. Therefore, we again focus on results with the F-9 prompt, which are shown in Table 5. We see that GPT-4o slightly outperforms Doubao-pro and performs the same as in English (Table 3). Qwen2.5-7b performs slightly worse than on the English benchmark.

6.1. Error analysis

Table 6 shows that many differences between human ratings and GPT-4o predictions involve confusion between cognitive relatedness/association and semantic similarity. This is similar to the English gpt-oss-20b results. Retrieved word pairs potentially related to this include [荣誉(honor), 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 distinguishing relatedness/association and semantic sim-

ilarity is an important factor in the Multi-SimLex dataset curation, the low mean human annotations scores (all below 1) for these word pairs reveal the success of these annotation instructions, while the much higher prediction scores from GPT-4o may reveal that the model does not clearly make this distinction either.

Another potential reason 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.

We do not have space for a full error analysis of Doubao-pro, but it's worth mentioning that 3 top errors from GPT-4o are also top errors of Doubao-pro: [收到(receive), 接受(accept)], [蜜蜂(bee), 蚂蚁(ant)] and [装作(pretend), 似乎(seem)], and two of these pairs were also top errors in English for gpt-oss-20b. All three LLMs seem to overestimate their semantic similarities across two typologically distinct languages.

For the Mandarin word pairs with the highest absolute human-model difference retrieved from Qwen2.5-7B-Instruct using the F-9 prompt, the mixture of cognitive relatedness/association and semantic similarity is again obvious. Word pairs such as [蝙蝠(bat), 吸血鬼(vampire)] (diff=3.78), [蜥蜴(lizard), 鳄鱼(crocodile)] (diff=3.75), [尖顶(spire), 教堂(church)] (diff=3.74), [床(bed), 毯子(blanket)] (diff=3.31), [骨头(bone), 牙齿(teeth)] (diff=3.23) and [男人(man), 武士(warrior)] (diff=3.05) are quite typical in this regard. The pairs of [衣服(clothes), 布料(fabric)] (diff=3.55), [生日(birthday), 日期(date)] (diff=3.55) are more nuanced, with the former one bearing a material/constitutive relation, while the latter can be attributed to hyponymy. Interestingly, the pairs of [收到(receive), 接受(accept)] and [装作(pretend), 似乎(seem)] are present again as examples of the model's over-generalisation of semantic similarities under certain contexts. These two pairs signal that this over-generalisation is rather universal across different LLMs.

R	Word 1	Word 2	SimL	Qwen
1	收到 *receive*	接受 *accept*	0.91	5.47
2	蝙蝠 *bat*	吸血鬼 *vampire*	0.55	4.33
3	蜥蜴 *lizard*	鳄鱼 *crocodile*	0.45	4.20
4	尖顶 *spire*	教堂 *church*	0.73	4.47
5	衣服 *clothes*	布料 *fabric*	1.45	5
6	生日 *birthday*	日期 *date*	1.45	5
7	床 *bed*	毯子 *blanket*	0.09	3.4
8	装作 *pretend*	似乎 *seem*	0.27	3.53
9	骨头 *bone*	牙齿 *teeth*	0.64	3.87
10	男人 *man*	武士 *warrior*	0.55	3.60

Table 7: The words with the highest absolute difference in human similarity score and model predicted similarity for Qwen2.5-7B-Instruct on Mandarin Chinese using the F-9 prompt, on a scale of 0 to 6.

7. Discussion

The comparison between gpt-3.5-turbo and text-embedding-3 shows that similarity scores reported through prompting better correlate with human judgements than cosine similarity scores of internal embeddings for two contemporary models. The comparison between Mistral-7b-v0.3-Instruct and Mistral-7b-v0.3 shows that instruction tuning is an important contributing factor to the high correlations with human semantic similarity ratings that we observe for prompted similarity scores. The strongest correlation with the human benchmark was seen with the open gpt-oss-20b model for English. This model outperformed the previous state-of-the-art for English, GPT-4o, as elicited by the experiment of Trott (2024).

However, despite very good human correlations, even the prompted scores exhibit issues in distinguishing the concepts of similarity and relatedness. While the RLHF tuning performed on these generative decoder LLMs may have enhanced their ability to produce semantic similarity scores compared to direct embedding extraction, the model still produces judgements that are different from human judgements and in line with typical representational errors of word embedding models. This raises concerns about using model-generated semantic similarity scores in place of human scores in psycholinguistic studies. Such scores should be checked in particular for errors that may arise from

assumptions of the distributional hypothesis.

7.1. Comparison with Previous Studies

While the different methods of obtaining semantic similarity scores vary between model architectures, we observed that prompting for scores is more effective than extraction across languages (Table 3 for English, Table 5 for Mandarin).

While not many other error analyses are found in the literature, Brans and Bloem (2024) observe that Dutch BERT-based models exhibit the largest differences with human ratings on pairs that exhibit antonym relations. They also observe more errors for pairs with longer and lower-frequency words, which we did not observe. They also observe issues with ambiguous words, which we observed for GPT-4o in the Mandarin error analysis. This suggests similar underlying issue for different architectures, though large-scale error analysis would be required for a comprehensive picture. Chronis and Erk (2020) also performed layer-wise analysis in their BERT-based approach, noting that the final layer optimizes relatedness, and layer 7 (a middle layer) optimizes similarity. If generative decoder models have similar encoding patterns, that would indeed yield the type of relatedness errors we observed (over-rating antonymy and co-hyponymy).

8. Conclusion

Overall, we have shown that while semantic similarity ratings elicited by prompting generative decoder LLMs strongly correlate with human ratings, deviations from human ratings do occur and follow the same patterns that are found in evaluations of earlier language modelling architectures. Models are likely to assign overly high scores to antonyms, presumably still mixing up the concepts of similarity and association despite far larger-scale training and tuning than older models.

We also observed that such scores are best elicited by prompting for them one-by-one in a few-shot setting, and that it does not seem to help to add elaborate instructions that explain synonymy and antonymy. However, these prompting findings are model dependent.

8.1. Future Work

In future work, it would be worth investigating typical error patterns for other lexical-semantic rating tasks as well, especially if model-elicited ratings are intended to be used in linguistic research. Follow-up research could also aim to investigate the source of these errors in more detail, for example by layer-wise embedding analysis of an open-source model, and by correlating embedding scores and prompted scores from the same open-source model.

This line of research would also benefit from more advanced text mining techniques to reliably extract the semantic similarities from model responses. The inclusion of a strict output format in our prompt is less than ideal as it does not resemble human conversations present in instruction tuning, but it was necessary to be able to analyse the result, and even then the outputs for some models were too inconsistent to process.

The SimLex-999 and MultiSimLex series of benchmarks provide human data for a wider range of languages than we have investigated, suggesting further rating elicitation possibilities for non-English languages. However, there is also a risk of English interference with ratings in other languages due to the fact that this language dominates the training data. Further work could examine whether multilingual LLMs directly compute semantic similarities between word pairs in under-resourced languages, or if the similarities exhibit English interference. This could be addressed by testing word pairs where semantic interference from English is likely, e.g. due to lexical gaps.

9. Limitations

There are many important limitations of the general idea of prompting LLMs for lexical-semantic ratings, potentially substituting human ratings, that our experiment did not cover but that are extensively discussed in other literature, such as Anglocentric bias (Atari et al., 2023), test set leakage (Trott, 2024) and uneven representation of language varieties (Grieve et al., 2025).

There are also well-established limitations of semantic similarity benchmarks more broadly. Scalar semantic similarity ratings probably do not capture all aspects of the semantic network in human cognition, limiting their use as a benchmark for semantic embedding spaces in language models. Human judgements can be affected by confounds of the experimental setup and the instructions (Gladkova and Drozd, 2016).

In a prompting setup, models might replicate these confounds (or be affected by training data contamination) rather than reflecting a learned semantic space. Furthermore, high correlation on these benchmarks does not always correlate with high performance on downstream tasks (Gladkova and Drozd, 2016).

9.1. Sample Size and Models

The OpenAI API charges various rates per token depending on the model, with GPT-4 and GPT-4o being more expensive than GPT-3.5 (OpenAI, 2024d). Due to resource constraints, the sample sizes were limited to 20 and 15 samples per prompt.

Relatedly, this study evaluates a limited range of models. Due to resource constraints, we were not able to evaluate the gpt-oss-120b model. As this is the larger counterpart of our best-performing model, it is likely to be the state-of-the-art.

9.2. Explainability of LLMs

Another limitation is the closed-source design and probabilistic nature of some of the models we investigated, including the best-performing model for Mandarin. While gpt-oss-20b is open-weight, gpt-3.5-turbo, GPT-4o and subsequent models used in the ChatGPT service are closed-source. Furthermore, we cannot be sure whether *gpt-3.5-turbo-0125* makes use of the *text-embedding-3-large* embeddings or has been trained on the same data for detailed architectural comparison, although both are related OpenAI products.

More broadly, it remains unclear whether the elicited similarity scores represent the internal representations of decoder LLMs, or generate plausible results that happen to correlate well with human ratings. It is also not clear whether multilingual LLMs compute the semantic similarities directly in a lower-resourced language like Dutch, or first internally translate the word pairs into English. This lack of transparency makes it challenging to understand the mechanisms that yield these scores in modern decoder LLMs.

This ambiguity relates to whether decoder LLMs utilize patterns learned during training, or actually gain an "understanding" of NLP tasks during the inference itself (Brown et al., 2020). "*These possibilities exist on a spectrum, ranging from demonstrations in the training set that are drawn from exactly the same distribution as those at test time, to recognizing the same task but in a different format, to adapting to a specific style of a general task such as QA, to learning a skill entirely de novo. Where GPT-3 is on this spectrum may also vary from task to task.*" (Brown et al., 2020, p. 34). We also do not know whether any explicit instruction on synonymy or semantic similarity was given during RLHF tuning.

Lastly, previous research from OpenAI concludes that data contamination in the training data of GPT-3.5 and GPT-4 has minimal effects on their performance (OpenAI, 2023). It is unclear whether the SimLex-999 benchmark is included in the training data of any of these large decoder models, but it is quite likely given its popularity. A further question is whether the different language variants are also included, and to what extent such data leakage affects the extracted similarity scores and whether it does so across languages. Follow-up experiments with newly rated items could address this, as well as dataset contamination tests.

9.3. Explicit Prompt Instructions

Because decoder LLMs are optimized for unconstrained generation, prompting for semantic similarities between word pairs results in responses with inconsistent formats. To reduce these inconsistencies, we had to include explicit instructions regarding the desired format of the response. Although this approach simplifies data processing, it diminishes the human-like aspects of the prompts and their similarities to SimLex-999 questionnaires. In future studies, the specific prompt instructions can be omitted to better represent human-like aspects, requiring more advanced data processing techniques to extract the similarity scores from LLM responses.

Although discussed for GPT-3 (Brown et al., 2020), the technical reports of GPT-4, GPT-4o and Doubao-pro lack details about their computational costs and ecological footprint. Besides the energy consumed during model training, the general usage of the model also consumes energy. This is a valid concern, particularly when assuming the sizes of these models continue to increase.

10. Bibliographical References

Katherine Abramski, Clara Lavorati, Giulio Rossetti, and Massimo Stella. 2024. LLM-generated word association norms. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, pages 3–12. IOS Press.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, et al. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Alaa Alzahrani, Wafa Aljuaythin, Hassan Alshumrani, Alaa Mamoun Saleh, and Mohamed M Mostafa. 2025. Kalimah norms: Ratings for 2,467 Modern Standard Arabic words on two scales. *Behavior Research Methods*, 57(7):1–20.

Stefan Arnold, Marian Fietta, and Dilara Yesilbas. 2024. Routing in sparsely-gated language models responds to context. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 15–22.

Mohammad Atari, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023. Which humans?

Siamak Barzegar, Brian Davis, Manel Zarrouk, Siegfried Handschuh, and Andre Freitas. 2018.

SemR-11: A multi-lingual gold-standard for semantic similarity and relatedness for eleven languages. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *Advances in Neural Information Processing Systems*, 3:1137–1155.

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.

Lizzy Brans and Jelke Bloem. in press. Multi-SimLex for Dutch: Benchmarking embedding- and prompt-based model performance on semantic similarity. In *Proceedings of the Fifteenth International Conference on Language Resources and Evaluation (LREC 2026)*.

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.

Marc Brysbaert, Gonzalo Martínez, and Pedro Reviriego. 2025. Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. *Behavior Research Methods*, 57(1):1–15.

Jose Camacho-Collados, Mohammad Taher Pilehvar, Nigel Collier, and Roberto Navigli. 2017. SemEval-2017 task 2: Multilingual and cross-lingual semantic word similarity. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, page 15–26.

Billy Chiu, Sampo Pyysalo, Ivan Vulić, and Anna Korhonen. 2018. Bio-SimVerb and Bio-SimLex: wide-coverage evaluation sets of word similarity in biomedicine. *BMC Bioinformatics*, 19(1).

Gabriella Chronis and Katrin Erk. 2020. When is a bishop not like a rook? when it's like a rabbi! multi-prototype BERT embeddings for estimating semantic relationships. *Proceedings of the 24th Conference on Computational Natural Language Learning*, page 227–244.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:4171–4186.

Doubao Team. 2024. 8 key moments of Doubao Big Model in 2024. Accessed: 2025-01-15.

Anton Ehrmanntraut, Thora Hagen, Leonard Konle, and Fotis Jannidis. 2021. Type- and token-based word embeddings in the digital humanities. *Workshop on Computational Humanities Research*.

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.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*.

Anna Gladkova and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 36–42.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. *Findings of the Association for Computational Linguistics: EMNLP 2023*, page 10136–10148.

Jack Grieve, Sara Bartl, Matteo Fuoli, Jason Grafmiller, Weihang Huang, Alejandro Jawerbaum, Akira Murakami, Marcus Perlman, Dana Roemling, and Bodo Winter. 2025. The sociolinguistic foundations of language modeling. *Frontiers in Artificial Intelligence*, 7:1472411.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Yoshihiko Hayashi. 2025. Evaluating LLMs' capability to identify lexical semantic equivalence: Probing with the word-in-context task. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6985–6998,

Abu Dhabi, UAE. Association for Computational Linguistics.

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

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

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.

Jaap Jumelet, Willem Zuidema, and Arabella Sinclair. 2024. Do language models exhibit human-like structural priming effects? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14727–14742, Bangkok, Thailand. 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.

Jonghyun Lee, Dojun Park, Jiwoo Lee, Hoekeon Choi, and Sung-Eun Lee. 2025. Exploring multimodal perception in large language models through perceptual strength ratings. *IEEE Access*, 13:176751–176769.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232.

Alessandro Lenci. 2018. Distributional models of word meaning. *Annual review of Linguistics*, 4(1):151–171.

Ira Leviant and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling.

Jeannette Littlemore, Paula Pérez Sobrino, David Houghton, Jinfang Shi, and Bodo Winter. 2018. What makes a good metaphor? A cross-cultural study of computer-generated metaphor appreciation. *Metaphor and Symbol*, 33(2):101–122.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? A study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bolei Ma. 2024. Evaluating lexical aspect with large language models. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 123–131.

Gonzalo Martínez, Javier Conde, Pedro Reviriego, and Marc Brysbaert. 2024. AI-generated estimates of familiarity, concreteness, valence, and arousal for over 100,000 Spanish words. *Quarterly Journal of Experimental Psychology*, page 17470218241306694.

Gonzalo Martínez, Juan Diego Molero, Sandra González, Javier Conde, Marc Brysbaert, and Pedro Reviriego. 2025. Using large language models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, 57(1):1–11.

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, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of ICLR (Workshop Track)*.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, Johannes Heidecke, Pranav Shyam, Boris Power, Tyna Eloundou Nekoul, Girish Sastry, Gretchen Krueger, David Schnurr, Felipe Petroski Such, Kenny Hsu, Madeleine Thompson, Tabarak Khan, Toki Sherbakov, Joanne Jang, Peter Welinder, and Lilian Weng. 2022. Text and code embeddings by contrastive pre-training.

OpenAI. 2022. Introducing ChatGPT. Retrieved April 8, 2024 from https://openai.com/blog/chatgpt#OpenAI.

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

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

OpenAI. 2024b. Models — OpenAI API. Retrieved April 8, 2024 from https://platform.openai.com/docs/models.

OpenAI. 2024c. New embedding models and API updates. Retrieved April 8, 2024 from https://openai.com/blog/new-embedding-models-and-api-updates.

OpenAI. 2024d. Pricing. Retrieved April 11, 2024 from https://openai.com/pricing.

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.

Mohammad Taher Pilehvar, Dimitri Kartsaklis, Victor Prokhorov, and Nigel Collier. 2018. Card-660: Cambridge Rare Word Dataset - a reliable benchmark for infrequent word representation models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 1391–1401.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Timo Schick and Hinrich Schütze. 2022. True few-shot learning with Prompts—A real-world perspective. *Transactions of the Association for Computational Linguistics*, 10:716–731.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

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

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

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.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344.

Qinyuan Ye, Maxamed Axmed, Reid Pryzant, and Fereshte Khani. 2023. Prompt engineering a prompt engineer.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models.

Yanlu Zhong, Simon Todd, Nicole Xu, and Laurel Brehm. 2025. Evaluating LLMs as proxies for humans in psycholinguistic ratings: A comparison of statistical knowledge. *Research Methods in Applied Linguistics*, 4(3):100274.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2022. Large language models are human-level prompt engineers.

A. Prompt templates (English)

ID	Category	Prompt
F-1	Zero-shot, default	Rate the semantic similarity of each word pair on a scale from 0 to 10, where 0 represents no semantic similarity, and 10 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-2	One-shot, default	Rate the semantic similarity of each word pair on a scale from 0 to 10, where 0 represents no semantic similarity and 10 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. — An example of a word pair and its semantic similarity score is: [('old', 'new', 1.58)].
F-3	Few-shot, default	Rate the semantic similarity of each word pair on a scale from 0 to 10, where 0 represents no semantic similarity, and 10 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. — Examples of word pairs and their semantic similarity scores are: [('old', 'new', 1.58), ('smart', 'intelligent', 9.20), ('hard', 'difficult', 8.77)].
F-4	Zero-shot, small scale (0-5)	Rate the semantic similarity of each word pair on a scale from 0 to 5, where 0 represents no semantic similarity, and 5 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-5	Zero-shot, categorical scale	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 10, where 0 represents no semantic similarity, and 10 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-7	Few-shot, detailed questionnaire	Two words are synonyms if they have very similar meanings. Synonyms represent the same type or category of thing. Here are some examples of synonym pairs: cup/mug, glasses/spectacles, envy/jealousy. In practice, word pairs that are not exactly synonymous may still be very similar. Here are some very similar pairs - we could say they are nearly synonyms: alligator/crocodile, love/affection, frog/toad. In contrast, although the following word pairs are related, they are not very similar. The words represent entirely different types of things: car/tyre, car/motorway, car/crash. Rate the semantic similarity of each word pair on a scale from 0 to 10, where 0 represents no semantic similarity, and 10 represents perfect semantic similarity. Remember, things that are related are not necessarily similar. If you are ever unsure, think back to the examples of synonymous pairs (glasses/spectacles), and consider how close the words are (or are not) to being synonymous. There is no right answer to these questions. It is perfectly reasonable to use your intuition or gut feeling as a native English speaker, especially when you are asked to rate word pairs that you think are not similar at all. The response should strictly adhere to the structure: [('word1', 'word2', <score>), ('word3', 'word4', <score>), ...]. Use two decimals. Do not provide additional explanations or context.
F-8	Few-shot, conversational	• system_content = "Rate the semantic similarity of each word pair on a scale from 0 to 10, where 0 represents no semantic similarity, and 10 represents perfect semantic similarity." • user_content = "[('old', 'new'), ('smart', 'intelligent'), ('hard', 'difficult)']" • assistant_content = "[('old', 'new', 1.58), ('smart', 'intelligent', 9.20), ('hard', 'difficult', 8.77)]"
F-9	Zero-shot, single-word pair	Rate the semantic similarity of the word pair: [(<word1>), (<word2>)] on a scale from 0 to 10, where 0 represents no semantic similarity, and 10 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 8: Description of prompt categories for English word pairs.