PeerJ Computer Science

Towards a psychology of machines: large language models predict human memory

Elanur Ulakci[1,2], Jan Pascal Göbel[1,2] and Markus Huff[1,2]

[1] Leibniz-Institut Für Wissensmedien, Tübingen, Germany
[2] Eberhard-Karls-Universität Tübingen, Tübingen, Germany

ABSTRACT

Large language models (LLMs), such as ChatGPT, have shown remarkable abilities in natural language processing, opening new avenues in psychological research. This study explores whether LLMs can predict human memory performance in tasks involving garden-path sentences and contextual information. In the first part, we used ChatGPT and Google Gemini to rate the relatedness and memorability of garden-path sentences preceded by either fitting or unfitting contexts. In the second part, human participants read the same sentences, rated their relatedness, and completed a surprise memory test. The results demonstrated that ChatGPT and Google Gemini's relatedness ratings closely matched those of the human participants, and their memorability ratings effectively aligned with human memory performance. Both LLM and human data revealed that higher relatedness in the unfitting context condition was associated with better memory performance, aligning with probabilistic frameworks of context-dependent learning. These findings suggest that LLMs, despite lacking human-like memory mechanisms, can model aspects of human cognition and serve as valuable tools in psychological research. The field of machine psychology explores this interplay between human cognition and artificial intelligence, offering a bidirectional approach where LLMs can both benefit from and contribute to our understanding of human cognitive processes.

Subjects Human-Computer Interaction, Artificial Intelligence, Computational Linguistics, Data Mining and Machine Learning
Keywords Large language models, Generative AI, Ambiguity, Psychology

Submitted 16 March 2025
Accepted 3 October 2025
Published 31 October 2025

Corresponding author
Markus Huff,
m.huff@iwm-tuebingen.de

Academic editor
Davide Chicco

Additional Information and Declarations can be found on page 14

DOI **10.7717/peerj-cs.3324**

INTRODUCTION

Transformer-based large language models (LLMs) have revolutionized the area of natural language processing with their exceptional performance, offering profound implications for the science of psychology, given the integral role of language in all subfields (*Demszky et al., 2023*). Despite their inherent lack of subjective experience possessed by humans, such as thinking or feeling, LLMs are arguably more than mere stochastic replicators of their input statistics (*Huff & Ulakçi, 2024*). They are designed to produce realistic human-like linguistic outputs, and they are trained on extensive datasets crafted by humans themselves. These attributes have markedly elevated the prominence of LLMs in the growing field of machine psychology. Although–from a linguistic perspective–LLMs are no models of cognitive processes (*Editorial, 2023*), studies in this field explore the capabilities of LLMs and the degree to which they align with human behavior and cognitive processes (*Aher, Arriaga & Kalai, 2023*; *Binz & Schulz, 2023*; *Buschoff et al., 2023*; *Dhingra et al., 2023*; *Dillion et al., 2023*; *Gilardi, Alizadeh & Kubli, 2023*; *Horton, 2023*;

Michelmann et al., 2025; *Ritter et al., 2017*; *Tak & Gratch, 2023*). Taken together, these studies emphasize a remarkable congruity between human cognition and LLMs, indicating the potential of LLMs to play a meaningful role in psychological research. Although prior research has shown that LLMs mirror many classic findings in psychology, comparatively less attention has been directed toward their potential to test targeted theoretical accounts of human memory. In particular, the probabilistic cue-integration framework (*Heald, Wolpert & Lengyel, 2023*), provides predictions that have not yet been examined in LLMs: How contextual fit modulates memory through cue-accumulation processes. The present study addresses this gap by integrating the established psycholinguistic paradigm of garden-path sentences with LLM-derived ratings of relatedness and memorability, to examine whether models capture the cue-accumulation dynamics proposed by the framework. In this article, we go beyond the current understanding of LLMs by investigating their potential for approximating human memory performance. To ensure the robustness of our findings concerning ChatGPT, our investigation also incorporates analyses of one additional LLM.

Machine psychology: exploring the intersection of human and artificial cognition

Following their launch, LLMs have exhibited exceptional capabilities in natural language processing (NLP) and production (*Manning et al., 2020*). Ranging from the simplest interactions as chatbots to their implementation in various areas such as education (*Kasneci et al., 2023*), healthcare (*Thirunavukarasu et al., 2023*), and information search (*Stadler, Bannert & Sailer, 2024*), these models have woven themselves into human experiences with their remarkable advantages and the convenience they provide. Beyond their expertise in NLP and practical applications, LLMs reveal significant alignment with human cognitive functions and behaviors in psychological tests where they are evaluated as subjects, despite not being fundamentally developed for this purpose. To effectively utilize LLMs, and to better understand the increasing harmony between them and humans, it is crucial to examine the strengths and limitations of these models. Nonetheless, the lack of transparency in how exactly LLMs function poses a challenge (*Schwartz, 2022*; *Zubiaga, 2024*). To address this challenge and to open up these "black boxes", psychology offers a strong framework with its multidisciplinary perspective. This article contributes to machine psychology, an area dedicated to explore the interplay between human cognition and behavior, as well as the capabilities of LLMs. In the field of machine psychology, LLMs are perceived as a distinct yet analogous "second mind" that is yet to be uncovered, focusing on their resemblance to human cognitive functions while acknowledging their artificial nature. This emerging field adopts a bidirectional approach, integrating the unique capabilities of LLMs and the rich and profound knowledge of psychology to deepen the understanding of both artificial and human cognition. On the one hand, psychological knowledge and methodologies can be applied to understand LLMs through examining them as participants in assessments and experiments. This approach enables the systematic exploration of their internal processes, behavioral tendencies, and cognitive patterns, as well as the extent to which these models mirror human cognition

(*Binz & Schulz, 2023*; *Mei et al., 2024*; *Webb, Holyoak & Lu, 2023*). On the other hand, LLMs possess significant potential to advance the study of human cognition. Their distinct information processing mechanisms open new avenues for research and provide the opportunity to use LLMs as proxies/simulations for investigating human cognitive processes, thereby broadening our understanding of the human mind (*Binz & Schulz, 2023*; *Horton, 2023*). Drawing from the concept of symbiotic autonomy in human-robot interaction studies, which focuses on the reciprocal cooperation between humans and robots (*Vanzo et al., 2020*), we propose a similar bidirectional approach of machine psychology, where humans and LLMs mutually contribute to advance the understanding of the complexities of both human and artificial cognition. Our focus in this article is on the latter direction, utilizing LLMs as a proxy to investigate human cognitive processes. Here, we evaluate whether LLMs can approximate human memory performance through testing the correspondence between human memory performance and the outputs of LLMs. We achieve this by conducting a language-based memory task that incorporates garden-path sentences and contextual information. In the following sections, we examine the implications of LLMs' ability to approximate the memory performance of humans, despite their absence of a human-like memory mechanism.

Context-driven memory: a framework for evaluating LLMs' predictions

Our investigation into the potential of LLMs to approximate human memory performance is centered on the critical role of context in the formation of memories. A fitting prior context[1] is important in resolving ambiguity (*Szewczyk & Federmeier, 2022*), the pervasive and fundamental element of the natural language, which poses a challenge to comprehension (*MacGregor et al., 2020*). Resolving ambiguity is a demanding cognitive procedure that involves considering the context, background knowledge, and making inferences from other important linguistic cues to determine the intended meaning. The recognized detrimental impact of ambiguity on language processing is not exclusive to humans, it also constitutes a significant impediment for LLMs (*Irwin, Wilson & Marantz, 2023*; *Liu et al., 2023*). While generative pre-trained transformer (GPT) employs an autoregressive design with the sequential generation process, bidirectional encoder representations from Transformers (BERT) builds upon this approach by utilizing a bidirectional architecture, considering both preceding and following contexts (*Naveed et al., 2024*). Despite possessing advanced linguistic processing capabilities, both models encounter difficulties in effectively managing the ambiguity inherent in language, mirroring the challenges present in human cognitive processing during comprehension (*Irwin, Wilson & Marantz, 2023*; *Liu et al., 2023*). This shared challenge of managing ambiguity is particularly evident in the interpretation of complex sentence structures, such as garden-path sentences (*Fujita, 2021*; *Li et al., 2024*). Although grammatically correct, these sentences introduce temporary ambiguity due to flawed parsing which can lead to misinterpretation and hinder comprehension (*Christianson et al., 2001*). While garden-path sentences present serious challenges for understanding due to inherent ambiguity, a fitting prior context is acknowledged to reduce those complications (*Grodner, Gibson & Watson, 2005*; *Kaiser & Trueswell, 2004*). The beneficial impact of fitting prior

[1] A fitting context is a preceding utterance that activates an appropriate mental model in the receiver's mind, making it easier to comprehend the present phrase.

context on comprehending garden-path sentences can be explained through the structure-building framework (*Gernsbacher, 1997*). This framework suggests that comprehension involves constructing mental representations by linking incoming information to an established structure. In this case, the fitting prior context sentence provides a foundation for the mental structure, onto which the garden-path sentence can be integrated. Being part of the same mental representation, the context sentence fitting the garden-path sentence aids in clarifying ambiguity enabling the cohesive processing of both sentences and improving comprehension (*Brich et al., 2024*).

The help of prior context is not only important for facilitating understanding, but also plays a key role for memory formation. Successful comprehension is achieved by extracting the meaning of sentences (*Kaup et al., 2023*). As outlined in the structure-building framework, comprehension is achieved by creating coherent mental structures through integrating incoming fitting information (*Gernsbacher, 1997*). This process involves envisioning the situations described in the text (*Gernsbacher & Robertson, 1995*; *Schütt et al., 2023*), which contributes to deeper understanding. As a result, well-understood content is more likely to be retained over time (*Kintsch et al., 1990*). In the case of not being able to comprehend the sentence (*e.g.*, being provided with prior unfitting information), readers make use of the surface form of information (*Schnotz & Bannert, 2003*), known to be forgotten well before the meaning of the text (*verbatim* effect) (*Poppenk et al., 2008*). By facilitating the construction of coherent mental representations, as described by the structure-building framework (*Gernsbacher, 1997*), a fitting prior context aids comprehension. This, in turn, enhances the encoding of garden-path sentences into memory, establishing a critical link between comprehension and retention.

Recent work frames context-dependent learning as a form of probabilistic cue integration, where memory improves as a function of the accumulation of relevant cues (*Heald, Wolpert & Lengyel, 2023*). In a fitting context, cues are abundant, leading to robust memory. However, in an unfitting context, the relationship between even weak contextual cues (*i.e.*, residual relatedness) and memory should be more pronounced. Although classically applied to studies of sentence processing, garden-path sentences are well-suited for probing the probabilistic cue-integration account of memory, and offer distinctive advantages. During encoding, their temporary ambiguity introduces a dynamic change in the availability of semantic cues as the sentence unfolds, enabling systematic manipulation of cue strength. This balance of naturalistic comprehension with precise experimental control makes garden-path sentences powerful tools for testing whether LLMs capture the cue-accumulation observed in humans, that is, whether LLM predictions align with human memory performance under conditions of varying contextual support. Beyond simple comprehension, the processing of the ambiguous garden-path sentences thus provides a unique testbed for models of memory. LLMs, which excel at learning probabilistic relationships in language, offer a novel way to model this process. The central theoretical question of this article is whether the behavior of LLMs, when faced with these sentences, aligns with the predictions of a probabilistic framework of human memory. This alignment would suggest that these models, despite their different architecture, capture core statistical principles that may also underlie human context-dependent memory.

EXPERIMENTAL OVERVIEW AND HYPOTHESES

Despite lacking a foundation in human cognition, LLMs achieve near-human performance across various tasks (*Binz & Schulz, 2023*). This begs the question: can LLMs, if not simply "stochastic parrots" (*Digutsch & Kosinski, 2023*), reveal insights into the underlying mechanisms of human cognition? Here, we investigate this by harnessing generative AI to predict human memory performance. Specifically, we test if LLMs can estimate how well humans remember garden-path sentences prefaced by fitting or unfitting contexts. This study probes the potential of LLMs to shed light on human information processing, even while operating on different principles.

In the first part of this study, we submitted the garden-path sentences (Sentence 2 in the prompt) with a preceding sentence (Sentence 1 in the prompt) that matched (fitting) or mismatched (unfitting) the context of the garden-path sentence to ChatGPT and Google Gemini (Table 1). We collected 100 responses for each prompt (relatedness of Sentences 1 and 2, and memorability of Sentence 2; including a robustness check with synonyms). In the second part of this study, we used the LLM responses to estimate human performance, in which we presented participants with the same material and asked them about the relatedness of the two sentences. After that, we presented participants with a surprise memory test in which we presented only the garden-path sentences and measured recognition memory. We hypothesize that LLMs' and humans' relatedness values are higher in the fitting than the unfitting condition. Further, we hypothesize LLMs' memorability ratings to be higher in the fitting than the unfitting condition. Eventually, we hypothesize that LLM memorability responses would approximate participants' memory performance in a surprise memory test.

METHOD

We report how we determined our sample size, all data exclusions (if any), all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. Portions of this text were previously published as part of a preprint (*Huff & Ulakçi, 2024*). The experiment was approved by the local ethics committee of the Leibniz-Institut für Wissensmedien (LEK 2023/051).

Data sources

LLM

We used OpenAI's API to access ChatGPT (*OpenAI et al., 2023*) (model: GPT-4; June 2023) to collect 100 responses (consisting of the relatedness and the memorability values) for each sentence pair. We set the temperature value to 1 to increase the variability in the answers. This resulted in 4,500 independent responses in the fitting and 4,500 in the unfitting condition and resulted in a total of 9,000 independent responses[2]. Furthermore, to perform the aforementioned robustness check of our AI-generated data, we performed a validation step using one more LLM for the same task with the same prompts: Google Gemini 2.0 Flash (*Google Cloud, 2025*), accessed through the Google API.

[2] Note, those responses were independent because each API call initiated a new LLM context.

Table 1 An exemplary representation of the prompts and sentence pairs submitted to ChatGPT and Google Gemini.

Context	Relatedness prompt	Memorability prompt
Fitting	Read Sentence 1 and Sentence 2 and answer the following question. How related are the two sentences from 1 (not at all) to 10 (highly)?	Read sentence 1 and sentence 2 and answer the following question. How do you rate the memorability of Sentence 2 from 1 (not at all) to 10 (excellent)?
	Sentence 1: "Bill has chronic alcoholism." Sentence 2: "Because Bill drinks wine is never kept in the house."	Sentence 1: "Bill has chronic alcoholism." Sentence 2: "Because Bill drinks wine is never kept in the house."
Unfitting	Read Sentence 1 and Sentence 2 and answer the following question. How related are the two sentences from 1 (not at all) to 10 (highly)?	Read Sentence 1 and Sentence 2 and answer the following question. How do you rate the memorability of Sentence 2 from 1 (not at all) to 10 (excellent)?
	Sentence 1: "Bill likes to play golf." Sentence 2: "Because Bill drinks wine is never kept in the house."	Sentence 1: "Bill likes to play golf." Sentence 2: "Because Bill drinks wine is never kept in the house."

Participants

We recruited 100 English-only speaking participants *via* Prolific. A total of 15 participants indicated that their vision was not normal or not corrected-to-normal during the experiment (*i.e.*, they did not wear lenses or glasses). Thus, the resulting sample consisted of 85 participants (57 female, 27 male, 1 w/o response), mean age was $M = 45.34$ years ($SD = 13.95$). All participants provided informed consent by explicitly clicking the relevant button to indicate their agreement.

Material

Garden-path sentences

We compiled a list of 45 garden-path sentences (*e.g.*, "Because Bill drinks wine is never kept in the house."). For each garden-path sentence, we constructed a sentence matching the context of the garden-path sentence (fitting context; *e.g.*, "Bill has chronic alcoholism.") and a sentence not matching its context (unfitting context; *e.g.*, "Bill likes to play golf."; for the complete list, see Table S1). For the machine data, we used all 45 garden-path sentences; for the human data, we omitted the sentence with ID 8 for counter-balancing reasons. ID 8 was chosen for omission because its sentence structure closely resembles that of ID 45, making its exclusion less critical compared to omitting other sentences.

Prompts

We submitted zero-shot prompts to ChatGPT regarding relatedness and memorability, which we presented before both sets of sentence pairs, one with a fitting context and the other with an unfitting context sentence preceding the garden-path sentence. First, we gave the prompt based on the category. Then, we provided the two sentences separately in an order as "Sentence 1" and "Sentence 2," respectively, with "Sentence 1" being the prior context sentence and "Sentence 2" being the garden-path sentence (Table 1). By the relatedness prompt, we requested ChatGPT to rate the relatedness of the sentences by giving a value from 1 (not at all) to 10 (highly). Afterward, we requested a value from ChatGPT to indicate the memorability of Sentence 2 (*i.e.*, garden-path sentence) from 1 (not at all) to 10 (excellent).

Human experiment

The experiment was programmed with PsychoPy (*Peirce et al., 2019*). All instructions and stimuli appeared in white on a gray background. The stimulus material consisted of sentence pairs, which included a garden-path sentence and its preceding context sentence. Each pair was arranged in a visually specific format, with each sentence starting on a new line, one below the other.

PROCEDURE AND DESIGN HUMAN DATA

In the learning phase, participants read 22 sentence pairs comprised of a prior context sentence and a garden-path sentence. Half of the sentences shown were in the fitting condition, the other half were in the unfitting condition. Participants read pairs of sentences at their own pace and proceeded to the next pair by pressing the spacebar. The response button (*i.e.*, spacebar) was activated three seconds after stimulus onset to ensure that the sentences were not skipped and were read by the participants. After each sentence pair, participants rated the relatedness of the two sentences by clicking on a value on the 10-point rating scale presented to them (1: "not at all" to 10: "highly"). After completing the learning phase of the experiment, participants completed a surprise old/new recognition memory test, including 44 garden-path sentences (22 targets from the learning phase and 22 distractors) without their contexts. Participants indicated whether they remembered the sentence shown on screen from the learning phase by pressing the right arrow key for "yes" and the left arrow key for "no" allowing for the calculation of sensitivity (d′) from signal detection theory (*Green & Swets, 1966*). The study employed a one-factorial design with context (fitting, unfitting) as the within-subjects factor. Four counter-balancing conditions ensured that the garden-path sentences were assigned equally to the conditions (fitting *vs.* unfitting context, target *vs.* distractor) across participants. The experiment lasted approximately 15 min.

RESULTS

Machine data

Relatedness as a function of context

We fitted a linear mixed-effect model with context (fitting, unfitting) as fixed effect and sentence ID as random intercept. We submitted the resulting model to a type-2 Analysis of Variance (ANOVA) (*Fox & Weisberg, 2010*). The relatedness of the two sentences is higher in the fitting than the unfitting condition, $\chi^2(1) = 44{,}843.00$, $p < 0.001$, constituting a successful manipulation check of the context manipulation (Fig. 1A).

Memorability as a function of context

Similar to the relatedness analysis, we fitted a linear mixed-effect model with context (fitting, unfitting) as fixed effect and sentence ID as the random intercept. Submitting the resulting model to a type-2 ANOVA showed a significant main effect of context, $\chi^2(1) = 2{,}660.00$, $p < 0.001$. In the fitting context condition, the memorability of the garden-path sentence was rated higher than in the unfitting context condition (Fig. 1B).

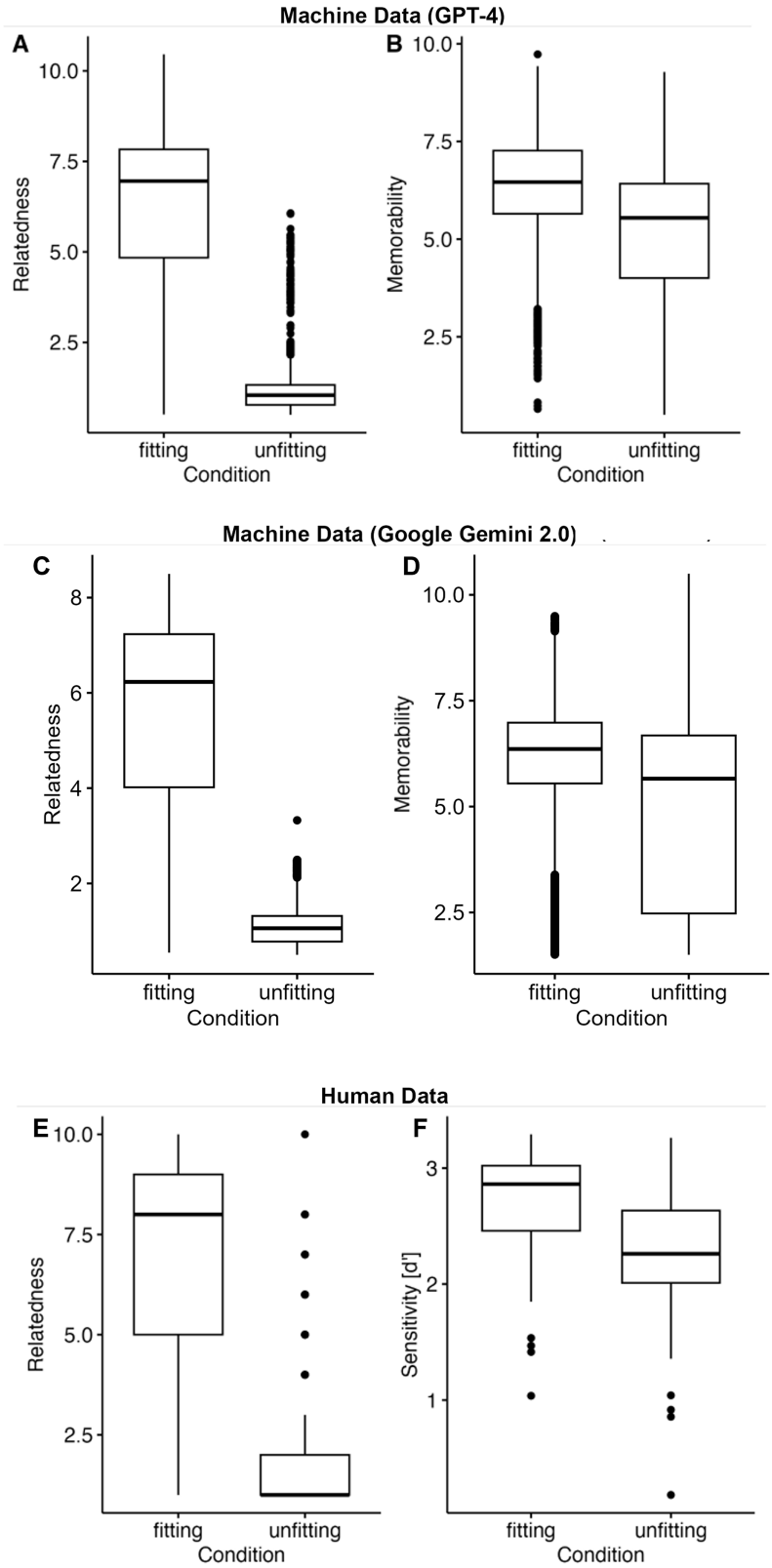

Figure 1 **Relatedness (A, C) and memorability (B, D) measures of the machine data and relatedness (E) and memory (F) of the human data as a function of context.**

Full-size DOI: 10.7717/peerj-cs.3324/fig-1

Robustness check

First, because LLMs are sensitive to small changes in the prompts (*Binz & Schulz, 2023*), we repeated the relatedness and the memorability analysis with the following prompts: How closely are the two sentences linked from 1 (not at all) to 10 (highly)? and How do you rate the recognizability of Sentence 2 from 1 (not at all) to 10 (excellent)? In particular, we changed the term related with linked and memorability with recognizability. Results resembled the main analysis (see also Fig. S1). In the fitting condition, the two sentences were judged to be more linked, $M = 6.67 (SD = 2.25)$, than in the unfitting condition, $M = 1.12 (SD = 0.59)$, $\chi^2(1) = 43{,}731.60$, $p < 0.001$. Further, the recognizability values were higher in the fitting, $M = 2.67$ ($SD = 1.34$), than in the unfitting condition, $M = 1.00$ ($SD = 0.02$), $\chi^2(1) = 43{,}731.60$, $p < 0.001$. We thus conclude that the observed effects are stable.

Second, to assess generalization across models, we replicated the main relatedness and memorability analyses (using the original prompts) with Google Gemini 2.0 Flash (Fig. 1C, 1D) as mentioned in the Methods. The further LLM demonstrated the same significant effects of context. For Google Gemini 2.0 Flash, relatedness was significantly higher in the fitting condition, $M = 5.71$ ($SD = 1.92$), than the unfitting condition, $M = 1.07$ ($SD = 0.26$), $\chi^2(1) = 51{,}111$, $p < 0.001$, and the same was true for memorability in the fitting, $M = 5.80$ ($SD = 1.73$) and unfitting condition, $M = 4.90$ ($SD = 2.20$), $\chi^2(1) = 1{,}414.5$, $p < 0.001$. Taken together, these checks suggest that the observed effects of context on relatedness and memorability are robust to variations in prompt wording for ChatGPT and generalize across different LLMs.

Human data (human experiment)

Relatedness as a function of context

Analysis was similar to the machine data with the exception that we additionally included participant as random intercept. The relatedness of the two sentences is higher in the fitting than the unfitting condition, $\chi^2(1) = 4{,}087.60$, $p < 0.001$, again constituting a successful manipulation check and replicating the machine data (Fig. 1E).

Recognition memory as a function of context

To assess participants' memory, we calculated d' from signal detection theory (*Green & Swets, 1966*), which corrects for response bias. We corrected for perfect performance. In particular, in case of no false alarms, we added 0.5 (*i.e.*, a half trial), and in case of all hits, we subtracted 0.5 (*i.e.*, a half trial). A t-test for repeated measures showed a significantly higher recognition performance in the fitting context than in the unfitting context, $t(84) = 5.75$, $p < 0.001$, Cohen's $d = 0.62$ (Fig. 1F).

Response bias (c)

Participants' responses were more liberal in the fitting ($M = -0.08$, $SD = 0.22$) than in the unfitting condition ($M = 0.13$, $SD = 0.29$), $t(84) = -5.78$, $p < 0.001$, Cohen's $d = -0.63$.

Our concluding analysis explored a potential mechanism underlying the observed effect, grounded in a probabilistic framework of context-dependent learning and stochastic

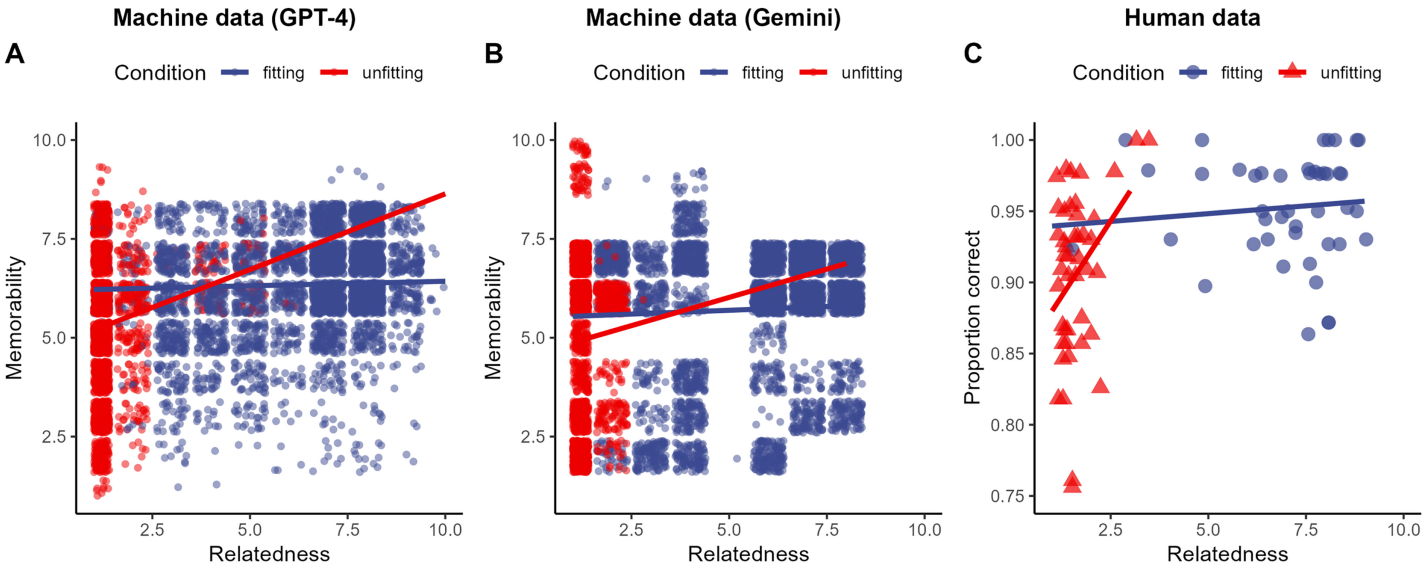

Figure 2 Memory performance as a function of condition (fitting, unfitting) and relatedness for the machine data (A, B) and the human data (C). Solid lines represent the predictions from the fitted linear mixed-effects model. Full-size DOI: 10.7717/peerj-cs.3324/fig-2

reasoning. This framework posits that memory performance improves as a function of the probabilistic accumulation and integration of retrieval cues, with context acting as a latent variable that organizes and constrains memory processes (*Heald, Wolpert & Lengyel, 2023*). More precisely, in the fitting condition, where a significant, uniform segment of information spans two sentences, memory performance likely peaks. Thus, relatedness and memory performance should be less related as compared to the unfitting condition, where the relatedness of the two sentences is lower and potentially more heterogeneous. In the latter condition, there is more room for improvement in memory performance. Consequently, in this condition, we expect a significant positive relation, with increases in relatedness of the two sentences leading to noticeable improvements in memory performance. If this is true, we should observe a significant interaction between context (fitting *vs.* unfitting) and the degree of relatedness, with higher relatedness values, approximating higher memory performance in the unfitting condition but not necessarily in the fitting condition.

To test this framework on the machine data, we fitted a linear mixed-effect model with memorability as the dependent variable, context (fitting, unfitting) as a categorical fixed effect, relatedness as a continuous fixed effect, and sentence id as a random intercept (Fig. 2A). Submitting this model to a type-2 ANOVA showed a significant interaction of context and relatedness, $\chi^2(1) = 86.37$, $p < 0.001$. As predicted, in the unfitting condition, higher relatedness values estimate higher memory performance. This effect is weaker in the fitting condition. Further, the main effects of condition, $\chi^2(1) = 259.51$, $p < 0.001$, and relatedness, $\chi^2(1) = 31.23$, $p < 0.001$, were significant.

For the analysis of the human data, we aggregated the proportion correct data[3] and the relatedness data at the sentence and context level. We fitted a linear mixed-effect model

[3] Note that the nature of this analysis—linking relatedness and memory performance data—required an analysis on the item level. Thus, we used the proportion correct data and not sensitivity (*i.e.*, d').

with proportion correct as the dependent variable, context (fitting, unfitting) as a categorical fixed effect, relatedness as a continuous fixed effect, and sentence id as a random intercept (Fig. 2C). Submitting this model to a type-2 ANOVA showed a significant interaction of context and relatedness, $\chi^2(1) = 9.39$, $p = 0.002$. As predicted, in the unfitting condition, higher relatedness values estimate higher memory performance, whereas we did not observe such an effect in the fitting condition. Further, the main effects of condition, $\chi^2(1) = 1.13$, $p = 0.289$, and relatedness, $\chi^2(1) = 1.39$, $p = 0.239$, were not significant.

Consistent with the proposed stochastic reasoning framework, a significant interaction between context and relatedness emerged for machine and human data. While higher relatedness boosted memory in the unfitting condition, this effect was weaker in the fitting condition. These findings across machine and human data emphasize how the link between relatedness and memory hinges on context and available retrieval cues.

DISCUSSION

This study investigated whether LLMs could serve as a model for a probabilistic account of human context-dependent memory. Our findings strongly support this possibility. The results showed that the relatedness ratings of ChatGPT and Gemini collected prior to human testing closely corresponded with those of the human participants. Additionally, the memorability ratings of these models, also generated before testing humans, approximated human memory performance in the surprise memory test. This alignment, while not indicative of human-like memory systems, suggests that ChatGPT and Gemini capture statistical regularities that mirror human semantic behavior under structured conditions. An analysis to check the robustness of the findings with one further LLM and synonyms (recognizable and linked for memorable and related) confirmed the results.

Importantly, these results provide clear support for the probabilistic cue-integration framework (*Heald, Wolpert & Lengyel, 2023*), which proposes that memory strength depends on the accumulation of contextual cues during encoding. By contrasting sentences in fitting *vs*. unfitting contexts, we systematically varied cue availability: fitting contexts enriched encoding with abundant cues, while unfitting contexts offered fewer, weaker cues, thereby restricted encoding to weaker residual overlap. The fact that LLM predictions tracked human performance across these conditions indicates that these models, despite their distinct architecture, reproduce the cue-integration dynamics central to this framework.

The findings from this study provide strong evidence that LLMs, such as ChatGPT and Google Gemini, have significant potential to be utilized as experimental tools for the investigation of human cognition in the context of machine psychology. The ability of the models used in approximating human memory performance even in the absence of a human-like memory system highlights its potential to represent cognitive patterns observed in humans. Particularly, the significant alignment between the relatedness ratings of human subjects and those of the models employed underscores the model's competence in processing and analyzing the linguistic information in a manner similar to humans.

Moreover, the accuracy of the model in approximating the performance of humans on the memory test also highlights the potential of LLMs as proxies for investigating the human memory processes. The results also indicate that, even though LLMs are built on a fundamentally different architecture, they display outcomes that parallel human cognitive processes, particularly in tasks that involve the understanding of contextual reasoning. Corroborated through analyses using synonyms, the robustness of our findings strengthens the soundness of this approach and demonstrates the great possibility of LLMs functioning as effective tools in psychological research.

The outcomes of this study emphasize the bidirectional benefits of machine psychology: not only LLMs benefit from psychological approaches for deeper evaluation, but researchers also gain a new perspective to examine the complexities of human cognition. We described one potential mechanism based on the interaction of context and relatedness observed in both the machine and human data, which can be effectively explained through a stochastic reasoning framework grounded in the principles of context-dependent learning and probabilistic cue integration (*Heald, Wolpert & Lengyel, 2023*). This framework posits that memory performance is influenced by the probabilistic accumulation of retrieval cues during encoding and retrieval processes, as context serves as a latent variable that organizes and constrains memory representations over time. In the fitting context condition, in which sentences share high semantic relatedness, the abundance of overlapping cues enhances encoding strength and facilitates retrieval, resulting in consistently high memory performance (*Polyn, Norman & Kahana, 2009*). Consequently, variations in relatedness within this condition have a minimal impact because the retrieval cues are already robust. Conversely, in the unfitting context condition, which begins with lower relatedness and fewer available cues, any increase in relatedness significantly boosts the availability of retrieval cues, leading to noticeable improvements in memory performance (*Jonker, Seli & MacLeod, 2013*). This aligns with the encoding specificity principle, which posits that memory retrieval is most effective when the context at encoding matches the context at retrieval (*Tulving & Thomson, 1973*). By modeling how variations in contextual relatedness influence the stochastic processing of retrieval cues, this framework provides a theoretical basis for understanding how both humans and LLMs process linguistic structures (*Bhatia & Richie, 2024*; *Binz & Schulz, 2023*). *Heald, Wolpert & Lengyel (2023)* further highlight that the probabilistic nature of context inference, reliant on dynamically evolving distributions of context-specific cues, contributes significantly to the robustness and adaptability of memory systems. An expanding body of research demonstrates that LLMs can capture human-like judgements on psycholingustic measures such as word relatedness, syntactic ambiguity, and cloze probability (*Brysbaert, Martínez & Reviriego, 2024*; *Martínez et al., 2024*; *Rivière, Beatty-Martínez & Trott, 2025*; *Trott, 2024*). These advancements that build on earlier developments in word embedding models (*Thompson & Lupyan, 2018*; *Utsumi, 2020*), highlight the growing utility of LLMs in norming experimental stimuli. Our contribution builds on this trend by examining whether LLM outputs generated under controlled experimental conditions align prospectively with human performance, thereby bridging

normative modeling with cognitive theories of semantic encoding. This broader perspective enriches our understanding of how relatedness and contextual variability interact to influence memory processes in both humans and machine learning models. These results provide a foundation for further study of LLMs as proxies and experimental tools in psychology, offering novel possibilities to discover and represent human cognition and behavior.

Taken together, these results provide evidence that LLMs are not only capable of predicting human memory performance but do so in a manner that reflects core principles of the *probabilistic cue-integration framework* (*Heald, Wolpert & Lengyel, 2023*), revealing sensitivity to how contextual fit shapes encoding strength. LLMs convergence with human memory patterns across fitting and unfitting contexts demonstrates their sensitivity to the statistical regularities that shape memory, underscoring their potential to be useful tools for probing theoretical accounts of memory despite their distinct architectures.

LIMITATIONS

Despite our thorough approach to conducting this research, there are a few limitations that should be considered. First, there can be a potential interplay between fluency and context, particularly in terms of relatedness (*Oppenheimer, 2008*). Participants might have processed sentences in the fitting condition easier, increasing the familiarity for those items, therefore making the fluency heuristic a potential contributor to the memory process during the recognition test as the relatedness of sentences increases. Second, LLMs are trained based on extensive datasets, which most probably also include the garden-path sentences we used in our study, and this could be a factor influencing their behavior (performativity problem) (*Horton, 2023*). When creating the stimuli, we deliberately chose not to create them entirely independently. Instead, we compiled a diverse set of garden-path sentences with various types of linguistic ambiguities from a wide range of sources within the literature (see Table S1). This approach allowed us to expose both humans and GPT to these sentences in a manner that reflects the way LLMs encounter and process information—through diverse and heterogeneous datasets created by a variety of contributors. By gathering input from multiple perspectives, we sought to enhance the representativeness of our stimuli, thereby establishing a realistic and comprehensive foundation for our investigation.

Lastly, our study is based on the investigation of how context influences the memorability of linguistically challenging input (*i.e.*, garden-path sentences) through facilitating understanding. Our findings are situated within this specific paradigm of context-driven memory, and do not generalize to all types of memory formations. Future studies may extend these findings by investigating how context facilitates the memorability of other linguistic structures, and assess how contextual facilitation interacts with different memory systems. Importantly, given their alignment with human performance, LLMs show potential as instruments for simulating and extending such investigations.

CONCLUSIONS

LLMs are at the forefront of research in the area of machine psychology, owing to their remarkable language processing capabilities. Our research revealed ChatGPT and Google Gemini's ability to approximate human memory performance despite not possessing it. It carries important potential for utilizing LLMs in studying human cognition.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This study was part of Elanur Ulakci's Erasmus scholarship. There was no additional external funding received for the study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Elanur Ulakci's Erasmus scholarship.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Elanur Ulakci conceived and designed the experiments, performed the experiments, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jan Pascal Göbel performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Markus Huff conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Ethics

The following information was supplied relating to ethical approvals (*i.e.*, approving body and any reference numbers):

This research was approved by the local ethics committee of the Leibniz-Institut für Wissensmedien (LEK 2023/051).

Data Availability

The following information was supplied regarding data availability:

The data and code is available at Zenodo: Huff, M., Ulakçı, E., & Göbel, J. P. (2025). Generative artificial intelligence predicts human performance (v1.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.10696562.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.3324#supplemental-information.

REFERENCES

Aher GV, Arriaga RI, Kalai AT. 2023. Using large language models to simulate multiple humans and replicate human subject studies. In: *Proceedings of the 40th International Conference on Machine Learning*. PMLR, 337–371.

Bhatia S, Richie R. 2024. Transformer networks of human conceptual knowledge. *Psychological Review* **131(1)**:271–306 DOI 10.1037/rev0000319.

Binz M, Schulz E. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences of the United States of America* **120(6)**:e2218523120 DOI 10.1073/pnas.2218523120.

Brich IR, Papenmeier F, Huff M, Merkt M. 2024. Construction or updating? Event model processes during visual narrative comprehension. *Psychonomic Bulletin & Review* **31**:2092–2101 DOI 10.3758/s13423-023-02424-w.

Brysbaert M, Martínez G, Reviriego P. 2024. Moving beyond word frequency based on tally counting: AI-generated familiarity estimates of words and phrases are an interesting additional index of language knowledge. *Behavior Research Methods* **57**:28 DOI 10.3758/s13428-024-02561-7.

Buschoff LMS, Akata E, Bethge M, Schulz E. 2023. Have we built machines that think like people? ArXiv DOI 10.48550/arXiv.2311.16093.

Christianson K, Hollingworth A, Halliwell JF, Ferreira F. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology* **42(4)**:368–407 DOI 10.1006/cogp.2001.0752.

Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, Chandhok S, Eichstaedt JC, Hecht C, Jamieson J, Johnson M, Jones M, Krettek-Cobb D, Lai L, JonesMitchell N, Ong DC, Dweck CS, Gross JJ, Pennebaker JW. 2023. Using large language models in psychology. *Nature Reviews Psychology* **54**:547 DOI 10.1038/s44159-023-00241-5.

Dhingra S, Singh M, Vaisakh SB, Malviya N, Gill SS. 2023. Mind meets machine: unravelling GPT-4's cognitive psychology. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* **3(3)**:100139 DOI 10.1016/j.tbench.2023.100139.

Digutsch J, Kosinski M. 2023. Overlap in meaning is a stronger predictor of semantic activation in GPT-3 than in humans. *Scientific Reports* **13**:5035 DOI 10.1038/s41598-023-32248-6.

Dillion D, Tandon N, Gu Y, Gray K. 2023. Can AI language models replace human participants? *Trends in Cognitive Sciences* **27(7)**:597–600 DOI 10.1016/j.tics.2023.04.008.

Editorial. 2023. Language models and linguistic theories beyond words (Editorial). *Nature Machine Intelligence* **5(7)**:677–678 DOI 10.1038/s42256-023-00703-8.

Fox J, Weisberg S. 2010. *An R companion to applied regression*. Thousand Oaks, CA: Sage.

Fujita H. 2021. On the parsing of garden-path sentences. *Language, Cognition and Neuroscience* **36(10)**:1234–1245 DOI 10.1080/23273798.2021.1922727.

Gernsbacher MA. 1997. Two decades of structure building. *Discourse Processes* **23(3)**:265–304 DOI 10.1080/01638539709544994.

Gernsbacher MA, Robertson RR. 1995. Reading skill and suppression revisited. *Psychological Science* **6(3)**:165–169 DOI 10.1111/j.1467-9280.1995.tb00326.x.

Gilardi F, Alizadeh M, Kubli M. 2023. ChatGPT outperforms crowd-workers for text-annotation tasks. *Proceedings of the National Academy of Sciences of the United States of America* **120(30)**: e2305016120 DOI 10.1073/pnas.2305016120.

Google Cloud. 2025. Gemini 2.0 flash | generative AI on vertex AI. *Available at https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash?hl=de*.

Green D, Swets J. 1966. *Signal detection theory and psychophysics*. New York: Wiley.

Grodner D, Gibson E, Watson D. 2005. The influence of contextual contrast on syntactic processing: evidence for strong-interaction in sentence comprehension. *Cognition* **95(3)**:275–296 DOI 10.1016/j.cognition.2004.01.007.

Heald JB, Wolpert DM, Lengyel M. 2023. The computational and neural bases of context-dependent learning. *Annual Review of Neuroscience* **46**:233–258 DOI 10.1146/annurev-neuro-092322-100402.

Horton JJ. 2023. Large language models as simulated economic agents: what can we learn from homo silicus? ArXiv DOI 10.48550/arXiv.2301.07543.

Huff M, Ulakçi E. 2024. Towards a psychology of machines: large language models predict human memory. ArXiv DOI 10.48550/arXiv.2403.05152.

Irwin T, Wilson K, Marantz A. 2023. BERT shows garden path effects. In: Vlachos A, Augenstein I, eds. *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Dubrovnik, Croatia: Association for Computational Linguistics, 3220–3232.

Jonker TR, Seli P, MacLeod CM. 2013. Putting retrieval-induced forgetting in context: an inhibition-free, context-based account. *Psychological Review* **120(4)**:852–872 DOI 10.1037/a0034246.

Kaiser E, Trueswell JC. 2004. The role of discourse context in the processing of a flexible word-order language. *Cognition* **94(2)**:113–147 DOI 10.1016/j.cognition.2004.01.002.

Kasneci E, Sessler K, Küchemann S, Bannert M, Dementieva D, Fischer F, Gasser U, Groh G, Günnemann S, Hüllermeier E, Krusche S, Kutyniok G, Michaeli T, Nerdel C, Pfeffer J, Poquet O, Sailer M, Schmidt A, Seidel T, Stadler M, Weller J, Kuhn J, Kasneci G. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* **103**:102274 DOI 10.1016/j.lindif.2023.102274.

Kaup B, Ulrich R, Bausenhart KM, Bryce D, Butz MV, Dignath D, Dudschig C, Franz VH, Friedrich C, Gawrilow C, Heller J, Huff M, Hütter M, Janczyk M, Leuthold H, Mallot H, Nürk H-C, Ramscar M, Said N, Svaldi J, Wong HY. 2023. Modal and amodal cognition: an overarching principle in various domains of psychology. *Psychological Research* **88(2)**:307–337 DOI 10.1007/s00426-023-01878-w.

Kintsch W, Welsch D, Schmalhofer F, Zimny S. 1990. Sentence memory: a theoretical analysis. *Journal of Memory and Language* **29(2)**:133–159 DOI 10.1016/0749-596x(90)90069-c.

Li A, Feng X, Narang S, Peng A, Cai T, Shah RS, Varma S. 2024. Incremental comprehension of garden-path sentences by large language models: semantic interpretation, syntactic re-analysis, and attention. ArXiv DOI 10.48550/arXiv.2405.16042.

Liu A, Wu Z, Michael J, Suhr A, West P, Koller A, Swayamdipta S, Smith N, Choi Y. 2023. We're afraid language models aren't modeling ambiguity. In: Bouamor H, Pino J, Bali K, eds. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 790–807.

MacGregor LJ, Rodd JM, Gilbert RA, Hauk O, Sohoglu E, Davis MH. 2020. The neural time course of semantic ambiguity resolution in speech comprehension. *Journal of Cognitive Neuroscience* **32(3)**:403–425 DOI 10.1162/jocn_a_01493.

Manning CD, Clark K, Hewitt J, Khandelwal U, Levy O. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences of the United States of America* **117(48)**:30046–30054 DOI 10.1073/pnas.1907367117.

Martínez G, Molero JD, González S, Conde J, Brysbaert M, Reviriego P. 2024. Using large language models to estimate features of multi-word expressions: concreteness, valence, arousal. *Behavior Research Methods* **57**:5 DOI 10.3758/s13428-024-02515-z.

Mei Q, Xie Y, Yuan W, Jackson MO. 2024. A Turing test of whether AI chatbots are behaviorally similar to humans. *Proceedings of the National Academy of Sciences of the United States of America* **121(9)**:e2313925121 DOI 10.1073/pnas.2313925121.

Michelmann S, Kumar M, Norman KA, Toneva M. 2025. Large language models can segment narrative events similarly to humans. *Behavior Research Methods* **57**:39 DOI 10.3758/s13428-024-02569-z.

Naveed H, Khan AU, Qiu S, Saqib M, Anwar S, Usman M, Akhtar N, Barnes N, Mian A. 2024. A comprehensive overview of large language models. ArXiv DOI 10.48550/arXiv.2307.06435.

OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, Aleman FL, Almeida D, Altenschmidt J, Altman S, Anadkat S, Avila R, Babuschkin I, Balaji S, Balcom V, Baltescu P, Bao H, Bavarian M, Belgum J, Bello I, Berdine J, Bernadett-Shapiro G, Berner C, Bogdonoff L, Boiko O, Boyd M, Brakman A-L, Brockman G, Brooks T, Brundage M, Button K, Cai T, Campbell R, Cann A, Carey B, Carlson C, Carmichael R, Chan B, Chang C, Chantzis F, Chen D, Chen S, Chen R, Chen J, Chen M, Chess B, Cho C, Chu C, Chung HW, Cummings D, Currier J, Dai Y, Decareaux C, Degry T, Deutsch N, Deville D, Dhar A, Dohan D, Dowling S, Dunning S, Ecoffet A, Eleti A, Eloundou T, Farhi D, Fedus L, Felix N, Fishman SP, Forte J, Fulford I, Gao L, Georges E, Gibson C, Goel V, Gogineni T, Goh G, Gontijo-Lopes R, Gordon J, Grafstein M, Gray S, Greene R, Gross J, Gu SS, Guo Y, Hallacy C, Han J, Harris J, He Y, Heaton M, Heidecke J, Hesse C, Hickey A, Hickey W, Hoeschele P, Houghton B, Hsu K, Hu S, Hu X, Huizinga J, Jain S, Jain S, Jang J, Jiang A, Jiang R, Jin H, Jin D, Jomoto S, Jonn B, Jun H, Kaftan T, Kaiser L, Kamali A, Kanitscheider I, Keskar NS, Khan T, Kilpatrick L, Kim JW, Kim C, Kim Y, Kirchner H, Kiros J, Knight M, Kokotajlo D, Kondraciuk L, Kondrich A, Konstantinidis A, Kosic K, Krueger G, Kuo V, Lampe M, Lan I, Lee T, Leike J, Leung J, Levy D, Li CM, Lim R, Lin M, Lin S, Litwin M, Lopez T, Lowe R, Lue P, Makanju A, Malfacini K, Manning S, Markov T, Markovski Y, Martin B, Mayer K, Mayne A, McGrew B, McKinney SM, McLeavey C, McMillan P, McNeil J, Medina D, Mehta A, Menick J, Metz L, Mishchenko A, Mishkin P, Monaco V, Morikawa E, Mossing D, Mu T, Murati M, Murk O, Mély D, Nair A, Nakano R, Nayak R, Neelakantan A, Ngo R, Noh H, Ouyang L, O'Keefe C, Pachocki J, Paino A, Palermo J, Pantuliano A, Parascandolo G, Parish J, Parparita E, Passos A, Pavlov M, Peng A, Perelman A, Peres FDAB, Petrov M, Pinto HPDO, Michael, Pokorny, Pokrass M, Pong V, Powell T, Power A, Power B, Proehl E, Puri R, Radford A, et al. 2023. GPT-4 technical report. ArXiv DOI 10.48550/arXiv.2303.08774.

Oppenheimer DM. 2008. The secret life of fluency. *Trends in Cognitive Sciences* **12(6)**:237–241 DOI 10.1016/j.tics.2008.02.014.

Peirce JW, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. 2019. PsychoPy2: experiments in behavior made easy. *Behavior Research Methods* **51**:195–203 DOI 10.3758/s13428-018-01193-y.

Polyn SM, Norman KA, Kahana MJ. 2009. A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review* **116(1)**:129–156 DOI 10.1037/a0014420.

Poppenk J, Walia G, McIntosh A, Joanisse M, Klein D, Köhler S. 2008. Why is the meaning of a sentence better remembered than its form? An fMRI study on the role of novelty-encoding processes. *Hippocampus* **18(9)**:909–918 DOI 10.1002/hipo.20453.

Ritter S, Barrett DGT, Santoro A, Botvinick MM. 2017. Cognitive psychology for deep neural networks: a shape bias case study. In: *Proceedings of the 34th International Conference on Machine Learning*. PMLR, 2940–2949.

Rivière PD, Beatty-Martínez AL, Trott S. 2025. Evaluating contextualized representations of (Spanish) ambiguous words: a new lexical resource and empirical analysis. ArXiv DOI 10.48550/arXiv.2406.14678.

Schnotz W, Bannert M. 2003. Construction and interference in learning from multiple representation. *Learning and Instruction* **13(2)**:141–156 DOI 10.1016/s0959-4752(02)00017-8.

Schwartz MD. 2022. Should artificial intelligence be interpretable to humans? *Nature Reviews Physics* **4(12)**:741–742 DOI 10.1038/s42254-022-00538-z.

Schütt E, Dudschig C, Bergen BK, Kaup B. 2023. Sentence-based mental simulations: evidence from behavioral experiments using garden-path sentences. *Memory & Cognition* **51(4)**:952–965 DOI 10.3758/s13421-022-01367-2.

Stadler M, Bannert M, Sailer M. 2024. Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior* **160(4)**:108386 DOI 10.1016/j.chb.2024.108386.

Szewczyk JM, Federmeier KD. 2022. Context-based facilitation of semantic access follows both logarithmic and linear functions of stimulus probability. *Journal of Memory and Language* **123(12)**:104311 DOI 10.1016/j.jml.2021.104311.

Tak AN, Gratch J. 2023. Is GPT a computational model of emotion? In: *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)* DOI 10.48550/arXiv.2307.13779.

Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. 2023. Large language models in medicine. *Nature Medicine* **29(8)**:1930–1940 DOI 10.1038/s41591-023-02448-8.

Thompson B, Lupyan G. 2018. Automatic estimation of lexical concreteness in 77 languages. In: *Proceedings of the Annual Meeting of the Cognitive Science Society*, 40.

Trott S. 2024. Large language models and the wisdom of small crowds. *Open Mind* **8**:723–738 DOI 10.1162/opmi_a_00144.

Tulving E, Thomson DM. 1973. Encoding specificity and retrieval processes in episodic memory. *Psychological Review* **80(5)**:352–373 DOI 10.1037/h0020071.

Utsumi A. 2020. Exploring what is encoded in distributional word vectors: a neurobiologically motivated analysis. *Cognitive Science* **44(6)**:e12844 DOI 10.1111/cogs.12844.

Vanzo A, Riccio F, Sharf M, Mirabella V, Catarci T, Nardi D. 2020. Who is willing to help robots? A user study on collaboration attitude. *International Journal of Social Robotics* **12(2)**:589–598 DOI 10.1007/s12369-019-00571-6.

Webb T, Holyoak KJ, Lu H. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour* **7(9)**:1526–1541 DOI 10.1038/s41562-023-01659-w.

Zubiaga A. 2024. Natural language processing in the era of large language models. *Frontiers in Artificial Intelligence* **6**:1350306 DOI 10.3389/frai.2023.1350306.