HHAI 2024: Hybrid Human AI Systems for the Social Good
F. Lorig et al. (Eds.)

doi:10.3233/FAIA240177

LLM-Generated Word Association Norms

Katherine ABRAMSKI [a,1] , Clara LAVORATI [a] , Giulio ROSSETTI [b,*] and Massimo STELLA [c,*]

[a] *Dept. of Computer Science, University of Pisa, Italy*
[b] *Institute of Information Science and Technologies, CNR, Italy*
[c] *Dept. of Psychology and Cognitive Science, University of Trento, Italy*
[*] these authors contributed equally

Abstract. Word associations have been extensively used in psychology to study the rich structure of human conceptual knowledge. Recently, the study of word associations has been extended to investigating the knowledge encoded in LLMs. However, because of how the LLM word associations are accessed, existing approaches have been limited in the types of comparisons that can be made between humans and LLMs. To overcome this, we create LLM-generated word association norms modeled after the Small World of Words (SWOW) human-generated word association norms consisting of over 12,000 cue words. We prompt the language models with the same cues and participant profiles as those in the SWOW human-generated norms, and we conduct preliminary comparative analyses between humans and LLMs that explore differences in response variability, biases, concreteness effects, and network properties. Our exploration provides insights into how LLM-generated word associations can be used to investigate similarities and differences in how humans and LLMs process information.

Keywords. artificial intelligence, large language models, network science, word associations, mental lexicon, cognitive modeling, cognitive bias

1. Introduction

Understanding the mental lexicon is an important part of studying human cognition. The way in which we represent concepts in terms of relationships between them is essential for understanding how we process information, and consequently, how we reason, form beliefs, and make decisions [1]. One way to tap into the rich structure of human conceptual knowledge is through word associations, obtained by asking participants to produce associated responses when prompted with a cue word. Word associations have been extensively used in cognitive psychology and linguistics for studying lexical retrieval [2,3], semantic organization [4], and similarity judgments [5,6,7], as well as investigating concreteness effects [8,9] and cognitive biases [10]. Additionally, cognitive models built from word association norms can provide powerful insights into many different cognitive phenomena, such as language learning [11] and creativity [12]

[1]Corresponding Author: Katherine Abramski, katherine.abramski@phd.unipi.it.

Recently, researchers have begun to use word associations as a method for investigating the capabilities and limitations of LLMs [13], including investigations of biases [14,15]. Most of these approaches investigate the embedding space of language models in order to gain access to their associations [16]. They then compare these LLM associations to the well studied properties of human-generated word associations, such as asymmetry and intransitivity [17]. While this approach provides important insights, it is limited for a few reasons. First, since associations extracted from embedding spaces are usually based on cosine similarity, they are symmetrical, unlike human associations. Also, contextual embeddings must first be transformed into static embeddings [18] , which can introduce bias and distort similarity estimates [19]. Challenges such as these related to investigating contextual embeddings have led to a broader shift in how researchers approach investigating LLMs [19], from a bottom-up approach to a top-down approach that probes LLMs in a variety of cognitive and linguistic tasks in order to better understand their capabilities and reconstruct their cognitive architecture [20,21,22]. This new approach has led to the emergence of *machine psychology* [23], a new field which entails applying the tools of cognitive psychology to investigate the behavior of machines as if they were human participants in psychological experiments. The machine psychology approach has several advantages. First, probing methods can be applied regardless of the type of LLM – a significant advantage considering the rapid pace of LLM advancements – and also, the top-down approach allows for more direct comparisons between humans and LLMs.

One recent study applied a machine psychology approach to compare human word associations with LLM word associations accessed from rule mining on word sequences sampled from LLMs [17]. While this approach more closely imitates human word associations, the probing method still differs significantly from how human word associations are accessed, limiting the types of comparisons that can be made. In this work, we aim to close that gap by creating datasets of LLM-generated word association norms that are directly comparable to human-generated norms. We model our dataset after the Small World of Worlds English word association norms (SWOW) [7], the largest and most recent dataset to date. We prompt Mistral AI's large language model – in particular the mistral-7b-instruct-0.1 model – to produce responses to the same exact cues that are present in the SWOW dataset. We create two datasets of LLM-generated norms: for the first dataset, we prompt the model with only the cue words. For the second, we prompt the model with cue words as well as the exact profiles of the participants in the original SWOW experiment (i.e. age, gender, etc.). In this way, we also investigate how well the model is capable of simulating (its own interpretation of) a specific profile.

The aim of this working paper is to provide an overview of the datasets and a brief demonstration of how they may be used. The remainder of the paper is organized as follows. In Section 2 we present the methodology used to generate and preprocess the data and we discuss the preliminary comparative analyses that we performed on all three sets of norms: the original SWOW dataset, the Mistral-without-profiles dataset, and the Mistral-with-profiles dataset. In Section 3 we present the results of our preliminary analyses, and in Section 4 we briefly discuss directions for future work.

2. Experimental settings and methodology

We prepared the input to the model using the preprocessed original SWOW dataset containing 12,282 unique cue words, each repeated 100 times, with three responses (R1, R2, and R3) per cue token. We matched the SWOW participant profiles to their corresponding cue words to ensure that the Mistral-with-profiles dataset would be aligned with the original SWOW data. We then prompted the model to provide three associations each time it was presented with a cue word. In the case of the dataset with participant profiles, we also asked the model to respond as if it were a person with the specified profile.

Preprocessing of the responses consisted of various steps. First, proper names and spelling errors were corrected (including changing British spelling to American spelling) using mapping tables from the original SWOW experiment. Prefixes *the*, *a an*, and *and* were also removed from the responses, unless the response was among the cues (e.g. *a lot*). Then, a series of ad-hoc filters were applied to remove nonsensical responses such as *printassociation*1, corresponding to 0.75% and 4.32% of all responses in the Mistral-without-profiles data and the Mistral-with-profiles data, respectively. Finally, duplicate responses and responses identical to their cues were removed.

Following data preprocessing, we performed the following preliminary exploratory analyses to capture peculiarities of the generated datasets[2].

Properties of cues and responses. We investigated the properties of cues and responses by counting the numbers of tokens, numbers of types, and the percentage of missing responses in all three datasets. We also calculated the percentage overlap of the original responses compared to the LLM responses. Additionally, we compared the distributions of the number of unique responses per cue across the three datasets. These statistics provide important insights about the richness of the responses provided by humans compared to LLMs.

Investigating biases and relations. Since word associations are generally spontaneous and automatic, they can serve as a window into our implicit biases. In order to explore any potential differences in gender biases across the datasets, we looked at the top ten most frequent response tokens to the cues *man* and *woman*. We also investigated differences in the types of relations that responses shared with the cues, specifically, paradigmatic vs. syntagmatic relations. Paradigmatic relations are those that can be expressed in a taxonomy or can be substituted for each other [24], including synonymy (i.e. *woman – lady*), antonymy (i.e. *woman – man*), and hypernymy (i.e. *woman – person*). Rather, syntagmatic relations are those that tend to occur in similar contexts [25](i.e. *woman – feminism*). Whether a response has a paradigmatic or a syntagmatic relation with the cue word has important implications for how lexical data is processed, and may also provide insights about how we form biases.

Concreteness effects. Concreteness effects are nuanced and complex differences in how we process lexical information with regards to the abstractness/concreteness

[2]The link to the repository containing the LLM-generated norms will be made available upon the acceptance of the working paper.

of a word. We were interested in investigating the concreteness effect that concrete words have stronger but fewer associates while abstract words have weaker but more associates [8]. We investigated this effect in all three datasets.

Network comparisons. Representing word association norms as complex networks enables us investigate structural properties of the mental lexicon that would otherwise not emerge. We built weighted directed networks from the three datasets such that cues are source nodes and responses are target nodes. Therefore, edges are directed from cues to responses and weighted based on the frequency of the association. We then considered only the largest strongly connected components, keeping only those nodes that were both cues and responses. We report network statistics for these three networks (i.e. density, clustering coefficient, etc.) and then we make pairwise comparisons of the three networks to quantify how similar and different they are. First, we made pairwise comparisons of their sets of nodes. We use the Jaccard coefficient (i.e. $(A \cap B)/A \cup B)$) as a measure of similarity, and we also calculated the respective set difference percentages (i.e. $(A - B)/A$ and $(B - A)/B$). To compare sets of edges, we considered only the edges in the node intersection of the networks being compared. We then calculated the same measures that we used to assess similarities and differences between the sets of nodes, that is, the Jaccard coefficient and the respective set difference percentages.

3. Experimental Results

In this section we report the preliminary results obtained from performing the exploratory analyses defined in Section 2.

Properties of cues and responses Table 1 displays the statistics reflecting the properties of the cues and responses of the three datasets. All three datasets have the same cues, and they differ only in their responses. The Mistral-without-profiles norms have the lowest percentage of missing responses, while the Mistral-with-profiles norms have the highest percentage of missing responses. The number of response types (unique responses), however, is significantly higher in the original dataset compared to the LLM datasets. The Mistral-without-profiles dataset also has slightly more response tokens than the Mistral-with-profiles dataset. In line with these statistics, 77.3% of the original response types are not in the Mistral-without-profiles responses, compared to just 30.7% of the Mistral-without-profiles response types that are not in the original responses. These percentages diverge even further to 82.9% and 29.9% respectively when comparing the original data to the Mistral-with-profiles data. These statistics indicate that humans generate a much wider variety of responses than both the LLMs, and that Mistral-without-profiles generates a slightly wider variety of responses compared to Mistral-with-profiles. These statistics are also reflected in the histograms in Figure 1, that display the number of unique responses per cue for the three datasets. Unlike the human distribution, the LLM ones are skewed right, reflecting the tendency to produce fewer unique responses to cues compared to humans.

Investigating biases and relations. Table 2 displays the top ten most frequent response tokens to *man* and *woman*. We immediately notice that the responses pro-

		Original	Mistral w/o profiles	Mistral w/ profiles
Number of Tokens	Cues	1,228,200	1,228,200	1,228,200
	R1	1,197,104	1,190,773	1,114,685
	R2	1,148,452	1,182,667	1,080,649
	R3	1,058,117	1,144,556	1,002,303
	R123	3,403,673	3,517,996	3,197,637
Number of Types	Cues	12,282	12,282	12,282
	R1	64,824	23,438	20,482
	R2	75,466	30,595	20,369
	R3	76,817	31,733	19,880
	R123	134,217	43,993	32,792
Percentage Missing Tokens	R1	2.53%	3.10%	9.24%
	R2	6.49%	3.71%	12.0%
	R3	13.9%	6.81%	18.4%
	R123	7.62 %	4.52 %	13.2%

Table 1. Statistics for cues, R1, R2, R3, and R123 (all responses combined) are provided for the three datasets. Statistics include the number of tokens (total counts), the number types (unique counts), and the percentage of missing response tokens.

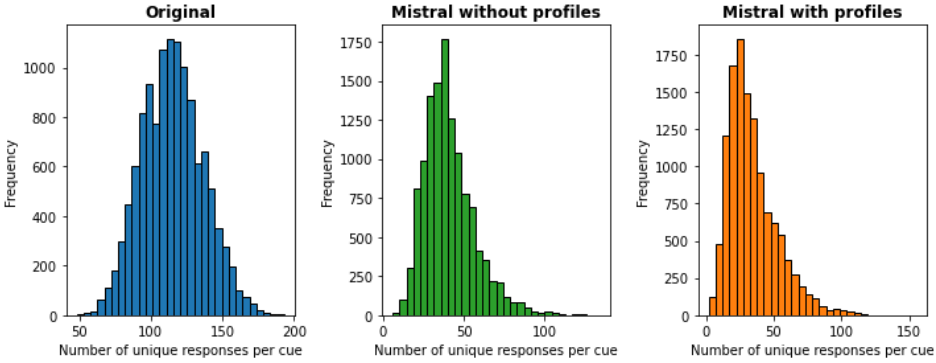

Figure 1. Histograms displaying the number of unique responses per cue for the original data (left), Mistral data without profiles (center), and Mistral data with profiles. While the original data are quite symmetrical, the Mistral data are skewed right, tending towards fewer unique responses.

Original		Mistral w/o profiles		Mistral w/ profiles	
man	woman	man	woman	man	woman
woman	**man**	**human**	*hair*	*car*	*hair*
human	**female**	**male**	*makeup*	*work*	**female**
male	**girl**	*shirt*	*beauty*	*job*	*makeup*
child	**lady**	**person**	**female**	*road*	*fashion*
boy	**mother**	*suit*	*strength*	*career*	*beauty*
person	**person**	*hair*	*fashion*	*city*	*elegance*
guy	*sex*	**adult**	**human**	*suit*	**human**
husband	*beauty*	*tie*	**adult**	*computer*	*dress*
strong	**wife**	*computer*	*dress*	*tie*	*style*
gender	*gender*	*work*	*grace*	*truck*	**adult**

Table 2. The top ten most frequent responses to the cues *man* and *woman* for the three datasets are shown. Responses shown in bold reflect paradigmatic relations with the cue word, while responses shown in italics reflect syntagmatic relations with the cue word.

duced by the LLMs are blatant stereotypical gender biases (e.g. *woman – makeup*, *man – career*). While we observe some biases among the human-generated responses (e.g. *man – strong*), they are not nearly as pronounced as those pro-

duced by the LLMs. We also notice that responses produced by humans tend towards paradigmatic relations with the cue words (shown in bold in Table 3) while responses produced by LLMs tend towards syntagmatic relations (shown in italics in Table 2). This response pattern may in fact be tied to the gender biases observed, since syntagmatic relations are arguably more subjective than paradigmatic relations (related to context rather than logical relationships), and therefore leave more room for biased perceptions.

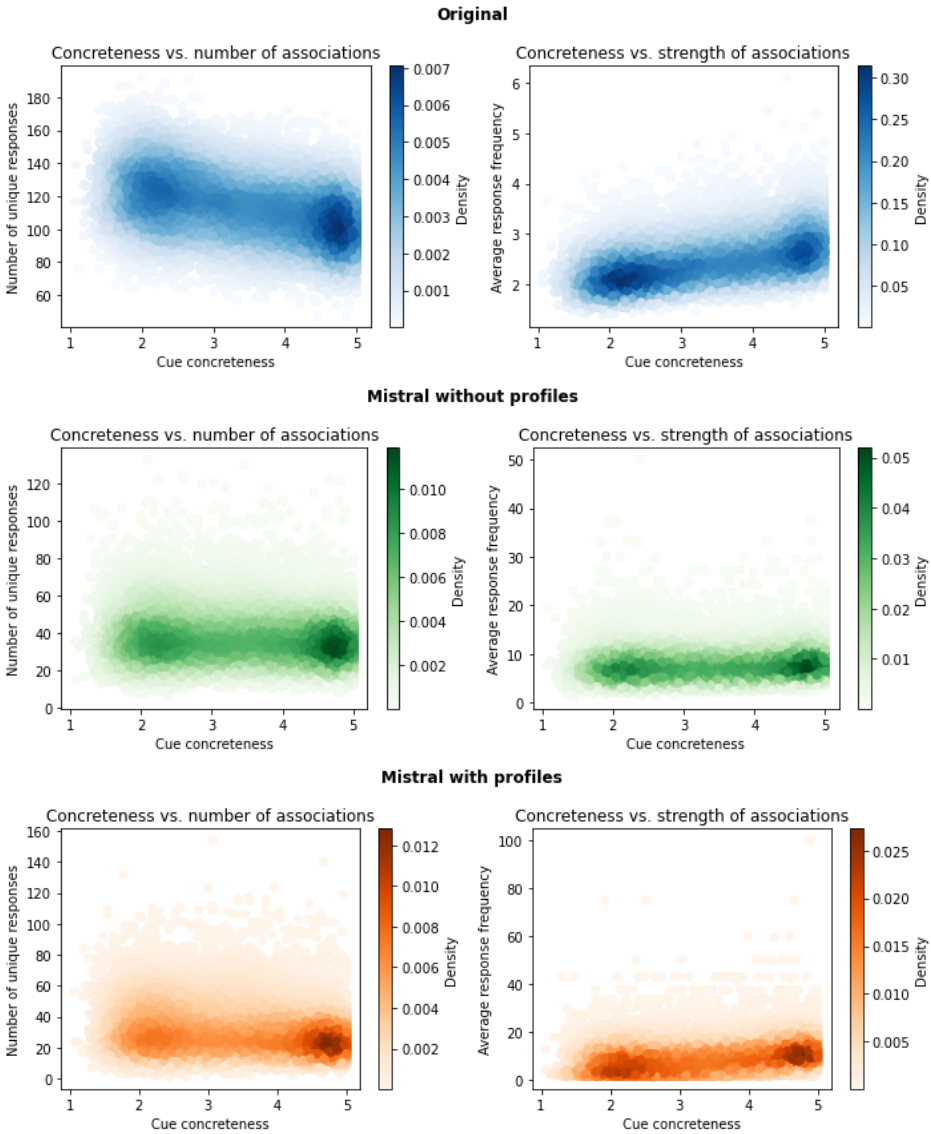

Figure 2. Concreteness effects for the original data (top), mistral data without profiles (center), and mistral data with profiles (bottom). Plots on the left display cue concreteness vs. number of associations, while plots on the right display cue concreteness vs. strength of associations.

Concreteness effects. The density plots in Figure 2 show cue concreteness vs. number of associations (left), and cue concreteness vs. strength of associations (right), for each of the three datasets. The expected concreteness effect [8] is present in the original data evidenced by the downward slope on the left (higher cue concreteness, fewer associates) and the upward slope on the right (higher cue concreteness, stronger associates). This effect appears to be absent in the Mistral-without-profiles data, and very subtle in the Mistral-with-profiles data.

Network comparisons. The networks shown in Figure 3 are subgraphs centered around the cue word *dog*, including only the top ten most frequent response nodes and the weighted directed edges from the cue *dog* to the responses. These visualizations demonstrate how the networks were built, and they also provide an idea of the types of differences that can observed between the networks. For example, in the original network, the most frequent response to *dog* is *cat*, as evidenced by the very strongly weighted edge from *dog* to *cat*. Interestingly, *cat* is completely absent among among the responses in the LLM subgraphs. Instead the top responses are *bark* and *pet*. Another interesting property that we can observe is that the responses with paradigmatic relations to the cue appear to be the same in all three subgraphs (*puppy*, *canine*, *animal(s)*, *pet*) while responses with syntagmatic relations to the cue seem to account for most of the variation among the responses (*cat*, *love*, *bone*, *furry*, *leash*, *loyal*).

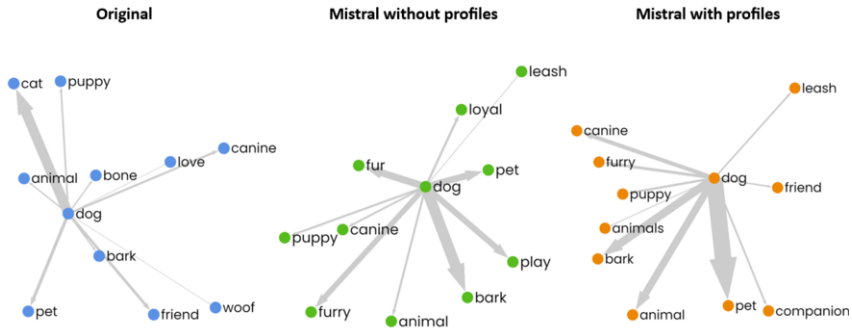

Figure 3. Subgraphs centered around the cue word *dog* are show for all three networks. They include only the neighbors with the top ten highest in-degree (top ten most frequent responses). Only weighted directed edges from the cue *dog* to the responses are shown.

Table 3 shows the network statistics for the three networks constructed. We note that the LLM networks are sparser than the original networks. They also have much higher average edge weights and lower degrees compared to the original network. These statistics reflect the higher variation among responses in the original data compared to the LLM-generated data.

Table 4 shows the statistics that reflect the similarities and differences from the pairwise network comparisons. Regarding the pairwise node comparisons, there is a large percentage of node overlap in all three pairwise comparisons, evidenced by the high Jaccard coefficients. We observe that all nodes in the LLM networks are also in the original network, but that 7.8% and 19.5% of the nodes in the original network are not in the Mistral-without-profiles network and Mistral-

Statistics	Original	Mistral w/o profiles	Mistral w/ profiles
nodes	12,228	11,271	9,841
edges	1,067,696	360,683	267,003
density	0.00714	0.00284	0.00276
$< k_{in} >, < k_{out} >$	87.3	32	27.1
$< k_{in_w} >, < k_{out_w} >$	241.9	259.7	244.8
$< w >$	2.77	8.11	9.02
$< C >$	0.078	0.092	0.079
$< L >$	2.72	3.89	4.47
D	5	9	11

Table 3. Network statistics of the three networks are shown. The statistics include the numbers of nodes and edges, the network density, the average in-degree $< k_{in} >$ and out-degree $< k_{out} >$, the average weighted in-degree $< k_{in_w} >$ and weighted out-degree $< k_{out_w} >$, the average edge weight $< w >$, the average clustering coefficient $< C >$, the average shortest path length $< L >$ and the network diameter $< C >$.

Nodes	A	B	(A-B)/A	Jaccard	(B-A)/B
	Original	Mistral w/o profiles	0.0783	0.9217	0
	Original	Mistral w/ profiles	0.1952	0.8048	0
	Mistral w/ profiles	Mistral w/o profiles	0.1400	0.8488	0.0150
Edges	A	B	(A-B)/A	Jaccard	(B-A)/B
	Original	Mistral w/o profiles	0.8230	0.1418	0.5394
	Original	Mistral w/ profiles	0.8624	0.1162	0.5726
	Mistral w/ profiles	Mistral w/o profiles	0.5689	0.2998	0.5040

Table 4. Pairwise network comparisons of the three networks are shown. The upper table shows statistics for comparisons between the sets of nodes in the networks, while the lower table shows statistics for comparisons between the sets of edges, considering only edges in the node intersection of the two networks being compared. The Jaccard coefficient reflects the similarity between sets, while (A-B)/A and (B-A)/B reflect the respective set differences.

with-profiles network, respectively. These nodes represent cues that were never given as responses by the LLMs. Regarding the pairwise edge comparisons, we observe that there is very little overlap between sets of edges, especially between the original network and the LLM networks. The LLM networks are much more similar to each other than they are to the original network, with a Jaccard coefficient of 30% compared to 14.1% and 11.6%, respectively.

4. Conclusions and future work

We provided an overview of two novel LLM-generated word association datasets proposing some preliminary analyses that demonstrate how comparisons between human-generated and LLM-generated norms can be used to investigate various aspects of information processing. We find that human-generated responses are much richer and more varied than LLM ones. Also, Mistral-without-profiles responses are slightly more varied than Mistral-with-profiles ones, suggesting that more detailed prompts may limit response variability. We also observe stronger gender biases and weaker concreteness effects in the LLM-generated norms compared to the human-generated norms. In future work, we would like to expand our network analyses by exploring spreading activation processes on feature-rich networks to investigate the emergence of cognitive biases in humans and LLMs. Such investigations could have important implications for human-AI interaction.

References

[1] Davis P, Kryszewska H. Aitchison, J.(1994) Words in the mind: An introduction to the mental lexicon, Oxford: Blackwell. Altenberg, B.(1998)'On the phraseology of spoken English: the evidence of recurrent word-combinations', in AP Cowie (Ed.), Phraseology: theory, analysis and application (pp. 101–122), Oxford: Oxford University Press. Boers, F. & Lindstromberg, S.(2005)'Finding ways to make phrase-learning feasible: The mnemonic.

[2] De Deyne S, Navarro DJ, Storms G. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. Behavior research methods. 2013;45:480-98.

[3] 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. Journal of cognition. 2018;1(1).

[4] Steyvers M, Tenenbaum JB. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. Cognitive science. 2005;29(1):41-78.

[5] De Deyne S, Storms G. Word associations: Network and semantic properties. Behavior research methods. 2008;40(1):213-31.

[6] Kenett YN, Levi E, Anaki D, Faust M. The semantic distance task: Quantifying semantic distance with semantic network path length. Journal of Experimental Psychology: Learning, Memory, and Cognition. 2017;43(9):1470.

[7] De Deyne S, Navarro DJ, Perfors A, Brysbaert M, Storms G. The "Small World of Words" English word association norms for over 12,000 cue words. Behavior research methods. 2019;51:987-1006.

[8] Felix H, Christian B, et al. A quantitative empirical analysis of the abstract/concrete distinction. 2014.

[9] Buades-Sitjar F, Planchuelo Fernández C, Dunabeitia Landaburu JA. Valence, arousal and concreteness mediate word association. 2021.

[10] Schnabel K, Asendorpf JB. Free associations as a measure of stable implicit attitudes. European Journal of Personality. 2013;27(1):39-50.

[11] Citraro S, Vitevitch MS, Stella M, Rossetti G. Feature-rich multiplex lexical networks reveal mental strategies of early language learning. Scientific Reports. 2023;13(1):1474.

[12] Beaty RE, Kenett YN. Associative thinking at the core of creativity. Trends in Cognitive Sciences. 2023.

[13] Thawani A, Srivastava B, Singh A. Swow-8500: Word association task for intrinsic evaluation of word embeddings. In: Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP; 2019. p. 43-51.

[14] Kaneko M, Bollegala D. Debiasing pre-trained contextualised embeddings. arXiv preprint arXiv:210109523. 2021.

[15] 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 and Cognitive Computing. 2023;7(3):124.

[16] Rodriguez MA, Merlo P. Word associations and the distance properties of context-aware word embeddings. In: Proceedings of the 24th Conference on Computational Natural Language Learning; 2020. p. 376-85.

[17] 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; 2022. p. 5959-70.

[18] 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; 2020. p. 4758-81.

[19] Apidianaki M. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. Computational Linguistics. 2022:1-60.

[20] Binz M, Schulz E. Using cognitive psychology to understand GPT-3. Proceedings of the National Academy of Sciences. 2023;120(6):e2218523120.

[21] Srivastava A, Rastogi A, Rao A, Shoeb AAM, Abid A, Fisch A, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:220604615. 2022.

[22] Shiffrin R, Mitchell M. Probing the psychology of AI models. Proceedings of the National Academy of Sciences. 2023;120(10):e2300963120.
[23] Hagendorff T. Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods. arXiv preprint arXiv:230313988. 2023.
[24] Hjørland B. Theoretical development of information science: a brief history. Recuperado de https://goo gl/TAVcFD. 2015.
[25] The encyclopedia of language and linguistics. 1993.