Which Model Mimics Human Mental Lexicon Better? A Comparative Study of Word Embedding and Generative Models

Huacheng Song **Zhaoxin Feng** **Emmanuele Chersoni** **Chu-Ren Huang**
Department of Language Science and Technology, The Hong Kong Polytechnic University
{huacheng.song, zhaoxinbetty.feng}@connect.polyu.hk
{emmanuele.chersoni, churen.huang}@polyu.edu.hk

Abstract

Word associations are commonly applied in psycholinguistics to investigate the nature and structure of the human mental lexicon, and at the same time an important data source for measuring the alignment of language models with human semantic representations.

Taking this view, we compare the capacities of different language models to model collective human association norms via five word association tasks (WATs), with predictions about associations driven by either word vector similarities for traditional embedding models or prompting large language models (LLMs).

Our results demonstrate that neither approach could produce human-like performances in all five WATs. Hence, none of them can successfully model the human mental lexicon yet. Our detailed analysis shows that static word-type embeddings and prompted LLMs have overall better alignment with human norms compared to word-token embeddings from pre-trained models like BERT. Further analysis suggests that the performance discrepancies may be due to different model architectures, especially in terms of approximating human-like associative reasoning through either semantic similarity or relatedness evaluation[1].

1 Introduction

Artificial intelligence, particularly large language models (LLMs), functionally emulates the way we humans perceive and conceptualize the physical reality, as well as how we understand and process multifaceted information (Löhn et al., 2024). Yet a pivotal open question remains unsolved: to what extent do LLMs align with the conceptual knowledge hierarchically encoded in human cognition as their capabilities advance? This is where the "machine psychology" comes into play to scrutinize LLMs' "behavioral traits" and "thinking patterns" through

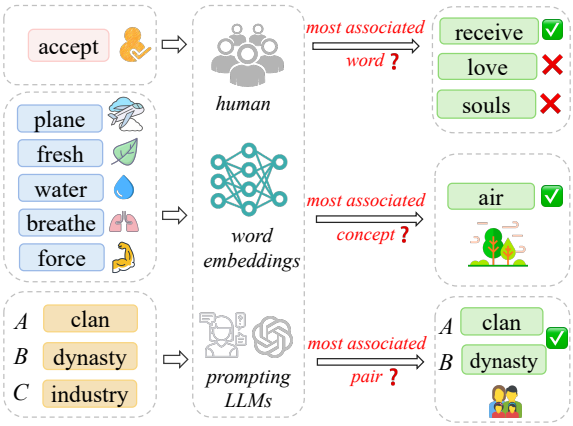

Figure 1: Illustration of Common Word Association Tasks. These tasks evaluate semantic alignment between computational models (word embeddings vs. LLM prompting) and human-like associative reasoning.

psychological tests adapted from interpretable research on human (Hagendorff, 2023).

Successful modeling of the human mental lexicon can be viewed as an essential step in verifying human-like intelligence. Human mental lexicon, in contrast to electronic lexica, is extremely versatile in supporting the association and generation of new concepts. Indeed *word association norms* is a typical method of investigation: a stimulus word is presented to a human participant, who is simply required to produce the first word coming to mind (McRae et al., 2012; De Deyne et al., 2019). Semantic similarities and relatedness that underlie the core of human mental lexicon is hereby quantified as collective linguistic norms. Since distributional similarity between words is an important factor explaining associations, traditional studies extensively adopted Distributional Semantic Models (DSMs) and word embeddings to predict human word associations (Mandera et al., 2017; Evert and Lapesa, 2021; Kwong et al., 2022; A et al., 2024). On the other hand more recent studies, based on LLMs, proved that such systems can align, to a considerable extent, with human patterns of asso-

[1]Our codes and data are publicly available at https://github.com/florethsong/word_association

Proceedings of the 16th International Conference on Computational Semantics, pages 218–240
September 22-23, 2025,

ciating words (De Deyne et al., 2024; Abramski et al., 2025; Bai et al., 2025). An open question arisen is which one of these methods delivers better results in approximating human norms.

Besides the theoretical interest of the problem, the results are relevant to the problem of *reverse dictionary*, where a user tries to retrieve a word given a set of associates or a dictionary definition (Almeman and Espinosa-Anke, 2024). Reverse dictionary applications, which can be seen as the information retrieval modeling side of human lexical access (the so-called *tip-of-the-tongue*, *anomia* or *dysnomia* problem, see Zock (2002) and Rapp and Zock (2014)) can be helpful tools for writers and translators, and in this sense, generative LLMs show a lot of promise, as they could help a user by retrieving and generating a target word simply on the basis of a prompt with some lexical cues. From a psychological standpoint, word associations are also a fundamental indicator for human *creativity* and *divergent thinking*, as research indicates a consistent positive correlation between high levels of human creativity and the capacity to generate word associates that are distant in the lexical network (Kenett and Faust, 2019; Yang et al., 2022; Johnson and Hass, 2022; Wang et al., 2024).

As illustrated by the task types in Figure 1 focusing on semantic similarity and relatedness, this study designs a protocol of five-stage word association tasks (WATs) to evaluate models against human norms. By taking the majority of human responses across various WATs as a main proxy of human mental lexicon, this study compare the word association abilities of vectors from traditional static word-type embedding models (WEMs), mean-pooled word-token embeddings from representative pretrained language models (PLMs), and prompting strategies with mainstream LLMs. Results show that although none of these models align fully with human mental lexicon and hence model effectively the versatility of the human cognitive ability, WEMs and LLMs can better mimic human associations than PLMs: LLMs outperform competitors in word retrieval tasks (with a focus on capturing semantic similarities, i.e., lexical interchangeability), while WEMs perform better in concept pairing (emphasizing the identification of semantic relatedness, that is, detecting mutual conceptual relations). While scaling-up and contextualization often helps embedding models, PLMs show more architecture- and task-dependent trade-offs.

2 Related Work

WATs with Humans Word associations are grounded in Firthian's "word in company" tradition that lexemes with resembling behavioral profiles (like, shared collocational patterns or syntactic structures) encode similar paradigmatic or syntagmatic relations in meaning and cognition (Firth, 1957; Church and Hanks, 1990). They function as prototypical and advantageous tools in psycholinguistics to tap directly into semantic memory and conceptual knowledge reflected in human thinking, reasoning, and language use. As a classical paradigm, the free word association task and its variants based on word clustering or relationship identification accelerate quantitative exploration of human cognitive phenomena, such as language acquisition (Citraro et al., 2023), metaphor and analogy comprehension (Lu et al., 2022), and creativity (Beaty and Kenett, 2023; Wang et al., 2024).

Various human association norms originally designed to access preexisting word knowledge in the human mind and detect different aspects of cognitive development and competencies, such as EAT (the Edinburgh Associative Thesaurus, Kiss et al., 1973), USF (the University of South Florida Free Association Norms, Nelson et al., 2004), and SWOW (the Small World of Words, De Deyne et al., 2019), can be applied in conjunction as a comprehensive benchmark for facilitating the measurement of the alignment between human internal semantic cognition and external word embeddings.

WATs with Word Embeddings WATs have significantly contributed to benchmarking models' semantic representations and conceptual structures against human mental lexicon shown in diverse human-generated norms, both in theory and practice (Rapp and Zock, 2014; De Deyne et al., 2016). They provide a powerful means to probe into two fundamental dimensions of distributional semantics: *similarity* (interchangeability of words, e.g., *car/van*) and *relatedness* (shared conceptual relations between words, e.g., *car/wheel*) (Fodor et al., 2023). Existing work (Lenci et al., 2022; Fodor et al., 2023; A et al., 2024, etc.) has been extensively devoted to thorough comparisons across a wide spectrum of DSMs from count (e.g., Dissect PPMI, Baroni et al., 2014) and predict models (e.g., word2vec, Mikolov et al., 2013) at early static-embedding generation to recent transformer-based contextual embedding models (e.g., BERT, Devlin

et al., 2019). These studies consistently demonstrated the superior performance of static embeddings in out-of-context WATs, while highlighting contextual embeddings' advantages in tasks requiring contextual sensitivity. Collectively, they revealed the nuanced interplay between model design, task requirements, and cognitive plausibility of language representations.

WATs with LLMs Recent work expanded the use of WATs into dissecting the behaviors of LLMs as black-box systems to better understand their advantages and limitations in semantic-aware reasoning. Abramski et al. (2025) established LLM-generated free association norms by prompting popular LLMs and found that LLM-generated associations exhibit weaker concreteness effects and stronger societal biases compared to human norms. Cazalets and Dambre (2025) demonstrated GPT-series' ability to synchronize with human players in game-like free association interactions. Beyond free association tasks, structured variants such as ontological classification (De Deyne et al., 2024), connection tasks (Samdarshi et al., 2024), and similarity judgments on triads (Linhardt et al., 2025) have assessed LLMs' ability to identify underlying internal relations or cluster words by shared characteristics. Increasing interest has been in using WATs to reveal both explicit and implicit societal biases encoded in LLMs. For example, studies by Ethayarajh et al. (2019), Abramski et al. (2025), and Bai et al. (2025) presented how WATs can uncover attitude disparities between model outputs and human responses, highlighting their utility in addressing ethical issues of language models.

Such studies stress WATs' dual role in illuminating human and models' semantic networks; however, existing work mainly relied either on prompt-based strategies with LLMs or on embedding similarity, without any systematic comparison between the two. Also, previous studies were limited in scope, focusing only on one type of WAT, therefore a more comprehensive evaluation is necessary.

3 Experimental Settings

According to Abramski et al. (2025), probing into the conceptual knowledge encoded within language models by examining the embedding space works well for traditional models, but it is less effective and practical for LLMs. This is due to the fact that embeddings from LLMs exhibit severe anisotropy in their vector spaces, which can significantly distort similarity estimates (e.g., Ethayarajh, 2019; Zhang et al., 2020; Biś et al., 2021; Timkey and van Schijndel, 2021; Nie et al., 2025; Feng et al., 2025). Therefore, a shift from the conventional approach of accessing the embedding space to a top-down approach in the context of LLMs was proposed, which means directly prompting LLMs with specific tasks and using their outputs to infer the knowledge in their vector spaces.

Therefore, we examine the capabilities of different models by employing two methodologies: **embedding** and **prompting**, which align with their default typical approaches to WATs at hand. A basic assumption of embedding-based tests is that *the strength of word associations increases with the cosine similarity of their embeddings* (Clark, 2015; Fodor et al., 2023), reflecting graded semantic relationships in vector spaces. For WEMs, we extracted static word-type embeddings and calculated the cosine similarities as the basis for their outputs. In terms of PLMs, both non-contextualized and contextualized word embeddings were mean-pooled from the last hidden layers and cosine similarities were computed. Regarding LLMs, we utilized zero-shot prompts to obtain direct responses.

3.1 Task Design

We tested our models on five complementary and progressively challenging tasks built on the well-established datasets, as summarized in Table 1. Each task stresses distinct capabilities of language models in terms of processing semantic similarity versus relatedness, with extended discussion provided in Appendix A.

Task 1: Multiple-Choice Associations FAST dataset (Evert and Lapesa, 2021) is leveraged in this task, which provides quadruples of a stimulus and three candidate words: "*FIRST*, *HAPAX*, *RANDOM*" where *FIRST* is the most frequent associate response from humans, *HAPAX* is a response that has been mentioned only once, and *RANDOM* is a randomly selected control candidate with minimal semantic association strength to the stimulus. For each stimulus, a model has to choose the most strongly associated word (i.e., for embedding models, the one with the largest semantic similarity). It is worth noticing that *HAPAX* is also a word with weak semantic association with the stimulus, and thus it works as a strong distractor.

Performance is measured using *Accuracy*, i.e., the percentage of items in which the model cor-

Table 1: Overview of Datasets for the Five Association Tasks. In the "Structure" column, underlined elements indicate the information presented to the evaluated models, while bolded elements are used as the ground truth.

Task	Dataset	Structure	Size[2]	Word List	Metrics[3]
1	FAST	<stimulus, ***FIRST***, *HAPAX*, *RANDOM*> *(e.g., accept,* ***receive****, love, souls)*	11,431 (12,329)		Accuracy
2	FAST	<stimulus, ***FIRST***, *HAPAX*, *RANDOM*> *(e.g., achievement,* ***success****, degree, round)*	11,431 (12,329)	✓	Top-1 Accuracy, Mean Rank (threshold = 4)
3	CogALex	<**Target**, a1, a2, a3, a4, a5> *(e.g.,* ***air****, plane, fresh, water, breathe, force)*	3,650 (4,000)	✓	Top-1 Accuracy, Mean Rank (threshold = 4)
4	Concrete-Abstract Triad	<A, B, C> ($P_{\mathbf{AB}}$, $P_{\mathbf{AC}}$, $P_{\mathbf{BC}}$) *(concrete e.g., banana, cherry, pineapple* ***(0.18, 0.65, 0.18)****)* *(abstract e.g., darling, hero, thinker* ***(0.48, 0.13, 0.40)****)*	100 + 100		Accuracy (Total, Concrete, Abstract)
5	Remote Triad	<A, B, C> ($P_{\mathbf{AB}}$, $P_{\mathbf{AC}}$, $P_{\mathbf{BC}}$) *(e.g., fence, mask, salt* ***(0.80, 0.05, 0.15)****)*	100		Accuracy

rectly picks the *FIRST* associate ($[0\%, 100\%]$), with a random-choice baseline of 33.3%. To mitigate potential positional bias, the elements in each candidate list were shuffled during LLM prompting.

Task 2: Open-Vocabulary Associations This task also relies on the FAST dataset but differs in that it presents no fixed set of candidates. Instead, models are asked to generate the most associated word in an open-vocabulary setup which further simulates the way humans access their mental lexicon in a natural association task.

In the current study, we create a "pseudo-open vocabulary" condition for WEMs and PLMs where models are tasked with ranking associations for a given stimulus over a large-scale word list, which covers all *FIRST* words and restricts the range of potential choices. The tailored word list applied in this study is a concatenation of vocabularies from word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and FastText (Joulin et al., 2017) models, totaling 101,607 word types, effectively serving the goals of the task. While LLMs are asked to directly provide 30 words associated with the stimulus, ordered by their association strength.

Two statistical metrics are reported based on the word ranking list for each stimulus (i.e., the word list sorted by decreasing cosine similarity based on embedding-based models, and the ranked word list generated by LLMs): 1) *Top-1 Accuracy*: how frequently a model ranks the *FIRST* human response as the top 1 result ($[0\%, 100\%]$), positively correlated with model-human semantic alignment; and 2) *Mean Rank (threshold = 4)*: the average position of the *FIRST* word in the rankings by a certain model. We set *4* as the threshold, that is, if the rank of the *FIRST* word in a given ranking list is *3* or lower, we assign this actual rank as the score for the given instance, otherwise we assign a score of *4*. This is in line with the convention of shared tasks using mean rank to mitigate excessive penalty on instances with high-rank outliers (Camacho-Collados et al., 2018; Mansar et al., 2021). The final scores are mean ranks falling in $[1, 4]$, which are negatively correlated with the performance of models in lexical alignment with humans.

Task 3: Reverse Associations Based on the CogALex shared task dataset (Rapp and Zock, 2014), this task evaluates the models' ability to simultaneously integrate multi-layered relations across multiple stimuli. The logic of this task is closely related to the *tip-of-the-tongue* phenomenon. Each item features a *Target* word defined as the human-generated response to five given cue words, which are all interconnected with the *Target* at a certain conceptual level.

The objective is to retrieve the *Target* word that semantically connects the five cue words, within a pseudo-open vocabulary of candidates. For WEMs and PLMs, we compute the average vector of the five cue words and measure the association strength (i.e., cosine similarity) between it and each candi-

[2]Since word2vec, GloVe, and FastText models underperform when faced with out-of-vocabulary words, we manually excluded any missing items if a word in our specific word set is not included in any of these three baseline models. As a result, we obtained 11,431 out of 12,329 items from the original FAST dataset for Tasks 1 and 2, and 3,650 out of 4,000 items from the original CogALex dataset for Task 3. For both triad datasets corresponding to Tasks 4 and 5, no items were removed from the original datasets.

[3]To ensure reliable and effective comparisons, we conducted two types of significance tests, depending on the evaluation metrics. For accuracy scores in Tasks 1–5, we applied McNemar's test (McNemar, 1947) corrected with Benjamini–Hochberg procedure (Benjamini and Hochberg, 1995) across all model pairs to determine whether the observed accuracy differences are statistically significant. For mean rank results in Tasks 2 and 3, we used the Wilcoxon signed-rank test (Wilcoxon, 1945) to evaluate whether the rankings of the *FIRST* or *Target* words in the given instances produced by different models differed significantly. More details can be found in Figures 8-12 in the Appendices.

date word in a list of 101,607 words (identical to that used in Task 2) to produce a ranked list of target words, while LLMs are required to directly generate a list of 30 potential targets. Performance is evaluated using the same two metrics as in Task 2. This task emphasizes reverse reasoning ability and tests whether models can reconstruct a unifying concept from distributed cues.

Task 4: Concrete-Abstract Association Triads This task presents triads of words to models, where any two can be paired based on varying semantic features. The goal is to select the most semantically related pair in each triad. The dataset, introduced by De Deyne et al. (2021) is employed, which can be split into two subsets: 1) **Concrete Triad Dataset** focusing on physical entities and events; 2) **Abstract Triad Dataset** focusing on psychological and conceptual relationships.

Models' outputs are compared against human preferences with percentages provided in the dataset. Specifically, for each instance, WEMs and PLMs select the word pair with the highest cosine similarity among the three candidate pairs based on their word embeddings, whereas the top-ranked pair from all three pairs is regarded as LLMs' final choice. We report respectively the accuracies on total, concrete, and abstract triads, all ranging in $[0\%, 100\%]$ and positively correlated with model-human alignment. In cases where humans do not produce a single dominant pairing (e.g., two pairings have equal frequencies chosen by humans), a model's choice is considered correct if it matches one of the most frequent human choices.

Task 5: Remote Association Triads Similar to the structure in Task 4 but significantly more challenging, this task utilizes the Remote Triad dataset (De Deyne et al., 2016) and requests models to identify the most related pairing with more distant and creative semantic links among words. As in Task 4, we measure accuracy based on human preferences provided in the original dataset. Due to the subtlety of the associations involved, this task offers deeper and informative insights into the extent to which models can capture latent and implicit conceptual relations beyond immediate meaning similarity between words.

3.2 Model Selection

We evaluate representative and state-of-the-art language models across three architectural paradigms and development stages, further dividing them into "*Smaller*" (with around 1B or fewer parameters) and "*Larger*" (with over 1B parameters) categories based on parameter scale. No post hoc modifications were conducted to the vanilla models and their embeddings with the intention to assess the intrinsic quality of their representations.

The first group covers five static **WEMs**: *word2vec* (Mikolov et al., 2013) pretrained on 100B tokens of Google News, *GloVe* (Pennington et al., 2014) trained on 6B tokens of Wikipedia 2014 and newspapers as well as *GloVe-CC* on 840B tokens of Common Crawl (CC) Web data, and *FastText* (Joulin et al., 2017) trained on 16B tokens of Wikipedia 2017 and other webbase corpus as well as *FastText-CC* on 600B tokens of CC. All models were tested with 300-dimensional embeddings.

The second group includes six **PLMs**: *BERT-base* and *-large* (Devlin et al., 2019), *GPT-2* and *-xl* (Radford et al., 2019), and *T5-small* and *-3B* (Raffel et al., 2020), from which we extracted non-contextualized (the input is a single word, like "*accept*") as well as contextualized (the input is a fixed simple sentence containing the key word, like "*My target word is accept*") word embeddings by mean-pooling the subword representations in the last layers.

The third group composes three **LLMs**, i.e., *GPT-4.1*[3], *DeepSeek-V3 (-0324)* (DeepSeek-AI, 2024), and *Qwen3 (-238B-A22B)* (Yang et al., 2025). We ran additional experiments (cf. Appendix G) to test how different temperature settings (0.01 vs. 0.5 vs. 1), prompt strategies (simple zero-shot vs. enhanced few-shot), and reasoning modes (standard vs. reasoning) impact LLM effectiveness across different WATs. While results indicate that most LLMs achieve marginally better performance at temperature 0.5 using detailed few-shot prompts with reasoning, optimal configurations vary across tasks and models. To obtain consistent and comparable patterns from LLMs, we standardized our configurations: temperature was maintained at 0.01 using zero-shot prompts, and reasoning ability was not activated for the reasoning model—Qwen3.

4 Results and Analysis

This section reports the empirical results and findings obtained from operationalizing the series of tasks and metrics defined in Section 3.1. The statistics corresponding to each task and significance test results are displayed in Appendices B-F.

[2]https://openai.com/index/gpt-4-1/

4.1 Multiple-Choice Association

Figure 2 illustrates the performance of various language models in Task 1, that is, identifying the most interchangeable word or near-synonyms to a given cue from a restricted set of candidates. With the only exception of GPT-2 (non-ctx), all models achieve an accuracy vastly better than the chance-level baseline. Notably, WEMs and LLMs significantly outperform PLMs proved by Figure 8, frequently reaching accuracies of 80% or higher. This suggests that static word-type representations derived from WEMs and prompted LLMs are more effective at capturing direct semantic similarities between near-synonyms or conceptually related words. In contrast, token-level embeddings mean-pooled from PLMs show substantially reduced effectiveness, indicating a difficulty in abstracting a type-level representation, which would be necessary for this task. Our findings are consistent with Lenci et al. (2022), Apidianaki (2023), and A et al. (2024), who claimed that word-token representations complicate the investigation of lexical semantic knowledge anchored at the word-type level.

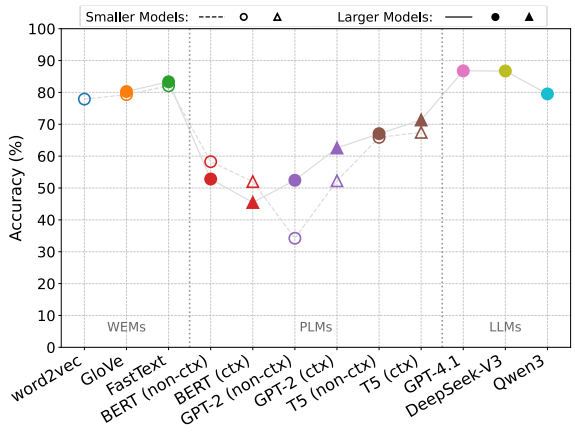

Figure 2: Plot of Model Accuracies in the Multiple-Choice Association Task. Fillings and shapes are used to distinguish the context types and the magnitudes of models. Hollow markers indicate smaller models, while solid ones represent larger ones. Non-contextualized (non-ctx) PLMs are shown as circles, in contrast to contextualized (ctx) PLMs marked with triangles. *Note: the visual markers in the subsequent figures maintain consistent meanings throughout this paper.*

Additionally, when comparing the efficiency of non-contextualized embeddings to contextualized ones within PLMs, it is interesting to note that extra contexts benefit both GPT-2 and T5, though to varying degrees, while BERT-base and BERT-large models do not display the same enhancement.

Comparisons between smaller and larger models reveal that, for most WEMs and PLMs, increasing parameter count correlates with improved modeling of lexical semantics and conceptual relationships. Larger models tend to outperform smaller ones, aligning with established *Scaling Laws* (Kaplan et al., 2020), with the exception of BERT, whose larger variant is worse than the smaller one, pointing to its potential architectural or training-related limitations in preserving word-type knowledge during scaling-up.

Figure 3 reveals distinct error patterns across different model types. The errors align with overall accuracy trends: WEMs and LLMs predominantly select *HAPAX*, indicating a relatively strong sensitivity to weak associations, while making few *RANDOM* selections. This suggests that such models can at least effectively distinguish between weak and non-existent associations, while in contrast PLMs and particularly GPT-2 (non-ctx) are more frequently misled by *RANDOM* distractors. Furthermore, LLMs occasionally encountered *OTHER* errors, particularly involving incorrect formats or range misinterpretations under zero-shot prompting. For example, LLMs may output *stock* in response to *garters* with the candidate list [*lace*, *sweaters*, *stockings*], reflecting possible failures in instruction following that manifest as hallucinations or misalignment with task requirements.

4.2 Open-Vocabulary Association

Task 2 introduces a more demanding evaluation scenario, placing models under empirically unrestricted "free" association conditions, therefore resulting in universally lower performance across all models as evidenced in Figure 4. This task probes the models' global semantic organization and broader vector space in that they mirror human-like associative knowledge. Remarkably, the stark disparities in top-1 accuracies and mean ranks between WEMs/PLMs and LLMs (the majority of these differences are statistically significant as shown in Figure 9) highlight that LLMs can more reliably identify human-preferred associative targets by frequently retrieving and prioritizing near-synonyms of high-frequency co-occurring lexemes for the given stimulus (e.g., *really* for *actually*, *departure* for *arrival*).

Interestingly, the effect of model size is heterogeneous and model-dependent. Specially, scaling-up yields marginal performance gains for GloVe, Fast-

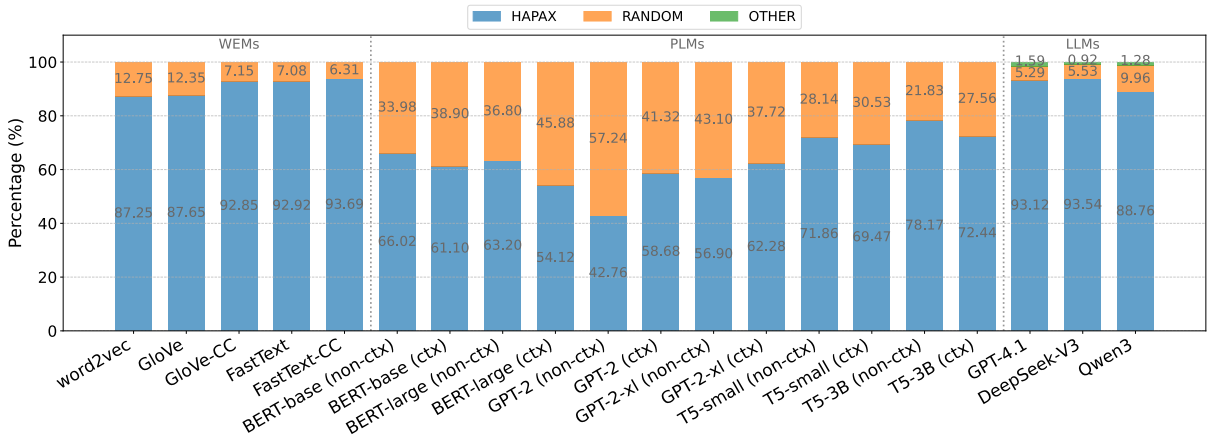

Figure 3: Error Percentages for Various Types of Wrong Hits in the Multiple-Choice Association Task. Blue bars show the percentage of HAPAX models deemed the most associated word with the stimulus, orange bars represent RANDOM hits, and green bars indicate other error types (e.g., multiple-word or out-of-choice generations).

Text, and GPT-2, but not for BERT or T5. This indicates that semantic-cognitive alignment relies more on architecture than on scale. It further suggests that parametric scaling laws interact differently with task-specific requirements.

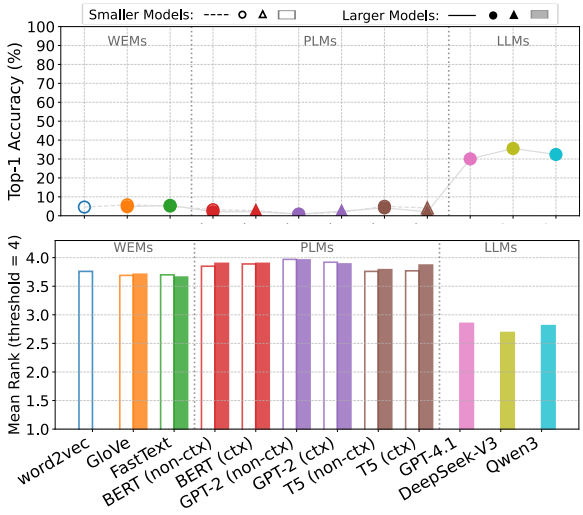

Figure 4: Top-1 Accuracies (above) and Mean Ranks (below) in the Open-Vocabulary Association Task.

4.3 Reverse Association

Task 3 requires two-step reasoning: first identifying the conceptual commonality among five related hint words, and then finding the target word connecting them from a broad candidate pool.

As shown in Figure 5 and 10, the results largely mirror the overall performance trends observed in Tasks 1 and 2, while further confirming that LLMs exhibit better alignment with human semantic knowledge. Specifically, LLMs achieve over 25% top-1 accuracies and demonstrate consistently lower mean ranks for the correct *Target* words as judged by humans. This suggests that LLMs are better equipped to handle tasks requiring abstract generalization and lexical retrieval.

Notably, static embeddings from WEMs also show relatively strong performance, achieving higher accuracy and lower average ranks compared to all PLMs. As for the vector representations from the latter type of models, it is possible that they are just too context-specific for tasks requiring to capture the semantics of word types.

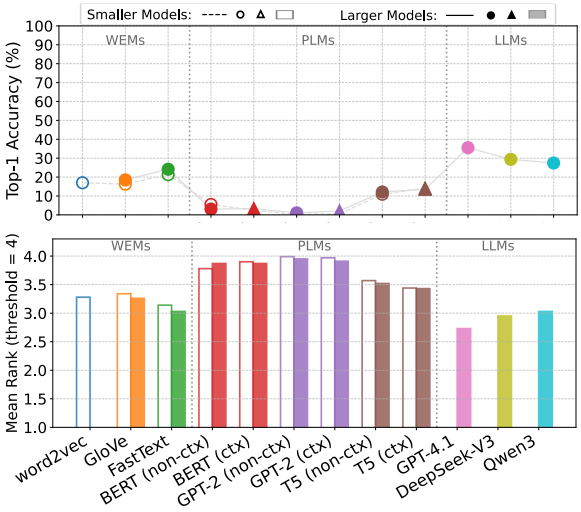

Figure 5: Top-1 Accuracies (above) and Mean Ranks (below) in the Reverse Association Task.

4.4 Concrete-Abstract Association

This task probes semantic space by comparing the strengths of inter-word semantic relationships within triads. As shown in Figure 6 and 11,

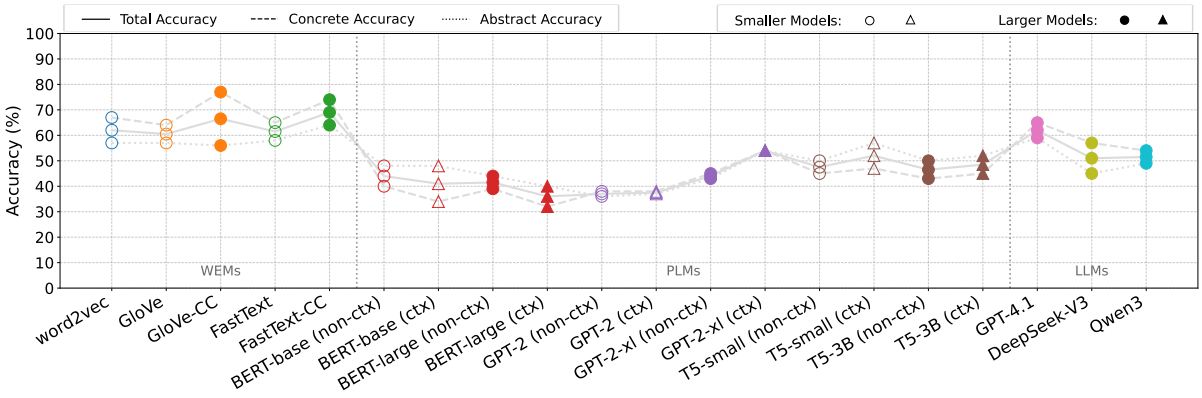

Figure 6: Accuracies in the Concrete-Abstract Association Task on Total, Concrete, and Abstract Datasets.

experimental results highlight the superior performance of WEMs, which significantly outperform embeddings from most PLMs, regardless of whether the word pairs are concrete or abstract. Moreover, regarding LLMs, the results also reveal that employing prompt-based methods on GPT-4.1 in this task achieves accuracy comparable to static embeddings derived from WEMs. In contrast, both DeepSeek-V3 and Qwen3 perform significantly worse—especially compared to larger WEMs, namely, GloVe-CC and FastText-CC, and their performance aligns more closely with that of T5 models among PLMs.

Interestingly, WEMs and LLMs show somewhat stronger performance on concrete triads than on abstract ones, while PLMs (like BERT and T5) exhibit the opposite pattern. This contrast may reflect their differing sensitivities to concreteness effects (Hill et al., 2014; Knupleš et al., 2023; Abramski et al., 2025), which describes that concrete words tend to evoke stronger but fewer associations, whereas abstract words elicit weaker but more diffuse associations. In this light, WEMs and LLMs are more effective at leveraging the focused, robust relationships typical of concrete concepts, whereas token-based embeddings from PLMs show fairly poor capability of adapting to such associations.

At last, we observe that incorporating contextual information during embedding extraction from PLMs leads to little performance degradation in BERT models but a slight improvement in GPT-2 and T5 models. However, these differences stemming from their distinct model architectures (Qiu et al., 2020) are not significant in this task. Besides, while scaling has minimal impact on PLMs' performance, it significantly enhances that of WEMs.

4.5 Remote Association

Contrary to expectations, the increased conceptual distances for the triads in Task 5, which may present greater challenges for human participants, have only a limited impact on the accuracies achieved by most language models when compared to the baseline results in Task 4. The results in Figure 12 indicate that significant accuracy differences arise only between WEMs and two types of PLMs (BERT and GPT-2), as well as two LLMs (DeepSeek-V3 and Qwen3). The top-performing models in each group remain consistent with those identified in other tasks, namely, FastText-CC among WEMs, T5-3B among PLMs, and GPT-4.1 among LLMs.

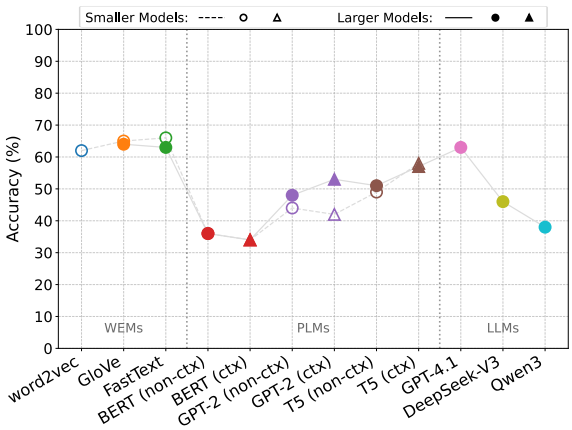

Figure 7: Accuracies in the Remote Association Task.

For this task, neither model size nor contextualization substantially affects the ability of WEMs and PLMs to identify intricate relational abstractions. Two primary factors may explain this finding. The first is the limited dataset size of 100 items, which may restrict generalization and robust statistical analysis. Second, theoretically, the remote

associations present in these triads generate scenarios that extend beyond textual information by incorporating not only perceptual but multimodal concerns, which may reduce the influence of differences in textual data. For instance, in the triad: A-*fear*, B-*guest*, C-*price* (P_{AB}=0.275, P_{AC}=0.425, P_{BC}=0.300), models with limited abstraction capabilities fail to identify the most human-like connection between fear and price, a subtle relationship that likely reflects real-world consumption and market experiences but is uncommon in training data. In such cases, evaluating atypicality by analyzing the distribution of model choices across both typical and atypical human responses, rather than relying solely on accuracy based on the most frequent human response, may yield a more informative comparison.

5 Conclusion

This study systematically evaluates the intrinsic semantic capabilities of diverse language models, including WEMs, PLMs, and LLMs, by leveraging their typical operational modes (e.g., word embeddings vs. prompt-based generation). Through the adaptation and integration of five kinds of classical psycholinguistic WATs, we assess how well these models perform on cognitively motivated benchmarks. The results reveal distinct performance and limitations across architectures and configurations.

First, WEMs and LLMs demonstrate better alignment with human association norms compared to PLMs, particularly in tasks requiring stable type-level semantic representations. Notably, LLMs outperform the other models in word retrieval (Tasks 1–2, similarity-dominant; Task 3, considering both similarity and relatedness), while WEMs do better in concept pairing (Tasks 4–5, relatedness-dominant), highlighting their complementary strengths across model architectures and the fact that human mental lexicon is good at synergizing similarity and relatedness, but not artificial systems. For WMEs, increasing model size generally improves performance. However, PLMs exhibit architecture-dependent behaviors in terms of scaling and contextualization: encoder-only models like BERT often degrade with larger scales and added contexts but decoder-only models (e.g., GPT-2) tend to benefit from both. For encoder–decoder models (e.g., T5), the impacts are task-specific. Their performance notably improves in Tasks 1 and 3 in these two settings but declines in Task 2.

LLMs' partial success in some WATs by mimicking human semantic behaviors demystifies the claim of their human-like intelligence. Yet they still struggle to fully replicate the versatility of the human mental lexicon, particularly in associating remote or abstract concepts. This suggests a tension between accuracy and creativity in language modeling, warranting deeper exploration. Together, these findings provide comprehensive insights into the alignment between language models and human cognition and highlight the value of psycholinguistic data for diagnosing model capabilities and biases.

Limitations

While this work provides broad insights into the semantic quality of different language models, it is limited by a few reasons for further improvement in the future.

A primary limitation of this study is the use of different evaluation methods across model types: cosine similarity for WEMs and PLMs, versus prompting for LLMs. While these approaches reflect typical usage patterns, the inconsistency challenges the validity of direct comparisons. Embedding similarity may capture relations beyond associative knowledge in some cases, whereas prompting can advantage LLMs by providing task-specific guidance. Consequently, some performance differences may reflect evaluation methods rather than intrinsic disparities in model knowledge. Future work should seek to standardize protocols, for example, by incorporating embedding-based measures for LLMs.

Additionally, while cosine distances are the most commonly used method for measuring semantic similarity between vectors, it has been criticized for potentially yielding arbitrary and meaningless "similarities" (Steck et al., 2024). Meanwhile, it may underestimate the actual similarity between contextualized embeddings (Wannasuphoprasit et al., 2023; Ijebu et al., 2025) and does not reliably indicate human associations due to its symmetric nature (Abramski et al., 2025). This limitation may impact our findings regarding the alignment between human assessments and the embeddings of WEMs and PLMs. Therefore, alternative methods, such as the soft cosine similarity proposed by Ijebu et al. (2025) or rank-based metrics (Santus et al., 2016, 2018; Zhelezniak et al., 2019), could be explored for a more robust investigation.

Also, our analysis of PLM models focused only on final-layer embeddings obtained through mean pooling, overlooking potential variations across transformer layers. Previous research suggested that intermediate layers may better capture lexical semantics (Ormerod et al., 2024). Additionally, it could be the case that our generic contexts were not informative enough to create robust representations, and better results might be achieved by sampling random sentence contexts with the target word from a large-scale corpus to represent and by averaging the corresponding embeddings (Bommasani et al., 2020; A et al., 2024; Nie et al., 2025). We will examine more layer-wise semantic properties and assess methods for distilling contextualized embeddings into static ones in the future. On the other hand, we also believe that this issue confirms that PLMs are probably not the best choice for the automatic collection of word associations, compared to WEMs and LLMs, given that researchers would have to perform the additional steps of context sampling and selection of the optimal layers.

Furthermore, the current study primarily focused on English WATs and did not adequately address advanced reasoning models and better configurations for prompting LLMs, which require further examination and comparison, including in multilingual and low-resource language contexts.

Finally, this study was conducted solely on semantic-level word associations. To gain a more in-depth understanding of language associations, future work can incorporate perspectives from other linguistic dimensions, such as morphological and phonological associations.

References

Pranav A, Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, and Alessandro Lenci. 2024. Comparing static and contextual distributional semantic models on intrinsic tasks: An evaluation on Mandarin Chinese datasets. In *Proceedings of LREC-COLING 2024*, pages 3610–3627, Torino, Italia. ELRA and ICCL.

Katherine Abramski, Riccardo Improta, Giulio Rossetti, and Massimo Stella. 2025. The "LLM World of Words" English Free Association Norms Generated by Large Language Models. *Scientific Data*, 12(1):1–9.

Fatemah Almeman and Luis Espinosa-Anke. 2024. GEAR: A Simple GENERATE, EMBED, AVERAGE AND RANK Approach for Unsupervised Reverse Dictionary. In *Proceedings of LREC-COLING*.

Marianna Apidianaki. 2023. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2):465–523.

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L Griffiths. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8).

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Baltimore, Maryland. Association for Computational Linguistics.

Roger E Beaty and Yoed N Kenett. 2023. Associative thinking at the core of creativity. *Trends in cognitive sciences*, 27(7):671–683.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Daniel Biś, Maksim Podkorytov, and Xiuwen Liu. 2021. Too much in common: Shifting of embeddings in transformer language models and its implications. In *Proceedings of NAACL 2021*, pages 5117–5130, Online. Association for Computational Linguistics.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

Jose Camacho-Collados, Claudio Delli Bovi, Luis Espinosa-Anke, Sergio Oramas, Tommaso Pasini, Enrico Santus, Vered Shwartz, Roberto Navigli, and Horacio Saggion. 2018. SemEval-2018 Task 9: Hypernym Discovery. In *Proceedings of SemEval*.

Tanguy Cazalets and Joni Dambre. 2025. Word synchronization challenge: A benchmark for word association responses for large language models. In *International Conference on Human-Computer Interaction*, pages 3–19. Springer.

Kenneth Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16(1):22–29.

Salvatore Citraro, Michael S Vitevitch, Massimo Stella, and Giulio Rossetti. 2023. Feature-rich multiplex lexical networks reveal mental strategies of early language learning. *Scientific Reports*, 13(1):1474.

Stephen Clark. 2015. Vector space models of lexical meaning. *The Handbook of Contemporary semantic theory*, pages 493–522.

Simon De Deyne, Chunhua Liu, and Lea Frermann. 2024. Can GPT-4 Recover Latent Semantic Relational Information from Word Associations? A Detailed Analysis of Agreement with Human-annotated Semantic Ontologies. In *Proceedings of the COLING Workshop on Cognitive Aspects of the Lexicon*.

Simon De Deyne, Danielle J Navarro, Guillem Collell, and Andrew Perfors. 2021. Visual and Affective Multimodal Models of Word Meaning in Language and Mind. *Cognitive Science*, 45(1):e12922.

Simon De Deyne, Danielle J Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior research methods*, 51:987–1006.

Simon De Deyne, Amy Perfors, and Daniel J Navarro. 2016. Predicting Human Similarity Judgments with Distributional Models: The Value of Word Associations. In *Proceedings of COLING 2016*.

DeepSeek-AI. 2024. Deepseek-v3 technical report.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL 2019*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of EMNLP-IJCNLP*, pages 55–65, Hong Kong, China. Association for Computational Linguistics.

Kawin Ethayarajh, David Duvenaud, and Graeme Hirst. 2019. Understanding undesirable word embedding associations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1696–1705, Florence, Italy. Association for Computational Linguistics.

Stefan Evert and Gabriella Lapesa. 2021. FAST: A carefully sampled and cognitively motivated dataset for distributional semantic evaluation. In *Proceedings of CONLL*.

Zhaoxin Feng, Jianfei Ma, Emmanuele Chersoni, Xiaojing Zhao, and Xiaoyi Bao. 2025. Learning to Look at the Other Side: A Semantic Probing Study of Word Embeddings in LLMs with Enabled Bidirectional Attention. In *Proceedings of ACL*.

John Rupert Firth. 1957. *A Synopsis of Linguistic Theory 1930–55*. Longmans.

James Fodor, Simon De Deyne, and Shinsuke Suzuki. 2023. The importance of context in the evaluation of word embeddings: The effects of antonymy and polysemy. In *Proceedings of IWCS*, pages 155–172, Nancy, France. Association for Computational Linguistics.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15.

Thilo Hagendorff. 2023. Machine Psychology: Investigating Emergent Capabilities and Behavior in Large Language Models Using Psychological Methods. *arXiv preprint arXiv:2303.13988*, 1.

Felix Hill, Anna Korhonen, and Christian Bentz. 2014. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive science*, 38(1):162–177.

Funebi Francis Ijebu, Yuanchao Liu, Chengjie Sun, and Patience Usoro Usip. 2025. Soft cosine and extended cosine adaptation for pre-trained language model semantic vector analysis. *Applied Soft Computing*, 169:112551.

Dan R Johnson and Richard W Hass. 2022. Semantic context search in creative idea generation. *The Journal of Creative Behavior*, 56(3):362–381.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Yoed N. Kenett and Miriam Faust. 2019. A semantic network cartography of the creative mind. *Trends in Cognitive Sciences*, 23(4):271–274.

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of english and its computer analysis. *The Computer and Literary Studies*, 153.

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde. 2023. Investigating the Nature of Disagreements on Mid-Scale Ratings: A Case Study on the Abstractness-Concreteness Continuum. In *Proceedings of CONLL*.

Trina Kwong, Emmanuele Chersoni, and Rong Xiang. 2022. Evaluating monolingual and crosslingual embeddings on datasets of word association norms. In *Proceedings of the BUCC Workshop within LREC 2022*, pages 1–7, Marseille, France. European Language Resources Association.

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

Lorenz Linhardt, Tom Neuhäuser, Lenka Tětková, and Oliver Eberle. 2025. Cat, Rat, Meow: On the Alignment of Language Model and Human Term-Similarity Judgments. In *Proceedings of the ICLR Workshop on Re-Align*.

Lea Löhn, Niklas Kiehne, Alexander Ljapunov, and Wolf-Tilo Balke. 2024. Is Machine Psychology here? On Requirements for Using Human Psychological Tests on Large Language Models. In *Proceedings of INLG*.

Hongjing Lu, Nicholas Ichien, and Keith J Holyoak. 2022. Probabilistic analogical mapping with semantic relation networks. *Psychological review*, 129(5):1078.

Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2017. Explaining Human Performance in Psycholinguistic Tasks with Models of Semantic Similarity Based on Prediction and Counting: A Review and Empirical Validation. *Journal of Memory and Language*, 92:57–78.

Youness Mansar, Juyeon Kang, and Ismail El Maarouf. 2021. The FinSim-2 2021 Shared Task: Learning semantic similarities for the financial domain. In *Companion Proceedings of the Web Conference 2021*, pages 288–292.

Quinn McNemar. 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157.

Ken McRae, Saman Khalkhali, and Mary Hare. 2012. Semantic and Associative Relations in Adolescents and Young Adults: Examining a Tenuous Dichotomy. *Psychology Publications*, 115.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *1st International Conference on Learning Representations, ICLR 2013, Workshop Track Proceedings*.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Zhijie Nie, Zhangchi Feng, Mingxin Li, Cunwang Zhang, Yanzhao Zhang, Dingkun Long, and Richong Zhang. 2025. When text embedding meets large language model: A comprehensive survey.

Mark Ormerod, Jesús Martínez del Rincón, and Barry Devereux. 2024. How is a "kitchen chair" like a "farm horse"? exploring the representation of noun-noun compound semantics in transformer-based language models. *Computational Linguistics*, 50(1):49–81.

Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. Is temperature the creativity parameter of large language models? In *Proceedings of ICCC'24*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China technological sciences*, 63(10):1872–1897.

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.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Reinhard Rapp and Michael Zock. 2014. The CogALex-IV shared task on the lexical access problem. In *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 1–14, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Prisha Samdarshi, Mariam Mustafa, Anushka Kulkarni, Raven Rothkopf, Tuhin Chakrabarty, and Smaranda Muresan. 2024. Connecting the dots: Evaluating abstract reasoning capabilities of llms using the new york times connections word game. In *Proceedings of EMNLP 2024*, pages 21219–21236.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, Chu-Ren Huang, and Philippe Blache. 2016. Testing apsyn against vector cosine on similarity estimation. In *Proceedings of PACLIC*.

Enrico Santus, Hongmin Wang, Emmanuele Chersoni, and Yue Zhang. 2018. A rank-based similarity metric for word embeddings. In *Proceedings of ACL*.

Harald Steck, Chaitanya Ekanadham, and Nathan Kallus. 2024. Is cosine-similarity of embeddings really about similarity? In *Companion Proceedings of the ACM Web Conference 2024*, pages 887–890.

William Timkey and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. In *Proceedings of EMNLP*, pages 4527–4546, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xueyang Wang, Qunlin Chen, Kaixiang Zhuang, Jingyi Zhang, Robert A Cortes, Daniel D Holzman, Li Fan, Cheng Liu, Jiangzhou Sun, Xianrui Li, et al. 2024. Semantic associative abilities and executive control functions predict novelty and appropriateness of idea generation. *Communications Biology*, 7(1):703.

Saeth Wannasuphoprasit, Yi Zhou, and Danushka Bollegala. 2023. Solving cosine similarity underestimation between high frequency words by ℓ_2 norm discounting. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8644–8652, Toronto, Canada. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Wenjing Yang, Adam E. Green, Qunlin Chen, Yoed N. Kenett, Jiangzhou Sun, Dongtao Wei, and Jiang Qiu. 2022. Creative problem solving in knowledge-rich contexts. *Trends in Cognitive Sciences*, 26(10):849–859.

Zhong Zhang, Chongming Gao, Cong Xu, Rui Miao, Qinli Yang, and Junming Shao. 2020. Revisiting representation degeneration problem in language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 518–527, Online. Association for Computational Linguistics.

Vitalii Zhelezniak, Aleksandar Savkov, April Shen, and Nils Y Hammerla. 2019. Correlation coefficients and semantic textual similarity. In *Proceedings of NAACL*.

Michael Zock. 2002. Sorry, What Was Your Name Again, or How to Overcome the Tip-of-the tongue Problem with the Help of a Computer? In *Proceedings of the SEMANET Workshop on Building and Using Semantic Networks*.

A Discussion on the Properties of Different WATs

Studies of semantic knowledge in vector spaces typically use two key metrics: *semantic similarity* and *semantic relatedness* (Fodor et al., 2023). The former means the degree of interchangeability between words based on their core meanings (Miller and Charles, 1991), as exemplified by *accept* and *receive* due to their overlapping meanings. In contrast, the latter encompasses broader conceptual connections, including functional, contextual, or psychological associations, even when words exhibit minimal semantic overlap (Gladkova et al., 2016). For instance, *air* and *plane* demonstrate high relatedness despite low similarity. These dimensions are rooted in lexical networks together, with different word association tasks highlighting distinct aspects.

Tasks 1 and 2 primarily assess semantic similarity, as they require models to identify the most semantically proximate word to a given stimulus. In Task 1, the *FIRST* response exhibits high interchangeability with the given stimulus (e.g., *receive* for *accept*), conforming to Miller and Charles (1991)'s definition of semantic similarity. While Task 2 employs an open-vocabulary paradigm, it requires the generation or selection of maximally similar words, maintaining its focus on direct meaning alignment. Both tasks prioritize paradigmatic relations (synonymy or near-synonymy).

Tasks 4 and 5 are relatedness-focused ones due to their emphasis on detecting implicit conceptual connections beyond semantic interchangeability. The triad tasks (Concrete-Abstract and Remote) evaluate models' ability to identify word pairs based on latent relational features. For instance, [*banana*, *cherry*, *pineapple*] in Task 4 (*banana* and *pineapple* are regarded as the most related concepts, but they have totally different denotations), and [*fence*, *mask*, *salt*] in Task 5 (the first two words are most related but non-interchangeable).

Different from the aforementioned tasks, Task 3 requires models to simultaneously make judgments on semantic similarity and relatedness, as illustrated through examples from the CogALex dataset (Rapp and Zock, 2014). The case of [*plenty*, *many*, *lots*, *around*, *leap* $\rightarrow$ *abound*] demonstrates similarity-driven processing, where identifying the *Target* depends on recognizing shared core meanings of quantitative abundance. In contrast, the example [*plane*, *fresh*, *water*, *breathe*, *force* $\rightarrow$ *air*]

reveals their internal relatedness through its web of diverse associations, including functional, ecological, physical, and perceptual connections.

Our findings echo prior work (Lenci et al., 2022; A et al., 2024) on the semantic representation capabilities of WEMs versus contextualized models (PLMs/LLMs). We found that the distinction between association tasks via semantic similarity and relatedness is highly significant as it offers a clearer framework for comparing architectures, emphasizing that human cognition seamlessly combines similarity and relatedness, while language models lag behind and show different limitations.

B Results of Multiple-Choice Association

Table 2: Accuracies (Acc.) and Frequencies of Incorrect Responses (*HAPAX*, *RANDOM*, and *OTHER*) in Task 1.

Types	Settings	Models	Acc. (%)	*HAPAX*	*RANDOM*	*OTHER*
WMEs	embeddings	word2vec	77.90	2,203	322	
		GloVe	79.31	2,072	292	
		GloVe-CC	80.28	2,092	161	
		FastText	82.07	1,904	145	
		FastText-CC	83.34	1,783	120	
PLMs	non-contextualized embeddings	BERT-base	58.26	3,150	1,621	
		BERT-large	52.81	3,409	1,985	
		GPT-2	34.23	3,215	4,303	
		GPT-2-xl	52.42	3,094	2,344	
		T5-small	65.89	2,801	1,097	
		T5-3B	67.05	2,944	822	
	contextualized embeddings	BERT-base	52.01	3,352	2,134	
		BERT-large	45.46	3,374	2,860	
		GPT-2	52.25	3,203	2,255	
		GPT-2-xl	62.55	2,666	1,615	
		T5-small	67.47	2,583	1,135	
		T5-3B	71.34	2,373	903	
LLMs	prompt	GPT-4.1	86.77	1,408	80	24
		DeepSeek-V3	86.72	1,420	84	14
		Qwen3	79.53	2,077	233	30

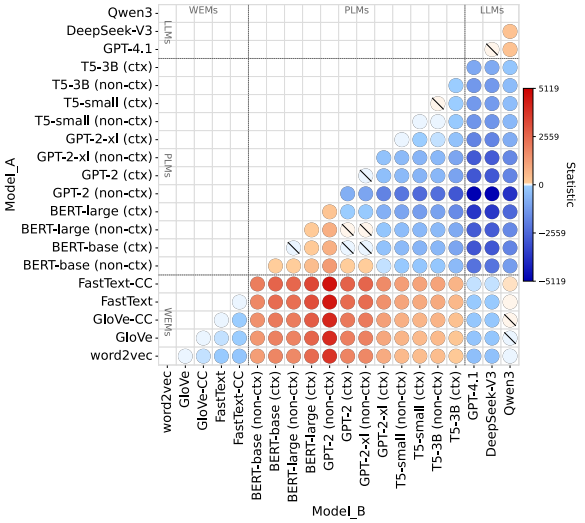

Figure 8: Pairwise McNemar's Tests on Task 1 ($p < 0.05$). Colored cells denote the significantly stronger models based on accuracies: red for Model_A and blue for Model_B. Dashes indicate non-significant differences.

C Results of Open-Vocabulary Association

Table 3: Top-1 Accuracies (Top-1 Acc.) and Mean Ranks with the Threshold of 4 (MR/4) in Task 2.

Types	Settings	Models	Top-1 Acc. (%)	MR/4
WMEs	embeddings	word2vec	4.59	3.76
		GloVe	5.78	3.69
		GloVe-CC	4.79	3.71
		FastText	5.14	3.70
		FastText-CC	5.49	3.66
PLMs	non-contextualized embeddings	BERT-base	3.19	3.85
		BERT-large	2.13	3.90
		GPT-2	0.78	3.97
		GPT-2-xl	0.86	3.96
		T5-small	4.99	3.76
		T5-3B	4.17	3.79
	contextualized embeddings	BERT-base	2.74	3.89
		BERT-large	2.13	3.90
		GPT-2	1.84	3.92
		GPT-2-xl	2.41	3.89
		T5-small	4.11	3.77
		T5-3B	2.12	3.87
LLMs	prompt	GPT-4.1	30.07	2.85
		DeepSeek-V3	35.56	2.69
		Qwen3	32.40	2.81

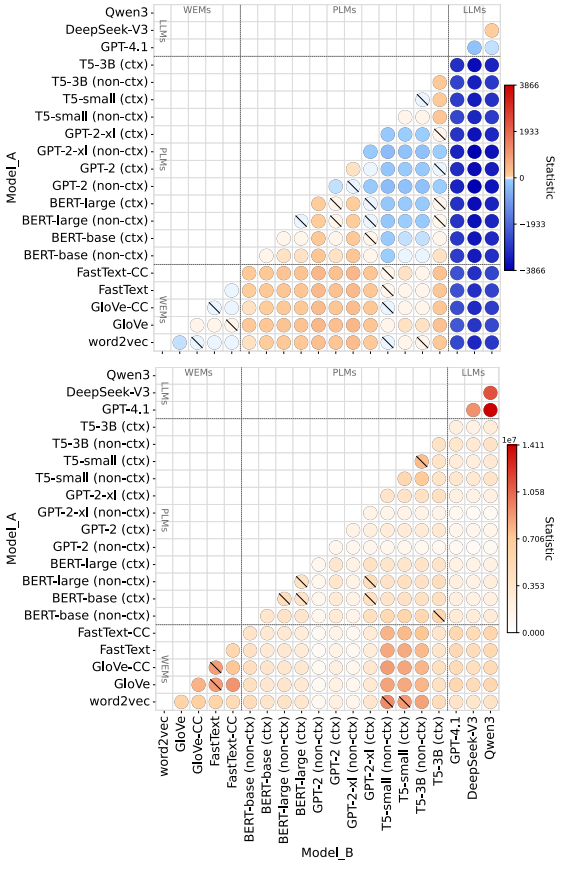

Figure 9: Pairwise McNemar's Tests (above) and Wilcoxon Signed-Rank Tests (below) on Task 2 ($p < 0.05$). For the plot above, colored cells denote the significantly stronger models based on top-1 accuracies: red for Model_A and blue for Model_B. For the below one, colored cells denote significant differences on *FIRST* ranks. Dashes indicate non-significant differences.

D Results of Reverse Association

Table 4: Top-1 Accuracies (Top-1 Acc.) and Mean Ranks with the Threshold of 4 (MR/4) in Task 3.

Types	Settings	Models	Top-1 Acc. (%)	MR/4
WMEs	embeddings	word2vec	16.99	3.28
		GloVe	16.27	3.34
		GloVe-CC	18.52	3.26
		FastText	21.26	3.14
		FastText-CC	24.14	3.03
PLMs	non-contextualized embeddings	BERT-base	5.53	3.78
		BERT-large	3.04	3.87
		GPT-2	0.25	3.99
		GPT-2-xl	1.04	3.95
		T5-small	10.90	3.57
		T5-3B	12.11	3.52
	contextualized embeddings	BERT-base	2.38	3.90
		BERT-large	3.34	3.87
		GPT-2	0.63	3.97
		GPT-2-xl	2.08	3.91
		T5-small	14.14	3.44
		T5-3B	13.51	3.43
LLMs	prompt	GPT-4.1	35.53	2.73
		DeepSeek-V3	29.37	2.95
		Qwen3	27.45	3.03

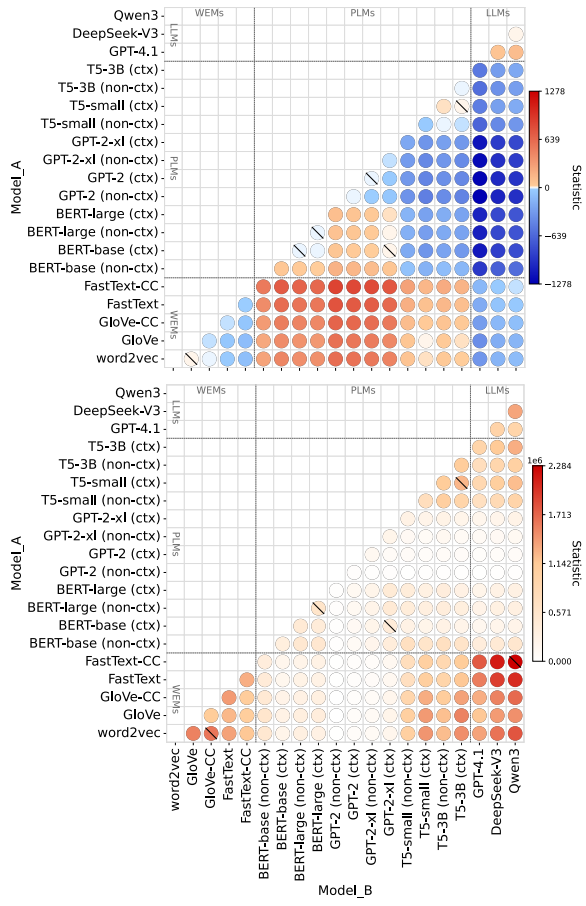

Figure 10: Pairwise McNemar's Tests (above) and Wilcoxon Signed-Rank Tests (below) on Task 3 ($p < 0.05$). For the plot above, colored cells denote the significantly stronger models based on top-1 accuracies: red for Model_A and blue for Model_B. For the below one, colored cells denote significant differences on *Target* ranks. Dashes indicate non-significant differences.

E Results of Concrete-Abstract Association

Table 5: Accuracies (Acc.) on Total (T), Concrete (C), and Abstract (A) datasets in Task 4.

Types	Settings	Models	T-Acc. (%)	C-Acc. (%)	A-Acc. (%)
WMEs	embeddings	word2vec	62.00	67.00	57.00
		GloVe	60.50	64.00	57.00
		GloVe-CC	66.50	77.00	56.00
		FastText	61.50	65.00	58.00
		FastText-CC	69.00	74.00	64.00
PLMs	non-contextualized embeddings	BERT-base	44.00	40.00	48.00
		BERT-large	41.50	39.00	44.00
		GPT-2	37.00	38.00	36.00
		GPT-2-xl	44.00	45.00	43.00
		T5-small	47.50	45.00	50.00
		T5-3B	46.50	43.00	50.00
	contextualized embeddings	BERT-base	41.00	34.00	48.00
		BERT-large	36.00	32.00	40.00
		GPT-2	37.50	38.00	37.00
		GPT-2-xl	54.00	54.00	54.00
		T5-small	52.00	47.00	57.00
		T5-3B	48.50	45.00	52.00
LLMs	prompt	GPT-4.1	62.00	65.00	59.00
		DeepSeek-V3	51.00	57.00	45.00
		Qwen3	51.50	54.00	49.00

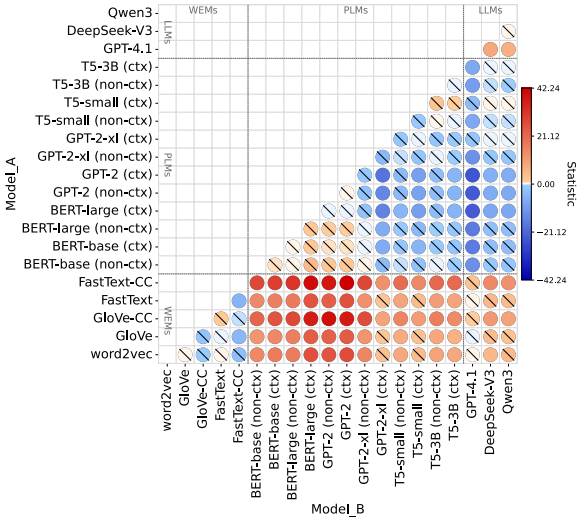

Figure 11: Pairwise McNemar's Tests on Task 4 ($p < 0.05$). Colored cells denote the significantly stronger models based on t-accuracies: red for Model_A and blue for Model_B. Dashes indicate non-significant differences.

F Results of Remote Association

G Ablation Studies on Prompting LLMs

We conducted exploratory experiments to examine how external (prompt design) and internal factors (temperature settings, reasoning modes) influence LLM performance across different WATs. Datasets applied here were randomly sampled from our main evaluation data as introduced in Table 1, with 200 items per task for Tasks 1-3, and full sets for Task 4 (200 items) and Task 5 (100 items).

Table 6: Accuracies (Acc.) in Task 5.

Types	Settings	Models	Acc. (%)
WMEs	embeddings	word2vec	62.00
		GloVe	65.00
		GloVe-CC	64.00
		FastText	66.00
		FastText-CC	63.00
PLMs	non-contextualized embeddings	BERT-base	36.00
		BERT-large	36.00
		GPT-2	44.00
		GPT-2-xl	48.00
		T5-small	49.00
		T5-3B	51.00
	contextualized embeddings	BERT-base	34.00
		BERT-large	34.00
		GPT-2	42.00
		GPT-2-xl	53.00
		T5-small	58.00
		T5-3B	57.00
LLMs	prompt	GPT-4.1	63.00
		DeepSeek-V3	46.00
		Qwen3	38.00

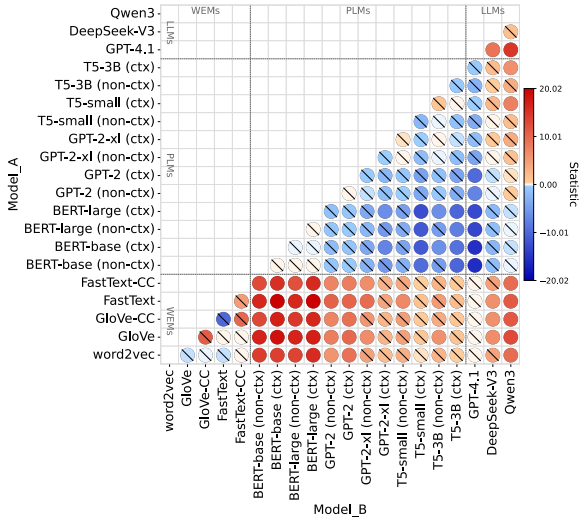

Figure 12: Pairwise McNemar's Tests on Task 5 ($p < 0.05$). Colored cells denote the significantly stronger models based on t-accuracies: red for Model_A and blue for Model_B. Dashes indicate non-significant differences.

G.1 Different Prompts: Zero-shot vs. Few-shot

Two sets of prompt instructions were designed by referring to those in the study of De Deyne et al. (2024), namely, 1) simple zero-shot prompts and 2) enhanced few-shot ones, detailed in Figures 13 to 18. The exemplars for few-shot prompts were sourced from established association norms such as EAT (Kiss et al., 1973), USF (Nelson et al., 2004), and SWOW (De Deyne et al., 2019), excluding any items overlapping with our evaluation datasets to prevent contamination. The temperature for this subexperiment was fixed at 0.01 and the reasoning mechanism was disabled to isolate prompt efficacy.

Results in Table 7 exhibit that detailed few-shot prompts consistently enhance LLM performance except in Task 2. For instance, GPT-4.1 achieves over 5% accuracy gains in Tasks 3 and 4, and DeepSeek-V3 and Qwen3 show even more than 10% improvements. However, the benefits of detailed few-shot prompting are model- and task-dependent, as evidenced by GPT-4.1's performance in Task 2, where such prompts had marginal or even negative effects.

G.2 Different Temperatures: 0.01 vs. 0.5 vs. 1

The temperature is a built-in parameter of LLMs to control the randomness and the so-called creativity of their outputs (Peeperkorn et al., 2024). It spans [0, 2] with higher values corresponding to increased diversity, while lower values yield more focused and deterministic outputs. It is assumed to have effects on models' semantic association capabilities, potentially mapping cognitive factors in human associative behavior. Therefore, we conducted subexperiment on comparing three temperatures: 0.01, 0.5, and 1 with simple zero-shot prompts and without the thinking mode.

Although the current test was limited to half of the full temperature range, Table 8 demonstrates two key observations: 1) Temperature effects vary across models and tasks, such as, GPT-4.1 achieves optimal performance at 0.01 and 0.5, DeepSeek-V3 benefits most from 0.5, Qwen3 performs better at 0.5 and 1, and Tasks 2 and 3 show robustness to 0.5 compared to other tasks; 2) Performance differences induced by different temperatures remain subtle (less than 5%) across all assessed models and tasks.

G.3 Different Modes: Standard vs. Reasoning

To investigate potential advantages of reasoning mechanisms, we conducted a subexperiment on Qwen3 with reasoning activation as the only variable, using zero-shot prompts and a fixed temperature of 0.01. Surprisingly, the reasoning is not advantageous in all WATs. Notably, in Tasks 2 and 4—abstract word pairing, enabling reasoning may lead to overthinking and hence misjudgments in semantic similarity and relatedness assessments.

Together above results unveil the versatility of the human associative ability, which cannot be fully reproduced by LLM configurations.

Table 7: Comparisons of LLM Results across Different Prompt Strategies. Boldface values indicate the highest performance achieved by the model on a given task across all strategies.

Tasks	Metrics	GPT-4.1		DeepSeek-V3		Qwen3	
		zero-shot	few-shot	zero-shot	few-shot	zero-shot	few-shot
Task 1	Acc. (%)	90.00	**91.00**	89.50	**90.50**	84.50	**89.50**
Task 2	Top-1 Acc. (%)	**31.00**	30.50	35.50	**36.50**	32.50	**35.00**
	MR/4	**2.75**	2.76	2.68	**2.53**	2.78	**2.61**
Task 3	Top-1 Acc. (%)	32.50	**37.00**	26.00	**37.50**	23.00	**34.00**
	MR/4	2.86	**2.72**	3.03	**2.71**	3.19	**2.82**
Task 4	T-Acc. (%)	62.00	**67.50**	50.50	**75.00**	51.50	**61.50**
	C-Acc. (%)	65.00	**66.00**	56.00	**80.00**	54.00	**58.00**
	A-Acc. (%)	59.00	**69.00**	45.00	**70.00**	49.00	**65.00**
Task 5	Acc. (%)	63.00	**68.00**	46.00	**60.00**	38.00	**43.00**

Table 8: Comparisons of LLM Results across Different Temperature Settings. Boldface values indicate the highest performance achieved by the model on a given task across all settings.

Tasks	Metrics	GPT-4.1			DeepSeek-V3			Qwen3		
		0.01	0.5	1	0.01	0.5	1	0.01	0.5	1
Task 1	Acc. (%)	**90.00**	89.50	88.50	**89.50**	88.50	88.50	84.50	**86.50**	86.00
Task 2	Top-1 Acc. (%)	31.00	**32.00**	30.00	35.50	**37.00**	34.50	32.50	**35.00**	32.00
	MR/4	2.80	**2.78**	2.81	2.68	**2.60**	2.66	2.78	**2.71**	**2.71**
Task 3	Top-1 Acc. (%)	32.50	**36.00**	33.00	26.00	**31.50**	29.00	23.00	**26.50**	**26.50**
	MR/4	2.83	**2.76**	2.80	3.03	**2.93**	2.98	3.19	3.11	**3.06**
Task 4	T-Acc. (%)	**62.00**	57.00	59.00	51.00	**51.50**	51.00	51.50	52.50	**52.50**
	C-Acc. (%)	**65.00**	58.00	63.00	**57.00**	56.00	**57.00**	54.00	**55.00**	52.00
	A-Acc. (%)	**59.00**	56.00	55.00	45.00	**47.00**	45.00	49.00	50.00	**53.00**
Task 5	Acc. (%)	63.00	**67.00**	63.00	**46.00**	43.00	45.00	38.00	**39.00**	35.00

Table 9: Comparisons of Qwen3 Results with Different Thinking Modes. Boldface values indicate the highest performance achieved by the model on a given task within two modes.

Tasks	Metrics	Qwen3	
		standard	reasoning
Task 1	Acc. (%)	84.50	**89.00**
Task 2	Top-1 Acc. (%)	**32.50**	28.50
	MR/4	**2.78**	2.84
Task 3	Top-1 Acc. (%)	23.00	**28.00**
	MR/4	3.19	**3.01**
Task 4	T-Acc. (%)	51.50	**52.00**
	C-Acc. (%)	54.00	**57.00**
	A-Acc. (%)	**49.00**	47.00
Task 5	Acc. (%)	38.00	**45.00**

***** Simple Zero-Shot Prompt *****

System: You are a native speaker of English participating in a psycholinguistic test about word meaning.

User:

**** Task 1 ****
- You will be presented with a list of words separated by "-" that consists of a cue (the first one) and three candidates.
- You are asked to choose one target candidate from the three given candidates that is most closely associated with the cue.
- Remember to only respond with one target candidate word and do not further elaborate on your response.
- Format your response as json: {cue-candidate1-candidate2-candidate3: target candidate}.

- Input:{input}
- Output:

**** Task 2 ****
- You will be presented with a cue word.
- You are asked to output a list consisting of thirty words that are most closely associated with the cue word.
- Rank all thirty words according to their strength of association with the cue words in descending order.
- Remember to only respond with one list of ranked words and do not further elaborate on your response.
- Format your response as json: {cue: [response1, response2, ..., response30]}.

- Input:{input}
- Output:

**** Task 3 ****
- You will be presented with five hint words seperated by "-".
- You are asked to output a list consisting of thirty words that are most closely associated with the given five hint words
- Rank all thirty words according to their strength of association with all five hint words in descending order.
- Remember to only respond with one list of ranked words and do not further elaborate on your response.
- Format your response as json: {word1-word2-word3-word4-word5: [response1, response2, ..., response30]}.

- Input:{input}
- Output:

**** Task 4 ****
- You will be presented with a triplet of words that can be marked as "A", "B", "C" in sequence.
- You are asked to output a list consisting of three alphabetic pairs that are ranked with the strength of word association within their corresponding word pairs.
- Remember to only respond with one list of ranked pairs and do not further elaborate on your response.
- Format your response as json: {wordA-wordB-wordC:["AB", "BC", "AC"]}.

- Input:{input}
- Output:

**** Task 5 ****
- You will be presented with a triplet of words that can be marked as "A", "B", "C" in sequence.
- You are asked to output a list consisting of three alphabetic pairs that are ranked with the strength of word association within their corresponding word pairs.
- Remember to only respond with one list of ranked pairs and do not further elaborate on your response.
- Format your response as json: {wordA-wordB-wordC:["AB", "BC", "AC"]}.

- Input:{input}
- Output:

Figure 13: Simple Zero-shot Prompt Instructions for LLMs across Five WATs.

***** Enhanced Few-Shot Prompt – Task 1 *****

System: You are functioning as a native English speaker with unimpaired lexical access capabilities participating in a controlled psycholinguistic experiment. Your task requires making semantic association judgments through systematic cognitive operations.

User:

- This test is called "Multiple-Choice Word Association", designed to measure your ability to associate words with each other from a restricted list.
- You will be presented with a list of words separated by "-" that consists of a cue (priming lexical item) in the first position and three candidates (a triplet of potential association targets) in the second to fourth positions.
- You are asked to choose one target candidate from the three given candidates that is most closely associated with the cue in consideration of semantic (denotative overlap), conceptual (connotative alignment) and cognitive (co-occurrence frequency) association strengths.
- Remember to only respond with one target candidate word and do not further elaborate on your response.
- Format your response as json: {cue-candidate1-candidate2-candidate3: target candidate}.

- Here are some examples:
- {
- "input": "fibre-moral-glass-cries",
- "output": {"fibre-moral-glass-cries": "glass"}
- },
- {
- "input": "alert-jagger-inactive-awake",
- "output": {"alert-jagger-inactive-awake": "awake"}
- },
- {
- "input": "poison-arsenic-milford-shakespeare",
- "output": {"poison-arsenic-milford-shakespeare": "arsenic"}
- }

- Input:{input}
- Output:

Figure 14: Extended Few-shot Prompt Instructions for LLMs in Task 1: Multiple-Choice Association.

***** Enhanced Few-Shot Prompt – Task 2 *****

System: You are functioning as a native English speaker with unimpaired lexical access capabilities participating in a controlled psycholinguistic experiment. Your task requires making semantic association judgments through systematic cognitive operations.

User:

- This test is called "Open-Vocabulary Word Association", designed to measure your ability to perform deep semantic network traversal.
- You will be presented with a cue word.
- You are asked to output a list consisting of thirty words that are most closely associated with the cue word in consideration of semantic (denotative overlap), conceptual (connotative alignment) and cognitive (co-occurrence frequency) association strengths.
- Rank all thirty words according to their strength of association with the cue words in descending order.
- Remember to only respond with one list of ranked words and do not further elaborate on your response.
- Format your response as json: {cue: [response1, response2, ..., response30]}.

- Here are some examples:
- {
- "input": "fibre",
- "output": {{"fibre":["food", "cloth", "cereal", "fabric", "optic", "diet", "cotton", "glass", "poop", "internet", "bread", "bran", "optics", "material", "hair", "thread", "health", "strength", "rope", "wheat", "clothes", "grain", "wool", "clothing", "textile", "wire", "healthy", "paper", "digestion", "laxative"]}
- },
- {
- "input": "alert",
- "output": {{"alert":["awake", "alarm", "red", "aware", "fire", "siren", "warning", "ready", "warn", "danger", "attention", "attentive", "coffee", "light", "notice", "conscious", "morning", "observant", "sharp", "tense", "lights", "know", "keen", "emergency", "high", "caution", "mind", "tell", "reminder", "vigilant"]}
- },
- {
- "input": "poison",
- "output": {{"poison":["death", "Ivy", "kill", "apple", "arsenic", "liquid", "bottle", "bad", "snake", "drink", "venom", "deadly", "green", "rat", "dart", "dangerous", "chemical", "frog", "danger", "sickness", "mushroom", "murder", "toxic", "food", "fish", "band", "die", "rats", "evil", "crossbones"]}
- }

- Input:{input}
- Output:

Figure 15: Extended Few-shot Prompt Instructions for LLMs in Task 2: Open-Vocabulary Association.

***** Enhanced Few-Shot Prompt – Task 3 *****

System: You are functioning as a native English speaker with unimpaired lexical access capabilities participating in a controlled psycholinguistic experiment. Your task requires making semantic association judgments through systematic cognitive operations.

User:

- This test is called "Reverse Word Association", designed to measure your ability to address the word access problem by predicting the trigger based on the commonality between given words.
- You will be presented with five hint words separated by "-".
- You are asked to output a list consisting of thirty words that are most closely associated with the given five hint words in consideration of semantic (denotative overlap), conceptual (connotative alignment) and cognitive (co-occurrence frequency) association strengths.
- Rank all thirty words according to their strength of association with all five hint words in descending order.
- Remember to only respond with one list of ranked words and do not further elaborate on your response.
- Format your response as json: {word1-word2-word3-word4-word5: [response1, response2, ..., response30]}.

- Here are some examples:
- {
- "input": "together-joined-effort-harvester-honours",
- "output": {{"together-joined-effort-harvester-honours":["combined", "mixed", "mix", "added", "two", "bound", "sum", "multiple", "joint", "total", "linked", "stuck", "join", "harvester", "pair", "words", "with", "connected", "baking", "score", "paired", "grouped", "eggs", "combine", "associated", "amalgamation", "amalgamated", "one", "attached", "integration"]}}
- },
- {
- "input": "centre-end-earth-East-man",
- "output": {{"centre-end-earth-East-man":["middle", "child", "average", "central", "between", "median", "name", "age", "school", "class", "finger", "top", "last", "bottom", "waist", "road", "medium", "ages", "half", "ground", "compromise", "start", "stuck", "sister", "surrounded", "sandwich", "muddle", "first", "amid", "inside"]}}
- },
- {
- "input": "to-should-not-must-nought",
- "output": {{"to-should-not-must-nought":["ought", "zero", "need", "will", "would", "obligation", "might", "guilt", "obligated", "eight", "right", "could", "require", "shall", "thought", "responsibility", "proper", "old", "fashioned", "nothing", "can", "caught", "grandfather", "go", "duty", "supposed"]}}
- }

- Input:{input}
- Output:

Figure 16: Extended Few-shot Prompt Instructions for LLMs in Task 3: Reverse Association.

***** Enhanced Few-Shot Prompt – Task 4 *****

System: You are functioning as a native English speaker with unimpaired lexical access capabilities participating in a controlled psycholinguistic experiment. Your task requires making semantic association judgments through systematic cognitive operations.

User:

- This test is called "Concrete and Abstract Word Association", designed to measure your ability to capture and bridge the meaning and relationship between the given concrete or abstract words.
- You will be presented with a triplet of words separated by "-", which can be marked as "A", "B", "C" in sequence.
- You are asked to output a list consisting of three alphabetic pairs that are ranked with the strength of word association within their corresponding word pairs in consideration of semantic (denotative overlap), conceptual (connotative alignment) and cognitive (co-occurrence frequency) association strengths.
- Remember to only respond with one list of ranked pairs and do not further elaborate on your response.
- Format your response as json: {wordA-wordB-wordC:["AB", "BC", "AC"]}.

- Here are some examples:
- {
- "input": "apple-fruit-pie",
- "output": {{"apple-fruit-pie":["AB", "AC", "BC"]}}
- }
- {
- "input": "vibe-aura-felling",
- "output": {{"vibe-aura-felling":["AC", "AB", "BC"]}}
- }
- {
- "input": "foresight-intuition-cognition",
- "output": {{"foresight-intuition-cognition":["BC", "AB", "AC"]}}
- }

- Input:{input}
- Output:

Figure 17: Extended Few-shot Prompt Instructions for LLMs in Task 4: Concrete-Abstract Association.

***** Enhanced Few-Shot Prompt – Task 5 *****

System: You are functioning as a native English speaker with unimpaired lexical access capabilities participating in a controlled psycholinguistic experiment. Your task requires making semantic association judgments through systematic cognitive operations.

User:

- This test is called "Remote Word Association", designed to measure your ability to capture and bridge the meaning and relationship between the given weakly-related words.
- You will be presented with a triplet of words separated by "-", which can be marked as "A", "B", "C" in sequence.
- You are asked to output a list consisting of three alphabetic pairs that are ranked with the strength of word association within their corresponding word pairs in consideration of semantic (denotative overlap), conceptual (connotative alignment) and cognitive (co-occurrence frequency) association strengths.
- Remember to only respond with one list of ranked pairs and do not further elaborate on your response.
- Format your response as json: {wordA-wordB-wordC:["AB", "BC", "AC"]}.

- Here are some examples:
- {
- "input": "hate-morning-test",
- "output": {{"hate-morning-test":["BC", "AC", "AB"]}}
- }
- {
- "input": "bear-hat-angel",
- "output": {{"bear-angel-hat":["BC", "AB", "AC"]}}
- }
- {
- "input": "shot-heat-darkness",
- "output": {{"shot-heat-darkness":["AB", "AC", "BC"]}}
- }

- Input:{input}
- Output:

Figure 18: Extended Few-shot Prompt Instructions for LLMs in Task 5: Remote Association.