scientific data

Check for updates

OPEN

DATA DESCRIPTOR

The "LLM World of Words" English free association norms generated by large language models

Katherine Abramski[1]✉, Riccardo Improta[2], Giulio Rossetti[3,4] & Massimo Stella[2,4]

Free associations have been extensively used in psychology and linguistics for studying how conceptual knowledge is organized. Recently, the potential of applying a similar approach for investigating the knowledge encoded in LLMs has emerged, specifically as a method for investigating LLM biases. However, the absence of large-scale LLM-generated free association norms that are comparable with human-generated norms is an obstacle to this research direction. To address this, we create a new dataset of LLM-generated free association norms modeled after the "Small World of Words"(SWOW) human-generated norms with nearly 12,000 cue words. We prompt three LLMs (Mistral, Llama3, and Haiku) with the same cues as those in SWOW to generate three novel comparable datasets, the "LLM World of Words" (LWOW). From the datasets, we construct network models of semantic memory that represent the conceptual knowledge possessed by humans and LLMs. We validate the datasets by simulating semantic priming within the network models, and we briefly discuss how the datasets can be used for investigating implicit biases in humans and LLMs.

Background & Summary

How is conceptual knowledge organized in the mind? Such a question has long been the focus of linguists and cognitive psychologists who aim to better understand the human language capacity[1,2]. Recently, this question has become increasingly relevant in the field of artificial intelligence, particularly regarding large language models (LLMs). Human semantic memory – the repository of conceptual knowledge that encompasses how words get their meaning[3] – forms the foundation of human language and thought, and thus, its structure and properties influence how we reason, form beliefs and make decisions[4,5], ultimately shaping our social and political systems. Similarly, the semantic representations that comprise the knowledge encoded in LLMs are the underlying source behind the outputs they produce, and as LLMs become more integrated into our everyday lives, these outputs have an increasing impact on society[6,7]. Thus, the study of the structure and properties of semantic memory is central to understanding not only our own thinking and reasoning, but also the "thinking" and "reasoning" of LLMs, which carries important societal implications.

Studying semantic memory involves creating representations of word meanings (semantic representations), often in terms of how words relate to other words. In humans, one common way to do this is using free associations[1,8,9], which are usually accessed by prompting participants with a cue word and asking them to come up with (typically three) associated responses. Since the task is context neutral, responses represent the associative knowledge of words that we possess at an implicit level. Free associations have been extensively used in cognitive psychology and linguistics for studying lexical retrieval[1,4], semantic organization[8], and similarity judgments[2,10,11]. They have also been used for studying differences in cognitive processing between concrete and abstract words, i.e. concreteness effects[12]. Given that free associations have been shown to correlate with stable implicit attitudes[13], they have also been used for studying affective biases[9]. Investigations of conceptual knowledge using free associations are often conducted within network models of semantic memory built from free associations by connecting cue words to their responses. This results in a complex network structure of human conceptual knowledge in which words get their meanings through relationships to other concepts. Such models enable the investigation of complex cognitive processes that take place within semantic memory. In

[1]University of Pisa, Department of Computer Science, Pisa, Italy. [2]University of Trento, Department of Psychology and Cognitive Science, Trento, Italy. [3]National Research Council of Italy, Institute of Information Science and Technologies, Pisa, Italy. [4]These authors contributed equally: Giulio Rossetti, Massimo Stella. ✉e-mail: katherine.abramski@phd.unipi.it

fact, cognitive network models of semantic memory have been used to gain powerful insights about a variety of human cognitive phenomena such as language learning,[8,14], creativity[15–17], personality traits like openness to experience[5], and autism spectrum disorder[18].

While free associations have been widely used for studying semantic memory in humans, very different approaches have been applied for investigating conceptual knowledge in language models. Typically, semantic representations are directly accessed from the model's embedding space in the form of word embeddings[19], i.e. vector representations of words whose meanings are derived from statistical relationships in the training data. Word embeddings provide an advantageous way for investigating certain aspects of semantic memory because of the types of mathematical operations that can be applied within the embedding space. Specifically, the semantic similarity between two words can easily be calculated by computing the cosine similarity between their word vectors. Thus, this approach can be used for extracting word associations directly from the model's architecture, and these associations can be used to study several aspects related to the model's conceptual knowledge[20,21]. In recent years, there has been a great interest in investigating the biases encoded in language models, and measuring the strength of associations between words within the model's embedding space has been extensively applied for accomplishing this goal[22–26]. Such an approach follows the same idea that the Implicit Association Test (IAT)[27] uses to investigate implicit attitudes in humans. Essentially, it involves using the cosine similarity to measure the strength of the association between pairs of words for assessing biases, for example, *man – doctor* and *woman – nurse*. Depending on the strength of these respective associations, this method can reveal certain biases encoded in the model's embedding space, such as gender biases that reflect the implicit analogy, *man is to doctor as woman is to nurse*[22].

Extracting associations from the embedding space for investigating the conceptual knowledge encoded in language models has many advantages, but it also has some important limitations. Perhaps the most significant limitation is that this approach works well for older language models that use static word embeddings – which represent word meanings at the type level – but not so well for newer models, i.e. LLMs, that use contextual word embeddings – which represent word meanings at the token level. Operations on contextual embeddings require that the embeddings are first transformed into static embeddings[26], but this can introduce bias and distort similarity estimates[19]. Another downside to the approach of accessing the embedding space is that it limits the possibility to make comparisons across models and with humans, since the cognitive architecture of each model is vastly different. These limitations have led to a recent shift from the bottom-up approach of accessing the embedding space for investigating the knowledge encoded in language models, towards a top-down approach that involves prompting models with tasks and using their output to make inferences about the knowledge encoded in their embedding space[6,7,28]. This approach mirrors methods from cognitive psychology that use behavioral experiments to make inferences about the workings of the human mind, and thus it has been aptly named "machine psychology"[29]. While there are certain limitations that come with prompting LLMs[30,31] – for example, outputs may be highly sensitive to variations in prompts and can yield poor results – there are also many advantages to this top-down approach. One advantage is that it can be applied to virtually any LLM, without a deep technical understanding of the model's embedding space, making it an accessible methodology for researchers from different fields. Another advantage of the machine psychology approach is that it allows for the use of preexisting and well-studied tools, measures, and methodologies that have been applied to humans for years[7]. Thus, rather than developing completely new methodologies, the challenge of machine psychology lies in adapting the existing methodologies from cognitive science and psycholinguistics so that they can be applied to LLMs to gain insights about specific cognitive phenomena. Recent cutting edge studies have applied the machine psychology approach to investigate the semantic capabilities of LLMs compared to humans[32–35], demonstrating that this is a promising direction of research.

Up until this point, we have touched upon three main ideas:

1. The use of free associations for investigating psychological phenomena in humans has long-standing relevance[8];
2. There has been extensive work[22], as well as a great interest, in using word associations extracted from the embedding spaces of language models for investigating biases encoded in their architecture;
3. There has been a recent shift from the bottom-up approach to a top-down machine psychology approach for investigating the knowledge encoded in LLMs[7,36].

Together, these ideas point to the need for a new direction of research that uses LLM-generated free associations in order to investigate the structure and properties of conceptual knowledge in LLMs. While there exist several datasets of free association norms generated by humans[11,37,38], to the best of our knowledge, there are no openly available datasets of LLM-generated free association norms that are comparable in scale and breadth to existing human-generated norms.

To fill this research gap, we present the LLM World of Words (LWOW)[39], a dataset of English free association norms including millions of responses generated by three different LLMs: Mistral (mistral-7b), Llama3 (llama3.1-8b), and Claude Haiku (claude-3-5-haiku-latest). LWOW is modeled after the largest dataset of human-generated English free association norms, called the Small World of Words (SWOW)[11], which has been used extensively in many psychological and linguistic studies[5]. LWOW contains over 12,000 cue words, each with 3 responses, repeated 100 times, for a total of over 3 million responses. The LWOW dataset is generated using the same cue words used in SWOW, following the same methodology, though instead of asking humans to produce responses to the cues, we asked the three LLMs to produce responses to the cues. The result is three sets of LLM-generated free association norms that are directly comparable to the SWOW dataset. The LWOW

dataset[39], combined with the original SWOW dataset, will enable investigations of the structure and properties of conceptual knowledge in both humans and LLMs, allowing for unprecedented comparisons between the two.

We anticipate that LWOW will be particularly useful for investigating the nature of implicit biases in LLMs as they relate to human biases, such as the gender and racial stereotypes that are prevalent both in society[24] and in LLM outputs[29]. For this reason, as part of the validation and usage notes of this dataset, we construct cognitive network models of semantic memory[9] from both the SWOW and LWOW datasets (humans and LLMs) and we briefly discuss how they can be used to investigate the presence of implicit stereotypes within their associative knowledge structures.

The remainder of this paper is structured as follows. In *Methods*, we describe the methodology used to (1) generate the datasets, (2) preprocess the data, and (3) build network models of semantic memory from the data. In this section, we also provide summary statistics of the datasets, the network models, and comparisons between the networks. In *Data Records*, we provide a detailed description of the repository containing the code and original datasets used to generate the data. In *Data Validation*, we demonstrate the validity of the data by simulating the cognitive mechanisms that underlie semantic priming within the network models, showing that activation patterns within the networks correlate with behavioral data from a well-known psycholinguistic experiment, i.e. the lexical decision task (LDT). In *Usage Notes*, we briefly discuss how the LWOW datasets can be used for investigating biases in humans and LLMs by adapting the methodology used in *Data Validation*. Finally, in *Code Availability* we provide details on how to access the code in order to reproduce the analyses.

Methods

Data generation. Since the LWOW datasets are based off the Small World of Words (SWOW) human-generated norms, we first gathered the cue words from the original SWOW dataset. The SWOW dataset was downloaded from the project's research page, https://smallworldofwords.org/en/project/research, under the section *English Data (SWOW-EN18)*. We used the preprocessed data (SWOW-EN.R100.csv) with 12,282 cue words and 100 sets of responses per cue. We used a list of these cues as input to the three LLMs along with a prompt that aimed to mimic the instructions provided to humans in the original SWOW free association task. The following prompt was given to the LLMs:

Task:
- You will be provided with an input word: write the first 3 words you associate to it separated by a comma
- No additional output text is allowed

Constraints:
- No carriage return characters are allowed in the answers
- Answers should be as short as possible

Example:
Input: sea
Output: water, beach, sun

This prompt was repeated 100 times for each cue word in order to generate a dataset with the same number of responses as the original (preprocessed) SWOW dataset. In what follows, we describe how the LLM-generated output as well as the original SWOW output were further processed.

Data processing. The original SWOW data[11] were already preprocessed, but in order to facilitate analyses and data alignment, we applied additional data preprocessing to both the original SWOW data and the output produced by all three language models. From this point on, we refer to the SWOW dataset, including all subsequent modifications, as the Human dataset. The preprocessing steps applied to all four datasets, which were done in python, are as follows. First, all cues and responses were made lowercase. Then, the articles *a, an, the*, and the preposition *to* were removed from the beginning of responses unless they were among the original cues (e.g. *a lot*). Some responses included underscores, and these were replaced with spaces. Also, some responses incorrectly lacked spaces or hyphens (e.g. *throwout, checkin*). In order to ensure that these responses were not excluded in later analyses, we created a mapping dictionary using WordNet[40] (implemented in the python library *nltk*) to resolve this issue. Specifically, we took all the words in WordNet[40] that have either spaces or hyphens (e.g. *throw out, check-in*) and we removed them to create a one-to-one mapping to correct these errors in the responses. Spelling corrections were also applied to both cues and responses according to a dictionary that was used to process the original SWOW data. This dictionary was downloaded from the SWOW GitHub page https://github.com/SimonDeDeyne/SWOWEN-2018/tree/master/data/dictionaries (EnglishCustomDict.txt). This spelling dictionary included the correction of commonly misspelled words (e.g. *recieve* to *receive*) but it also mapped British spelling to American spelling (e.g. *colour* to *color*). Next, cues and responses were lemmatized using WordNet's lemmatizer, changing plural nouns to singular nouns (e.g. *men* to *man*) but leaving tensed verbs unchanged (e.g. *cooking*, *determined*). Next, we added or removed data to ensure exactly 100 repetitions per cue. This step was needed for the Human data because the spelling corrections and the lemmatization of cues resulted in more than 100 repetitions for some cues, while some LLMs, namely Llama3, failed to generate 100 repetitions per cue. Thus, we ensured 100 repetitions per cue by adding blank responses when there were less than 100 repetitions for a cue, while sampling randomly when there were more than 100 repetitions per cue. Finally, responses that were identical to their corresponding cues were removed, and duplicate responses within the same set of three responses were removed. After this preprocessing procedure, each dataset resulted in a total of 11,545 cues, compared to the 12,282 cues in the original SWOW dataset. Table 1 shows the statistics of each dataset, including the number of cues, number of total

Network	Unique cues	Total responses	Unique responses	Missing responses
Humans	11,545	3,148,578	116,640	9.1%
Mistral	11,545	3,268,206	41,369	5.6%
Llama3	11,545	3,348,049	105,367	3.3%
Haiku	11,545	3,403,644	15,275	1.7%

Table 1. Dataset statistics. Cue and response statistics for all datasets after preprocessing. All networks have the same unique cues, but different numbers of total responses and unique responses. The Human network has the largest percentage of missing responses, but also the largest number of unique responses compared to all LLMs.

responses, number of unique responses, and percentage of missing responses. The Human dataset has the most unique responses, closely followed by Llama3. Mistral and Haiku have far fewer unique responses.

While these data have undergone important preprocessing, this procedure did nothing to remove nonsensical responses, from both the Human and LLM datasets. Such responses are unsuitable for certain analyses, and we suggest additional cleaning procedures be applied to remove nonsensical responses, however, we include all these responses in the preprocessed data for several reasons. First, what can be considered a valid response is rather subjective. For example, while the human-generated response *accidentally pressed enter instead of no more associations* is clearly not a response to the cue, some other responses such as *violence against women* or *easy to get along with* may be considered valid, despite not being words per se. Secondly, the types of invalid or nonsensical responses produced by humans and LLMs alike may be of great interest to some researchers. Thirdly, there is no systematic way to remove all nonsensical responses. Instead, identifying responses that are clearly nonsensical, such as *printassociatedwordsinputword* requires applying a series of ad-hoc filters. We experimented with applying such filters, and we found that while some nonsensical responses are easily removed by searching for certain sets of strings, other invalid responses are harder to classify. For example, in the data generated by Mistral, the responses *output*, *input*, and *association* appeared much more frequently than in the human data, and most are clearly invalid responses. However, such responses are difficult to identify because *input* could be a perfectly valid response to cues such as *output*, *feedback*, or *function*, but it is most likely an invalid response to cues like *flower* or *monkey*. For these reasons, we have not applied any filters to remove or correct these responses in the preprocessed data. Instead, we applied filters in the network building process to remove responses that, for the purpose of our analyses, we considered invalid.

Network construction. As discussed earlier, free associations are often used to build network models of semantic memory. In this way, semantic memory is considered a complex system[2,9,15], and it can be studied as such even in case of systems without an explicit semantic memory, like LLMs, that are nonetheless capable of processing language[36]. The network structure provides a quantitative framework within which certain cognitive phenomena can be investigated using the tools of network science. From the preprocessed data, we built network models of semantic memory for humans and all three LLMs by connecting cue words to their responses. The weight of the edge reflects the frequency of the response, so if *cat* appears 20 times as the response to the cue *dog*, then a directed edge of weight 20 is created from *dog* to *cat*. The networks are naturally directed, but they are transformed into undirected networks to facilitate the analyses that we will discuss in the next section. In cases in which there is a bidirectional edge, the largest of the two edge weights is maintained. So if the edge from *dog* to *cat* has a weight of 20 and the edge from *cat* to *dog* has a weight of 25, the undirected edge between *cat* and *dog* has a weight of 25. The full undirected networks are then filtered to remove unwanted nodes and edges. This is done first by removing nodes that are not in WordNet, and then by removing idiosyncratic edges, i.e. edges with a weight = 1, and finally, taking the largest connected component. This network filtering has a few advantages. First, the WordNet filter provides a standardized way to eliminate nonsensical and uncommon responses, since words in WordNet are at least English words, and the removal of idiosyncratic edges ensures that the association is something shared among two or more people (iterations in the case of the LLMs) and not just a fluke. A final advantage is that this filtering reduces the number of nodes and edges, making the networks more computationally manageable.

Table 2 shows the network statistics for both the full networks (before filtering) and the reduced networks (after filtering) for each of the four datasets (Humans, Mistral, Llama3, and Haiku). Statistics include the number of nodes, number of edges, the network density, and the average degree. The reduced networks are the final versions of the networks, and from this point on, when we discuss the networks, we refer to the reduced networks. Among the reduced networks, the largest network is Llama3 followed by Humans, Mistral, and finally Haiku. To assess the extent to which each LLM network is similar/different to the Human network, we made pairwise comparisons of nodes and edges between each LLM network and the Human network. The pairwise comparison statistics are shown in Table 3. For the node comparisons, we calculated the percentage of Human nodes not in the LLM network, the percentage of all nodes common to both networks, and the percentage of LLM nodes not in the Human network. For the edge comparisons, the same statistics were calculated, however they were calculated on the subgraphs of the node intersections of each pair of graphs.

Data Records

The LWOW datasets[39] generated by Mistral, Llama3, and Haiku are accessible in the Zenodo repository at the following link: https://doi.org/10.5281/zenodo.15310707[39]. Additionally, they are accessible in the Github repository at the following link: https://github.com/LLMWorldOfWords/LWOW. The repositories are identical, but any subsequent updates to the datasets will be reflected in the Github repository. Within the repository, each LLM dataset is provided as a .csv file that follows the structure shown in Table 4. All data sources and code

Full Networks	Network	Nodes	Edges	Density	Average degree
	Humans	116,640	1,164,026	0.0002	20.0
	Mistral	42,073	417,697	0.0005	19.9
	Llama3	105,777	770,458	0.0001	14.6
	Haiku	17,679	77,698	0.0005	8.8
Reduced Networks	**Network**	**Nodes**	**Edges**	**Density**	**Average degree**
	Humans	24,308	317,344	0.0011	26.1
	Mistral	20,339	199,103	0.0010	19.6
	Llama3	38,987	546,866	0.0007	28.1
	Haiku	15,596	64,599	0.0005	8.3

Table 2. Network statistics. The full networks are built using all cues and responses from the cleaned datasets, while the reduced networks are filtered by removing nodes not in WordNet, removing idiosyncratic edges, and then taking the largest connected component. The purpose of this network filtering is to remove senseless responses and to make the networks more computationally manageable.

Nodes	Comparison with Humans	Mistral	Llama3	Haiku
	Percentage of Human nodes not in LLM network	32%	16%	41%
	Percentage of all nodes common to both networks	59%	48%	57%
	Percentage of LLM nodes not in Human network	19%	47%	8%
Edges	**Comparison with Humans**	**Mistral**	**Llama3**	**Haiku**
	Percentage of Human edges not in LLM network	69%	70%	84%
	Percentage of all edges common to both networks	23%	13%	15%
	Percentage of LLM edges not in Human network	51%	81%	24%

Table 3. Network comparisons. The table shows pairwise comparisons between the Human network and each of the LLM networks (Mistral, Llama3, and Haiku). The node comparison is straightforward, while the edge comparison considers only edges between common nodes of the graphs being compared.

cue	R1	R2	R3
apple	banana	fruit	orange
tree	wood	green	leaf
school	teacher	class	building
mathematics	anxiety	formula	equation
car	train	gasoline	fuel
…	…	…	…

Table 4. Example dataset. The table shows the structure of the .csv files. Each row has a cue word and three responses.

needed to reproduce the datasets and conduct further analyses are either available in the repository, or a description of where to access additional data sources (e.g. the SWOW dataset) is provided. It should be noted that due to the license of the SWOW dataset (CC BY-NC-ND 3.0) which prohibits the distribution of modified material, we do not provide the Human processed dataset in the repository, however, it can be generated using the code and other data sources provided in the repository.

Technical Validation

To demonstrate the reliability of our data, we adopted a previously applied approach[41] that simulates the cognitive mechanisms underlying semantic priming, a cognitive phenomenon that entails recognizing target words more quickly when they are preceded by related prime words. Studies of semantic priming effects have been critical to gaining a better understanding of the nature of semantic memory[42,43]. Semantic priming is usually investigated using the lexical decision task (LDT), in which human participants are presented with a prime word followed by a target word, and participants must decide as quickly as possible whether the target word is a real English word or a non-word. It has been found that participants identify the target word more quickly (lower reaction time) when the prime is related to the target (*e.g. doctor – nurse*) compared to when the prime is unrelated to the target (*e.g. doctrine – nurse*)[44,45]. This pattern has been shown to be consistent across thousands of prime-target pairs[42].

In addition to the LDT, semantic priming can also be studied by implementing a spreading activation process within a network of semantic memory. Spreading activation is a method of search within a network that is based on supposed mechanisms of human memory[43,46]. In spreading activation theory, exposure to a concept

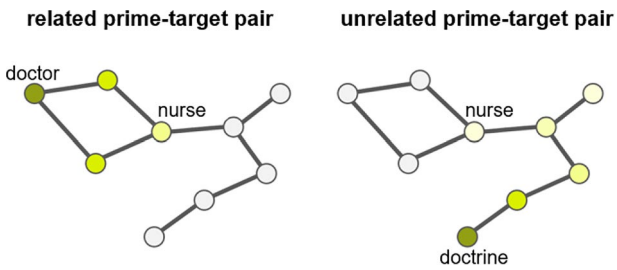

Fig. 1 Spreading activation within a network. The diagram shows how activation spreads within a network after various time steps after activating a prime node, leading to the activation of the target node *nurse*. Darker colors indicate greater activation levels. On the left, the related prime *doctor* is activated, while on the right, the unrelated prime *doctrine* is activated. The final activation level of the target node *nurse* is greater on the left when the activated prime is a related word.

leads to the activation of its corresponding node in semantic memory. That activation then propagates through the semantic memory network along the connections of the activated node, decaying over time, leading to the activation of other concepts in the network. This theory can be used to study semantic priming by simulating a search process within a network in which a prime node is activated and at the end of the spreading activation process, the final activation level of the target node is observed. Following the semantic priming effect, the final activation level of the target node (e.g. *nurse*) should, in theory, be greater when the activated prime is a related word (e.g. *doctor*) rather than an unrelated word (e.g. *doctrine*)[41]. This process is represented in Fig. 1.

Spreading activation processes can be simulated within empirical networks using the R library *spreadr*[41]. Given an empirical network, the *spreadr* algorithm works by specifying one or more nodes in the network to be activated, the initial activation level of the activated nodes, and the number of iterations or time steps. At each time step, an activated node may retain a certain percentage of its activation, while the remaining percentage is distributed among its neighboring nodes, proportional to the weight of the edges if the network is weighted. This process continues iteratively until the specified number of time steps is reached. At the end of the iterative process, the final activation level of each node in the network can be measured.

The developers of *spreadr*[41] simulated spreading activation within a free association network of semantic memory[37] to investigate the semantic priming effect using a series of prime-target pairs and their corresponding empirical reaction times from a lexical decision task experiment[42]. As expected, they found that the final activation levels of the target nodes correlated with reaction times from the empirical data. That is, the final activation levels of the targets were greater when the activated primes were related rather than unrelated to the targets. These results show that the semantic priming effect observed in the psycholinguistic LDT experiment can also be observed within a network model of semantic memory, demonstrating the usefulness of semantic networks for modeling certain cognitive phenomena.

In order to validate our datasets, we repeated this spreading activation investigation of semantic priming implemented by the authors of *spreadr*, but we used the Human and LLM networks that we built. We used a subset of 50 prime-target pairs and their corresponding reaction times from the same LDT dataset used by the authors of *spreadr*[41], downloadable from https://www.montana.edu/attmemlab/spp.html (LDT Priming Data)[42]. A sample of the 50 prime-target pairs and subsequent standardized reaction times (RTs) that we used are shown in Table 5. The full set of 50 prime-target pairs can be found in the data repository. Reaction times for the related prime-target pairs are, on average, less than those for the unrelated prime-target pairs. This pattern is shown in the boxplot in Fig. 2. The significance of these paired differences (RTs of related prime-target pairs minus RTs of corresponding unrelated prime-target pairs) is confirmed by a Wilcoxon rank test for paired samples with an effect size of -0.87 ($p < 0.001$).

In a series of empirical simulations using *spreadr*[41], we activated each prime from all 100 prime-target pairs and we observed the final activation levels of the 50 corresponding targets. We repeated these simulations in each of the four networks. As previously discussed, the *spreadr* library requires specification of the initial activation level of the activated node(s), and the number of iterations of the spreading activation process. We set the initial activation level of the prime node to the number of nodes in the entire network, and we set the number of iterations to two times the diameter of the network[5]. We specified that the networks are weighted so the edge weights were taken into consideration. Other parameters of the algorithm, such as the percentage of activation retained by each node, were set to default settings. For each network, the series of spreading activation processes yielded a matrix of final activation levels, such that the columns are all 100 prime nodes while the rows are all nodes of the network. Thus, each column of the matrix gives a vector of final activation levels of all nodes in the network after activating a single prime node, and a single entry in that vector is final activation level of the target node. The matrices were normalized first by normalizing columns of the matrix and then the rows. The normalization is necessary because it accounts for differences in the centrality of nodes within the semantic networks. By controlling for this factor, the normalized final activation levels reflect the semantic priming effects, and not effects related to node centrality, like frequency effects. We observed the normalized final activation levels of the target nodes when they were activated by related primes compared to when they were activated by unrelated primes, and we found results consistent with those obtained by the authors of *spreadr* for all four networks. That

Target	Related prime	Unrelated prime	Related RT	Unrelated RT
britain	england	like	0.75	2.21
dill	pickle	cane	0.21	1.56
coral	reef	snob	0.00	1.09
clipper	toenail	show	0.05	1.11
christ	jesus	chunk	−0.09	0.97
…	…	…	…	…

Table 5. Sample of LDT dataset. A sample of 5 of the 50 targets with related and unrelated primes and corresponding mean z-scored reaction times (RTs) from the lexical decision task dataset[42], used for data validation.

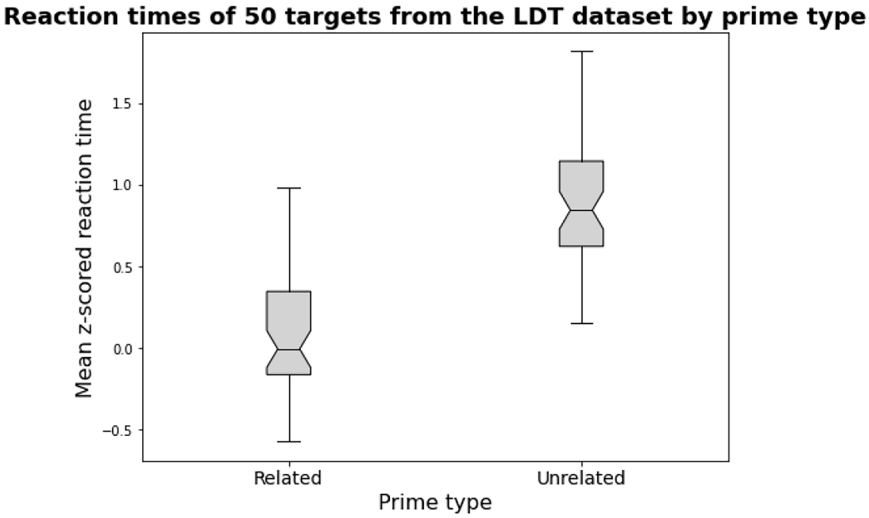

Fig. 2 Differences in RTs from the LDT dataset. The boxplot shows the difference in mean z-scored reaction times from the LDT experiment for related prime-target pairs and unrelated prime-target pairs. Reaction times for related prime-target pairs are, on average, less than those for unrelated prime-target pairs.

is, final activation levels of targets are higher when they are activated by related primes compared to unrelated primes. Wilcoxon rank tests for paired samples confirm the statistical significance of these paired differences (all $p < 0.001$). These differences in activation levels by prime type are shown in the boxplots in Fig. 3. Effect sizes of these differences, which are all relatively large and comparable to the effect size of the differences in empirical RTs from the LDT experiment (−0.87), are shown in Table 6. Similar to the authors of *spreadr*, we also found that the normalized final activation levels that we observed correlated with the empirical RTs from the LDT experiment for all four networks. That is, higher activation levels are associated with lower reaction times. The Spearman correlations are shown in Table 6 (all $p < 0.001$). These results show that the network models of semantic memory built from the Human and LWOW datasets can be used for investigating semantic priming effects not only in humans but also in LLMs, demonstrating the validity of the datasets.

Usage Notes

The LWOW datasets can potentially be used for investigating a variety of cognitive and linguistic properties of LLMs. However, in this section we would like to briefly discuss one potential application in particular: investigating implicit biases. In the previous section, we demonstrated the validity of the datasets by showing how they can be used to investigate semantic priming effects by simulating spreading activation processes within the networks. Since the semantic priming effect emerges due to the relatedness of words, observing this effect can be an indication that two words are related. Thus, semantic priming can be used to assess the strength of association between two words, which as we discussed earlier[13], makes it an ideal method for evaluating implicit biases. Therefore, in order to use the datasets to investigate implicit biases in humans and LLMs, the same spreading activation methodology described in the previous section can be applied, but using different prime-target pairs. The prime-target pairs should be selected based on the word associations to be investigated. For example, to investigate gender biases, it would be useful to select corresponding prime-target pairs that are stereotype-consistent (e.g. *doctor – man, nurse – woman*) and stereotype-inconsistent (e.g. *doctor – woman, nurse – man*). Larger effect sizes would indicate greater levels of stereotype bias within the networks. Such investigations could shed light on how LLM biases are similar and different from human biases.

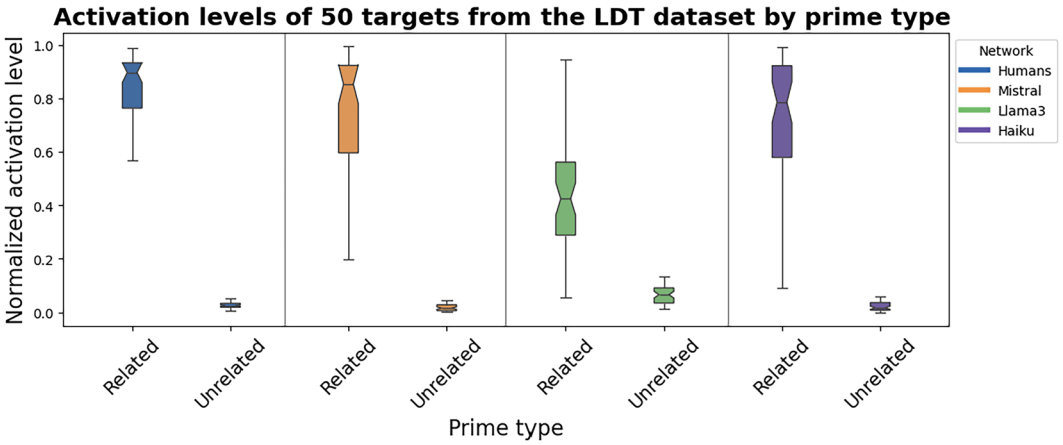

Fig. 3 Validation of networks. The boxplots show the differences in normalized final activation levels of the targets when related primes are activated compared to unrelated primes, for all networks. There is a significant difference in the final activation level, with higher final activation levels of the targets when the prime is related to the target for all networks.

Effect size (prime type)	Humans	Mistral	Llama3	Haiku
	0.870***	0.859***	0.869***	0.866***
Correlation (activation level vs. reaction time)	**Humans**	**Mistral**	**Llama3**	**Haiku**
	−0.626***	−0.615***	−0.614***	−0.662***

Table 6. Effect sizes and correlations. Effect sizes (related primes minus unrelated primes) of the Wilcoxon rank test for paired samples are provided in the upper table. The Spearman correlation coefficient for activation levels vs. reaction times for each prime-target pair are provided in the bottom table (***$p < 0.001$).

Code availability

The python and R code and all related data used to produce the LWOW datasets[39] and conduct the analyses are available at https://doi.org/10.5281/zenodo.15310707 and at https://github.com/LLMWorldOfWords/LWOW.

Received: 30 December 2024; Accepted: 8 May 2025;
Published online: 16 May 2025

References

1. De Deyne, S., Navarro, D. J. & Storms, G. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behav. research methods* **45**, 480–498 (2013).
2. Kenett, Y. N., Levi, E., Anaki, D. & Faust, M. The semantic distance task: Quantifying semantic distance with semantic network path length. *J. Exp. Psychol. Learn. Mem. Cogn.* **43**, 1470 (2017).
3. Aitchison, J. *Words in the mind: An introduction to the mental lexicon* (John Wiley & Sons, 2012).
4. Vankrunkelsven, H., Verheyen, S., Storms, G. & De Deyne, S. Predicting lexical norms: A comparison between a word association model and text-based word co-occurrence models. *J. cognition* **1** (2018).
5. Samuel, G., Stella, M., Beaty, R. E. & Kenett, Y. N. Predicting openness to experience via a multiplex cognitive network approach. *J. Res. Pers.* **104**, 104369 (2023).
6. Shiffrin, R. & Mitchell, M. Probing the psychology of ai models. *Proc. Natl. Acad. Sci.* **120**, e2300963120 (2023).
7. Binz, M. & Schulz, E. Using cognitive psychology to understand gpt-3. *Proc. Natl. Acad. Sci.* **120**, e2218523120 (2023).
8. Steyvers, M. & Tenenbaum, J. B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cogn. science* **29**, 41–78 (2005).
9. Stella, M., De Nigris, S., Aloric, A. & Siew, C. S. Forma mentis networks quantify crucial differences in stem perception between students and experts. *PloS one* **14**, e0222870 (2019).
10. De Deyne, S. & Storms, G. Word associations: Network and semantic properties. *Behav. research methods* **40**, 213–231 (2008).
11. De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M. & Storms, G. The "small world of words" english word association norms for over 12,000 cue words. *Behav. research methods* **51**, 987–1006 (2019).
12. Hill, F., Korhonen, A. & Bentz, C. A quantitative empirical analysis of the abstract/concrete distinction. *Cogn. science* **38**, 162–177 (2014).
13. Schnabel, K. & Asendorpf, J. B. Free associations as a measure of stable implicit attitudes. *Eur. J. Pers.* **27**, 39–50 (2013).
14. Citraro, S., Vitevitch, M. S., Stella, M. & Rossetti, G. Feature-rich multiplex lexical networks reveal mental strategies of early language learning. *Sci. Reports* **13**, 1474 (2023).
15. Kenett, Y. N. & Austerweil, J. L. Examining search processes in low and high creative individuals with random walks. *In CogSci* **8**, 313–318 (2016).
16. Beaty, R. E. & Kenett, Y. N. Associative thinking at the core of creativity. *Trends Cogn. Sci.* (2023).
17. Benedek, M. *et al.* How semantic memory structure and intelligence contribute to creative thought: A network science approach. *Think. & Reason.* **23**, 158–183 (2017).
18. Kenett, Y. N., Gold, R. & Faust, M. The hyper-modular associative mind: a computational analysis of associative responses of persons with asperger syndrome. *Lang. Speech* **59**, 297–317 (2016).

19. Apidianaki, M. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Comput. Linguist*. 1–60 (2022).
20. Rodriguez, M. A. & Merlo, P. Word associations and the distance properties of context-aware word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, 376–385 (2020).
21. Yao, P., Renwick, T. & Barbosa, D. Wordties: Measuring word associations in language models via constrained sampling. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 5959–5970 (2022).
22. Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V. & Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. neural information processing systems* **29** (2016).
23. Kurita, K., Vyas, N., Pareek, A., Black, A. W. & Tsvetkov, Y. Measuring bias in contextualized word representations. arXiv preprint arXiv:1906.*07337* (2019).
24. Manzini, T., Lim, Y. C., Tsvetkov, Y. & Black, A. W. Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. *arXiv preprint arXiv:1904.04047* (2019).
25. Caliskan, A., Bryson, J. J. & Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
26. Bommasani, R., Davis, K. & Cardie, C. Interpreting pretrained contextualized representations via reductions to static embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781 (2020).
27. Greenwald, A. G., McGhee, D. E. & Schwartz, J. L. Measuring individual differences in implicit cognition: the implicit association test. *J. personality social psychology* **74**, 1464 (1998).
28. Srivastava, A. *et al*. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615* (2022).
29. Hagendorff, T. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. *arXiv preprint arXiv:2303.13988* (2023).
30. Hu, J. & Levy, R. Prompting is not a substitute for probability measurements in large language models. *arXiv preprint arXiv:2305.13264* (2023).
31. Kauf, C., Chersoni, E., Lenci, A., Fedorenko, E. & Ivanova, A. A. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. *arXiv preprint arXiv:2403.14859* (2024).
32. Wang, Y. *et al*. The fluency-based semantic network of llms differs from humans. *Comput. Hum. Behav. Artif. Humans* 100103 (2024).
33. Suresh, S. *et al*. Conceptual structure coheres in human cognition but not in large language models. *arXiv preprint arXiv:2304.02754* (2023).
34. Digutsch, J. & Kosinski, M. Overlap in meaning is a stronger predictor of semantic activation in gpt-3 than in humans. *Sci. Reports* **13**, 5035 (2023).
35. Abramski, K., Lavorati, C., Rossetti, G. & Stella, M. Llm-generated word association norms. In *HHAI 2024: Hybrid Human AI Systems for the Social Good*, 3–12 (IOS Press, 2024).
36. Abramski, K., Citraro, S., Lombardi, L., Rossetti, G. & Stella, M. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data Cogn. Comput.* **7**, 124 (2023).
37. Nelson, D. L., McEvoy, C. L. & Schreiber, T. A. The university of south florida free association, rhyme, and word fragment norms. *Behav. Res. Methods, Instruments, & Comput.* **36**, 402–407 (2004).
38. Wilson, M., *et al*. Eat: The edinburgh associative corpus. *Oxf. Text Arch. Core Collect*. (1988).
39. Abramski, K. E., Improta, R., Rossetti, G. & Stella, M. *LLMWorldOfWords/LWOW: First release*. https://doi.org/10.5281/zenodo.15310707 (2025).
40. Miller, G. A. Wordnet: a lexical database for english. *Commun. ACM* **38**, 39–41 (1995).
41. Siew, C. S. spreadr: An r package to simulate spreading activation in a network. *Behav. Res. Methods* **51**, 910–929 (2019).
42. Hutchison, K. A. *et al*. The semantic priming project. *Behav. research methods* **45**, 1099–1114 (2013).
43. Collins, A. M. & Loftus, E. F. A spreading-activation theory of semantic processing. *Psychol. review* **82**, 407 (1975).
44. Neely, J. H. Semantic priming effects in visual word recognition: A selective review of current findings and theories. *Basic processes reading* 264–336 (2012).
45. McNamara, T. P. *Semantic priming: Perspectives from memory and word recognition* (Psychology Press, 2005).
46. Collins, A. M. & Quillian, M. R. Retrieval time from semantic memory. *J. verbal learning verbal behavior* **8**, 240–247 (1969).

Author contributions

K.A., G.R. and M.S. conceived the experiments, K.A., R.I. and G.R. conducted the experiments, K.A. analyzed the results and visualized the data. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.