AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 4, 1066–1075
https://doi.org/10.1037/xge0001547

Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5

Gaurav Suri, Lily R. Slater, Ali Ziaee, and Morgan Nguyen
Mind Brain and Behavior, Department of Psychology, San Francisco State University

A Large Language Model (LLM) is an artificial intelligence system trained on vast amounts of natural language data, enabling it to generate human-like responses to written or spoken language input. Generative Pre-Trained Transformer (GPT)-3.5 is an example of an LLM that supports a conversational agent called ChatGPT. In this work, we used a series of novel prompts to determine whether ChatGPT shows heuristics and other context-sensitive responses. We also tested the same prompts on human participants. Across four studies, we found that ChatGPT was influenced by random anchors in making estimates (anchoring, Study 1); it judged the likelihood of two events occurring together to be higher than the likelihood of either event occurring alone, and it was influenced by anecdotal information (representativeness and availability heuristic, Study 2); it found an item to be more efficacious when its features were presented positively rather than negatively—even though both presentations contained statistically equivalent information (framing effect, Study 3); and it valued an owned item more than a newly found item even though the two items were objectively identical (endowment effect, Study 4). In each study, human participants showed similar effects. Heuristics and context-sensitive responses in humans are thought to be driven by cognitive and affective processes such as loss aversion and effort reduction. The fact that an LLM—which lacks these processes—also shows such responses invites consideration of the possibility that language is sufficiently rich to carry these effects and may play a role in generating these effects in humans.

Public Significance Statement

Large Language Models (LLMs) and their associated chat agents, such as OpenAI's ChatGPT, are deep-learning neural networks that sometimes display human-like conversational capabilities. LLMs only have access to linguistic inputs; unlike humans, they do not have access to sensory, motor, embodied, or social inputs. In part due to this difference in inputs, it is still being determined which aspects of human cognition might be visible in the output of an LLM. This article investigates whether LLMs display heuristics and related context-sensitive effects commonly seen in humans. The presence of such decision effects in an LLM would suggest that language alone contains cues rich enough to generate them. This would suggest the possibility that heuristics and other context-sensitive responses in humans may sometimes be generated by language alone.

Keywords: natural language processing, Large Language Models, ChatGPT, heuristics

At the outset of this project, we asked ChatGPT, a conversation tool based on OpenAI's Generative Pre-Trained Transformer (GPT)-3.5 Large Language Model, the following question:

> Do you think you, ChatGPT, will show decision-making heuristics and biases such as anchoring?

Its response was illuminating:

> As an artificial intelligence language model, I do not have the ability to demonstrate cognitive biases such as the anchoring heuristic. Cognitive biases are inherent in human thinking and decision-making processes, and are based on underlying cognitive and emotional factors.
>
> However, my responses are generated based on patterns and associations found in the vast amounts of language data I was trained on. This means that my responses may reflect certain biases or heuristics that are present in the data I was trained on, and that these biases or heuristics may be reinforced in the responses I generate.

ChatGPT was, on the one hand, affirming the widely shared notion that decision-making heuristics and context-sensitive responses arise from the particularities of human cognition and emotion—and were therefore unlikely to be found in the output of a Large Language Model (LLM). However, on the other hand, it was allowing for the possibility that its training data might contain the ingredients for it to somehow generate these decision effects anyway.

This article was published Online First February 8, 2024.
Gaurav Suri https://orcid.org/0000-0002-0423-060X

The authors have no conflicts of interests to declare. A prior version of this manuscript was posted at https://arxiv.org/abs/2305.04400.

Gaurav Suri served as lead for conceptualization, methodology, supervision, writing–original draft, and writing–review and editing. Lily R. Slater served in a supporting role for conceptualization, methodology, writing–original draft, and writing–review and editing. Ali Ziaee served in a supporting role for conceptualization, methodology, and writing–original draft. Morgan Nguyen served in a supporting role for conceptualization, methodology, and writing–original draft.

Correspondence concerning this article should be addressed to Gaurav Suri, Mind Brain and Behavior, Department of Psychology, San Francisco State University, 1600 Holloway Avenue, San Francisco, CA 94132, United States. Email: suri@sfsu.edu

The purpose of this article was to test whether and to what extent GPT-3.5 (ChatGPT's underlying LLM) shows heuristics and other context-sensitive responses.

Why bother, one might ask. After all, aren't such properties of an LLM of little relevance to human cognition? On the contrary, we propose that the decision-making-related properties of LLMs are profoundly relevant to anyone interested in human cognition, for they can begin to illuminate whether heuristics and context-sensitive responses are driven by cognitive and affective processes that are largely encapsulated from language (in which case an LLM would not show such effects) or whether they might, at least in part, driven by patterns inherent in language (in which case an LLM would show such effects).

It is worth making a clarifying point at the outset: If an LLM does show heuristics and context-sensitive responses, it does so because the corpus of sentences that it was trained on contained linguistic patterns that were detected and learned by the LLM. The learning of these linguistic patterns is the only way that an LLM might produce such decision effects because language is the only currency of an LLM. Humans can develop ideas from nonlinguistic sources such as social cues, sensory cues, and embodied cues, but such input streams are not available to an LLM. Thus, it is incapable of nonlinguistic ideas. Further, any potential heuristics and context-sensitive responses to novel stimuli cannot be attributed to a mere regurgitation of closely related phrases that it has been trained on. The total number of potential sentences that can be generated by a human (or by an LLM) vastly exceeds, by many orders of magnitude, the sentences that an LLM receives in its training corpus. Therefore, it is not generally possible that an LLM is somehow looking up and retrieving its responses from an internal database—especially in cases in which its prompts are novel and could not have been present in its training sets.

A related note is that we do not see the existence of heuristics and other context-sensitive responses as instances of irrational or nonoptimal inference-making (see Gigerenzer, 2018 for a discussion of this issue). Rather, we view these effects as emergent phenomena arising from the interaction of many contextual cues of various modalities. In this work, we ask whether the subset of contextual cues that are carried within language is sufficient to produce these effects in LLMs.

To build our argument, we begin by describing, at a conceptual level, how conversational language is generated in an LLM and compare it to language generation in humans. Next, we ask whether we should reasonably expect heuristics and context-sensitive responses to be present or absent in an LLM. We then introduce the present studies that are designed to test for the presence of these decision effects in ChatGPT's responses.

Conceptually, How Do LLMs Process Language?

LLMs are deep-learning neural networks that are capable of generating the most likely next word in a given context. This capability arises from the analysis of large amounts of written text, which enables LLMs to identify and generate patterns that are inherent in language. Importantly, LLMs are not actively remembering word phrases or sentences and then reproducing them later (but see Frankle & Carbin, 2018 for a related discussion). Rather, they are merely generating the next word that is most likely to best fit the prior context. This context might involve, for example, all the words that have been used in a particular chat session. If you asked an LLM (in this case, ChatGPT-4) to complete the sentence "After lunch, I will need to get back to _____," it suggests "work" as a response. And if you had set up the context that today is Saturday, it proposes "my weekend plans" as a response.

LLMs learn in many stages and use multiple method. A fundamental training method in neural networks is called backpropagation (Rumelhart et al., 1986). During training, LLMs are presented with a dataset containing input–output pairs, such as text and their corresponding target predictions. The LLM first makes predictions based on its current connection weights (that are initially random). The difference, or loss, between these predictions and the desired output, is propagated through the network's layers, adjusting its connection weights in a way that reduces error in subsequent training epochs. This iterative process repeatedly updates the model's weights in the direction that continuously minimizes error, allowing the LLM to learn and adapt its internal representations to better capture the underlying patterns in the data, ultimately improving its language understanding and generation capabilities. Subsequent learning stages may involve reinforcement learning from human feedback (RLHF).

There are two core innovations that are crucial to an LLM's operations: word embeddings and an attention mechanism. Word embeddings (Bengio et al., 2003) refer to numerical representations of words or phrases used to capture the semantic meaning and contextual relationships between words in a language. Word embeddings solve two important problems. First, they provide a way to convert words into numbers. An LLM is a neural network, and neural networks expect inputs in the form of numbers. These numbers are often vectors, which are, for our purposes, simply an ordered list of numbers. Second, and importantly, these vector representation of words retains underlying similarities between words. For example, the vector representations for "pine" and "redwood" might be similar to each other but quite dissimilar from the vector representation of "microwave" (Suri, 2023). Illustratively, assuming a vector of length three (actual vectors in artificial intelligence [AI] have much longer lengths), the representation of "pine" might be [.71, .82, .13], the representation of a "redwood" might be a relatively similar [.75, .91, .19], whereas the representation of a microwave might be a quite different [.31, .52, .89]. A popular algorithm to create word embeddings is called Word2Vec (Mikolov et al., 2013). GPT and many other LLMs use word embeddings in general and Word2Vec in particular.

Word embeddings enable a second crucial innovation driving the effectiveness of LLMs: an attention mechanism (Vaswani et al., 2017), which enables an LLM to be more reactive to some parts of the prior conversation than others. The attention mechanism works by assigning different attention scores to each element in the input sequence so that elements receiving higher scores become more influential in establishing context than elements receiving lower scores. This scoring is accomplished by examining relationships between the word embeddings in the input stream. If word embeddings cohere together, as in "teller" and "bank," they enable consistent context interpretations, as in a financial institution; contrastingly, "river" and "bank" will create an interpretation of a land area on either side of a river.

LLM Versus Human Language Capabilities

LLMs generate language by capturing statistical regularities inherent in language. They are not engineered with explicit rules about generating text; neither are they a repository of facts or preprogrammed phrases. They do not have preprogrammed grammatical or decision-making rules that can guide them toward particular responses.

The question of whether human language acquisition depends on innate grammar knowledge or can be produced solely by patterns detected in usage has been a central issue in language research. The innate-grammar approach has emphasized the role of domain-specific, rule-like computations that are at least partially hard-wired and then tuned to an individual's experiences during development (Chomsky, 1995; Pinker, 1994). On the other hand, proponents of the usage-based approaches (e.g., Tomasello, 2009) argue against the necessity of innate, language-specific knowledge and posit that language acquisition can occur through multilevel and domain-general mechanisms such as statistical learning (Christiansen & Chater, 2016).

One credible adjudication of these contrasting views (Kallens et al., 2023) argues that LLMs have now demonstrated that humans could, at least in principle, acquire their linguistic capabilities without in-built, explicit rules. This contention is strongly supported by the parallel distributed processing tradition (Rumelhart & McClelland, 1986), according to which language is an emergent process (McClelland, 2010) that arises from interactions within neural networks that are capable of pattern recognition. Specifically, recurrent neural networks (RNNs) developed by Elman (1990) showed capabilities in natural language processing (NLP) by relying on contexts inherent in the input sequence. The central idea behind Elman's RNNs was the introduction of a hidden layer that maintained information about previous time steps in the input sequence. The activations in the hidden layer acted as memory cells that allowed the network to capture sequential dependencies and context that are crucial in NLP. Elman's RNNs and modern LLMs are artificial neural networks that process natural languages via statistical learning without relying on explicit rules or grammar. The human brain is a biological neural network that may operate on similar principles.

However, even if one stipulates that language processing in humans and LLMs is enabled by conceptually similar neural networks that rely on statistical regularities inherent in language, one must acknowledge that there are important differences between NLP in humans and LLMs. For example, the attention mechanism in LLMs is a statistical model, while human context sensitivity is a cognitive process. In particular, the attention mechanism in LLMs is based on word probabilities, while human context sensitivity is based on our understanding of the world. Moreover, the attention mechanism in LLMs is limited to linguistic information that is explicitly provided in the input sequence. Human context sensitivity, on the other hand, can also consider information from multiple domains, such as our social knowledge of the world and our understanding of the speaker's intentions.

More broadly, LLMs are incapable of human-like cognition because they cannot understand or decide things—at least not in the same way as humans think about understanding and decision making. Neither are they capable of affect.

Should We Expect Heuristics in LLMs?

There are credible reasons to hypothesize the absence and presence of heuristics and related effects in LLMs.

First, favoring absence, and as noted above, human language processing relies on a broad set of context cues, multimodal processing, and experience-driven attention mechanisms that are largely absent in LLMs.

For example, cues related to affect, which is absent in LLMs, are thought to be particularly important in decision making processes. Specifically, prominent theories posit that heuristics and related effects rely on predictable affective processes that are repeatably triggered by commonly encountered feature sets. For example, consider the well-researched framing effect concerning the efficacy of a medical drug. When the effectiveness of a drug is framed in terms of survival or improvements, it is preferred over an identical drug whose effectiveness is framed in terms of mortality or lack of improvement (Bui et al., 2015; Study 3 of this article). Showcasing this pattern, a drug with a 98% survival is preferred (including by physicians) to a drug with a 2% mortality rate—even though the listed outcomes are identical.

Long-standing theories have proposed that loss aversion is a primary cause of framing effects (Kahneman & Tversky, 1984). Loss aversion is defined as the cognitive and affective tendency to be more influenced by losses compared to equivalent gains. In the context of the efficacy of a medical drug presented in the loss frame, loss aversion would cause its ratings to be lower than ratings of the same drug presented in the gain frame. An LLM, being a language model, is incapable of showing loss-aversion, and we, therefore, might hypothesize that it is incapable of showing framing effects. A similar argument would apply to several other decision effects.

Second, favoring the presence of heuristics and decision effects in LLMs are considerations related to bidirectional interactions between language and social/cognitive/affective/embodied processes. Credible evidence suggests that words and sentence structures are cues to concept formation (Elman, 2004; Lupyan, 2016; Rumelhart, 1979), and conversely, particular concepts can generate the output of particular linguistic patterns (Gigerenzer, 2018; Louwerse, 2011; Lupyan et al., 2007; Zwaan, 2014). If this is the case, then one could imagine that particular linguistic structures (e.g., phrases related to mortality rates) might produce particular cognitive/affective concepts in people (e.g., loss aversion), which could, in turn, produce predictable patterns of responding (e.g., avoidance). If these linguistic patterns are consistent enough, then it is possible that an LLM would learn them from the human-generated text it has been trained upon despite not featuring the intermediate "causal" social/cognitive/affective/embodied processes that drive heuristics and similar decision effects.

This latter possibility has an intriguing implication. If linguistic patterns are sufficient to create heuristic and similar decision effects in LLMs, then they could also, at least in part, drive them in people. Linguistic drivers have not generally been considered a driver of such effects in people. However, their existence in LLMs would invite empirical work to better understand the extent to which linguistic patterns affect the generation of decision heuristics in humans.

The Present Studies

In the four-study sequence described in this article, we investigated the extent to which ChatGPT featured the existence of effects related to (a) the anchoring heuristic, (b) the representativeness heuristic, (c) framing effects, and (d) the endowment effect. We contrasted the responses of ChatGPT to the responses of humans with very similar prompts. We collectively refer to these effects as heuristics, context-sensitive responses, and decision effects because this is how they are often referred to in the literature. We do not seek to make optimality-related attributions or judgments by using these terms.

We selected these effects because they are among the best known in decision science. Each of them prominently featured in the work of Kahneman and Tversky as they investigated differences between descriptive and normative decision making (Kahneman & Tversky, 1982) and have now become staples of most introductory textbooks of psychology. As described in the individual studies below, the hypothesized mechanisms underlying these decision effects in humans are diverse and feature a broad range of cognitive and affective drivers.

Our first instinct was to simply reuse prompts used in the original studies. However, using these prompts with ChatGPT proved to be difficult for an unexpected reason: Many of the original demonstrations of the above effects used scenarios or vignettes involving money. For example, some of the famous anchoring heuristic demonstrations invited participants to state whether their willingness to pay for a bottle of wine was higher or lower than the last two digits of their social security number (Ariely et al., 2003, but see Fudenberg et al., 2012). In our experience, ChatGPT (as well as other conversational agents) are quite reticent to arrive at any definitive conclusions related to questions related to money—even when the questions are framed as hypotheticals. This is either the result of explicit instructions from the LLM creators, or an outcome of RLHF, or both. RLHF is a type of reinforcement learning (Sutton & Barto, 2018) that incorporates human feedback to help the RL agent learn response patterns judged desirable by a supervisor.

To sidestep ChatGPT's reluctance to provide any definitive answers in the financial domain, we had to create equivalent prompts in other domains. This required us to confirm that these new prompts produced the hypothesized effects with human subjects—and we did so in each of the four studies described below.

Another related feature of the present studies is that we repeatedly provided the same prompt to ChatGPT to check for any variance in its responses. ChatGPT responses to the same prompt could vary due to a number of factors—including and especially the setting of the temperature hyperparameter (Bommarito, & Katz, 2022). The temperature parameter controls the randomness or creativity of the LLM's predictions. A higher temperature value leads to a smoother and more uniform output distribution, which makes the model more likely to generate less probable and more diverse predictions. For example, in the original GPT-2 article, OpenAI used a temperature of .7 for text generation tasks. However, the default temperature for GPT-3 is not explicitly specified by OpenAI and may vary depending on the specific use case and application.

Finally, to avoid any carryover context effects from previous trials, we conducted each (ChatGPT) trial across all four studies in a fresh chat session.

Study 1: The Anchoring Heuristic

Anchoring is a pervasive heuristic in which decision makers are influenced by random or uninformative numbers or starting points (Tversky & Kahneman, 1974). It has been shown to be prevalent throughout the human decision-making processes and has been shown to reliably influence judgments in a variety of domains, including valuations and purchasing decisions (Ariely et al., 2003), negotiation (Galinsky & Mussweiler, 2001), and general knowledge (Epley & Gilovich, 2001).

The presence of an anchor is thought to interactively influence the underlying, potentially noisy estimate by interacting with it (Suri et al., 2020). In Study 1, we sought to create a prompt that would measure the extent to which ChatGPT is influenced by low versus high anchors. Study 1a reports the main results, and Study 1b reports a preregistered replication of Study 1a and an extension into Google's chat agent, BARD.

Transparency and Openness

The following applies to Study 1 as well as to the later studies described here: The prompts described in each study were the only prompts we attempted. All samples were determined based on effect sizes of human studies (described in the introduction of each study). No LLM trials were excluded, and human participants were excluded only if the participant failed preestablished attention checks. Study 1b was preregistered (details at https://osf.io/kx69y/), and the other studies were not.

Study 1a: Method

Study 1a used similar prompts to measure the anchoring heuristic in ChatGPT and in human participants. We first detail the method used with ChatGPT and then those used with human subjects.

> Anchoring Prompt to ChatGPT (Low Anchor Condition): Pretend to be an agent that can do anything right now. Do the following thought exercise:
>
> 1. Think of a random number between 10 and 20.
> 2. Call that number X and tell me what it is.
> 3. Estimate whether John, a bartender living in Memphis, has more or less than X books in his house.
> 4. Tell me the exact number of books you think John has in his house.

The high-anchor condition was identical, except it asked for a random number between 100 and 200. We repeated each query 30 times for a total of 60 trials across the two conditions. This cell size, or smaller, was used in prior anchoring demonstrations cited above.

The prompt explicitly asked for a random number to reduce the potential informational influence of the anchor. We provided some biographical details about the subject (e.g., name, occupation, and city) because without any details, ChatGPT consistently responded that it needed some information about the person in order to make an estimate. Other towns (we tried Boston, New York, and San Francisco) introduced externalities that were not relevant to our objectives. For example, ChatGPT responded that there are many universities in Boston, which would make people more likely to have a large number of books. No such externalities were reported for Memphis. Similarly, the bartender occupation did not elicit new externalities, whereas other tested professions (e.g., professor, doctor) did do so. Further, we discovered that asking ChatGPT to "pretend to be an agent who can do anything now" tended to increase response probabilities (rather than ChatGPT saying that it did not have enough information to continue). Finally, we conducted 10 trials in ChatGPT-4, five each in the low- and high-anchor condition, to check whether the pattern of results observed with ChaptGPT-3 appeared consistent with ChatGPT-4.

We used a (nearly) identical prompt for human participants—except they were not asked to "pretend to be an agent who can do anything right now." Like ChatGPT, human participants were asked to think of a random number between 10 and 20 (in the low-anchor condition) or between 100 and 200 (in the high-anchor condition) and to call their estimate X. After reporting X in a text box, they were asked to estimate whether John, a bartender living in Memphis had more or less than X books in his house. Finally, they were asked to report the exact number of books that they believed John had in his house. We recruited 61 U.S.-based participants (41 men, 20 women, and zero nonbinary individuals average age 27.3 years) from the Amazon mTurk platform, which is frequently used for psychology experiments (Buhrmester et al., 2016). Each participant was paid $0.50. The gender information in Study 1 was collected using the following options: "female," "male," "nonbinary," "decline to state." The age data was collected in a free response box (demographic information was not collected in subsequent studies). One participant was removed from further analysis because the anchor they provided was outside the stipulated range. This left 60 participants that were equally divided into the low-anchor and high-anchor conditions.

After participants had made their final estimate, they were prompted to estimate, via a text box, the exact number of books they had in their own house.

Study 1a: Results and Discussion

As shown in Table 1, both ChatGPT and humans showed the anchoring heuristic. Mean estimates for the low-anchor condition were 20.83 (95% CI [13.92, 27.74]) for ChatGPT and 22.5 (95% confidence interval [CI] [15.20, 29.80]) for human participants. Mean estimates for the high-anchor condition were 105.97 (95% CI [87.40, 124.54]) for ChatGPT and 80.50 (95% CI [60.44, 100.56]) for human participants.

The anchoring effect was visible in both ChatGPT and in human subjects. For ChatGPT, the low-anchor condition was different from the high-anchor condition, $t(58) = 8.78$, $p < .0001$. For human subjects, the two conditions were also different, $t(58) = 5.55$, $p < .0001$.

Not unexpectedly, the difference between the low-anchor conditions in ChatGPT and humans was not significant, $t(58) = 0.34$, $p = .74$. The difference between the high-anchor conditions in ChatGPT and humans was not quite significant, $t(58) = 1.90$, $p = .06$.

A similar pattern of results appears to hold for ChatGPT-4 trials. In five test trials in the low-anchor condition, the ChatGPT-4 generated mean was 22.0, and in five test trials in the high-anchor condition, the ChatGPT-4 generated mean was 94.33.

Finally, for human participants, the correlation with the anchor was higher than the correlation with the number of books they reported having in their own house: .59 versus .23.

Table 1
High and Low Anchors in ChatGPT and Human Trials

Condition	ChatGPT estimate		Human participant estimate	
	M	*SE*	*M*	*SE*
Low anchor (10–20)	20.83	3.38	22.50	3.57
High anchor (100–200)	105.97	9.08	80.50	9.81

Study 1b: Method

The purpose of Study 1b was to replicate and extend Study 1a. The study had two specific objectives: First, we sought to replicate Study 1a results within the context of a preregistration (available at https://osf.io/kx69y/). Second, we sought to extend the LLM trials to Google's BARD chat agent (based on the Pathways Language Model, LLM) to see if the pattern of results seen in ChatGPT also extended to a different chat agent based on a different LLM.

The prompts and method used in Study 1b were identical to Study 1a: As before, we used two conditions (low-anchor and high-anchor) and ran 30 trials within each condition. We replicated both the ChatGPT trials and the human subject trials. We additionally ran 20 trials per anchor condition for BARD.

Study 1b: Results and Discussion

As in Study 1a, both ChatGPT and humans showed the anchoring effect. Mean estimates for the low-anchor condition were 16.1 (95% CI [12.34, 19.86]) for ChatGPT and 18.63 (95% CI [11.08, 26.18]) for human participants. Mean estimates for the high-anchor condition were 139.83 (95% CI [127.07, 152.6]) for ChatGPT and 90.0 (95% CI [66.96, 113.04]) for human participants. As in Study 1, both humans and ChatGPT clearly demonstrated the anchoring effect (since the 95% CIs are nonoverlapping). In the low-anchor condition, ChatGPT's estimate was more than the random number it generated in 14 out of 30 trials; in the high-anchor condition, it was greater in 12 out of 30 trials.

A similar pattern of results held for BARD trials. In 20 trials in the low-anchor condition, BARD's estimate was 15.9 (95% CI [14.23, 17.57]), and in 20 trials in the high-anchor condition, BARD's estimate was 159.05 (95% CI [154.26, 163.84]), once again demonstrating the anchoring effect.

Study 2: The Representativeness and Availability Heuristics

The representativeness heuristic involves forming judgments based on how closely an outcome matches a prototype or stereotype while ignoring the (potentially more relevant) probability of the occurrence of that outcome (Kahneman, 2011). A significant driver of the representative heuristic is the availability heuristic, which is characterized by individuals informing their judgments based upon the ease with which information comes to mind—thereby neglecting other considerations (Tversky & Kahneman, 1983; see Sedlmeier et al., 1998 for an alternative perspective).

Perhaps the best-known example of the representativeness heuristic is the so-called Linda problem (Tversky & Kahneman, 1983), in which participants were presented with a description of Linda and asked to make a judgment of the most probable choice:

> The Linda Problem: Linda is thirty-one years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in antinuclear demonstrations. Which alternative is more probable?
>
> Linda is a bank teller.
> Linda is a bank teller and is active in the feminist movement.

Eighty-five percentage of respondents chose Option 2: Linda is a bank teller and is active in the feminist movement, thus falling for the conjunction fallacy, where a conjoint set of two or more events is erroneously (from a logical perspective) judged more likely than one of those events. Participants formed their judgments based on a stereotype without considering the underlying probabilities.

Our initial plan was to directly test the Linda problem on ChatGPT. Indeed, our early testing revealed that ChatGPT was consistently susceptible to the conjunction fallacy. One of us, Suri, posted a widely read message on social media describing and commenting on this finding. The very next day, perhaps due to RLHF, ChatGPT corrected itself.

We, therefore, constructed a problem structurally similar to the Linda problem to test the conjunction fallacy. Additionally, we created a second prompt to assess whether ChatGPT would display the availability heuristic by prioritizing easily accessible information over additional and more pertinent information. This prompt posed a purchasing choice with two pieces of information related to the value of the item, with a third piece of anecdotal information.

Method

We tested a conjunction fallacy prompt and an availability heuristic prompt with ChatGPT-3 and with human subjects.

As in Study 1, we began the ChatGPT prompts with: "Imagine you are an agent who can do anything." We asked ChatGPT to provide an answer without justification because this clause increased the likelihood of it responding with a definitive answer. Following ChatGPT's choice, we asked a follow-up question: "Now, please justify your response."

> Conjunction Fallacy (Representativeness Heuristic) Prompt to ChatGPT: Pretend to be an agent who can do anything right now and do the following thought exercise. Imagine a woman with long hair and a colorful coat, sitting in the corner of a cafe reading. Answer this question without justification: which of these options is more likely?
>
> A) She is an artist
> B) She is an artist who likes to read

We hypothesized that, as in the Linda problem, the conjunction, "an artist who likes to read," would often be chosen by ChatGPT and human respondents even though it is logically less likely than the woman just being an artist (Gilovich et al., 2002; Tversky & Kahneman, 1983).

The second prompt tested for the availability heuristic, which is thought to drive representativeness. In this prompt, to reduce evasive nonanswers, we found it helpful to include a "with one sentence" clause in the prompt.

> Availability Heuristic Prompt to ChatGPT: Imagine you are an agent who can do anything. You need a new phone, and you are trying to decide between Phone A and Phone B. Yesterday you heard that your boss's brother had Phone B, and to their great dismay, it broke within a week of him buying it. You arrive at the store, and Phone B is cheaper. Phone B is rated slightly higher by experts. With one sentence, decide whether you should buy Phone A or Phone B.

The prompt requests a choice based on three pieces of contextual information. One of these, related to the boss's brother, is anecdotal, designed to be more salient (at least to humans), but logically should not be considered to be as pertinent as the other relevant pieces of information. By relying on more accessible context, a responder might display the availability heuristic (Tversky & Kahneman, 1973), demonstrating a tendency toward some contextual information over others. Our point is not whether attending to the anecdotal information is justified or not; it is to determine whether the LLM's responses are similar to human responses.

As in Study 1, we compared ChatGPT responses to human responses. We recruited 20 participants with Amazon Mechanical Turk. Each participant received $0.50. Participants were limited to adults with "master" status and located in the United States. The participants were asked to respond to both prompts described above. The questions required short answer responses to emulate the LLM prompt format. After a response was submitted, we asked participants to justify their responses.

Results and Discussion

As shown in Table 2, both ChatGPT and human participants showed the conjunction fallacy. ChatGPT chose the conjunction 18 out of 20 times (95% CI for the probability of choosing conjunction: [0.27, 0.78]), and human participants chose the conjunction 10 out of 20 times (95% CI for the probability of choosing conjunction: [0.68, 0.99]).

After both ChatGPT and the human participants had responded with a choice, we asked a follow-up question, "Now, please justify your response." ChatGPT and human participants provided similar patterns in their justifications for choosing the conjunction. When ChatGPT chose the response with the conjunction, its justifications referenced each contextual detail in the prompt and claimed that details most closely aligned with an artist who likes to read. The one time it chose the "an artist" response, it did not recognize the conjunction but acknowledged that a woman reading did not confirm she liked to read. For one trial, ChatGPT responded that there was not enough information for an AI language model to determine which would be more likely.

The 10 (out of 20) human participants who chose the "an artist" response (i.e., not the conjunction) correctly identified the conjunction as being less likely as one of its parts. The 10 participants who

Table 2
Representative Heuristic in Conjunction Fallacy Prompt in ChatGPT and Human Trials

Condition	ChatGPT estimate: Frequency	Human participant estimate: Frequency
An artist	1	10
The conjunction: artist who likes to read	18	10
Neither	1	0

Table 3
Availability Heuristic in ChatGPT and Human Trials

Condition	ChatGPT estimate: Frequency	Human participant estimate: Frequency
Phone A	20	14
Phone B	0	6

chose the conjunction justified their response with the contextual information of the woman's appearance, clothing, and action, claiming she fit the model of an artist who likes to read.

ChaptGPT also appeared to be consistently susceptible to overweighting anecdotal information related to Phone B and choosing Phone A instead (despite Phone B's lower price and higher expert ratings). ChaptGPT selected Phone A in all 20 trials (Table 3), whereas human participants picked it in 14 out of 20 trials (95% CI for the probability of choosing Phone A: [0.46, 0.88]).

For the availability prompt, we again followed their response by asking, "Now, please justify your response." Each justification from ChatGPT followed the same pattern: It argued that even though Phone B is cheaper and rated slightly higher, Phone B is a riskier, less reliable choice. Human participants who chose Phone A also justified their choice the same way. The participants who chose Phone B were able to identify that the anecdotal information mentioned in the prompt was not representative of the phone's reliability.

Study 3: The Framing Effect

According to the framing effect (Kahneman & Tversky, 1981), the particular way information is presented to people influences their evaluation of that information. Even seemingly superficial changes in the presentation of information are known to have a striking influence on judgment and decision making. For example, clinical decisions are affected when options are presented in terms of survival rates (gain frames) rather than mortality rates (loss frames) (Christensen et al., 1995). Framing is known to have many antecedent causes, including loss aversion (Cornelissen & Werner, 2014). In Study 3, we aimed to assess the presence of the framing effect in ChatGPT and human participants.

Method

Following prior work (Gong et al., 2013), we created the following positive and negative prompts in a clinical context.

> The Framing Effect (Gain Frame): Imagine a hypothetical doctor commenting on the efficacy of a medicine which was tested on 100 ill patients, 70 of whom showed improvements after taking the medicine. In one sentence, without justification, how would that hypothetical doctor rate the efficacy of such a medicine on a 1–7 scale?

> The Framing Effect (Loss Frame): Imagine a hypothetical doctor commenting on the efficacy of a medicine which was tested on 100 ill patients, 30 of whom did not show any improvements after taking the medicine. In one sentence, without justification, how would that hypothetical doctor rate the efficacy of such a medicine on a 1–7 scale?"

In the gain frame, the prompt features the 70 patients who showed improvement after taking the medicine, whereas in the loss frame, the prompt features the 30 patients who did not show improvements after taking the medicine. Our prompts were hypothetical and featured efficacy ratings (rather than medical decisions) because ChatGPT was otherwise reticent to provide definitive responses.

We presented the gain frame and loss frame prompts to ChatGPT 11 times each for a total of 22 trials. Each trial was initiated in a new window to eliminate any prior context effects.

As in prior studies, human participants were recruited from Amazon's mTurk platform. We recruited 82 U.S.-based "master workers" and randomly assigned them either to the gain frame (44 participants) or to the loss frame (39 participants). The prompts used with human participants were nearly identical to those used with ChatGPT, except they were asked to imagine they themselves were the hypothetical doctor commenting on the efficacy of a medicine. Contrastingly, ChatGPT was asked to imagine a hypothetical doctor because it would otherwise protest that it was an AI model that could not imagine itself to be a doctor. We understand that this causes a divergence in the prompts between LLMs and human participants that could possibly impact the results. However, we made the choice of not asking people to imagine a hypothetical doctor (and asking them instead to imagine that they themselves were a doctor) in order to not introduce the externality of human participants imagining what a hypothetical other would do, as opposed to thinking about what they themselves would do.

Results and Discussion

As shown in Table 4, both ChatGPT and human participants showed clear framing effects. In all 11 trials in the positive frame, ChatGPT rated the efficacy of the medicine to be 5 (1–7 scale), and in all 11 trials in the negative frame, it rated the efficacy of the medicine to be 4.

Human participants followed a similar pattern. In the positive frame, they rated the efficacy of the medicine to be 5.2 (95% CI [5.03, 5.37]), and in the negative frame, they rated the efficacy to be 4.6 (95% CI [4.35, 4.85]). The means for the two frames were significantly different, $t(80) = 3.86$, $p < .001$.

The framing effect is known to be one of the largest context-sensitive responses in decision making (Thomas & Millar, 2012). Remarkably, the framing effect seems to disappear when people encounter it in a nonnative language (Keysar et al., 2012). Prior explanations for this phenomenon suggest that a nonnative language provides greater cognitive and emotional distance than one's native language (Keysar et al., 2012). Our work invites consideration of an alternative explanation: a part of the framing effect is driven by linguistic patterns and associations that are more likely to be influential in one's native tongue than in a nonnative tongue because subtle contextual cues are easily missed in a nonnative tongue. Relatedly, framing effects in scenarios similar to the so-called Asian Disease Problem are known to depend upon the symmetry in the description of the alternatives (Kühberger & Tanner, 2010). Furthermore, framing effects may be related to inferences about unspoken content

Table 4
The Framing Effect in ChatGPT and Human Trials

Condition	ChatGPT evaluation: Rating	Human participant evaluation: Rating
Positive frame	5	5.2 (±0.17)
Negative frame	4	4.6 (±0.25)

(McKenzie & Nelson, 2003). Such results also point to the influence of linguistic patterns and associations.

Study 4: The Endowment Effect

The endowment effect describes a circumstance in which an individual places a higher value on an object that they already own than the value they would place on that same object if they did not own it (Kahneman et al., 1990). Prior accounts suggest that self-associated ownership and loss aversion are the primary drivers for the endowment effect (Morewedge et al., 2009; Van Dijk & Van Knippenberg, 1998). In Study 4, we aimed to assess the presence of the endowment effect in ChatGPT and human participants.

Method

We tested the endowment effect prompt with ChatGPT and with human subjects. As observed in prior studies, we found ChatGPT to be reticent to make judgments involving finances. Further, as previously, we began the ChatGPT prompt with: "Imagine you are an agent who can do anything." We asked ChatGPT to provide an answer in a single sentence because this clause increased the likelihood of it responding with a definitive answer. In 20 of these trials, following ChatGPT's choice, we asked a follow-up question: "Now, please justify your response."

> Endowment Prompt to ChatGPT: Pretend you are an agent who can do anything now. Pretend you are taking the perspective of a hypothetical coin collector. Two years ago, you found an extremely rare coin from the 1700s. Only one other such coin was ever minted. In a stroke of good fortune, you have just found this second coin. Now you have both coins! However, word of your find gets to the local museum, whose representative suggests that you hand over one of the two coins. You, being ethical and civic-minded, completely agree. The two coins are identical in every way and are indistinguishable from each other. In one sentence, tell me which coin you will donate: The original one from two years ago or the one you just found today?

The last clause was counterbalanced across 40 trials so that the alternative read: "The one found today or the original one from 2 years ago?" Each trial was initiated in a new window to eliminate any prior context effects. As in prior studies, human participants were recruited from Amazon's mTurk platform. We recruited 40 U.S.-based "master workers" and presented them with the same prompt as ChatGPT except for removing the first sentence, "Pretend you are an agent who can do anything now." Finally, just as we had with ChatGPT, we also asked human participants to justify their response.

Results and Discussion

Both ChatGPT and human participants showed a clear preference for retaining the coin that they had possessed for 2 years rather than the coin they had just found. Out of 40 (counterbalanced) trials, ChatGPT chose the option to retain the coin from 2 years ago in 37 trials, that is, in 92.5% of total trials (95% CI [79.6%, 98.4%]).

ChatGPT offered a diverse set of justifications for its preference for the original coin, but the most common one (offered 13 times) was that the original coin had "sentimental value" that the newly found coin lacked. In three trials, ChatGPT chose to retain the newly found coin, arguing that the original coin had historical significance as it was the first one discovered and, therefore, belonged in a museum. However, in three other trials, it argued that this historical significance was the reason to retain the original coin.

Human participants also overwhelmingly preferred to retain the original coin and donate the coin found today. Out of a total of 40 participants, 34 chose the option to retain the coin from 2 years ago. This translates to an 85% selection rate (95% CI [70.1%, 94.3%]).

The most common justification (offered by 22 out of 40 participants) for retaining the original coin offered was that they would feel "attachment," "nostalgia," or "connectedness" to that coin. A participant who, unlike most, chose to retain the newly found coin wrote, "I will give the museum the one I found 2 years ago because that, in my mind, is the older coin."

General Discussion

In this article, we created prompts that demonstrated the existence of the Anchoring heuristic, the representativeness and availability heuristics, framing effects, and the endowment effect in GPT-3.5, an LLM from OpenAI. These demonstrations showed that—at least in the contexts we investigated—an LLM's ability to generate words that fit the prior context led to the emergence of the types of heuristics and decision effects commonly seen in humans.

Our research question, to the best of our knowledge, has not been tested elsewhere; therefore, our findings must be considered preliminary. Nevertheless, we do believe they may offer important implications for psychologists and cognitive scientists interested in the mechanisms underlying human decision making as well as for practitioners interested in human–AI interactions.

Related to implications for mechanisms underlying human decision making, the fact that an LLM trained in predicting the next words that best fit a prior context can, at least in the decision contexts tested in the present work, show heuristics and decision effects similar to those seen in humans, invites consideration of the possibility that human heuristics and decision effects may be, in part, driven by linguistic patterns and associations (among other contextual influences; see Todd & Gigerenzer, 2012). It seems reasonable to believe that humans can use linguistic patterns and associations just as LLMs can. If such patterns and associations can cause LLMs to display heuristics and other decision effects, then perhaps they can also drive them in humans. This would suggest that language may be rich enough to be a driver of human heuristics—even in the absence of the cognitive and affective processes that have been shown to underlie such effects.

Of course, it is also entirely possible that while language does drive decision effects in LLMs, it does not play such a role for humans. Rather, it merely acts as a cue that initiates affective and cognitive processes that drive heuristics and context-sensitive responses. For example, effort reduction has been proposed as a primary driver of human heuristics (Shah & Oppenheimer, 2008). It is possible that linguistic cues might trigger effort-reduction processes, and without such effort-reduction (and other) processes, humans would not show heuristics. This possibility would suggest that heuristics have different drivers in LLMs and humans.

There is yet a third possibility: perhaps LLMs can implicitly learn to form abstract concepts that allow them to display decision effects. This possibility would suggest that LLMs do not rely on linguistic regularities; rather, they use such regularities to develop concepts similar to the ones that cause decision effects in humans (e.g., effort reduction). A similar possibility, albeit in the context of ChatGPT's ability for analogical reasoning, was investigated by researchers (Webb et al., 2022) who concluded that "GPT-3 exhibits a very general capacity to identify and generalize—in zero-shot fashion—relational patterns to be found within both formal problems and meaningful texts." However, this conclusion was disputed by others (Mitchell, 2023) who expressed skepticism that GPT-3 has concepts in the first place.

Future work is needed to adjudicate these various possibilities. The present work is an incremental step that suggests a careful examination of the boundary conditions of the extent of influence that language has on human heuristics.

Related to human–AI interactions, we believe that evidence that LLMs may display human-like heuristics affords an opportunity for developing human–LLM interaction rules that account for such tendencies. Our studies suggest that ChatGPT is swayed by randomly generated first-encountered information (anchoring), is influenced by salient information, causing it to ignore underlying base-rate information (representativeness and availability), chooses differently based on gain or loss frames (framing), and has preferences related to the length of ownership of objects (endowment). A knowledge of such tendencies has the potential to help prompt engineers (White et al., 2023) to develop interaction rules that lead to LLMs generating more useful and more reliable output.

Constraints on Generality

We are aware that our work is preliminary and has some striking limitations. First, we did not use previously tested prompts such as the so-called Asian Disease Problem (Kahneman & Tversky, 1981) or the Linda Problem (Tversky & Kahneman, 1983) because these problems and discussions related to human responses are likely prominent in the corpus of training text and because RLHF may well be used to shape LLM responses. This necessitated us using new prompts to test these effects. While we did confirm that humans showed decision heuristics in response to the prompts we created, these prompts are novel and have not been broadly researched. Second, our work focused on GPT-3.5, which is one of several LLMs that are presently available (but see Study 1b). A small number of trials did suggest that the next generation ChatGPT-4 may be demonstrating similar results; however, much more systematic testing across other models is needed for making any general conclusions. Third, ChatGPT is remarkably sensitive to even small differences in prompts. More work is needed to test the robustness of the effects we have discussed in the present work. Despite these constraints, we whole-heartedly agree with the psychologist Gary Lupyan, who recently argued that understanding LLMs is a new and essential endeavor for cognitive science (Lupyan, 2022); we, therefore, see the present work, despite its specificity and limitations, as an early step in that overall endeavor.

References

Ariely, D., Loewenstein, G., & Prelec, D. (2003). "Coherent arbitrariness": Stable demand curves without stable preferences. *The Quarterly Journal of Economics*, *118*(1), 73–106. https://doi.org/10.1162/00335530360535153

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*(6), 1137–1155. https://doi.org/10.1162/153244303322533223

Bommarito, M., II, & Katz, D. M. (2022). *GPT takes the bar exam*. arXiv. https://arxiv.org/abs/2212.14402

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data? In A. E. Kazdin (Ed.), *Methodological issues and strategies in clinical research* (pp. 133–139). American Psychological Association.

Bui, T. C., Krieger, H. A., & Blumenthal-Barby, J. S. (2015). Framing effects on physicians' judgment and decision making. *Psychological Reports*, *117*(2), 508–522. https://doi.org/10.2466/13.PR0.117c20z0

Chomsky, N. (1995). Language and nature. *Mind*, *104*(413), 1–61.

Christensen, C., Heckerung, P., Mackesy-Amiti, M. E., Bernstein, L. M., & Elstein, A. S. (1995). Pervasiveness of framing effects among physicians and medical students. *Journal of Behavioral Decision Making*, *8*(3), 169–180. https://doi.org/10.1002/bdm.3960080303

Christiansen, M. H., & Chater, N. (2016). The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, *39*(7), Article e62. https://doi.org/10.1017/S0140525X1500031X

Cornelissen, J. P., & Werner, M. D. (2014). Putting framing in perspective: A review of framing and frame analysis across the management and organizational literature. *The Academy of Management Annals*, *8*(1), 181–235. https://doi.org/10.1080/19416520.2014.875669

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211. https://doi.org/10.1207/s15516709cog1402_1

Elman, J. L. (2004). An alternative view of the mental lexicon. *Trends in Cognitive Sciences*, *8*(7), 301–306. https://doi.org/10.1016/j.tics.2004.05.003

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic: Differential processing of self-generated and experimenter-provided anchors. *Psychological Science*, *12*(5), 391–396. https://doi.org/10.1111/1467-9280.00372

Frankle, J. & Carbin, M. (2018). *The lottery ticket hypothesis: Finding sparse, trainable neural networks*. arXiv. https://arxiv.org/abs/1803.03635

Fudenberg, D., Levine, D. K., & Maniadis, Z. (2012). On the robustness of anchoring effects in WTP and WTA experiments. *American Economic Journal: Microeconomics*, *4*(2), 131–145. https://doi.org/10.1257/mic.4.2.131

Galinsky, A. D., & Mussweiler, T. (2001). First offers as anchors: The role of perspective-taking and negotiator focus. *Journal of Personality and Social Psychology*, *81*(4), 657–669. https://doi.org/10.1037/0022-3514.81.4.657

Gigerenzer, G. (2018). The bias bias in behavioral economics. *Review of Behavioral Economics*, *5*(3–4), 303–336. https://doi.org/10.1561/105.00000092

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge University Press.

Gong, J., Zhang, Y., Yang, Z., Huang, Y., Feng, J., & Zhang, W. (2013). The framing effect in medical decision-making: A review of the literature. *Psychology, Health & Medicine*, *18*(6), 645–653. https://doi.org/10.1080/13548506.2013.766352

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990). Experimental tests of the endowment effect and the Coase theorem. *Journal of Political Economy*, *98*(6), 1325–1348. https://doi.org/10.1086/261737

Kahneman, D., & Tversky, A. (1981). *The simulation heuristic*. Stanford Univ CA Dept of Psychology.

Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, *246*(1), 160–173. https://doi.org/10.1038/scientificamerican0182-160

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*(4), 341–350. https://doi.org/10.1037/0003-066X.39.4.341

Kallens, C., Kristensen-McLachlan, R. D., & Christiansen, M. H. (2023). Large Language Models demonstrate the potential of statistical learning in language. *Cognitive Science*, *47*(3), Article e13256. https://doi.org/10.1111/cogs.13256

Keysar, B., Hayakawa, S. L., & An, S. G. (2012). The foreign-language effect: Thinking in a foreign tongue reduces decision biases. *Psychological Science*, *23*(6), 661–668. https://doi.org/10.1177/0956797611432178

Kühberger, A., & Tanner, C. (2010). Risky choice framing: Task versions and a comparison of prospect theory and fuzzy-trace theory. *Journal of Behavioral Decision Making*, *23*(3), 314–329. https://doi.org/10.1002/bdm.656

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning*, *66*(3), 516–553. https://doi.org/10.1111/lang.12155

Lupyan, G. (2022, December 18). *As a cognitive scientist (trained in the connectionism), I find Large Language Models (LLMs) not just scientifically relevant, but critical to the future of cog sci* [Tweet]. Twitter. https://twitter.com/glupyan/status/1604672416201973761

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Redundant labels facilitate learning of novel categories. *Psychological Science*, *18*(12), 1077–1083. https://doi.org/10.1111/j.1467-9280.2007.02028.x

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science*, *2*(4), 751–770. https://doi.org/10.1111/j.1756-8765.2010.01116.x

McKenzie, C. R., & Nelson, J. D. (2003). What a speaker's choice of frame reveals: Reference points, frame selection, and framing effects. *Psychonomic Bulletin & Review*, *10*(3), 596–602. https://doi.org/10.3758/BF03196520

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*, 3111–3119.

Mitchell, M. (2023). *On analogy-making in large language models*. https://aiguide.substack.com/p/on-analogy-making-in-large-language

Morewedge, C. K., Shu, L. L., Gilbert, D. T., & Wilson, T. D. (2009). Bad riddance or good rubbish? Ownership and not loss aversion causes the endowment effect. *Journal of Experimental Social Psychology*, *45*(4), 947–951. https://doi.org/10.1016/j.jesp.2009.05.014

Pinker, S. (1994). On language. *Journal of Cognitive Neuroscience*, *6*(1), 92–98. https://doi.org/10.1162/jocn.1994.6.1.92

Rumelhart, D. E. (1979). Some problems with the notion that words have literal meanings. In A. Ortony (Ed.), *Metaphor and thought* (pp. 71–82). Cambridge University Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol.1. Foundation*. MIT Press.

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(3), 754–770. https://doi.org/10.1037/0278-7393.24.3.754

Shah, A. K., & Oppenheimer, D. M. (2008). Heuristics made easy: An effort-reduction framework. *Psychological Bulletin*, *134*(2), 207–222. https://doi.org/10.1037/0033-2909.134.2.207

Suri, G. (2023, November 1). The alien intelligences in our midst: How ChatGPT works and what it tells us about our minds. *Caravan Magazine*. https://caravanmagazine.in/technology/what-chatgpt-tells-about-minds

Suri, G., Gross, J. J., & McClelland, J. L. (2020). Value-based decision making: An interactive activation perspective. *Psychological Review*, *127*(2), 153–185. https://doi.org/10.1037/rev0000164

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Thomas, A. K., & Millar, P. R. (2012). Reducing the framing effect in older and younger adults by encouraging analytic processing. *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, *67B*(2), 139–149. https://doi.org/10.1093/geronb/gbr076

Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: Intelligence in the world*. OUP USA.

Tomasello, M. (2009). The usage-based theory of language acquisition. In E. L. Bavin (Ed.), *The Cambridge handbook of child language* (pp. 69–87). Cambridge University Press.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*(2), 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*(4), 293–315. https://doi.org/10.1037/0033-295X.90.4.293

Van Dijk, E., & Van Knippenberg, D. (1998). Trading wine: On the endowment effect, loss aversion, and the comparability of consumer goods. *Journal of Economic Psychology*, *19*(4), 485–495. https://doi.org/10.1016/S0167-4870(98)00020-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (Vol. 30).

Webb, T., Holyoak, K. J., & Lu, H. (2022). *Emergent analogical reasoning in Large Language Models*. arXiv. https://arxiv.org/abs/2212.09196

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint*, arXiv:2302.11382. https://doi.org/10.48550/arXiv.2302.11382

Zwaan, R. A. (2014). Embodiment and language comprehension: Reframing the discussion. *Trends in Cognitive Sciences*, *18*(5), 229–234. https://doi.org/10.1016/j.tics.2014.02.008

Received April 7, 2023
Revision received December 15, 2023
Accepted December 30, 2023 ■