Fine-tuned Large Language Models predict human word associations and improve performance across lexical and semantic processing tasks.

Simon De Deyne (simon.dedeyne@unimelb.edu.au)[1] **Sukhai Huang**[2] **Lea Frermann**[3] **Chunhua Liu**[3]

[1]School of Psychological Sciences, University of Melbourne, Australia

[2]Faculty of Information Technology, Monash University, Australia

[3]School of Computing and Information Systems, University of Melbourne, Australia

Abstract

Recent studies have shown that LLMs can predict various aspects of semantic cognition when prompted with instructions that closely mimic those of humans. This study examines to what degree LLMs can predict human word associations, asking two previously underexplored questions: (1) Does fine-tuning LLMs on a limited amount of human data increase prediction quality? And (2) Does a more human-like association prediction also facilitate predicting other behaviors such as lexical processing and similarity judgments? Our findings suggest that the answer to both questions is yes. We first show that word associations can be predicted more accurately after fine-tuning. This was especially evident for the strongest associates, as the median rank of the predicted first associate was 1, a substantial improvement over previous approaches. Fine-tuning also enhances reflection of human response biases (e.g., favoring short, frequent responses). This close alignment with human associations further benefited the prediction of lexical processing and similarity judgments. Our results have implications on using LLMs fine-tuned on word associations in studies of a wide range of cognitive behavior, and on capturing factors that drive diversity in human associations.

Keywords: word associations, Large Language Models, biases, similarity, lexical decision

Introduction

Word associations have long fascinated cognitive psychologists, as the simple task of listing words that come to mind in response to a cue reveals how meaning is structured and constrains theories of semantic representation. Word association norms, in the form of semantic networks or other measures, also explain a wide range of human behaviors related to how central words are in lexical processing tasks, how related meanings are, and what factors determine cued recall, recognition, and false memories (e.g., Nelson, McKinney, Gee, & Janczura, 1998; Roediger, Watson, McDermott, & Gallo, 2001).

An enduring question is how word associations relate to language, both in how a person's associations shape language over time and in how statistical associations in the linguistic environment are encoded in the mental lexicon. There's a broad consensus that word associations are the product of direct or indirect co-occurrence in language (e.g., Deese, 1965; Landauer & Dumais, 1997; Griffiths, Steyvers, & Tenenbaum, 2007), but whether this account is sufficient and to what degree non-linguistic factors like mental imagery play a role remains unclear (De Deyne, Navarro, Collell, & Perfors, 2021).

Evidence for the role of language comes from studies that determine to what degree word associations can be predicted from how words co-occur in language corpora by encoding this information in text-based word embeddings. Cattle and Ma (2017), for example, report a correlation between human and model estimates of word association strength of .36. Others combined embedding models with extra lexical features and predicted human associations using supervised learning, resulting in a correlation of .42 (Hayashi, 2016). Another line of work compared topic models to predict associations. The median rank of the associations produced by topic models in Griffiths et al. (2007) to match the first human association was 18. In Nematzadeh, Meylan, and Griffiths (2017), the highest correlation among several models was .27, with the best performance showing that the median rank of the first human response was 11, a result closely replicated by Matusevych and Stevenson (2018), who reported a rank of 12 for embedding models, which outperformed topic models. Together, these findings demonstrate the role of language in word associations, although much of the variance is unaccounted for.

Recent work has shown that Large Language Models (LLMs) trained on massive datasets that capture longer contexts outperform traditional word embeddings and topic models in predicting multiple aspects of semantic cognition (Trott, 2024). Notably, LLMs also predict judgments often assumed to depend on perceptual or affective experience, such as color and affective meaning (Trott, 2024; Marjieh, Sucholutsky, van Rijn, Jacoby, & Griffiths, 2023). This is consistent with theories that regard language as a reliable proxy for perceptual, sensorimotor, and affective information (Louwerse, 2011). This suggests that LLMs might fill a deficiency in previous association prediction studies, as word-based embeddings and topic models only partially capture the imagery and affective information encoded in word associations (De Deyne et al., 2021). Crucially, strong LLM performance does not provide direct insight in whether LLMs and humans rely on similar internal representations, or answers the question about the contribution of linguistic and non-linguistic information, as proposed by dual-coding and embodiment accounts (Paivio, 1986; Barsalou, Simmons, Barbey, & Wilson, 2003). However, if it turns out that associations can be near-perfectly predicted by an LLM, this still provides valuable insights into the upper bounds of association knowledge encoded in language.

The superior performance of LLMs on a variety of human semantic judgements has inspired several studies to generate word associations by prompting Large Language Models

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

(LLMs) using instructions that matched the human task. For example, Abramski, Improta, Rossetti, and Stella (2025) investigated whether graphs built from these associations exhibited a similar structure to empirical graphs from human associations. Xiao, Duan, Haslett, and Cai (2025) used a similar approach to derive an LLM-based semantic network to explain lexical processing and semantic similarity. Despite encouraging results, questions remain about how these models perform compared to previous approaches, as word association prediction was not the primary focus of these studies. Furthermore, neither study controlled the parameters that generate responses, such as model temperature, which might decrease prediction accuracy when responses are too diverse due to a high temperature or inflate it when only a small number of responses are predicted due to low temperature values. An additional complication is that the most promising results are obtained with proprietary, closed-source models such as GPT-4, making a systematic investigation of model parameters difficult and costly. Another limitation of existing work is that prompting without additional tuning, although successful across a wide range of semantic tasks in previous work (Trott, 2024), might fail to capture important biases that characterize how humans generate associations, such as the principle of least effort, which leads to a preference for short, concrete, and common words. Similarly, previous work has shown that most LLMs have their own biases (see Bender, Gebru, McMillan-Major, & Shmitchell, 2021), which, due to their training, might not entirely reflect the demographic composition of the participants in large-scale crowd-sourced word association studies like the Small World of Words project (De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019).

The goal of this study is to determine how close we can approach human-level word association generation by optimizing LLMs with task-specific fine-tuning and parameter optimization. We also study whether a close match translates to improved prediction on external tasks. To do this, we will use English word association data from the Small World of Words (SWOW) project and replicate the results in the original study by De Deyne et al. (2019). To ensure that our results are not due to data leakage, we supplement our analyses with new, unpublished data.

Study 1: Predicting Human Word Associations

Study 1 aims to evaluate the effect of fine-tuning relative to a prompting-only baseline on predicting word associations, while matching the number and variety of responses by calibrating the model temperature. The results will cover both predicting word association responses and determining to what degree human biases that characterize associative responses are acquired through the process of fine-tuning.

Method and Procedure.

Word association data. The word association data consisting of 12,282 cues used for training were taken from the English Small World of Words project (De Deyne et al., 2019), henceforth referred to as SWOW-2018. In this study, participants performed a continued association task providing three responses (R1,R2,R3) to a list of cues. To reduce bias in response variability, only associations by participants who lived in the US and speak American English were included which reduced the data from 3.4M responses to 1.7M (50.1% of the data). The data were split into disjoint sets of test and training cues. The training set consisted of a random sample of 20% (2,456) cues from SWOW-2018. The training data consisted of the all the raw participant-level responses (R1,R2,R3) to a list of cues.[1] To address the possibility of LLMs being exposed to these data during pre-training, we also included unpublished recent data from the ongoing data collection of the SWOW project (SWOW-2023) for a set of 2,163 cues and 331,451 responses to be used for further testing.

Fine-tuning The fine-tuned model (FT) was based on Llama-3.1-8B-Instruct, which was originally trained on +15 trillion tokens from several online sources up to December 2023, followed by Reinforcement Learning from Human Feedback (RLHF) optimisation (Huang et al., 2024). We performed supervised fine-tuning of Llama using next-token prediction conditioned on the preceding context (Ouyang et al., 2022). During fine-tuning, next-token prediction was performed on inputs formatted to match the original SWOW prompts (see next paragraph), so that, given a cue, the model learned to generate the same three associated responses produced by human participants.

Specifically, we used Low-Rank Adaptation (LoRA; Hu et al., 2022) for parameter-efficient fine-tuning, which keeps the original weights fixed while learning a small set of additional "adapter" parameters that nudge the model's behavior. We used standard LoRA hyperparameters (rank $r = 64$, scaling $\alpha = 256$), an AdamW gradient-descent optimizer, and a cosine learning rate scheduler. The learning rate was set to 10^{-5} (increasing from 0 to 10^{-5} at the first 10% training steps). Fine-tuning was performed on a single A100 GPU with a batch size 18 and gradient accumulation over two steps. Since the validation loss plateaued early, we trained the model for one epoch, which took 218 minutes to complete.

Association generation for the fine-tuned and baseline model. Word associations were generated for all SWOW-2018 and SWOW-2023 cues using the fine-tuned model (FT) and a baseline (Base), corresponding to the same LLama-3.1-8-Instruct model without any fine-tuning. At inference, following the same procedure used in training (three responses per cue) and in the human study (100 participants per cue), we prompted each LLM to generate three associations per cue across 100 independent repetitions. The generation process used stochastic sampling with variable temperature to ensure that outputs across repetitions were not identical and approximated the variability observed in large human samples. The

[1] Training with more than 20% of the cues was also performed but for reasons of space, we will only cover this briefly in the discussion.

prompts matched those given to human participants:

> **Background:** On average, an adult knows about 40,000 words, but what do these words mean to people like you and me? You can help scientists understand how meaning is organized in our mental dictionary by playing the game of word associations. This game is easy: Just give the first three words that come to mind for a given cue word.
>
> **Output Format:** Output your response in the following format: response1, response2, response3 Do not provide any additional context or explanations.
>
> **Cue word:** `{cue_word}`

Temperature calibration. Because stochastic decoding in LLMs is controlled by the temperature parameter, we generated association responses for multiple temperatures (0.5, 1.0, 1.5, and 2.0) to assess which values provided the best match with the human data. To do so, we compared the human vocabulary growth curve derived from the SWOW-2018 responses (test and training) which measures the expected vocabulary size as a function of the number of tokens and compared this curve with LLM vocabulary curves for different temperature values (0.5,1,1.5,2). Based on initial fits, we refined the range of temperatures and added 1.25 and 2.5 for the Base model and 1.1, 1.2, 1.3 and 1.4 for the FT model. As can be seen from the curves in Figure 1, temperature had a different effect in the Base and FT models, resulting in the best matching values being 2 for the Base model and 1.4 for the FT model, which will be used in all subsequent analyses.

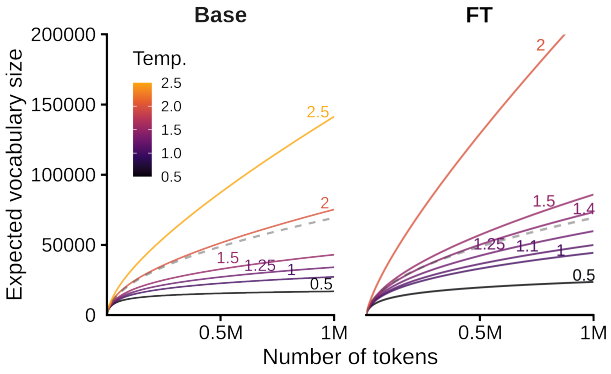

Figure 1: Human vocabulary growth curves from human associations (dashed line) compared to LLM-generated curves for the baseline (Base) and fine-tuned model (FT) at different temperatures.

Results

Responses were screened to retain only valid outputs (comma-separated, exactly three associations, no refusals or extraneous text). A total of 18 cues, mostly taboo words, with over 10% of missing response data, were removed. Spelling errors were rare, and further processing was limited to lowercasing both human and generated responses. Performance in the following sections excluded any cue for training and was based on test cues only. Examples that illustrate the effect of fine-tuning are shown in Table 1.

Agreement with human associations To aid comparison with previous work, we include the median rank of the first associate based on the test cues and matching the number of responses per cue. In cases where a response was not available in the set of responses in the LLM, the rank was set to the maximum rank of that cue + 1 (e.g. a cue with 50 different responses would get rank 51). Dense ranks were used, assigning identical ranks to tied values and avoiding gaps in rank. Across all cues, the median rank for the first association was 3 for the Base model and 1 for the FT model. The same results were obtained when split by cue-set (SWOW-2018, SWOW-2023), except for the FT model, where the median rank was 2 for the smaller SWOW-2023 dataset.

Next, we calculated the average correlations between the response-strength distributions. For each cue, we merged all responses and replaced missing responses in both the human and LLM data to have strengths of 0. For the Base model the correlations with SWOW-2018 and SWOW-2023 were respectively $r(9821) = .453$ (Q1 = .270, Q3 = .650) and $r(2162) = .462$ (Q1 = .286, Q3 = .660). The corresponding FT values were $r(9821) = .691$ (Q1 = .595, Q3 = .821) and $r(2162) = .646$ (Q1 = .536, Q3 = .787).

For interpretability, we assessed cue-level reliability using the Spearman–Brown split-half method, based on two randomly sampled partitions of human association responses, which provides an upper bound on explainable variance. For SWOW-2018, the average reliability over cues was $r_{sb} = .718$ for both trained and test sets, nearly identical to the reliability obtained for SWOW-2023 ($r_{sb} = .719$) and close the values obtained for the FT model.

Distributional and Lexico-semantic bias Word associations are subject to several types of human biases, which provide insight into the processes and underlying representations, such as the principle of least effort, developmental stages, and varying use of both verbal and imagery-driven responding. For example, responses tend to be more concrete, more frequent, and more often nouns than what would be expected from language statistics (Cramer, 1968). To determine whether fine-tuning acquires these biases, we summarised the response properties of each cue in terms of their valence and arousal (Warriner, Kuperman, & Brysbaert, 2013), concreteness (Brysbaert, Warriner, & Kuperman, 2014), age-of-acquisition (AoA, Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012), word length, log-transformed word frequency, and proportion of noun responses using the SUBTLEX-US norms for the latter two (Brysbaert, New, & Keuleers, 2012). Distributional properties included response entropy H, out-degree (k_{out}), and two network metrics: clustering coefficient (CC) and the average shortest direct path length (incoming paths: $ASLP_{in}$, outgoing paths: $ASLP_{out}$) for each cue (see Abramski et al., 2025, for a similar approach). Similar to

Table 1: Top five associates for two concrete and two abstract example cues to illustrate how fine-tuning (FT) improves over the Base model. Color indicates the rank of each response in the human distribution (black = smallest rank). For the Base model, all correlations with SWOW response distributions are smaller than .10, for the FT model, they are larger than .75.

Rank	**food**			**raven**			**anti**			**justice**		
	Base	FT	SWOW	Base	FT	SWOW	Base	FT	SWOW	Base	FT	SWOW
1	pizza	eat	eat	intelligence	bird	black	vaccine	against	against	fairness	judge	court
2	salad	meal	good	mystery	black	bird	virus	negative	establishment	equality	court	law
3	restaurant	hungry	sustenance	midnight	crow	crow	antibody	opposite	opposite	morality	fair	fair
4	apple	hunger	meal	intelligent	poe	poe	gravity	war	con	law	law	judge
5	sandwich	drink	cook	death	murder	poem	bacterial	social	no	laws	right	peace

De Deyne et al. (2019), the network metrics were derived from a directed weighted graph that included nodes in the largest strongly connected component, and where source nodes represent cues and target nodes represent responses that were among the set of cue words. Edge-weights represent conditional association strength. To be able to compare distributional properties directly, the number of LLM-generated responses per cue were matched to take into account incomplete data due to participants not knowing a word or being unable to generate R2 or R3.

The results in Table 2 show the averages over SWOW-2018 and SWOW-2023 test cues $\bar{X}$ and the Pearson correlation between the model values per cue with those obtained from human associations. To summarise agreement, we used ρ_C, Lin's concordance correlation coefficient (Lawrence & Lin, 1989). This measure accounts for both precision (the correlation between human and model-based values) and accuracy (whether LLM averages match), with 1 indicating perfect concordance. Lin's ρ_C was higher for FT across all measures except the clustering coefficient *CC*, which was on par with the baseline model (.41 Base, .40 FT), suggesting that the process of fine-tuning results in better-aligned human biases.

To further demonstrate how the FT model captured key human biases, we replicated the analysis of the top-10 network hubs reported in De Deyne et al. (2019), defined as the words with the highest in-strength calculated by summing the strengths for each response word across all cues. Importantly, these hubs are distinct from the most frequent words in language but are near-universal across word associations in different languages. Across all words, and consistent with Table 2, the correlation with SWOW-based instrength was higher for in-strength from the FT model $r(91002) = .972$, compared to the Base model $r(91002) = .855$. Focusing on the hubs, the human top-10, ranked by in-strength (*money, water, food, car, music, green, red, sad, sex, old*) were closely align with those obtained from FT (*money, water, food, car, music, red, school, love, green, sad*). Notably, the first five items are identical and preserve the exact rank ordering. In contrast, the Base model captures only some of these network hubs (*family, music, pain, water, money, ocean, computer, beach, car, happiness*).

Table 2: Agreement between the Llama-base (Base) and Llama-FT (FT) and SWOW (SW) associations across lexico-semantic, distributional and graph-based measures. $\bar{X}_{SW}$ and $\bar{X}_{LLM}$ denote mean values. Results show the means, Pearson correlation r and ρ_c Lin's concordance coefficient (width of 95% CI for r and ρ_c all below .04, highest value in bold).

		Base			FT		
Measure	$\bar{X}_{SW}$	$\bar{X}_{LLM}$	r	ρ_c	$\bar{X}_{LLM}$	r	ρ_c
Word length	6.19	7.22	.71	.43	6.03	.83	**.80**
Noun prob	0.67	0.57	.80	.70	.59	.91	**.91**
log(WF)	3.03	2.71	.72	.51	3.13	.85	**.81**
AoA	6.62	7.38	.82	.66	6.46	.89	**.88**
Valence	5.30	5.45	.92	.87	5.30	.96	**.94**
Arousal	3.30	3.46	.91	.88	3.32	.96	**.94**
Concreteness	2.86	2.92	.91	.88	2.92	.96	**.94**
H	5.29	4.60	.29	.17	4.32	.70	**.27**
k_{out}	46.1	67.6	.32	.16	55.9	.67	**.47**
CC	0.11	0.16	.61	**.41**	0.19	.75	.40
$ASPL_{in}$	2.97	3.65	.52	.24	3.38	.70	**.47**
$ASPL_{out}$	2.98	3.66	.28	.01	3.38	.64	**.04**

Study 2: Predicting human behavior with LLM-generated word association networks

Mental representations encoded as association-based semantic networks explain various aspects of lexical and semantic processing (Kumar, Steyvers, & Balota, 2022). However, it is unclear whether LLM-based associations show the same level of consistency with these behavioral tasks. To address this question, we replicated the findings described in De Deyne et al. (2019) by considering both lexical processing and similarity judgements to determine how fine-tuning affects model and an empirical SWOW network from De Deyne et al. (2019) predict these external (but related) tasks.

Lexical Reaction Time Tasks

A robust finding in psycholinguistics is that certain words are recognized or named faster than others because they are more frequent, shorter, or otherwise more salient (Balota et al., 2007). As mentioned before, some words are more central because they have a larger number of links with other words,

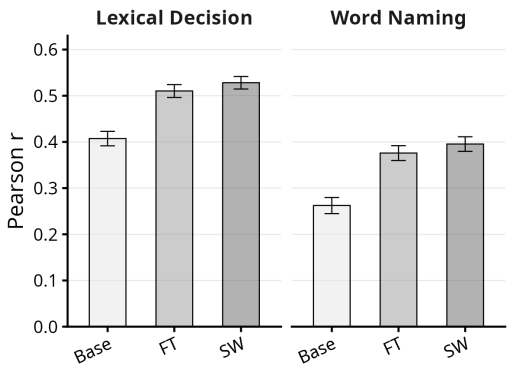

Figure 2: Correlations with lexical decision and word naming (negative) RTs for the Base, FT and SWOW (SW) models. Steiger Z-tests for dependent overlapping correlations showed all pairwise correlations were significantly different.

and previous work suggests that this independently predicts reaction times in naming and lexical decision (De Deyne et al., 2019). Similar to De Deyne et al. (2019), we focus on in-strength s_{in} centrality calculated by summing the associative over all responses to predict lexical decision times and naming times. Words and average response times were taken from the E-lexicon project (Balota et al., 2007). Both dependent and independent variables were log-transformed after adding the minimum to the z-scored RTs and a constant of 1 to the centrality measures. As before, cues from the training set were not used in the analysis resulting in a set of 10,914 words shared between the association cues and E-lexicon stimulus words. The results in Figure 2 show a consistent pattern, with significantly higher correlations for the FT model compared to Base across both tasks and correlations for the FT model that were approximating those obtained by using human associations ($r_{sw} - r_{ft}$ = .015 for Naming and .014 for LDT).

Similarity Judgements

To evaluate how word meaning was encoded in human and LLM-based semantic networks, the similarity datasets from De Deyne et al. (2019) were re-analysed, and supplemented with recent work. Since our goal is to compare different models based on test cues only, we only considered datasets with at least 100 pairs: the relatedness judgements from MEN (Bruni, Uijlings, Baroni, & Sebe, 2012), MTURK-771 (Halawi, Dror, Gabrilovich, & Koren, 2012) and Silberer-2014 (Silberer & Lapata, 2014) We included the strict similarity judgements from Simlex-999 (Hill, Reichart, & Korhonen, 2016) and added SimVerb-3500 dataset, which extends it to verbs (Gerz, Vulić, Hill, Reichart, & Korhonen, 2016). To gain insight into potential qualitative differences between the models, we replaced WordSim-353 with a more recent replication by Kliegr and Zamazal (2018) that carefully manipulated the instructions to distinguish thematic relatedness (WS Them) from interchangeability under cloze instructions (WS Cloze) and strict similarity replicating the SimLex instructions (WS Tax).

To allow for direct comparisons with SWOW-base measures, all results were based on Base and FT matching the exact number of human responses per cue. Similar to De Deyne et al. (2019), results are given for models where associative strength is weighted by positive pointwise mutual information (PPMI), and similarity is determined by the cosine overlap of the direct associations between two words. Indirect relatedness was measured by implementing a spreading activation as a random walk (RW) on the PPMI-weighted response distributions (see De Deyne et al., 2019, for details). As in the original study, the value of α, which determines the decay of longer paths over shorter ones, was set to a default of 0.75.

Table 3: Pearson correlations (r) between model similarities and human similarity ratings. Within each measure, orange indicates a correlation significantly larger than both others, and purple indicates a correlation significantly smaller than both others (pairwise dependent Steiger's Z tests, $p < .05$).

		PPMI			RW		
Dataset	n	Base	FT	SW	Base	FT	SW
MEN	1,793	.664	.725	.691	.818	.834	.804
MTURK	478	.610	.696	.668	.730	.771	.764
Silberer14	4,085	.750	.786	.761	.818	.848	.815
SimLex	630	.592	.637	.634	.654	.662	.694
SimVerb	2,227	.488	.558	.557	.540	.607	.637
WS Cloze	222	.607	.641	.636	.668	.691	.735
WS Taxonomic	222	.708	.760	.721	.791	.826	.843
WS Thematic	222	.675	.737	.683	.779	.818	.823

First, we calculated the correlation between the similarity measures for approximately 10,000 word pairs represented across all datasets. Llama models and SWOW were highly correlated, but the agreement was significantly higher for both similarity measures using the FT model, PPMI: $r_{base}(9936)$ = .818 $r_{FT}(9936)$ = .878, Steiger Z = -25.1, $p < .001$, RW: $r_{base}(9936)$ = .872 $r_{FT}(9936)$ = .928, Steiger Z = -37.39, $p < .001$. Next, we compared which models provided the best prediction of the human similarity and relatedness judgements. For the PPMI model Table 3 shows that, in all cases, the Base model was worse than the FT and SWOW models, and these differences were significant in all studies with larger sample sizes. In contrast, the FT model performed on par with the SWOW measures in most datasets, and outperformed SWOW in two relatedness judgement datasets (MEN and Silberer14). RW similarity improved performance for the Base model, although it remained significantly worse for two datasets.

How similarity was operationalized affected models differently. Fine-tuning improved relatedness correlations but not strict-similarity performance, leaving systematic differences with the SWOW measures. WordSim showed no performance interactions across models, with consistently high taxonomic and thematic correlations and lower cloze correlations.

Discussion

In this paper, we fine-tuned Llama 3.1 using the participant-level responses for about 2,500 cues taken from the English

Small World of Word association norms (De Deyne et al., 2019). Across multiple evaluation metrics, this fine-tuned model accurately predicted English word associations and these results were replicated using unpublished association data. Our results show excellent prediction for the first association with a median rank of 1, which vastly improves upon previous findings, and response-strength correlations close to the maximum of variance that can be explained in the data and significantly better than a baseline without fine-tuning.

Further analyses showed that the FT model acquired several human biases characterising word associations, such as a tendency to respond with simple (frequent, short, concrete) nouns. While a fine-tuned model might improve word association prediction, a priori, it is unclear whether capturing human biases, including task demands, facilitates or impairs the prediction of other tasks. LLMs such as Llama are already exposed to a broad range of semantic similarity benchmarks during pretraining (Xu, Wang, Fan, & Liu, 2024). However, across tasks, we found that aligning LLMs with human associations leads to systematic and noticeable performance gains.

Comparison with previous work

Findings from previous work include results for other models than Llama and shed some light on how model choice affects results. In Abramski et al. (2025) the median ranks for the first associate were predicted by prompting three different LLMs. They reported a rank of 15 for Llama 3, a rank of 2 for Haiku and 3 for Mistral. However, both Haiku and Mistral generated only a small number of distinct associates (respectively 8 and 39 on average), which suggests that their ranks are likely to be underestimated.

Our analyses of lexical processing and semantic similarity can be compared with Xiao et al. (2025), who used the same Llama 3.1-70b-instruct model next to GPT-4o. Their results showed that GPT-4o systematically outperformed Llama. In terms of in-strength, they reported correlations of .79 for GPT-4o and .67 for Llama-3.1 with SWOW-2018 in-strength. Our results, based on only 50% from US English speakers were .76 for the Base model and .98 for the FT model. In-strength measured from the FT model also predicted word processing noticeably better, with effect sizes that were nearly as good as those obtained using SWOW estimates (see Figure 2 vs .39 and .25 for GPT-4o in Xiao et al. (2025)). Differences with similarity estimates were less pronounced, especially for GPT-4o, which is consistent with our findings showing that RW compensates for the limited overlap when only direct associates are considered.[2]

Factors that determine performance

Besides the choice of LLM, the main factor that determines performance is the temperature parameter. While performance was smooth for different temperature values, both low and high temperatures had a clear detrimental effect, as they either over- or under-emphasized the most common responses. Notably, choosing the default Llama temperature (0.6) as was done in several previous studies (cf Xiao et al., 2025), might explain the suboptimal performance of Llama3.1 compared to GPT-4o. While beyond the scope of this paper, we also investigated how much training data is needed. Our findings show that a set of 2,456 cues with on average 1139 valid responses per cue represents a good trade-off between amount of data needed and prediction performance. However, additional training further improves results, although the gains are minor compared to tuning the temperature. For example when using a set of nearly 10,000 cues and testing on the remaining test cues in SWOW-2018, the correlation with human response strengths, $r = .686$, only slightly higher for the same set of cues but trained on 2,456 cues, $r = .669$. Finally, the current work explicitly matched the number of responses to make it possible to compare across human and LLM-based measures. However, when such matching isn't required, generating larger samples will provide more reliable estimates of associative strength, particularly for high-entropy cues such as abstract words, and result in more accurate centrality and similarity measures.

Extensions

The way LLMs are trained is not representative in how humans acquire meaning, which limits our ability to go beyond prediction and explain what words come to mind in an association task. However, when the goal is to provide human-like lexico-semantic norms, LLMs can reliably supplement human judgments of lexical properties such as familiarity, valence, concreteness, and semantic features (e.g., Martínez, Conde, Reviriego, & Brysbaert, 2025; Peng, Hsu, Chersoni, Qiu, & Huang, 2025). Our results extend this to word association norms and suggest that LLMs might be useful for predicting strong associates and for deriving aggregate measures (e.g., in-strength or random-walk–based similarity). A more ambitious approach is determining whether fine-tuning could also capture how associations change over time (Varnum, Baumard, Atari, & Gray, 2024), or vary across demographics (Garimella, Banea, & Mihalcea, 2017). To capture this kind of variation, existing word association norms might be insufficient, especially since the crowd-sourced approach of collecting these data introduced its self-selection biases (most tend to be female speakers, for example).

More broadly, fine-tuning LLMs to human association data might be beneficial in social and (cross)-cultural domains as well. One promising direction is fine-tuning LLMs with associates from specific people to reduce the misrepresentation of social groups (Wang, Morgenstern, & Dickerson, 2025). The potential of such approaches is already highlighted in work showing that word associations help steering models to be better aligned with culture-specific values in languages other than English (Liu, Shrestha, & Huang, 2025; Dai, Zhou, Wang, & Li, 2025). Given the limited amount of training needed to align associations, and the availability of human associations in diverse languages in projects like SWOW suggests several ways to test these ideas.

[2] RW similarities for MEN, MTURK, and SimLex were generally lower than those in Table 3 for LLaMA and mixed for GPT; however, direct comparison is limited as training cues were excluded.

Acknowledgments

This work was supported by Discovery Project DP240101873 funded by the Australian Research Council and by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative. We thank Kabir Manandhar Shrestha for valuable discussions.

References

Abramski, K., Improta, R., Rossetti, G., & Stella, M. (2025). The "LLM World of Words" English free association norms generated by large language models. *Scientific Data*, *12*, 803.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English Lexicon Project. *Behavior Research Methods*, *39*, 445–459.

Barsalou, L. W., Simmons, K. W., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Science*, *7*, 84–91.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 acm conference on fairness, accountability, and transparency* (pp. 610–623).

Bruni, E., Uijlings, J., Baroni, M., & Sebe, N. (2012). Distributional semantics with eyes: Using image analysis to improve computational representations of word meaning. In *Proceedings of the 20th ACM international conference on Multimedia* (pp. 1219–1228). ACM.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, *44*(4), 991–997.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*, 904–911.

Cattle, A., & Ma, X. (2017). Predicting word association strengths. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 1283–1288).

Cramer, P. (1968). *Word Association*. Academic Press.

Dai, X., Zhou, L., Wang, B., & Li, H. (2025). From word to world: Evaluate and mitigate culture bias via word association test. *arXiv preprint arXiv:2505.18562*.

De Deyne, S., Navarro, D. J., Collell, G., & Perfors, A. (2021). Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, *45*(1), e12922.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The Small World of Words English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*, 987-1006.

Deese, J. (1965). *The structure of associations in language and thought*. Johns Hopkins Press.

Garimella, A., Banea, C., & Mihalcea, R. (2017). Demographic-aware word associations. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2285–2295).

Gerz, D., Vulić, I., Hill, F., Reichart, R., & Korhonen, A. (2016). Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Halawi, G., Dror, G., Gabrilovich, E., & Koren, Y. (2012). Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1406–1414).

Hayashi, Y. (2016). Predicting the evocation relation between lexicalized concepts. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1657–1668).

Hill, F., Reichart, R., & Korhonen, A. (2016). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, *41*, 665-695.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Huang, W., Zheng, X., Ma, X., Qin, H., Lv, C., Chen, H., ... Magno, M. (2024). An empirical study of LLaMA3 quantization: from LLMs to MLLMs. *Visual Intelligence*, *2*, 36.

Kliegr, T., & Zamazal, O. (2018). Antonyms are similar: Towards paradigmatic association approach to rating similarity in SimLex-999 and WordSim-353. *Data & Knowledge Engineering*, *115*, 174-193.

Kumar, A. A., Steyvers, M., & Balota, D. A. (2022). A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science*, *14*(1), 54–77.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, *44*(4), 978–990.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's Problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lawrence, I., & Lin, K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 255–268.

Liu, C., Shrestha, K. M., & Huang, S. (2025). Align: Word association learning for cultural alignment in large language models. *arXiv preprint arXiv:2508.13426*.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*, 273–302.

Marjieh, R., Sucholutsky, I., van Rijn, P., Jacoby, N., & Griffiths, T. (2023). What language reveals about perception: Distilling psychophysical knowledge from large language models. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 45).

Martínez, G., Conde, J., Reviriego, P., & Brysbaert, M. (2025). Ai-generated estimates of familiarity, concreteness, valence, and arousal for over 100,000 spanish words. *Quarterly Journal of Experimental Psychology*, *78*(10), 2272–2283.

Matusevych, Y., & Stevenson, S. (2018). Analyzing and modeling free word associations. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 40).

Nelson, D. L., McKinney, V. M., Gee, N. R., & Janczura, G. A. (1998). Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological review*, *105*(2), 299.

Nematzadeh, A., Meylan, S. C., & Griffiths, T. L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th annual meeting of the Cognitive Science Society* (pp. 859–864).

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., ... Lowe, R. (2022). Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Paivio, A. (1986). *Mental Representations: A dual coding approach.* Oxford University Press.

Peng, B., Hsu, Y.-y., Chersoni, E., Qiu, L., & Huang, C.-R. (2025). Multilingual prediction of semantic norms with language models: a study on english and chinese. *Language Resources and Evaluation*, *59*(4), 3911–3937.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic bulletin & review*, *8*(3), 385–407.

Silberer, C., & Lapata, M. (2014). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 721–732).

Trott, S. (2024). Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, *56*(6), 6082–6100.

Varnum, M. E. W., Baumard, N., Atari, M., & Gray, K. (2024). Large language models based on historical text could offer informative tools for behavioral science. *Proceedings of the National Academy of Sciences*, *121*(42), e2407639121.

Wang, A., Morgenstern, J., & Dickerson, J. P. (2025). Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 1–12.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, *45*, 1191–1207.

Xiao, B., Duan, X., Haslett, D. A., & Cai, Z. G. (2025). Human-likeness of LLMs in the Mental Lexicon. In *Proceedings of the 29th Conference on Computational Natural Language Learning* (pp. 586–601).

Xu, R., Wang, Z., Fan, R.-Z., & Liu, P. (2024). Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.