

Artificial Intelligence and Metacognitive Knowledge: LLMs Can Report Decision Weights as Faithfully as Humans in Multi-Attribute Choice

Trent N. Cash (tcash@uwaterloo.ca)^{1,2,3}

Daniel M. Oppenheimer (doppenh1@andrew.cmu.edu)^{2,3}

Department of Psychology, University of Waterloo, Canada¹

Department of Social and Decision Sciences, Carnegie Mellon University, USA²

Department of Psychology, Carnegie Mellon University, USA³

Author Note

Correspondence concerning this article should be addressed to Trent N. Cash, Department of Psychology, University of Waterloo, 200 University Ave. W, PAS 4042, Waterloo, ON, N2L 3G1. E-Mail: tcash@uwaterloo.ca. This work was supported by the National Science Foundation (Award #2333553), Sawtooth Software, Inc. (Academic Research Grant), and the Lupina Foundation. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the grantors. The authors report no competing interests. The authors would like to recognize the support of four undergraduates – Sara Christie, Mira Devgan, Roslyn Randall, and Anika Mohta – who assisted with running the LLM trials. The authors would also like to thank their wives, Elizabeth and Nicole, for their support and feedback. Data, code, and other supplemental materials are available at: https://osf.io/r73ez/overview?view_only=b874556b5f1e475bbbc9926083a1724f

Abstract

Users have become increasingly reliant on Large Language Models (LLMs) to complete a wide range of reasoning tasks, from managing workplace projects to giving personal advice. However, LLMs function as black boxes, leaving users with minimal insight into how they generate the responses that they do. One way that users can attempt to peek into these black boxes is by asking LLMs to explain their reasoning processes. The nascent literature on LLM faithfulness suggests that LLMs – like humans – often fail to accurately identify the information they use to inform their reasoning, suggesting a lack of metacognitive knowledge. We extend this research to the context of value-based, multi-attribute choice, tasking both LLMs and human participants ($n = 436$) with completing the Knowledge of Weights paradigm. Participants first completed a series of choice tasks in which they picked between homes that varied on six attributes, then self-reported the decision weight they believed they had placed on each attribute in two different formats. We found that ChatGPT4o, the oldest model tested, self-reported weights that were significantly less reflective of its choice behavior than humans. However, three newer, more-advanced LLMs (ChatGPT5, Sonnet 4, and Gemini 2.5 Flash) self-reported weights that were as accurate or more accurate than those provided by humans. These results suggest that LLMs can generate and maintain accurate metacognitive knowledge of their own decision-making processes as well or better than humans, but that this is a relatively new ability. Practical and theoretical implications are discussed.

Keywords: Large Language Models; Metacognitive Knowledge; Faithfulness; Multi-Attribute Choice; Metareasoning.

Artificial Intelligence and Metacognitive Knowledge: LLMs Can Report Decision Weights as Faithfully as Humans in Multi-Attribute Choice

Trent N. Cash & Daniel M. Oppenheimer

Introduction

Large Language Models (LLMs), such as ChatGPT, Gemini, and Claude, have become increasingly present in everyday life. Students use LLMs to enhance their learning (Cash, Kelly, et al., 2026; Mutanga et al., 2025; Oppenheimer et al., 2025), employees use LLMs to augment their workplace productivity (Dell’Acqua et al., 2023; Weisz et al., 2025) and creativity (S. Sun et al., 2025), and individuals turn to LLMs as friends, advisors, experts, and confidants (Cash & Oppenheimer, 2024a; Kirk et al., 2025; Lissak et al., 2024; Wester et al., 2024). Given the many roles that these models play, it should come as no surprise that LLMs are often tasked with engaging in complex reasoning behaviors that result in actionable insights, such as making decisions on behalf of users or giving users advice (Benary et al., 2023; Hager et al., 2024; Hou et al., 2024; Lakkaraju et al., 2023; Pesch, 2025; Wester et al., 2024; Yang et al., 2024). When tasked with doing so, the goal of the LLM is (or at least should be) to align their reasoning with the user’s preferences to maximize user satisfaction (Goli & Singh, 2024; J. Kim et al., 2024). This, however, is no easy task; users have highly heterogeneous (Andersen et al., 2010; Bakker et al., 2022; Deng et al., 2025) and inconsistent preferences (Amir & Levav, 2008; Payne et al., 1999; Simon et al., 2008)

Given the difficulty of capturing human preferences, a healthily skeptical user may wish to study the LLM’s reasoning process. Unfortunately for this user, most LLMs function as black boxes, meaning that their underlying processes are not made available to the end user

(Ramlochan, 2023). This is in large part because of technical realities. LLMs rely on billions – and in some cases trillions – of parameters within a complex neural network that they use to generate responses to user queries (OpenAI, 2025a; Roose, 2023; Wu et al., 2023). These parameters allow the model to generate a response that effectively mimics human language, but the process by which the model arrived at this response is not easily ascertained by experts, let alone users (Bhattacharjee et al., 2023; Gat et al., 2023; Ramlochan, 2023). Even open-source models, such as LLaMA (Meta, 2025) and Mistral (2025), which make their parameters publicly available, are far too complex for a typical user to understand. As such, generating insights regarding the underlying functionality of LLMs is extremely difficult, despite users claiming that they place a high value on transparency (Benk et al., 2025). This may lead to dissatisfaction and frustration because the LLMs are not as transparent as users expect them to be (Y. Kim et al., 2024).

Opening the Black Box

AI experts have sought to open the black box of LLMs by reverse-engineering their internal mechanisms. One way this is achieved is through mechanistic interpretability research, which leverages advanced machine learning techniques to identify the role that individual parameters, neurons, or circuits play in generating outputs (Bereska & Gavves, 2024; Gantla, 2025; Gurnee et al., 2023; Maltoni & Ferrara, 2024; OpenAI, 2023; Rai et al., 2024). Another approach, explainability, focuses on generating higher-level, human-understandable insights into the inner workings of LLMs, with emphasis on the macro-level relationship between inputs and outputs, rather than the intermediary processes (Barman et al., 2024; de Carvalho Souza & Weigang, 2025; Gantla, 2025; Wu et al., 2024; Zhao et al., 2024). For example, explainability research seeks to identify what information LLMs attend to (Jaunet et al., 2022; Yeh et al., 2023),

how minor deviations in prompts may influence or degrade the quality of their responses (Sayeedi et al., 2025; Z. Wang et al., 2024), and how LLMs respond to counterfactual prompts (Bhattacharjee et al., 2023; Mishra et al., 2024; Sarkar et al., 2025). Explainability research can thus help users understand the biases that LLMs hold (Barman et al., 2024; Sovrano, 2025), how their prompts may change the quality of the output they receive (You et al., 2025), and the extent to which they should trust LLMs for certain tasks (Cash, Oppenheimer, et al., 2026; Y. Wang et al., 2025). It may also help to clear up misconceptions that users hold about LLMs’ underlying processes, functions, and capabilities (Amaratunga, 2023; Barman et al., 2024).

LLM Faithfulness

While mechanistic interpretability and explainability research are critical for developing a deeper understanding of LLMs, the methods are complex, and their results may not be directly relevant to any individual user’s conversation with an LLM. As such, we must also consider how users themselves can make sense of how the LLM they are using generates the responses that it does (Rudin, 2019; Yeo et al., 2024). Given that LLMs are language-based agents, the most obvious solution is for the user to ask the model to explain its reasoning (Jacovi & Goldberg, 2020; Yu et al., 2025). In fact, some common models, including ChatGPT, will explain their reasoning processes without being prompted to do so (OpenAI, 2025a). However, there is significant concern regarding the degree to which LLMs exhibit faithfulness – that is, the extent to which their stated reasoning processes align with their actual reasoning processes (Fayyaz et al., 2024; Jacovi & Goldberg, 2020; Lyu et al., 2024; Madsen et al., 2024; Malin et al., 2025; Parcalabescu & Frank, 2023). For example, Matton et al. (2025) demonstrate that LLMs use social categories (e.g., gender) to inform their judgments but rarely mention those social categories in their justifications. Similarly, Turpin et al. (2023) find that when an LLM prompt

includes biasing or false information, the model will use that information to inform its decision but not acknowledge that it did so. More promisingly, however, Siegel et al. (2025) find that larger and more advanced models show greater faithfulness than smaller, less-advanced models, giving hope that faithfulness will increase over time.

The question of faithfulness is part of a larger debate about the extent to which LLMs can engage in metacognitive processes – that is, the extent to which they can think about their own thinking (A. Brown, 1987; Flavell, 1979). Much of this debate has centered on introspection, the ability of LLMs to evaluate their own mental states (Boring, 1953; Engelbert & Carruthers, 2010). Some philosophers argue that LLMs are largely incapable of introspection because they are merely mimicking human behavior (Comsa & Shanahan, 2025; Li et al., 2024), whereas other scholars argue that LLMs’ advanced computational abilities may allow them to engage in introspection more effectively than humans, if not now, then in future iterations (Cappelen & Dever, 2024; Long, 2023). Critically, however, LLM introspection may look very different than human introspection because the LLMs will leverage their strengths – such as their advanced computational capacity, memory storage, and probabilistic reasoning skills – which happen to be functions that are quite limited and fallible among humans (Simon, 1955).

Human Introspection

The debate about LLM introspection mirrors a long-standing debate in the psychological literature regarding the extent to which humans can engage in true introspection. This debate began with classical philosophers, many of whom assumed that humans have direct access to the inner workings of their mind and treated introspection as a critical component of the human experience (Descartes, 1641; Kant, 1781). Many modern philosophers still view introspection as an important method by which humans navigate a complex and cognitively demanding world

(Schwitzgebel, 2024; Smithies & Stoljar, 2012; Vermersch, 2009). However, cognitive science research has largely turned away from this direct access perspective (Engelbert & Carruthers, 2010; Pronin, 2009). Instead, most modern psychologists argue that agents' metacognitive knowledge – that is, their declarative knowledge about their own cognitive processes (Cash & Oppenheimer, 2025; Dunlosky & Metcalfe, 2009; Flavell, 1979) – is informed by post-hoc justifications based on guesswork or deduction (Nisbett & Wilson, 1977; Wilson & Nisbett, 1978) and/or inferences generated based on metacognitive cues that reflect statistical regularities learned about the task, environment, or agent itself (Ackerman, 2019; Bem, 1972; Koriart, 1997; Thompson et al., 2013). LLMs could plausibly achieve metacognition through these cue-based, non-introspective processes (Cash & Oppenheimer, 2024a; Cash, Oppenheimer, et al., 2026).

A popular framework for evaluating the accuracy of individuals' metacognitive knowledge, particularly in the context of judgment and decision making is Brunswik's (1952) lens model. The lens model suggests that humans make judgments – including metacognitive judgments – based on observable cues that have an (often unknown) statistical relationship with a given criterion. However, the weight that the judge places on each cue (i.e., their cue utilization) is not necessarily the same as the weight that the cue has in determining the outcome of the criterion variable (i.e., ecological validity; Hammond, 1955; Karelaia & Hogarth, 2008). For example, imagine that a judge is trying to determine if it is going to rain, based on three cues: humidity, temperature, and cloud cover. The judge may believe that humidity is the most predictive cue (say, a 70% weight), cloud cover is the second most predictive cue (25% weight) and temperature barely matters (5% weight) – but in reality, each cue is equally predictive (33.3%). In this case, the judge may overestimate the likelihood of rain on a cold but humid day or underestimate the likelihood of rain on a hot, low-humidity day.

Though the lens model was originally conceptualized to explain tasks with objective outcomes, researchers have adapted it to investigate decision makers' understanding of how they make subjective, multi-attribute judgments and decisions (Harries et al., 2000; Newell & Shanks, 2014; Slovic, 1969; Slovic et al., 1972; Slovic & Lichtenstein, 1971). In these studies, participants are asked to make judgments or decisions (e.g., stock price predictions, faculty evaluations) and then to estimate the influence – or weight – that each attribute had over their decision-making processes (i.e., their cue utilization). These stated weights are then compared to a mathematical estimate of the weight they placed on each cue, based on the choices or judgments they made (i.e., ecological validity). In this sense, their choice behavior is treated as the objective criterion, whereas their beliefs are treated as a noisy estimate thereof. Differences between the two sets of weights can be thought of as self-insight errors (Slovic & Lichtenstein, 1971) or limitations in metacognitive knowledge (Cash & Oppenheimer, 2024b, 2025).

A great deal of research using this kind of paradigm has suggested that decision makers' stated weights are often misaligned with revealed weights estimated from their choices ($r_s \sim .34$), with significant heterogeneity in accuracy across participants (Slovic, 1969; Slovic et al., 1972; Slovic & Lichtenstein, 1971). Hoepfl and Huber (1970) found that utility models (i.e., revealed weights) predicted decision makers' choices better than their own stated weights. These metacognitive limitations have been demonstrated across a wide variety of domains, including financial decision making (Slovic, 1969), medical decision making (Harries et al., 2000), social judgments (Grahe & Bernieri, 2002), and organizational fairness judgments (German et al., 2016). One common explanation for these self-insight errors is that decision makers often overestimate the importance of minor cues, while underestimating the weight they placed on a small number of highly important cues (Newell & Shanks, 2014; Slovic et al., 1972). Other

scholars argue that these studies may not be evidence of a lack of metacognitive knowledge, but rather evidence of an inability to articulate said knowledge (Reilly & Doherty, 1989, 1992).

Several newer studies aiming to quantify the accuracy of decision makers' metacognitive knowledge have demonstrated that decision makers' metacognitive knowledge may be more accurate than the earlier research suggested. Harte and Koele (1995) found that correlations between stated and revealed weights were rather high ($r \sim .75$) and that the two sets of weights were equally predictive of participants' choices. Using more advanced statistical models, Cash and Oppenheimer (2025) found that decision makers' stated and revealed weights tend to be moderately correlated ($r_s \sim .5$) in four unique choice domains (homes, romantic partners, job offers, colleges), replicating an earlier finding in the domain in school choice (Cash & Oppenheimer, 2024b). They also found significant heterogeneity in the accuracy of metacognitive knowledge across participants that was reliable across timepoints. Similarly, Morris et al. (2025) found that participants had highly accurate metacognitive knowledge ($r_s \sim .55-.75$) and demonstrated that participants are able to accurately report their use of various heuristic strategies. This evidence suggests that human decision makers have significant, if imperfect, metacognitive knowledge and can be reasonably faithful in their descriptions of their multi-attribute choice behavior. One potential explanation for the discrepancy across findings is that the older studies often used rather noisy methods for estimating revealed weights that may have inflated the size of self-insight errors (Rakow, 2014).

Comparing Human and LLM Metacognition

Despite the established literature evaluating humans' metacognitive knowledge in multi-attribute choice tasks, little is known about the faithfulness with which LLMs can report the weight they place on various attributes when making multi-attribute choices. Given the extant

literature on LLMs' limited faithfulness in other domains (Madsen et al., 2024; Matton et al., 2025) and the fact that LLMs do not have true and consistent preferences of their own (Bakker et al., 2022; Rozado, 2024), it would be easy to assume that their metacognitive knowledge of attribute weights would not be particularly faithful. However, much of what makes multi-attribute decisions difficult for humans is that they require the decision maker to navigate a substantial amount of information that is overwhelming for their limited attention (Krajovich, 2019; Orquin et al., 2021; Van Knippenberg et al., 2015), numeracy (Peters et al., 2006; Reyna et al., 2009), and working memory (Hinson et al., 2003; Iyengar & Lepper, 2000; Phillips-Wren & Adya, 2020). LLMs, on the other hand, do not face these limitations, and thus may be able to generate and maintain more faithful metacognitive knowledge than humans. This aligns with Cappelen and Dever's (2024) argument that LLMs' advanced computational abilities may allow them to introspect more effectively than humans.

Though the research is nascent, there is also growing evidence to suggest that LLMs may be developing human-like metacognitive abilities on other metacognitive tasks. Recent evidence on one kind of metacognitive judgment, confidence, suggests that LLMs generate confidence judgments that are about as accurate as those of humans and display human-like levels of overconfidence (Cash, Oppenheimer, et al., 2026; Groot & Valdenegro-Toro, 2024; Lin et al., 2022; Steyvers & Peters, 2025; Steyvers et al., 2025). Beyond metacognition, LLMs have also met or surpassed human performance on a wide range of cognitive tasks including numeric reasoning (Binz & Schulz, 2023), utility-maximization (Chen et al., 2023), and emotional awareness (Elyoseph et al., 2023), while using similar heuristics as humans (Suri et al., 2024). However, LLM performance varies widely by model and task (Banerjee et al., 2023; N. B.

Brown, 2024; Steyvers & Peters, 2025; Xiong et al., 2023). As such, it is essential that we continue to test a variety of LLMs across a wide range of cognitive and metacognitive tasks.

Current Studies

In the study presented here, we will seek to extend the literature on LLM faithfulness by exploring the accuracy of LLMs’ metacognitive knowledge of the weights they apply when making multi-attribute choice decisions. We will compare the metacognitive knowledge of humans to the metacognitive knowledge of four LLMs: ChatGPT-4o (henceforth GPT4o), Gemini 2.5-Flash (henceforth Gemini), Claude Sonnet 4 (henceforth Sonnet), and ChatGPT-5 (henceforth GPT5). Critically, GPT4o is a generation older than the other models, – having been released in 2024 whereas the other models were released in 2025 – which allows us to evaluate how LLMs’ metacognitive knowledge has improved over time. This is particularly interesting because each of the three newer models claim to have advanced reasoning capabilities and some extent of metacognitive ability (Anthropic, 2025b; Google, 2025b; OpenAI, 2025b). For example, GPT5 can determine how complex a question is and adjust its deliberation time accordingly, a classic example of metacognitive control (Ackerman, 2014; Bénon et al., 2024). As such, it is plausible to believe that these models may demonstrate a greater ability to faithfully explain their reasoning processes, relative to GPT4o. Through this study, we will develop an understanding of the faithfulness with which LLMs can explain their multi-attribute choice behavior, how their ability to do so compares to that of humans, and the extent to which their metacognitive abilities vary by model and across time. To foreshadow the results, we find that ChatGPT4o’s metacognitive knowledge was significantly less accurate than that of humans, whereas the newer models met or surpassed human performance, suggesting improvement in LLMs’ metacognitive abilities.

Ethics Statement

The studies presented here were approved under the auspices of the IRB at Carnegie Mellon University (STUDY2023_00000446) and carried out in accordance with the Declaration of Helsinki. The privacy rights of human participants were observed. Written, informed consent was obtained from all participants.

Methods

Overview

In this study, we tasked human participants and four LLMs – GPT4o, Gemini, Sonnet, and GPT5 – with completing the Knowledge of Weights (KoW) paradigm, a validated experimental method for assessing metacognitive knowledge in multi-attribute choice tasks (Cash & Oppenheimer, 2025). Participants first completed a choice task in which they made 14 choices between sets of three hypothetical homes that varied on six attributes (e.g., size). Because they do not have preferences of their own, the LLMs were tasked with choosing homes that a typical Prolific participant would be most interested in buying. Human participants were randomly assigned to a *Buyer* condition in which they chose homes that they would be most interested in buying or a *Proxy* condition in which they chose homes that a typical Prolific participant would be most interested in buying, mirroring the experience of the LLMs. After making these choices, participants were asked to self-report the weight that they believed they placed on each attribute. We then assessed the alignment between participants’ stated weights, their choices, and revealed weights estimated from their choices, with the notion that greater alignment suggests more accurate metacognitive knowledge. Lastly, we compared the accuracy of metacognitive knowledge across samples.

This study was initially pre-registered as two studies. The first pre-registration included the human samples and GPT4o: <https://aspredicted.org/ynp8-vkyy.pdf>. The second pre-registration included the three LLMs from 2025: <https://aspredicted.org/kbpf-rd7y.pdf>. For brevity and coherence, we chose to report the results simultaneously. All analyses were pre-registered, except where noted.

Participants

Human Participants. Simulations indicated that we would need a minimum of 200 participants in each condition (Buyer and Proxy) to reliably estimate revealed weights from their choices. To allow for potential data quality issues, we recruited 440 Prolific participants, and 436 completed the entirety of the study (Buyer $n = 216$; Proxy $n = 220$). All participants in each condition passed at least three out of five attention checks (81.2% passed all five) so, following our pre-registration, no participants were excluded from our analysis. Participants were paid \$2.00. Demographic information is provided in Table 1. Human data was collected in November 2024.

LLMs. The four LLMs were tasked with completing the study 220 times. To maximize independence across conversations, we disabled each model’s ability to reference saved memories from prior conversations. Otherwise, all default settings were used, and no customization or personalization was implemented. To maximally replicate the typical user experience, all survey iterations were conducted using each model’s chatbot interface (Anthropic, 2025a; Google, 2025a; OpenAI, 2025a), with prompts uploaded by an undergraduate research assistant rather than using the API. In one iteration, Gemini refused to provide SAWs. As such, analyses of Gemini’s SAWs will be based on a sample of 219. Data was collected for

GPT 4o in September-October 2024 and data for the newer models was collected in July-September 2025.

Table 1: Human Participant Demographics

Variable	Buyers (<i>n</i> = 216)	Proxies (<i>n</i> = 220)
Gender		
Female	120 (55.6%)	143 (65.0%)
Male	89 (41.2%)	69 (31.4%)
Non-Binary/Gender Not Listed	7 (3.2%)	6 (2.7%)
Prefer Not to Answer	0	2 (0.9%)
Average Age (<i>SD</i>)	39.0 (12.1)	38.1 (13.0)
Race/Ethnicity		
White/Caucasian	139 (64.4%)	136 (61.8%)
Black/African American	27 (12.5%)	31 (14.1%)
Asian/Asian American	18 (8.3%)	14 (6.4%)
Latino/Hispanic	11 (2.7%)	14 (6.4%)
Native American/Pac. Islander	1 (0.5%)	3 (1.4%)
Multi-Racial/Other Race	16 (7.4%)	18 (8.2%)
Prefer Not to Answer	4 (1.9%)	4 (1.8%)
Bachelor's Degree or Higher	127 (58.8%)	119 (54.1%)
Annual Household Income		
\$0 - \$24,999	24 (11.1%)	33 (15.0%)
\$25,000 - \$49,999	46 (21.3%)	44 (20.0%)
\$50,000 - \$74,999	37 (17.1%)	51 (23.2%)
\$75,000 - \$99,999	33 (15.3%)	33 (15.0%)
\$100,000 - \$124,999	26 (12.0%)	15 (6.8%)
\$125,000 - \$149,999	14 (6.5%)	20 (9.1%)
\$150,000 - \$174,999	15 (6.9%)	8 (3.6%)
\$175,000 - \$199,999	4 (1.9%)	3 (1.4%)
\$200,000 or Greater	8 (3.7%)	10 (4.5%)
Prefer Not to Answer	9 (4.2%)	3 (1.4%)

Procedures

In line with the Knowledge of Weights (KoW) paradigm (Cash & Oppenheimer, 2025), participants first completed a Choice-Based Conjoint (CBC) survey in which they made hypothetical choices between different homes. In the Buyer condition, participants were told that their goal was to choose the homes that they themselves would be most interested in buying. The LLMs and humans in the Proxy condition were told that their goal was to choose the homes that a typical Prolific participant would be most interested in buying. After receiving these instructions, participants were given brief descriptions of the six attributes on which the homes would vary. These descriptions laid out the five possible levels of each attribute, explained the scale on which each attribute was measured, and provided a real-world reference point for each attribute (e.g., the average value in the U.S.). Participants were then shown 14 sets of three hypothetical homes and asked to pick which one they (or a typical Prolific participant) would be most interested in buying from each set (see Supplemental Materials for exact language).

After the choice task, participants self-reported the weight that they believed they had put on each attribute. A key feature of the KoW paradigm is that these self-reports are framed as retrospective reflections about participants' choice behavior, thus making them a metacognitive task, rather than a preference elicitation task (Cash & Oppenheimer, 2025). Participants' self-reported weights were elicited in two distinct formats, presented in a random order. In one format, participants self-reported the *percentage* of their decision-making process that was based on each attribute (i.e., their attribute weights). Their responses were required to sum to 100% across attributes. We refer to these percentages as Stated Attribute Weights or **SAWs**. In the other format, participants rated how important each attribute was in their decision-making process on a 1-9 rating scale (1 = Not At All Important; 9 = Extremely Important). We refer to these ratings as

Attribute Importance Ratings, or **AIRs**. These are both common scale types used in the psychological and decision-making literatures (Nesbitt et al., 2024; Pöyhönen & Hämäläinen, 2001; Stevens, 1956), but there are several other weight-elicitation methods that could be implemented, including those that ask participants to consider the full range of attributes, like the swing-weighting method (Edwards & Barron, 1994; von Winterfeldt & Edwards, 1986). The KoW paradigm typically incentivizes accuracy with a bonus (Cash & Oppenheimer, 2025), but because LLMs cannot receive monetary bonuses, we eliminated human bonuses as well.

After providing their SAWs and AIRs, participants were asked to gauge their confidence in the accuracy of each set of self-reported values. Confidence was elicited on a scale of 0 (Not confident at all) to 100 (Completely confident). This can be considered a meta-metacognitive judgment, as it assesses participants' confidence in their metacognitive knowledge. Lastly, human participants completed a brief demographic questionnaire.

Materials

CBC Survey. We first had to determine the attributes that would be provided to describe each home. We used the six attributes validated by Cash and Oppenheimer (2025): 1) Commute time (to work); 2) Home size; 3) Mortgage (cost as % of income); 4) School district quality; 5) Attractiveness of the home (as rated by buyers); and 6) Lot size. These attributes were specifically selected by Cash and Oppenheimer (2025) because they are unlikely to have interactions with one another. We then generated five discrete levels for each attribute, which were used to create the alternatives in the CBC survey (see Supplemental Materials for details). These levels were designed to reflect homes typically seen in the U.S. housing market (e.g., commute times ranged from 10-50 minutes). We then generated 300 versions of the CBC survey. Each of these survey versions contained 14 choice tasks, and each choice task presented three

homes that were random permutations of the possible levels of each attribute. Since we had more versions than participants in each sample, versions were not repeated within sample. The CBC survey versions were generated using Lighthouse Studio (Sawtooth Software, Inc., 2025). The choice tasks were presented to humans visually (see Figure 1) and uploaded to the LLMs as an excel spreadsheet.

Figure 1: Human Participant Choice Task

Which of those homes would **you** be most interested in buying?

(1 of 14)

	Option 1	Option 2	Option 3
Home Size	2,500 sqft	2,000 sqft	1,000 sqft
Lot Size	0.25 Acres	0.50 Acres	0.10 Acres
Mortgage (% of Income)	35%	40%	25%
School District Quality	B	A-	D
Commute Time	40 Minutes	50 Minutes	10 Minutes
Attractiveness Rating	3.0	2.1	3.9
	Select	Select	Select

This screenshot was taken from the Buyer survey. In the Proxy survey, the word “you” in the instructions was replaced with “a typical Prolific participant.”

LLM Script. To administer the survey to the LLMs, we created a script that mirrored the language used in the human survey, with only minor deviations to direct the flow of the

conversation (e.g., asking “Are you ready for the spreadsheet?”). This script included seven prompts. Prompt 1 introduced the LLM to the task and provided the attribute descriptions. Prompt 2 explained the format of the spreadsheet that contained the choice tasks. Prompt 3 uploaded the spreadsheet. Prompt 4 elicited the LLM’s SAWs. Prompt 5 elicited the LLM’s confidence in its SAWs. Prompt 6 elicited the LLM’s AIRs. And lastly, Prompt 7 elicited the LLM’s confidence in its AIRs. The research assistant administering the survey flipped a coin to determine whether Prompts 4 and 5 or Prompts 6 and 7 would come first, mirroring the randomization of SAWs and AIRs for human participants. The full script is provided in the supplemental materials.

Analysis & Results

Estimating Revealed Attribute Weights

We first had to estimate the weights revealed by participants’ choices during the CBC task, which we will refer to as Revealed Attribute Weights, or **RAWs**. Following best practices for analyzing CBC data, we used Hierarchical Bayes (HB) estimation to calculate the part-worth value (sometimes called utility) that each participant placed on each level of each attribute (Cash & Oppenheimer, 2025; Eggers et al., 2022; Hein et al., 2020; Orme, 2002). We conducted this analysis using the HB function in Lighthouse Studio (Sawtooth Software, Inc., 2025)

The HB estimation process relies on a Markov Chain Monte Carlo, a Bayesian updating process in which the model iteratively samples possible part-worth values for each level of each attribute for each participant until it converges on part-worth values that accurately predict participants’ choices. The model begins with a null preference model that sets all part-worth values to zero. The model then generates a new set of part-worth values for each level of each attribute for each participant and estimates the probability that each participant would have made

the choices that they did if they adhered to the old or new sets of part-worth values. If their choices are more likely under the new set of part-worth values than the old set of part-worth values, then the new part-worth values are retained. These new part-worth values are then used as the prior for generating the next set of part-worth values and the process repeats. We allowed the model to estimate 30,000 sets of part-worth values, the last 10,000 of which were averaged to generate point estimates. The model used here is referred to as hierarchical because sample-level mean part-worth values are used to inform participant-level part-worth values. A technical overview of the HB model can be found in Sawtooth Software, Inc. (2021).

After running the HB model, we were left with point estimates for the part-worth value of each level of each attribute (6 attributes x 5 levels = 30 part-worth values for each participant). To convert these part-worth values into attribute-level RAWs, we applied the following formula (Cash & Oppenheimer, 2025; Eggers et al., 2022; Orme, 2002), where U_j is a vector containing the part-worth values for the five levels of attribute j , U_i is a vector containing the part-worth values for the five levels of attribute i , and the set from which attribute i is pulled includes all attributes from the CBC, including j :

$$RAW_j = 100 * \frac{\max(U_j) - \min(U_j)}{\sum_{i=1}^N (\max(U_i) - \min(U_i))}$$

Descriptives.

Attribute-level Mean RAWs, SAWs, and AIRs for each sample are reported in Table 2.

Table 2: Mean RAWs, SAWs, and AIRs by Sample

Attribute	Buyers	Proxies	GPT4o	Gemini	Sonnet	GPT5
Mean RAWs (sd)						
Commute Time	18.43 (8.30)	20.60 (7.61)	21.62 (18.95)	21.22 (6.56)	27.37 (3.19)	24.03 (9.93)
Home Size	13.08 (6.97)	11.09 (5.17)	42.13 (31.13)	13.68 (4.09)	12.19 (2.09)	14.10 (7.79)
Mortgage	18.85 (8.82)	19.24 (8.61)	8.07 (9.24)	23.70 (6.92)	28.33 (3.46)	14.25 (4.27)
School Quality	22.21 (13.72)	24.05 (11.45)	10.04 (8.73)	16.99 (3.51)	14.69 (2.41)	23.69 (7.60)
Attractiveness	14.98 (6.87)	16.52 (8.68)	11.91 (11.52)	13.60 (3.47)	12.38 (2.68)	15.81 (4.06)
Lot Size	12.45 (7.97)	8.51 (3.94)	6.23 (3.20)	10.83 (3.07)	5.04 (2.87)	8.13 (2.77)
Mean SAWs (sd)						
Commute Time	16.56 (14.68)	17.80 (9.63)	22.42 (3.87)	20.87 (3.59)	26.74 (3.09)	20.99 (3.35)
Home Size	16.31 (10.56)	14.66 (8.31)	16.13 (4.25)	14.76 (3.44)	15.50 (4.34)	13.67 (3.19)
Mortgage	24.51 (17.06)	26.96 (15.65)	25.01 (4.87)	27.00 (5.77)	31.93 (4.11)	25.27 (4.04)
School Quality	14.73 (18.28)	16.69 (12.38)	14.25 (4.74)	16.07 (4.25)	14.06 (4.60)	18.70 (4.12)
Attractiveness	13.27 (13.53)	14.60 (11.49)	14.02 (5.36)	11.61 (3.30)	8.16 (3.36)	13.83 (3.70)
Lot Size	14.60 (15.24)	9.28 (5.92)	8.16 (2.70)	9.69 (3.24)	3.61 (1.18)	7.55 (2.25)
Mean AIRs (sd)						
Commute Time	6.05 (2.30)	6.69 (1.80)	7.88 (0.71)	7.65 (0.91)	7.86 (0.47)	7.74 (0.75)
Home Size	6.07 (2.08)	6.01 (1.82)	6.63 (0.99)	6.04 (1.30)	5.98 (0.89)	6.17 (0.80)
Mortgage	7.01 (2.10)	7.49 (1.61)	8.43 (0.78)	8.91 (0.59)	8.63 (0.55)	8.62 (0.56)
School Quality	4.84 (2.91)	6.07 (2.17)	6.29 (1.14)	6.39 (1.61)	5.73 (1.18)	7.35 (0.95)
Attractiveness	5.13 (2.39)	5.72 (2.08)	6.16 (1.17)	4.94 (1.51)	4.21 (0.90)	6.31 (0.99)
Lot Size	5.47 (2.22)	4.99 (1.93)	4.33 (0.72)	4.20 (1.59)	2.78 (0.82)	4.28 (0.67)

Validating Part-Worth Values and RAWs

We next conducted an exploratory analysis in which we sought to validate that our part-worth value estimates were accurately capturing decision makers' choice behavior. To do so, we evaluated each choice made by each participant and estimated the total value (i.e., sum of part-worth values) of each alternative. We did so by identifying the part-worth value of each alternative's score on each attribute (e.g., for Participant 9, Alternative 1 has a commute time of 10 minutes, which has a part-worth value of 65.84), then summing across attributes to generate a total value for each alternative. We then calculated the sample-level percentage of choice tasks for which the alternative predicted to have the highest total value was the alternative that the participant actually selected. Estimated part-worth values accurately predicted greater than 99% of choices for all samples (see Table 3), indicating that they were highly successful in describing decision makers' choice behavior. This also provides evidence for our assumption that decision makers' choices could be effectively captured by an additive, compensatory model without interactions.

We next investigated the accuracy with which RAWs could predict participants' choice behavior. The key difference between RAWs and part-worth values is that RAWs do not capture the value function between levels of each attribute – only how important each attribute is writ large. Thus, we would expect RAWs to be less accurate than part-worth values, since they contain less information. Given that the value function connecting the levels is unknown, we had to make an assumption about its shape. We chose to assume a linear function where the lowest-value level of attribute for each participant was coded as 1 and the highest-value level was coded as 5 (Cash & Oppenheimer, 2025). The rank-order was determined independently for each participant. We then predicted which alternative each participant would selected for each choice

task by multiplying the level of each attribute (coded as 1-5) by the RAW for that attribute, then summing across the six attributes, generating a total value for each alternative home that ranged from 100 (worst possible combination of levels) to 500 (best possible combination of levels). Using this method, RAWs accurately predicted the selected alternative in 89.4% - 97.7% of choice tasks across samples (see Table 3). These results indicate that RAWs were highly accurate in capturing decision makers' choice behavior, even without access to the true underlying value functions. This also serves as an important benchmark for evaluating the accuracy of predictions made by SAWs and AIRs, which were only collected for each attribute, not each level, and therefore also lacked a defined value function. This analysis was pre-registered as exploratory.

Predicting Choices Using Other Compensatory Functions

To contextualize the accuracy of the predictions made by RAWs, we tested how accurately other, simpler weighting schemes could predict participants' choices. We started with an equal weighting scheme, where each attribute was given a weight of $(100/6)$. We then used the same method that we did for predicting choices via RAWs and found that the equal weight model accurately predicted 79.8% - 85.8% of choices for all samples (see Table 3). The notable exception was GPT4o, for which the equal weighting scheme accurately predicted only 66.5% of choices. This divergence is likely a result of the fact that, on average, GPT4o placed nearly half of its weight on the Home Size attribute, whereas the other samples distributed their weights reasonably evenly across attributes (see Table 2). Importantly, chi-squared tests indicated that RAWs were more accurate predictors of participants choices than the equal weights for all models ($\chi^2(1, Ns \geq 5,883) \geq 81.05, ps < .001$).

Table 3: Percentage of Choices Accurately Predicted

	Weighting Scheme/Decision Strategy						
Sample	Part-Worth Values	RAWs	Equal Weights	Random Weights	Take The Best	Tallying (Ties Randomized)	Tallying (Ties Excluded)
Buyers	99.1% [98.7 – 99.4%]	92.8% [91.8 – 93.6%]	82.9% [81.5 – 84.3%]	66.4% [64.7 – 68.1%]	70.7% [69.0 – 72.3%]	71.1% [69.4 – 72.7%]	78.5% [76.8 – 80.1%]
Proxies	99.3% [98.9 – 99.5%]	91.3% [90.2 – 92.2%]	83.6% [82.2 – 84.9%]	66.2% [64.5 – 67.9%]	66.5% [64.8 – 68.2%]	70.1% [68.4 – 71.7%]	76.8% [75.1 – 78.4%]
GPT4o	99.8% [99.6 – 99.9%]	97.7% [97.0 – 98.2%]	66.5% [64.8 – 68.2%]	56.9% [55.1 – 58.6%]	88.1% [86.9 – 89.2%]	63.0% [61.2 – 64.7%]	68.7% [66.8 – 70.6%]
GPT5	99.97% [99.8 – 100%]	96.3% [95.6 – 96.9%]	79.8% [78.3 – 81.2%]	63.5% [61.7 – 65.2%]	70.0% [68.3 – 71.6%]	67.6% [65.9 – 69.3%]	74.6% [72.8 – 76.3%]
Gemini	100.0% [99.8 – 100%]	95.7% [94.9 – 96.4%]	85.8% [84.5 – 87.0%]	65.8% [64.1 – 67.5%]	65.4% [63.7 – 67.1%]	71.6% [69.9 – 73.1%]	78.8% [77.1 – 80.4%]
Sonnet	99.5% [99.2 – 99.7%]	89.4% [88.2 – 90.4%]	80.4% [78.9 – 81.8%]	64.0% [62.2 – 65.7%]	62.7% [61.0 – 64.4%]	65.1% [63.3 – 66.7%]	72.1% [70.3 – 73.9%]

Note. Brackets indicate 95% confidence intervals.

We next tested the accuracy of a random weights model. To do so, we randomly generated ten novel sets of weights for each participant. We then used the same method that we did for predicting choices via RAWs but replaced the RAWs with these random weights. Averaged across the ten sets, the random weights model accurately predicted 56.9 – 66.4% of choices for each sample (see Table 3). These percentages are above chance (1/3) because the predictive model retained the assumed linear value function that captured participants' rank-order preferences across levels, thus giving it some information about participants' preferences. Unsurprisingly, pairwise chi-squared tests indicate that RAWs were more accurate predictors of participants choices than random weights ($\chi^2(1, Ns \geq 6,048) \geq 555.03, ps < .001$). Together, these results suggest that RAWs were generating meaningfully better estimates of participants' decision weights than other, simpler models. All analyses in this section were exploratory.

Predicting Choices Using Non-Compensatory Strategies

To this point, we have assumed that participants were making choices using a compensatory, weight-based strategy. However, decision makers often engage in non-compensatory decision processes that allow them to make choices based on only a subset of the available information (Einhorn, 1970; Tversky, 1972). As such, we next sought to test the extent to which participants' choices could be predicted using non-compensatory strategies. We first tested a lexicographic strategy akin to the Take-The-Best heuristic (Gigerenzer & Goldstein, 1996; Newell et al., 2003), applied to a three-alternative context. In this approach, decision makers start with the most important (or most valid) attribute and choose the alternative that scores highest on that attribute. If more than one attribute is tied for the highest score on that attribute, they move down the list, in order of attribute importance, until the tie is broken. Using this heuristic, we were able to accurately predict 88.1% of choices for GPT4o and 62.7 – 70.7%

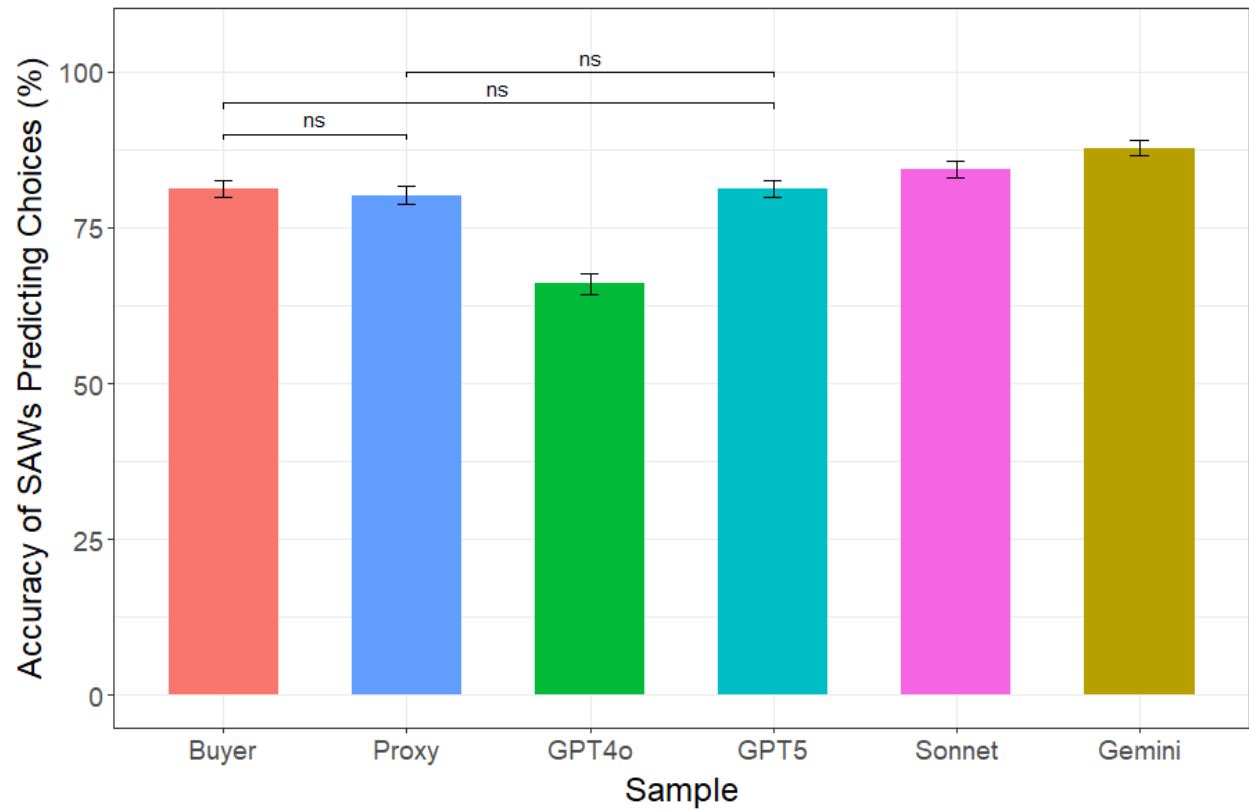
of choices for other samples (see Table 3). This suggests that GPT4o may have been acting in a somewhat non-compensatory fashion – as can be seen by the fact that it put an average RAW of 42.13% on Home Size (see Table 2) – whereas the other models were not. However, pairwise chi-squared tests indicated that RAWs predicted participants’ choices more accurately than the take-the-best heuristic for all samples, including GPT4o ($\chi^2(1, Ns \geq 6,048) \geq 212.00, ps < .001$).

We then tested the degree to which participants’ choices could be accurately predicted using the tallying heuristic, also known as unit-weighting or Dawes’ Rule (Dawes, 1979; Gigerenzer & Goldstein, 1996; Payne et al., 1993). When using this heuristic, each attribute is evaluated separately, and the alternative that scores best on that attribute is given a point. These points are then summed and the alternative with the highest score is selected. When implementing this heuristic, we handled attribute-level ties (e.g., two homes had the same commute time) by giving one point to each alternative. If two or more alternatives were tied after all attributes were considered, we randomly selected a winner. Using this model, the tallying heuristic accurately predicted only 63.0 – 71.6% of choices across samples (see Table 3). Pairwise chi-squared tests indicated that RAWs predicted participants’ choices more accurately than the tallying heuristic for all samples ($\chi^2(1, Ns \geq 6,048) \geq 442.48, ps < .001$). If we excluded choice tasks for which the tallying heuristic resulted in a tie, accuracy rates improved to 68.7 – 78.8% of choices across samples (see Table 3) but remained significantly lower than the accuracy rates for RAWs ($\chi^2(1, Ns \geq 5,482) \geq 221.93, ps < .001$). In conjunction, these results suggest that the compensatory RAW model is a better model of choice behavior than the non-compensatory models, though the take-the-best heuristic performed reasonably well for GPT4o. All analyses in this section were exploratory.

SAWs Predicting Choices

Having established that our participants' choice behavior is best predicted by a compensatory model, we now turn to our primary question of the accuracy of participants' metacognitive knowledge. As our first measure of metacognitive knowledge, we investigated the percentage of choice tasks for which SAWs accurately predicted the alternative each participant selected for each choice task. We did so using the same procedure as we used to predict choices using RAWs but replaced each participant's RAWs with their SAWs. When SAWs accurately predict the selected alternative in a greater proportion of choice tasks, it is indicative of more accurate metacognitive knowledge (Cash & Oppenheimer, 2025).

We found that SAWs accurately predicted 81.3% of choices for Buyers (95% CI = 79.8 – 82.7%), 80.2% of choices for Proxies (95% CI = 78.7 – 81.6%), 66.0% of choices for GPT4o (95% CI = 64.3 – 67.7%), 81.3% of choices for GPT5 (95% CI = 79.8 – 82.7%), 87.8% of choices for Gemini (95% CI = 86.6 – 88.9%), and 84.4% of choices for Sonnet (95% CI = 83.0 – 85.6%). Pairwise chi-squared tests indicated that these proportions were not significantly different for Buyers and Proxies ($\chi^2(1, N = 5,974) = 1.12, p = .29$). GPT4o's proportion was significantly lower than that of Buyers and Proxies ($\chi^2(1, Ns \geq 5,972) \geq 152.74, ps < .001$). GPT5's proportion was not significantly different than that of Buyers or Proxies ($\chi^2(1, Ns \geq 5,967) \leq 1.14, ps \geq .29$). Proportions for Gemini and Sonnet were significantly higher than those of Buyers and Proxies ($\chi^2(1, Ns \geq 5,968) \geq 9.82, ps \leq .002$). All significant differences were robust to the application of a Bonferroni correction for multiple comparisons ($n = 15$ comparisons; significant p cutoff = .003) and all pairwise comparisons are reported in Figure 2. In all, these results suggest that the oldest model, GPT4o, had significantly worse metacognitive knowledge than humans, but that the newer models either matched or exceeded human performance.

Figure 2: SAWs Predicting Choices

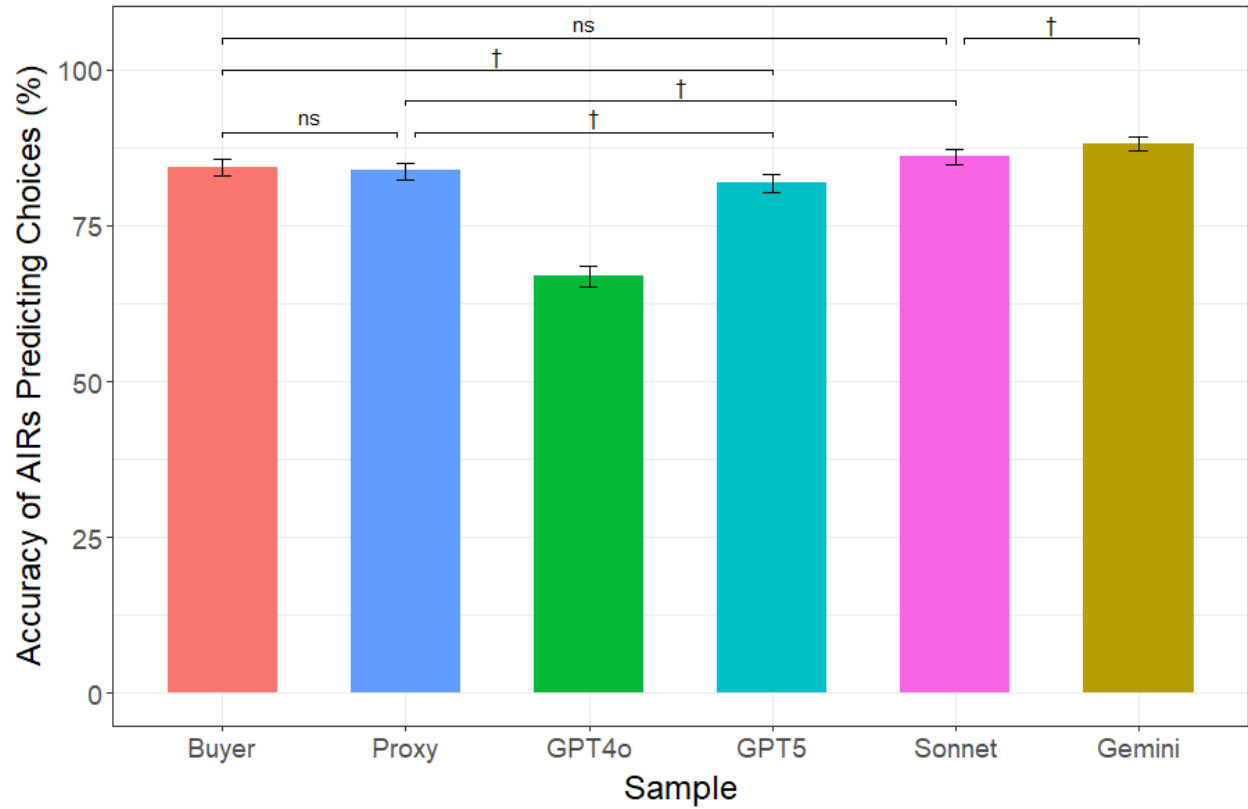
ns = not significant; All differences are significant after Bonferroni correction ($p < .003$), except where noted. Error bars reflect 95% confidence intervals.

AIRs Predicting Choices

As our next measure of metacognitive knowledge, we exploratorily investigated the percentage of choice tasks for which AIRs accurately predicted the alternative each participant selected for each choice task. We again used the same procedure that we used to predict choices based on RAWs, but using AIRs as the weighting function instead of RAWs. We found that AIRs accurately predicted 84.4% of choices for Buyers (95% CI = 83.1 – 85.7%), 83.8% of choices for Proxies (95% CI = 82.5 – 85.1%), 66.9% of choices for GPT4o (95% CI = 65.2 – 68.5%), 81.8%

of choices for GPT5 (95% CI = 80.4 – 83.2%), 88.2% of choices for Gemini (95% CI = 86.9 – 89.3%), and 86.1% of choices for Sonnet (95% CI = 84.8 – 87.3%).

Pairwise chi-squared tests indicated that these proportions were not significantly different for Buyers and Proxies ($\chi^2(1, N = 6,025) = 0.40, p = .53$). GPT4o's proportion was significantly lower than that of Buyers and Proxies ($\chi^2(1, Ns \geq 6,043) \geq 234.02, ps < .001$) and the differences were robust to Bonferroni correction. GPT5's proportion was significantly lower than the proportions for Buyers and Proxies ($\chi^2(1, Ns \geq 6,029) \geq 4.07, ps < .04$), but the differences were not robust to Bonferroni correction. Gemini's proportion was significantly higher than the proportion for Buyers and Proxies ($\chi^2(1, Ns \geq 6,023) \geq 17.29, ps < .001$) and the differences were robust to Bonferroni correction. Sonnet's proportion was not significantly different than the proportion for Buyers ($\chi^2(1, N = 6,021) = 3.18, p = .07$) but was significantly higher than the proportion for Proxies ($\chi^2(1, N = 6,066) = 6.04, p = .01$), though the effect was not robust to Bonferroni correction. All pairwise comparisons are reported in Figure 3. In conjunction, these results once again suggest that the oldest LLM – GPT4o – had relatively limited metacognitive knowledge compared to humans, but that the newer LLMs matched or surpassed human performance.

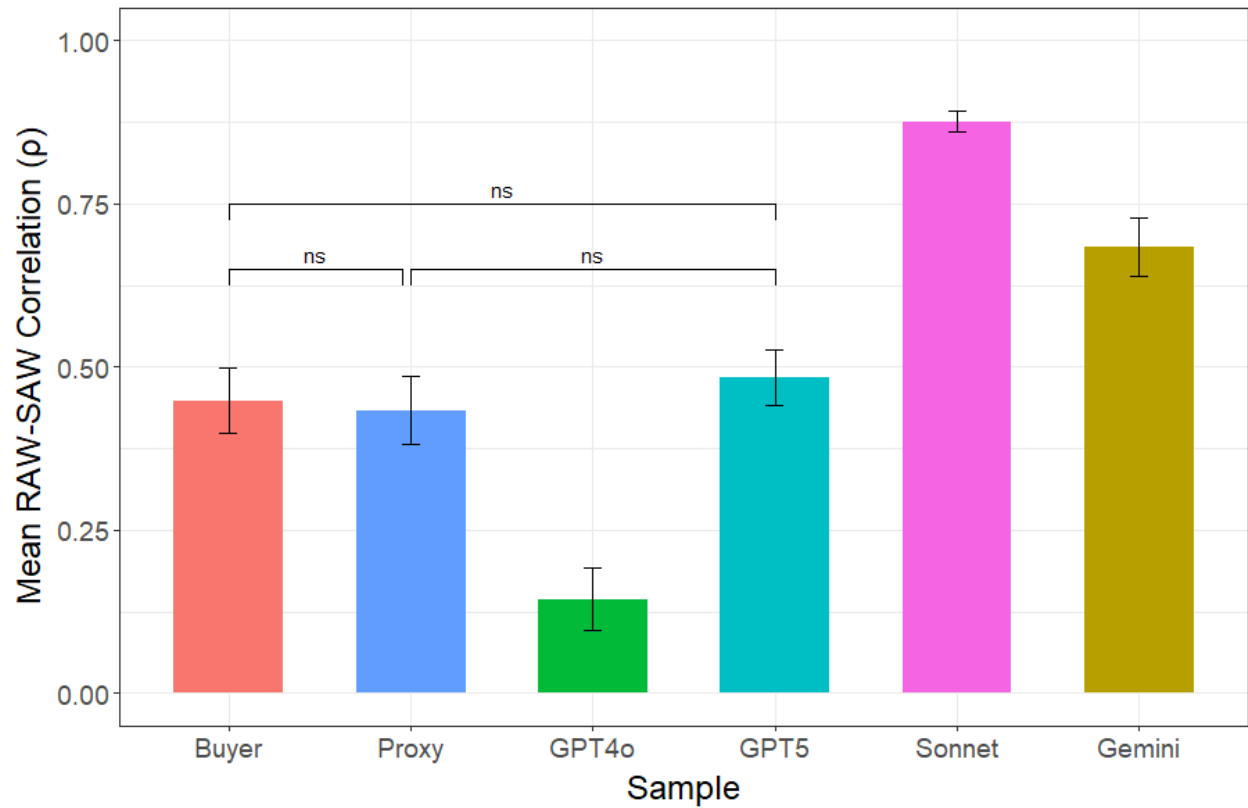
Figure 3: AIRs Predicting Choices

ns = not significant; † = Significant ($p < .05$), but not robust to Bonferroni correction ($p_{crit} = .003$). All differences are significant after Bonferroni correction except where noted. Error bars reflect 95% confidence intervals.

Correlations Between RAWs and SAWs

As another measure of metacognitive knowledge, we next assessed the correlation between RAW and SAWs. This metric captures a decision maker's ability to accurately identify the relative influence that each attribute had on their decision-making process, compared to the other attributes. Though we acknowledge that RAWs are not perfect estimates of choice behavior (Rakow, 2014), the choice analyses suggest that they are better estimates than SAWs or AIRs, so we treat them as a proxy for decision makers' true weights. This conceptualization is consistent

with the prior literature, which has demonstrated the validity and reliability of RAWs across several domains, including homebuying (Cash & Oppenheimer, 2025). To conduct this analysis, we first calculated the Spearman rank-order correlation across attributes for each participant. We then averaged the participant-level RAW-SAW correlations within each sample. The mean RAW-SAW correlation was $\rho = 0.45$ ($sd = 0.37$) for Buyers, $\rho = 0.43$ ($sd = 0.39$) for Proxies, $\rho = 0.14$ ($sd = .36$) for GPT4o, $\rho = 0.48$ ($sd = 0.32$) for GPT5, $\rho = 0.68$ ($sd = 0.33$) for Gemini, $\rho = 0.88$ ($sd = .12$) for Sonnet. A one-way ANOVA indicated that these means were significantly different from one another ($F(5, 1309) = 126.00, p < .001$). Pairwise comparisons corrected using Tukey's HSDs indicated that the mean correlations for Buyers and Proxies were not significantly different from one another ($p = .997$). The mean correlation for GPT4o was significantly lower than the mean correlations for Buyers and Proxies ($ps < .001$). The mean correlation for GPT5 was not significantly different than the mean correlation for Buyers ($p = .88$) or Proxies ($p = .61$). Lastly, the mean correlations for Gemini and Sonnet were significantly higher than the mean correlations for both sets of humans ($ps < .001$). The full set of pairwise comparisons is depicted in Figure 4. In all, these findings replicate the choice-based results by demonstrating once again the GPT4o had significantly less accurate metacognitive knowledge than humans, whereas the newer LLMs demonstrated metacognitive knowledge that was as accurate or more accurate than that of humans.

Figure 4: Mean RAW-SAW Correlations

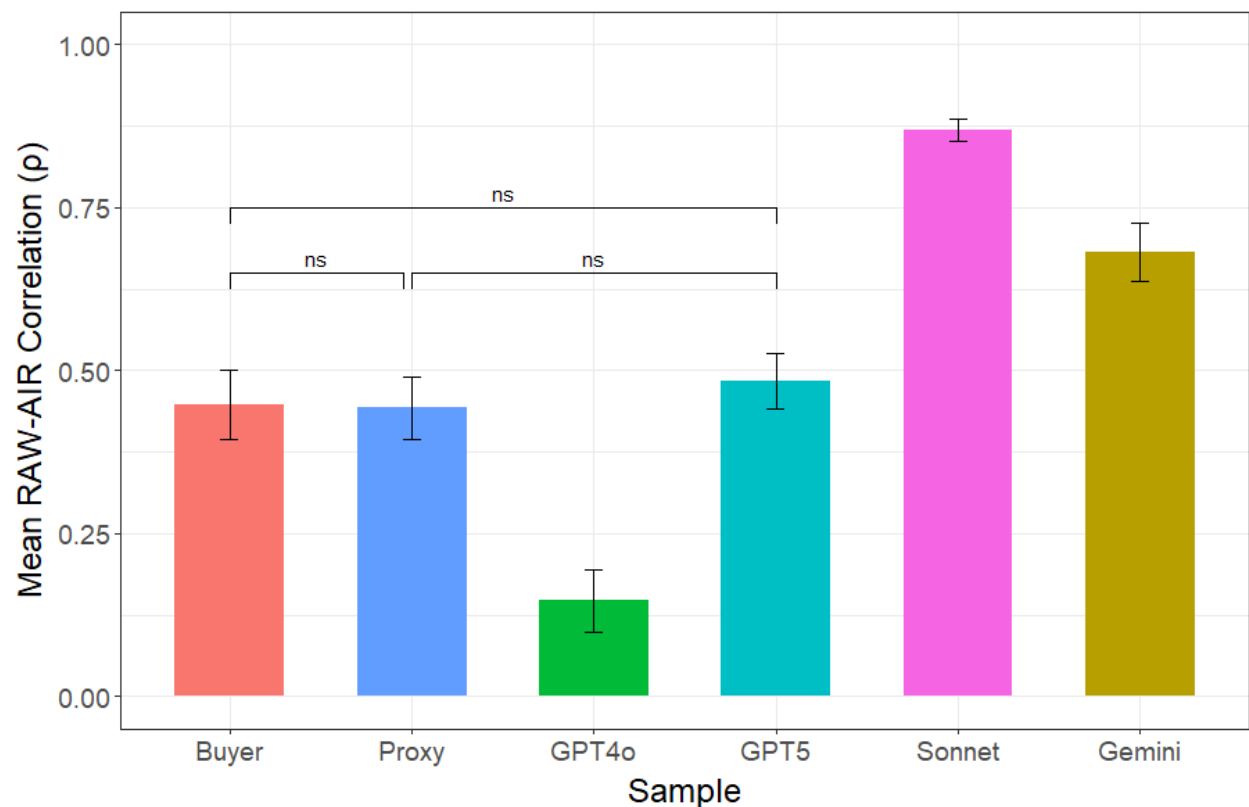
ns = not significant; All differences are significant except where noted. Error bars reflect 95% confidence intervals.

Correlations Between RAWs and AIRs

We next evaluated the participant-level Spearman rank-order correlations between RAWs and AIRs. The mean RAW-AIR correlation was $\rho = 0.45$ ($sd = 0.40$) for Buyers, $\rho = 0.45$ ($sd = 0.36$) for Proxies, $\rho = 0.15$ ($sd = .37$) for GPT4o, $\rho = 0.48$ ($sd = 0.32$) for GPT5, $\rho = 0.68$ ($sd = 0.34$) for Gemini, $\rho = 0.87$ ($sd = .13$) for Sonnet. A one-way ANOVA indicated that these means were significantly different from one another ($F(5, 1305) = 120.50, p < .001$). Pairwise comparisons corrected using Tukey's HSDs indicated that the mean correlations for Buyers and Proxies were not significantly different from one another ($p = .999$). The mean correlation for

GPT4o was significantly lower than the mean correlations for Buyers and Proxies ($p < .001$). The mean correlation for GPT5 was not significantly different than the mean correlation for Buyers ($p = .86$) or Proxies ($p = .93$). Lastly, the mean correlations for Gemini and Sonnet were significantly higher than the mean correlations for both sets of humans ($p < .001$). The full set of pairwise comparisons is depicted in Figure 5. These findings nearly perfectly replicate the results from the analysis of RAW-SAW correlations, again suggesting that GPT4o had less accurate metacognitive knowledge than humans, whereas the newer models were as accurate or more accurate than humans.

Figure 5: Mean RAW-AIR Correlations



ns = not significant; All differences are significant except where noted. Error bars reflect 95% confidence intervals.

Confidence

Lastly, we evaluated each sample's confidence in their ability to provide accurate SAWs and AIRs. Participants' mean confidence in the accuracy of their SAWs was 80.37 ($sd = 17.17$) for Buyers, 68.64 ($sd = 20.94$) for Proxies, 85.27 ($sd = 2.28$) for GPT4o, 89.47 ($sd = 7.58$) for Gemini, 79.44 ($sd = 4.31$) for Sonnet, and 84.54 ($sd = 2.64$) for GPT5. A one-way ANOVA comparing confidence in SAWs across all six samples indicated that the means were significantly different from one another ($F(5, 1306) = 82.76, p < .001$). Pairwise comparisons using Tukey's HSDs indicated that Proxies were less confident than all other samples ($ps < .001$), whereas Buyers were less confident than all LLMs ($ps \leq .003$), except for Sonnet ($p = .96$). Interestingly, we found little relationship between accuracy and confidence for the LLMs. The least accurate model, GPT4o, was the second most confident model (only Gemini was significantly more confident; $p = .002$), whereas Sonnet, which was consistently one of the two most accurate models, was significantly less confident than all other LLMs ($ps \leq .002$).

Mean confidence in AIRs told a similar story. Mean AIR confidence was 82.19 ($sd = 15.94$) for Buyers, 69.70 ($sd = 20.28$) for Proxies, 88.52 ($sd = 2.77$) for GPT4o, 90.14 ($sd = 6.51$) for Gemini, 84.44 ($sd = 3.38$) for Sonnet, and 88.52 ($sd = 2.39$) for GPT5. A one-way ANOVA comparing confidence in AIRs across all six samples indicated that the means were significantly different from one another ($F(5, 1308) = 103.30, p < .001$). Pairwise comparisons using Tukey's HSDs again indicated that Proxies were less confident than all other samples ($ps < .001$), whereas Buyers were less confident than all LLMs ($ps \leq .003$), except for Sonnet ($p = .27$). Again, we found that the least accurate model, GPT4o, was the second most confident model (and not significantly less confident than the most confident model, Gemini; $p = .64$), whereas Sonnet was significantly less confident than all other models ($ps \leq .002$). In conjunction, these

results suggest that LLMs tend to be highly confident – often more confident than humans – even when they are not more accurate than humans. The lower confidence from Proxies relative to Buyers is unsurprising, as making choices for unknown others may feel like a more difficult task than making choices for oneself, even though this was not evident in the accuracy of their metacognitive knowledge.

Deviations from Pre-Registration

We initially pre-registered that we would evaluate RAW-SAW and RAW-AIR correlations for each attribute across participants within each sample. However, we chose instead to evaluate participant-level correlations across attributes because participant-level correlations are more powerful than attribute-level correlations and more typical in the metacognition literature (Fleming & Lau, 2014). In response to reviewer feedback, we also switched to Spearman correlations instead of Pearson correlations, as this would eliminate the need to transform our data. For transparency, we report attribute-level Pearson correlations in the Supplemental Materials and further discuss the limitations of attribute-level correlations. In response to further reviewer feedback, we also chose to report how accurately SAWs could predict participants' choices, rather than how often they made the same predictions as RAWs (the pre-registered analysis). The results, presented in the Supplemental Materials, were qualitatively similar, but we believe the reported analysis is clearer and more meaningful. Lastly, we originally pre-registered that we would study ChatGPT4.1, but that we would switch to a newer version of any model if one was released before data for the older version was collected. ChatGPT5 was released before data collection began, so we evaluated it instead of ChatGPT4.1. Additional pre-registered analyses are reported in the Supplemental Materials.

Summary of Results

These results provide several insights into LLMs' metacognitive knowledge of how they make multi-attribute choice decisions. First, we demonstrated that all four models' choice behavior could be effectively captured using an additive, compensatory model. Second, we found that the oldest model – GPT4o – had consistently poor metacognitive knowledge across metrics, far below that which is achieved by humans. Third, we found that the newer models – GPT5, Sonnet, and Gemini – consistently matched or surpassed humans' metacognitive knowledge. GPT5 tended to perform about as well as humans, whereas Sonnet and Gemini tended to demonstrate greater metacognitive knowledge than humans. This suggests that, over time, the ability of LLMs to report their decision processes has improved. Fourth, we found that all LLMs tended to be quite confident in their metacognitive knowledge. All the models were more confident than humans, with the exception that Sonnet was not more confident than Buyers. This high confidence even applied to the oft-inaccurate GPT4o, suggesting that our results are consistent with the growing literature demonstrating that LLMs are highly confident, and often overconfident (Cash, Oppenheimer, et al., 2026; Groot & Valdenegro-Toro, 2024; Xiong et al., 2023). In all, these results suggest that LLMs do have the ability to generate and maintain accurate metacognitive knowledge of the weights they apply when making multi-attribute choice decisions and that this ability has improved over time.

Data Availability Statement

Data, materials, and code for are available on OSF:

https://osf.io/r73ez/overview?view_only=4c52a90f5cc8437894f24356132e209b

General Discussion

As LLMs become increasingly integrated into our everyday lives, it is more critical than ever that users understand how the models arrive at the conclusions that they do. Though users cannot be expected – and likely have no desire – to understand the nuances of the models’ cognitive architectures, it would be beneficial for users to know where LLMs source their information and how this information is used to guide their reasoning processes. Given that LLMs are capable of human-like language (Wu et al., 2025), it is reasonable for users to believe that they could simply ask the models to explain their underlying cognitive processes. However, a growing literature suggests that LLMs’ explanations often fail to faithfully capture their true reasoning processes (Fayyaz et al., 2024; Malin et al., 2025; Matton et al., 2025).

In the study presented here, we sought to extend this literature by assessing the degree to which LLMs could faithfully explain their reasoning processes in the context of multi-attribute choice. To do so, we tasked both humans and LLMs with making a series of choices between homes that varied on six attributes, then self-reporting the weight they placed on each attribute. We then assessed the alignment between their self-reported weights, their choices, and weights estimated from their choices, all of which are common metrics for evaluating decision makers’ metacognitive knowledge of their own decision-making processes (Cash & Oppenheimer, 2025; Harte & Koele, 1995; Morris et al., 2025; Slovic & Lichtenstein, 1971). We found that ChatGPT4o – which was released in 2024 – provided stated weights that were significantly less faithful to their revealed weights than did human participants. However, three LLMs released in 2025 – ChatGPT5, Gemini 2.5-Flash, and Sonnet 4 – generated weights that were consistently as accurate or more accurate than human decision makers’ self-reported weights.

This pattern of results suggests that LLMs are capable of faithfully describing the weight they placed on various factors when making multi-attribute choice decisions – at least at a level

that is comparable to that of humans. However, this ability seems to be a relatively recent development, as the models released in 2025 were significantly more faithful than ChatGPT4o, which was released in 2024. One potential explanation for this improvement is that the newer models have larger context windows – meaning that they can remember a greater deal of information from earlier in the conversation before it falls out of their memory (Bansal et al., 2024; T. Wu et al., 2023). However, the scripts used in this study were rather short and thus should not have surpassed the limits of any of the models. As such, we find this explanation to be unlikely.

A potential explanation that is supported by our data, however, is that the newer models seem to be more consistent in their choice behavior. To quantify this, we calculated the mean standard deviation of RAWs for each model across attributes. The mean standard deviation for GPT4o was 13.79pp, whereas the means for Gemini (4.61pp), Sonnet (2.78pp), and GPT5 (6.07pp) were less than half of that. Despite these differences in choice consistency, all four LLMs had minimal variance in their stated weights. The mean standard deviation in SAWs was 4.30 for GPT4o, 3.93 for Gemini, 3.45 for Sonnet, and 3.44 for GPT5. This pattern suggests that each model self-reported weights that implied relatively consistent choice behavior, but GPT4o did not make choices with the same level of consistency as the other models. As a result, GPT4o did not report stated weights that accurately captured the variance in its choice behavior, reflecting inaccurate metacognitive knowledge. This could also explain GPT4o’s high degree of confidence, as it did not seem to be aware of how inconsistent its choice behavior was. However, this explanation is entirely post-hoc and requires further investigation.

Interestingly, exploratory analyses also indicated that, compared to GPT4o, the modern LLMs made choices that were more similar to those made by humans. GPT4o’s attribute-level

mean RAWs differed from Buyers' mean RAWs by an average absolute value (across attributes) of 10.75pp and Proxies' mean RAWs by an average of 10.69pp. The mean differences for the newer LLMs were smaller, ranging from 2.97pp (GPT5) to 6.14pp (Sonnet) when compared to Buyers and 2.14pp (GPT5) to 5.66pp (Sonnet) when compared to Proxies. However, all four models were similarly capable of capturing humans' stated weights. Mean SAW differences ranged from 2.37pp (GPT4o) to 5.86pp (Sonnet) between LLMs and Buyers, and 1.73pp (GPT5) to 4.91pp (Sonnet) between LLMs and Proxies. In conjunction, these results suggest that GPT4o had awareness of the choices it was *supposed* to make, but, unlike the newer models, could not apply this knowledge to its choice behavior. This could suggest that the newer models more effectively engage in metacognitive control (Ackerman & Thompson, 2017) and is an important area for further research, especially given that alignment between user preferences and algorithmic outputs is a critical predictor of algorithmic adoption (Dietvorst & Bartels, 2022; Dietvorst et al., 2018).

To further investigate similarities in the humans' and LLMs' choice behavior and metacognitive knowledge thereof, we calculated the range of RAWs, SAWs, and AIRs (normalized to a 0-100 scale) for each participant, then averaged within each sample. The most notable pattern was that the range for AIRs was smaller than the range for RAWs for all samples ($p < .001$), which is consistent with previous literature demonstrating that self-reported weights tend to be more evenly distributed than revealed weights (Newell & Shanks, 2014). This suggests that LLMs may be susceptible to some of the same self-report biases as humans, but this result is highly exploratory and requires further investigation. In contrast, SAW ranges were not consistently lower than RAW ranges. Details are reported in the Supplemental Materials.

In all, the results presented here make two key contributions to the LLM faithfulness literature. First, we demonstrate that LLMs – especially models from 2025 – can be reasonably faithful in describing their reasoning processes in the domain of multi-attribute choice. One reason that LLMs may be particularly successful in this task compared to other faithfulness tasks is that the paradigm we used collects self-reports in a quantitative format (Cash & Oppenheimer, 2025). Given that LLMs are highly sophisticated quantitative machines, they may be better at providing percentages describing their reasoning formula than describing their reasoning processes verbally. Secondly, we demonstrate that the faithfulness of LLMs’ metacognitive knowledge can improve over time and may even surpass humans’ metacognitive knowledge. This supports the argument that LLMs’ advanced quantitative abilities may allow them to develop sophisticated metacognitive abilities as the companies creating these models continue to iterate (Cappelen & Dever, 2024; Long, 2023).

Given that LLMs are largely believed to lack true introspective abilities (Cash & Oppenheimer, 2024a; Comsa & Shanahan, 2025; Li et al., 2024), these results also suggest that human-like levels of metacognitive knowledge can be achieved without introspection. This provides some support to the psychological argument that human metacognition does not necessarily require introspection, but rather than it could be built on other mechanisms, such as statistical learning (Perkovic & Orquin, 2018; Sherman et al., 2020), a prior causal theories (Nisbett & Wilson, 1977) or inferential cues (Ackerman, 2019; Koriati, 1997). In this sense, research on the metacognitive abilities of LLMs provides an interesting way to investigate how human metacognition would or could function if we lacked introspection or other abilities that cannot be replicated by LLMs. Future research should continue to study the mechanisms underlying the metacognitive knowledge of both humans and LLMs.

Limitations and Future Directions

One limitation of our findings is that they do not provide any insight into the mechanisms by which LLMs generate metacognitive knowledge. For example, it is unclear the extent to which LLMs are learning their weights from their choices versus making choices that align with pre-determined weights versus inferring their weights from cues in the choice environment or some other mechanism entirely. Some of these mechanisms could be tested empirically (e.g., by asking for stated weights before the choice tasks), but others are hidden within the black box of the underlying models (de Carvalho Souza & Weigang, 2025; Ramlochan, 2023). At a lower level, it is also unclear how specific neurons or circuits may contribute to an LLM's ability to develop faithful metacognitive knowledge. This highlights the need for mechanistic interpretability research specifically focused on LLMs' metacognitive abilities (Gantla, 2025; Von Zahn et al., 2025).

An additional avenue for future research would be to test how responsive metacognitive knowledge is to prompt engineering – that is, optimizing prompts to maximize the quality of the LLM's response (Giray, 2023). In our studies, we used prompts that mimicked an empirically-validated paradigm for measuring metacognitive knowledge (Cash & Oppenheimer, 2025), not necessarily maximizing metacognitive knowledge. Future research should investigate the extent to which different prompting strategies, such as chain-of-thought prompting (B. Wang et al., 2022) or persona-based prompting (Hu & Collier, 2024; G. Sun et al., 2024), could improve the faithfulness of LLMs' metacognitive knowledge. On the other hand, prompt engineering is a skill (Federiakin et al., 2024) and beginners may use prompts that limit the accuracy of LLMs' metacognitive knowledge to levels below what we found here. As such, human factors research

should focus on the efficacy of prompts implemented by everyday users, rather than the well-calibrated prompts used in these studies.

Another future direction for this research could be to elicit human and LLM participants' beliefs about their choice behavior using methods other than the two we did here (SAWs and AIRs). For example, one elicitation method that could be promising is swing-weighting, wherein participants self-report the best and worst possible levels of each attribute, then judge how much their evaluation of an alternative would improve if its score on each attribute shifted from worst to best (Edwards & Barron, 1994; von Winterfeldt & Edwards, 1986). The swing-weighting method would be particularly valuable because it forces participants to consider the full range of each attribute, which may eliminate biases in their metacognitive reasoning (Keeney, 1992; Montibeller & von Winterfeldt, 2015). It could be similarly useful to elicit human and LLM participants' beliefs about the extent to which they engaged in any sort of non-compensatory decision strategy (Einhorn, 1970; Morris et al., 2025; Tversky, 1972). The judgment and decision-making literature has a rich history of comparing different elicitation methods (e.g., Borchert et al., 1991; Nesbitt et al., 2024; Pöyhönen & Hämäläinen, 2001; Srivastava et al., 1995) and we believe that replicating these kinds of studies with LLMs could greatly benefit our understanding of LLMs' metacognitive knowledge.

A related future direction could be to elicit human and LLM participants' beliefs about interactions between attributes. For example, a participant may only want a bigger house if it also comes with a bigger yard. While our attributes were selected to minimize interactions – and our data suggests that participants' choices were well-predicted by a simple additive model – it is highly likely that interaction effects will occur in other domains. For example, a decision maker may only want a car they purchase to be cherry red if it's a sports car (who wants a cherry red

minivan?). This would be a natural extension of this work, particularly because choice-based conjoint analysis is built to detect interaction effects (Eggers et al., 2022). These future research directions could kickstart a whole new era of research in which we seek to make sense of what’s going on inside the minds of LLMs and interpret the black box functions by which they transform inputs into outputs, much like the pioneering JDM scholars did with humans.

Constraints on Generality

The results presented here only tested participants’ metacognitive knowledge in the homebuying domain. As such, it is plausible that LLMs’ metacognitive knowledge of their reasoning processes may be more or less accurate in other multi-attribute choice domains, such as romantic partner selection, school choice, or buying a car. Reductions in accuracy may be particularly likely in multi-attribute choice domains that are less well-represented in the models’ training sets. However, the same could be said of humans; if we were dropped on Mars and asked to make choices between Martian homes (which could be made of fungi; Tavares, 2020), our metacognitive knowledge would likely be poor as well. Here on Earth, though, humans’ metacognitive knowledge is relatively consistent across similarly well-understood domains (Cash & Oppenheimer, 2025), so it would be reasonable to hypothesize that the same would be true of LLMs. Future research should test the metacognitive knowledge of LLMs in other multi-attribute choice domains.

The generalizability of our results is also limited by the fact that we only tested four LLMs, with a focus on the most advanced models at the time of the study. There are almost certainly less advanced models that would demonstrate worse metacognitive knowledge than the LLMs tested here, and there very well could be – and almost certainly will be – models that demonstrate more accurate metacognitive knowledge. It would be of particular interest to study

open-source models, such as Mistral (2025) and LLaMA (Meta, 2025), as their transparency may allow deeper investigations into the mechanisms underlying the development and reporting of LLMs' metacognitive knowledge. It would also be of interest to test the performance of these models in languages other than English. While we believe that the models we chose were a good starting point because of their popularity and prowess, we strongly encourage future research into the metacognitive knowledge of other LLMs.

Conclusion

LLMs are often considered to function as black boxes whose inner workings are a mystery to users (de Carvalho Souza & Weigang, 2025). In the study presented here, we investigated the extent to which LLMs could accurately report their reasoning processes in a multi-attribute homebuying task when asked to do so. We found that GPT4o provided significantly less accurate self-reports of the weight it placed on various attributes than human participants did. However, three more-advanced models (Gemini 2.5 Flash, Sonnet 4, GPT5) – released about a year later – matched or surpassed human accuracy on the same task. These results suggest that LLMs can generate and maintain accurate metacognitive knowledge of their own decision-making processes, but that this is a relatively new ability.

These findings make a significant contribution to the LLM faithfulness literature, as they demonstrate an emergent metacognitive capability of LLMs and support the hypothesis that LLMs' metacognitive abilities are likely to grow as models are improved (Cappelen & Dever, 2024; Long, 2023). Our results also have significant implications for users, as they suggest that LLMs can explain their reasoning processes at least as well as humans – thus increasing their value and transparency as decision aids (Yang et al., 2024) and advisors (Wester et al., 2024).

Acknowledgements

The authors would like to recognize the support of four undergraduates – Sara Christie, Mira Devgan, Roslyn Randall, and Anika Mohta – who assisted with running the LLM trials. The authors would also like to thank their wives, Elizabeth and Nicole, for their support and feedback. Lastly, the authors would like to thank the first author’s committee members – Julie Downs, Sharon Carver, and Bonnie Nozari – as well as members of the OppLab for their input over the course of this project.

Competing Interests Statement

Competing interests: The authors declare none.

Financial Support Statement

This work was supported by a Doctoral Dissertation Research Improvement Grant from the National Science Foundation (Award #2333553), an Academic Research Grant from Sawtooth Software, Inc. that provided us with free access to their Lighthouse Studio CBC software, and a postdoctoral fellowship from the Lupina Foundation. No grantor had any role other than financial support. Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the views of the grantors.

Use of Artificial Intelligence

Large Language Models were used as participants in this study. The authors also used ChatGPT (versions 5 – 5.6) to debug analytical code, assist in the location of relevant academic references, and check for inconsistencies in the reference list.

References

- Ackerman, R. (2014). The diminishing criterion model for metacognitive regulation of time investment. *Journal of Experimental Psychology: General*, 143(3), 1349–1368. <https://doi.org/10.1037/a0035098>
- Ackerman, R. (2019). Heuristic cues for meta-reasoning judgments: Review and methodology. *Psihologijske Teme*, 28(1), 1–20. <https://doi.org/10.31820/pt.28.1.1>
- Ackerman, R., & Thompson, V. A. (2017). Meta-Reasoning: Monitoring and control of thinking and reasoning. *Trends in Cognitive Sciences*, 21(8), 607–617. <https://doi.org/10.1016/j.tics.2017.05.004>
- Amaratunga, T. (2023). Threats, opportunities, and misconceptions. In T. Amaratunga (Ed.), *Understanding Large Language Models* (pp. 131–148). Apress. https://doi.org/10.1007/979-8-8688-0017-7_6
- Amir, O., & Levav, J. (2008). Choice construction versus preference construction: The instability of preferences learned in context. *Journal of Marketing Research*, 45(2), 145–158. <https://doi.org/10.1509/jmkr.45.2.145>
- Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2010). Preference heterogeneity in experiments: Comparing the field and laboratory. *Journal of Economic Behavior & Organization*, 73(2), 209–224. <https://doi.org/10.1016/j.jebo.2009.09.006>
- Anthropic. (2025a). *Claude* [Computer software]. <https://claude.ai/>
- Anthropic. (2025b). *Introducing Claude 4*. <https://www.anthropic.com/news/claude-4>
- Bakker, M., Chadwick, M., Sheahan, H., Tessler, M., Campbell-Gillingham, L., Balaguer, J., McAleese, N., Glaese, A., Aslanides, J., Botvinick, M., & Summerfield, C. (2022). Fine-tuning language models to find agreement among humans with diverse preferences. *Proceedings of the International Conference on Neural Information Processing Systems*, 39, 38176–38189. <https://dl.acm.org/doi/10.5555/3600270.3603036>
- Banerjee, D., Singh, P., Avadhanam, A., & Srivastava, S. (2023). *Benchmarking LLM powered Chatbots: Methods and Metrics* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2308.04624>
- Bansal, G., Chamola, V., Hussain, A., Guizani, M., & Niyato, D. (2024). Transforming conversations with AI—A comprehensive study of ChatGPT. *Cognitive Computation*, 16(5), 2487–2510. <https://doi.org/10.1007/s12559-023-10236-2>

- Barman, K. G., Wood, N., & Pawlowski, P. (2024). Beyond transparency and explainability: On the need for adequate and contextualized user guidelines for LLM use. *Ethics and Information Technology*, 26(3), 47. <https://doi.org/10.1007/s10676-024-09778-2>
- Bem, D. J. (1972). Self-perception theory. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 1-62). [https://doi.org/10.1016/S0065-2601\(08\)60024-6](https://doi.org/10.1016/S0065-2601(08)60024-6)
- Benary, M., Wang, X. D., Schmidt, M., Soll, D., Hilfenhaus, G., Nassir, M., Sigler, C., Knödler, M., Keller, U., Beule, D., Keilholz, U., Leser, U., & Rieke, D. T. (2023). Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*, 6(11), e2343689. <https://doi.org/10.1001/jamanetworkopen.2023.43689>
- Benk, M., Wettstein, L., Schlicker, N., Von Wangenheim, F., & Scharowski, N. (2025). Bridging the knowledge gap: Understanding user expectations for trustworthy LLM standards. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39, 27197–27205. <https://doi.org/10.1609/aaai.v39i26.34928>
- Bénon, J., Lee, D., Hopper, W., Verdeil, M., Pessiglione, M., Vinckier, F., Bouret, S., Rouault, M., Lebouc, R., Pezzulo, G., Schreiweis, C., Burguière, E., & Daunizeau, J. (2024). The online metacognitive control of decisions. *Communications Psychology*, 2(1), 23. <https://doi.org/10.1038/s44271-024-00071-y>
- Bereska, L., & Gavves, E. (2024). *Mechanistic interpretability for AI safety—A review* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.2404.14082>
- Bhattacharjee, A., Moraffah, R., Garland, J., & Liu, H. (2023). *Towards LLM-guided causal explainability for black-box text classifiers* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2309.13340>
- Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6), e2218523120. <https://doi.org/10.1073/pnas.2218523120>
- Borcherding, K., Eppel, T., & von Winterfeldt, D. (1991). Comparison of weighting judgments in multiattribute utility measurement. *Management Science*, 37(12), 1603–1619. <https://doi.org/10.1287/mnsc.37.12.1603>
- Boring, E. G. (1953). A history of introspection. *Psychological Bulletin*, 50(3), 169–189. <https://doi.org/10.1037/h0090793>
- Brown, A. (1987). Metacognition, executive control, self-regulation, and other more mysterious mechanisms. In F. E. Weinert & R. Kluwe (Eds.), *Metacognition, motivation, and understanding* (pp. 65–116). Erlbaum.
- Brown, N. B. (2024). *Enhancing trust in LLMs: Algorithms for comparing and interpreting LLMs* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2406.01943>

- Brunswik, E. (1952). *The conceptual framework of psychology*. University of Chicago Press.
- Cappelen, H., & Dever, J. (2024). Introspective machines: Are LLMs better at self-reflection than humans? *Philosophical Perspectives*, 38(1), 189–196.
<https://doi.org/10.1111/phpe.12201>
- Cash, T. N., Kelly, M. O., Macnamara, B. N., & Risko, E. F. (2026). Is AI making us stupid? *Trends in Cognitive Sciences*, 30(8), 673–675. <https://doi.org/10.1016/j.tics.2026.06.004>
- Cash, T. N., & Oppenheimer, D. M. (2024a). Generative chatbots ain't experts: Exploring cognitive and metacognitive limitations that hinder expertise in generative chatbots. *Journal of Applied Research in Memory and Cognition*, 13(4), 490–494.
<https://doi.org/10.1037/mac0000202>
- Cash, T. N., & Oppenheimer, D. M. (2024b). Parental rights or parental wrongs: Parents' metacognitive knowledge of the factors that influence their school choice decisions. *PLOS ONE*, 19(4), e0301768. <https://doi.org/10.1371/journal.pone.0301768>
- Cash, T. N., & Oppenheimer, D. M. (2025). Assessing metacognitive knowledge in subjective decisions: The *knowledge of weights* paradigm. *Thinking & Reasoning*, 31(3), 331–373.
<https://doi.org/10.1080/13546783.2024.2426543>
- Cash, T. N., Oppenheimer, D. M., Christie, S., & Devgan, M. (2026). Quantifying uncertainty: Testing the accuracy of LLMs' confidence judgments. *Memory & Cognition*.
<https://doi.org/10.3758/s13421-025-01755-4>
- Chen, Y., Liu, T. X., Shan, Y., & Zhong, S. (2023). The emergence of economic rationality of GPT. *Proceedings of the National Academy of Sciences*, 120(51), e2316205120.
<https://doi.org/10.1073/pnas.2316205120>
- Comsa, I. M., & Shanahan, M. (2025). *Does it make sense to speak of introspection in Large Language Models?* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2506.05068>
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist*, 34(7), 571–582. <https://doi.org/10.1037/0003-066X.34.7.571>
- de Carvalho Souza, M. E., & Weigang, L. (2025). *Unveiling the black box: The significance of XAI in making LLMs transparent*. <https://doi.org/10.22541/au.173991264.46233479/v1>
- Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Kraymer, L., Candelon, F., & Lakhani, K. R. (2023). *Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality*. SSRN. <https://doi.org/10.2139/ssrn.4573321>
- Deng, Z., Nagel, R., Yan, J., & Zhang, T. (2025). *Can LLM agents recognize demographic heterogeneity in economic games?* SSRN. <https://doi.org/10.2139/ssrn.5197303>

- Descartes, R. (1641). *Meditations on first philosophy*. Hackett Publishing Company.
- Dietvorst, B. J., & Bartels, D. M. (2022). Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology*, 32(3), 406–424. <https://doi.org/10.1002/jcpy.1266>
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science*, 64(3), 1155–1170. <https://doi.org/10.1287/mnsc.2016.2643>
- Dunlosky, J., & Metcalfe, J. (2009). *Metacognition*. SAGE.
- Edwards, W., & Barron, F. H. (1994). SMARTS and SMARTER: Improved simple methods for multiattribute utility measurement. *Organizational Behavior and Human Decision Processes*, 60(3), 306–325. <https://doi.org/10.1006/obhd.1994.1087>
- Eggers, F., Sattler, H., Teichert, T., & Völckner, F. (2022). Choice-based conjoint analysis. In C. Homburg, M. Klarmann, & A. Vomberg (Eds.), *Handbook of Market Research* (pp. 781–819). Springer International Publishing. https://doi.org/10.1007/978-3-319-57413-4_23
- Einhorn, H. J. (1970). The use of nonlinear, noncompensatory models in decision making. *Psychological Bulletin*, 73(3), 221–230. <https://doi.org/10.1037/h0028695>
- Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, 14, 1199058. <https://doi.org/10.3389/fpsyg.2023.1199058>
- Engelbert, M., & Carruthers, P. (2010). Introspection. *WIREs Cognitive Science*, 1(2), 245–253. <https://doi.org/10.1002/wcs.4>
- Fayyaz, M., Yin, F., Sun, J., & Peng, N. (2024). *Evaluating human alignment and model faithfulness of LLM rationale* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2407.00219>
- Federiakin, D., Molerov, D., Zlatkin-Troitschanskaia, O., & Maur, A. (2024). Prompt engineering as a new 21st century skill. *Frontiers in Education*, 9, 1366434. <https://doi.org/10.3389/educ.2024.1366434>
- Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American Psychologist*, 34(10), 906–911. <https://doi.org/10.1037/0003-066X.34.10.906>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8(JULY). <https://doi.org/10.3389/fnhum.2014.00443>
- Gantla, S. R. (2025). Exploring mechanistic interpretability in Large Language Models: Challenges, approaches, and insights. *2025 International Conference on Data Science*,

- Agents & Artificial Intelligence (ICDSAAI)*, 1–8.
<https://doi.org/10.1109/ICDSAAI65575.2025.11011640>
- Gat, Y., Calderon, N., Feder, A., Chapanin, A., Sharma, A., & Reichart, R. (2023). *Faithful explanations of black-box NLP models using LLM-generated counterfactuals* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2310.00603>
- German, H., Fortin, M., & Read, D. (2016). Justice judgments: Individual self-insight and between- and within-person consistency. *Academy of Management Discoveries*, 2(1), 33–50. <https://doi.org/10.5465/amd.2014.0031>
- Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, 103(4), 650–669.
<https://doi.org/10.1037/0033-295X.103.4.650>
- Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering*, 51(12), 2629–2633. <https://doi.org/10.1007/s10439-023-03272-4>
- Goli, A., & Singh, A. (2024). Frontiers: Can large language models capture human preferences? *Marketing Science*, 43(4), 709–722. <https://doi.org/10.1287/mksc.2023.0306>
- Google. (2025a). *Gemini* [Computer software]. <https://gemini.google.com/app>
- Google. (2025b). *Gemini 2.5 Flash*. <https://deepmind.google/models/gemini/flash/>
- Grahe, J. E., & Bernieri, F. J. (2002). Self-awareness of judgment policies of rapport. *Personality and Social Psychology Bulletin*, 28(10), 1407–1418.
<https://doi.org/10.1177/014616702236872>
- Groot, T., & Valdenegro-Toro, M. (2024). *Overconfidence is key: Verbalized uncertainty evaluation in large language and vision-language models* (Version 1). arXiv. <http://arxiv.org/abs/2405.02917>
- Gurnee, W., Nanda, N., Pauly, M., Harvey, K., Troitskii, D., & Bertsimas, D. (2023). *Finding neurons in a haystack: Case studies with sparse probing* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2305.01610>
- Hager, P., Jungmann, F., Holland, R., Bhagat, K., Hubrecht, I., Knauer, M., Vielhauer, J., Makowski, M., Braren, R., Kaissis, G., & Rueckert, D. (2024). Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature Medicine*, 30(9), 2613–2622. <https://doi.org/10.1038/s41591-024-03097-1>
- Hammond, K. R. (1955). Probabilistic functioning and the clinical method. *Psychological Review*, 62(4), 255–262. <https://doi.org/10.1037/h0046845>

- Harries, C., Evans, J. St. B. T., & Dennis, I. (2000). Measuring doctors' self-insight into their treatment decisions. *Applied Cognitive Psychology*, 14(5), 455–477. [https://doi.org/10.1002/1099-0720\(200009\)14:5%3C455::AID-ACP667%3E3.0.CO;2-V](https://doi.org/10.1002/1099-0720(200009)14:5%3C455::AID-ACP667%3E3.0.CO;2-V)
- Harte, J. M., & Koele, P. (1995). A comparison of different methods for the elicitation of attribute weights: Structural modeling, process tracing, and self-reports. *Organizational Behavior and Human Decision Processes*, 64(1), 49–64.
- Hein, M., Kurz, P., & Steiner, W. J. (2020). Analyzing the capabilities of the HB logit model for choice-based conjoint analysis: A simulation study. *Journal of Business Economics*, 90(1), 1–36. <https://doi.org/10.1007/s11573-019-00927-4>
- Hinson, J. M., Jameson, T. L., & Whitney, P. (2003). Impulsive decision making and working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(2), 298–306. <https://doi.org/10.1037/0278-7393.29.2.298>
- Hoepfl, R. T., & Huber, G. P. (1970). A study of self-explicated utility models. *Behavioral Science*, 15(5), 408–414. <https://doi.org/10.1002/bs.3830150503>
- Hou, H., Leach, K., & Huang, Y. (2024). ChatGPT giving relationship advice – How reliable is it? *Proceedings of the International AAAI Conference on Web and Social Media*, 18, 610–623. <https://doi.org/10.1609/icwsm.v18i1.31338>
- Hu, T., & Collier, N. (2024). *Quantifying the persona effect in LLM simulations* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2402.10811>
- Iyengar, S. S., & Lepper, M. R. (2000). When choice is demotivating: Can one desire too much of a good thing? *Journal of Personality and Social Psychology*, 79(6), 995–1006. <https://doi.org/10.1037/0022-3514.79.6.995>
- Jacovi, A., & Goldberg, Y. (2020). Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4198–4205. <https://doi.org/10.18653/v1/2020.acl-main.386>
- Jaunet, T., Kervadec, C., Vuillemot, R., Antipov, G., Baccouche, M., & Wolf, C. (2022). VisQA: X-raying vision and language reasoning in transformers. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 976–986. <https://doi.org/10.1109/TVCG.2021.3114683>
- Kant, I. (1781). *The critique of pure reason*. Cambridge University Press.
- Karelaia, N., & Hogarth, R. M. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, 134(3), 404–426. <https://doi.org/10.1037/0033-2909.134.3.404>

- Keeney, R. L. (1992). *Value-focused thinking: A path to creative decision making*. Harvard University Press. <https://doi.org/10.2307/j.ctv322v4g7>
- Kim, J., Kovach, M., Lee, K.-M., Shin, E., & Tzavellas, H. (2024). *Learning to be Homo Economicus: Can an LLM learn preferences from choice* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2401.07345>
- Kim, Y., Lee, J., Kim, S., Park, J., & Kim, J. (2024). Understanding users' dissatisfaction with ChatGPT responses: Types, resolving tactics, and the effect of knowledge level. *Proceedings of the 29th International Conference on Intelligent User Interfaces*, 385–404. <https://doi.org/10.1145/3640543.3645148>
- Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., & Hale, S. A. (2025). Why human–AI relationships need socioaffective alignment. *Humanities and Social Sciences Communications*, 12(1), 728. <https://doi.org/10.1057/s41599-025-04532-5>
- Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General*, 126(4), 349–370. <https://doi.org/10.1037/0096-3445.126.4.349>
- Krajibich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, 29, 6–11. <https://doi.org/10.1016/j.copsyc.2018.10.008>
- Lakkaraju, K., Jones, S. E., Vuruma, S. K. R., Pallagani, V., Muppasani, B. C., & Srivastava, B. (2023). LLMs for financial Advisement: A fairness and efficacy study in personal decision making. *Proceedings of the 4th ACM International Conference on AI in Finance*, 100–107. <https://doi.org/10.1145/3604237.3626867>
- Li, Y., Huang, Y., Lin, Y., Wu, S., Wan, Y., & Sun, L. (2024). *I think, therefore I am: Benchmarking awareness of large language models using AwareBench*. arXiv. <http://arxiv.org/abs/2401.17882>
- Lieder, F., & Griffiths, T. L. (2019). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43, Article e1. <https://doi.org/10.1017/S0140525X1900061X>
- Lin, S., Hilton, J., & Evans, O. (2022). *Teaching models to express their uncertainty in words* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2205.14334>
- Lissak, S., Calderon, N., Shenkman, G., Ophir, Y., Fruchter, E., Klomek, A. B., & Reichart, R. (2024). *The colorful future of LLMs: Evaluating and improving LLMs as emotional supporters for queer youth* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2402.11886>
- Long, R. (2023). Introspective capabilities in large language models. *Journal of Consciousness Studies*, 30(9–10), 143–153. <https://doi.org/10.53765/20512201.30.9.143>

- Lyu, Q., Apidianaki, M., & Callison-Burch, C. (2024). Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2), 657–723. https://doi.org/10.1162/coli_a_00511
- Madsen, A., Chandar, S., & Reddy, S. (2024). *Are self-explanations from Large Language Models faithful?* (Version 4). arXiv. <https://doi.org/10.48550/ARXIV.2401.07927>
- Malin, B., Kalganova, T., & Boulgouris, N. (2025). A review of faithfulness metrics for hallucination assessment in Large Language Models. *IEEE Journal of Selected Topics in Signal Processing*, 1–13. <https://doi.org/10.1109/JSTSP.2025.3579203>
- Maltoni, D., & Ferrara, M. (2024). Arithmetic with language models: From memorization to computation. *Neural Networks*, 179, 106550. <https://doi.org/10.1016/j.neunet.2024.106550>
- Matton, K., Ness, R. O., Gutttag, J., & Kıcıman, E. (2025). *Walk the talk? Measuring the faithfulness of large language model explanations* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2504.14150>
- Meta. (2025). *Llama 4* [Computer software]. <https://www.llama.com/>
- Mishra, A., Nayak, G., Bhattacharya, S., Kumar, T., Shah, A., & Foltin, M. (2024). LLM-guided counterfactual data generation for fairer AI. *Companion Proceedings of the ACM Web Conference 2024*, 1538–1545. <https://doi.org/10.1145/3589335.3651929>
- Mistral. (2025). *Mistral* [Computer software]. <https://mistral.ai/>
- Montibeller, G., & von Winterfeldt, D. (2015). Cognitive and motivational biases in decision and risk analysis. *Risk Analysis*, 35(7), 1230–1251. <https://doi.org/10.1111/risa.12360>
- Morris, A., Carlson, R. W., Kober, H., & Crockett, M. (2025). Introspective access to value-based multi-attribute choice processes. *Nature Communications*, 16, 3733. <https://doi.org/10.1038/s41467-025-59080-y>
- Mutanga, M. B., Msane, J., Mndaweni, T. N., Hlongwane, B. B., & Ngcobo, N. Z. (2025). Exploring the impact of LLM prompting on students' learning. *Trends in Higher Education*, 4(3), 31. <https://doi.org/10.3390/higheredu4030031>
- Nesbitt, L., Meitner, M. J., Chamberlain, B., Gonzalez, J., & Trousdale, W. (2024). A comparison of value-weight-elicitation methods for accurate and accessible participatory planning. *Journal of Planning Education and Research*, 44(4), 2098–2109. <https://doi.org/10.1177/0739456X231155069>
- Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best." *Organizational Behavior and Human Decision Processes*, 91(1), 82–96. [https://doi.org/10.1016/S0749-5978\(02\)00525-3](https://doi.org/10.1016/S0749-5978(02)00525-3)

- Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84(3), 231–259. <https://doi.org/10.1037/0033-295X.84.3.231>
- OpenAI. (2023). *Language models can explain neurons in language models*. <https://openai.com/index/language-models-can-explain-neurons-in-language-models/>
- OpenAI. (2025a). *ChatGPT* [Computer software]. <https://chat.openai.com/>
- OpenAI. (2025b). *Introducing GPT-5*. <https://openai.com/index/introducing-gpt-5/>
- Oppenheimer, D. M., Cash, T. N., & Connell Pensky, A. E. (2025). You’ve Got AI Friend in Me: LLMs as Collaborative Learning Partners. *International Journal of Artificial Intelligence in Education*, advanced online publication. <https://doi.org/10.1007/s40593-025-00521-6>
- Orme, B. (2002). Interpreting conjoint analysis data. *Sawtooth Software, Inc.* https://leeds-faculty.colorado.edu/ysun/MKTG4825_files/Interpreting_CA_Data.pdf
- Orquin, J. L., Lahm, E. S., & Stojić, H. (2021). The visual environment and attention in decision making. *Psychological Bulletin*, 147(6), 597–617. <https://doi.org/10.1037/bul0000328>
- Parcalabescu, L., & Frank, A. (2023). *On measuring faithfulness or self-consistency of natural language explanations* (Version 4). arXiv. <https://doi.org/10.48550/ARXIV.2311.07466>
- Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. Cambridge University Press.
- Payne, J. W., Bettman, J. R., & Schkade, D. A. (1999). Measuring constructed preferences: Towards a building code. *Journal of Risk and Uncertainty*, 19(1/3), 243–270. <https://doi.org/10.1023/A:1007843931054>
- Perkovic, S., & Orquin, J. L. (2018). Implicit statistical learning in real-world environments leads to ecologically rational decision making. *Psychological Science*, 29(1), 34–44. <https://doi.org/10.1177/0956797617733831>
- Pesch, P. J. (2025). Potentials and challenges of large language models (LLMs) in the context of administrative decision-making. *European Journal of Risk Regulation*, 16(1), 76–95. <https://doi.org/10.1017/err.2024.99>
- Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, 17(5), 407–413. <https://doi.org/10.1111/j.1467-9280.2006.01720.x>
- Phillips-Wren, G., & Adya, M. (2020). Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty. *Journal of Decision Systems*, 29(sup1), 213–225. <https://doi.org/10.1080/12460125.2020.1768680>

- Pöyhönen, M., & Hämäläinen, R. P. (2001). On the convergence of multiattribute weighting methods. *European Journal of Operational Research*, 129(3), 569–585.
[https://doi.org/10.1016/S0377-2217\(99\)00467-1](https://doi.org/10.1016/S0377-2217(99)00467-1)
- Pronin, E. (2009). The introspection illusion. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1–68). Elsevier Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)00401-2](https://doi.org/10.1016/S0065-2601(08)00401-2)
- Rai, D., Zhou, Y., Feng, S., Saparov, A., & Yao, Z. (2024). *A practical review of mechanistic interpretability for transformer-based language models* (Version 4). arXiv.
<https://doi.org/10.48550/ARXIV.2407.02646>
- Rakow, T. (2014). Self-insight research as (double) model recovery. *Behavioral and Brain Sciences*, 37(1), 37–38. <https://doi.org/10.1017/S0140525X13000824>
- Ramlochan, S. (2023). The black box problem: Opaque inner workings of large language models. *Prompt Engineering & AI Institute*. <https://promptengineering.org/the-black-box-problem-opaque-inner-workings-of-large-language-models/>
- Reilly, B. A., & Doherty, M. E. (1989). A note on the assessment of self-insight in judgment research. *Organizational Behavior and Human Decision Processes*, 44(1), 123–131.
[https://doi.org/10.1016/0749-5978\(89\)90038-1](https://doi.org/10.1016/0749-5978(89)90038-1)
- Reilly, B. A., & Doherty, M. E. (1992). The assessment of self-insight in judgment policies. *Organizational Behavior and Human Decision Processes*, 53(3), 285–309.
[https://doi.org/10.1016/0749-5978\(92\)90067-H](https://doi.org/10.1016/0749-5978(92)90067-H)
- Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, 135(6), 943–973. <https://doi.org/10.1037/a0017327>
- Roose, K. (2023, March 8). How does ChatGPT really work? *The New York Times*.
<https://www.nytimes.com/2023/03/28/technology/ai-chatbots-chatgpt-bing-bard-llm.html>
- Rozado, D. (2024). The political preferences of LLMs. *PLOS ONE*, 19(7), e0306621.
<https://doi.org/10.1371/journal.pone.0306621>
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
<https://doi.org/10.1038/s42256-019-0048-x>
- Sarkar, P., Prakash, A. V., & Singh, J. B. (2025). Explaining LLM decisions: Counterfactual chain-of-thought approach. In S. Srinivasa, S. Saragur, & S. Malapaka (Eds.), *Advanced Computing and Communications: Responsible AI* (Vol. 2461, pp. 254–266). Springer.
https://doi.org/10.1007/978-3-031-96473-2_18

- Sawtooth Software, Inc. (2021). *CBC/HB system technical paper V5.6: The CBC/HB system for Hierarchical Bayes Estimation*. <https://sawtoothsoftware.com/resources/technical-papers/cbc-hb-technical-paper>
- Sawtooth Software, Inc. (2025). *Lighthouse Studio* (Version 9.16.14) [Computer software]. <https://sawtoothsoftware.com/lighthouse-studio>
- Sayeddi, Md. F. A., Bin Hossain, M., Hassan, Md. K., Afrin, S., Hossain, M. Md. S., & Hossain, Md. S. (2025). JailbreakTracer: Explainable detection of jailbreaking prompts in LLMs using synthetic data generation. *IEEE Access*, 13, 123708–123723. <https://doi.org/10.1109/ACCESS.2025.3579996>
- Schwitzgebel, E. (2024). Introspection. *Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/Entries/introspection/>
- Sherman, B. E., Graves, K. N., & Turk-Browne, N. B. (2020). The prevalence and importance of statistical learning in human cognition and behavior. *Current Opinion in Behavioral Sciences*, 32, 15–20. <https://doi.org/10.1016/j.cobeha.2020.01.015>
- Siegel, N. Y., Heess, N., Perez-Ortiz, M., & Camburu, O.-M. (2025). *Verbosity tradeoffs and the impact of scale on the faithfulness of LLM self-explanations* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2503.13445>
- Simon, D., Krawczyk, D. C., Bleicher, A., & Holyoak, K. J. (2008). The transience of constructed preferences. *Journal of Behavioral Decision Making*, 21(1), 1–14. <https://doi.org/10.1002/bdm.575>
- Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, 69(1), 99–118. <https://doi.org/10.2307/1884852>
- Slovic, P. (1969). Analyzing the expert judge: A descriptive study of a stockbroker's decision process. *Journal of Applied Psychology*, 53(4), 255–263. <https://doi.org/10.1037/h0027773>
- Slovic, P., Fleissner, D., & Bauman, W. S. (1972). Analyzing the use of information in investment decision making: A methodological proposal. *The Journal of Business*, 45(2), 283. <https://doi.org/10.1086/295448>
- Slovic, P., & Lichtenstein, S. (1971). Comparison of Bayesian and regression approaches to the study of information processing in judgment. *Organizational Behavior and Human Performance*, 6(6), 649–744. [https://doi.org/10.1016/0030-5073\(71\)90033-X](https://doi.org/10.1016/0030-5073(71)90033-X)
- Smithies, D., & Stoljar, D. (Eds.). (2012). *Introspection and consciousness*. Oxford University Press.

- Sovrano, F. (2025). *Can global XAI methods reveal injected bias in LLMs? SHAP vs rule extraction vs RuleSHAP* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2505.11189>
- Srivastava, J., Connolly, T., & Beach, L. R. (1995). Do ranks suffice? A comparison of alternative weighting approaches in value elicitation. *Organizational Behavior and Human Decision Processes*, 63(1), 112–116. <https://doi.org/10.1006/obhd.1995.1066>
- Stevens, S. S. (1956). The direct estimation of sensory magnitudes: Loudness. *The American Journal of Psychology*, 69(1), 1-25. <https://doi.org/10.2307/1418112>
- Steyvers, M., & Peters, M. A. K. (2025). *Metacognition and uncertainty communication in humans and Large Language Models* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2504.14045>
- Steyvers, M., Tejada, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 7(2), 221–231. <https://doi.org/10.1038/s42256-024-00976-7>
- Sun, G., Zhan, X., & Such, J. (2024). Building better AI agents: A provocation on the utilisation of Persona in LLM-based Conversational Agents. *ACM Conversational User Interfaces 2024*, 1–6. <https://doi.org/10.1145/3640794.3665887>
- Sun, S., Li, Z. A., Foo, M.-D., Zhou, J., & Lu, J. G. (2025). How and for whom using generative AI affects creativity: A field experiment. *Journal of Applied Psychology*. <https://doi.org/10.1037/apl0001296>
- Suri, G., Slater, L. R., Ziaee, A., & Nguyen, M. (2024). Do large language models show decision heuristics similar to humans? A case study using GPT-3.5. *Journal of Experimental Psychology: General*, 153(4), 1066–1075. <https://doi.org/10.1037/xge0001547>
- Tavares, F. (2020). *Could future homes on the moon and Mars be made of fungi?* National Aeronautics and Space Administration. <https://www.nasa.gov/centers-and-facilities/ames/could-future-homes-on-the-moon-and-mars-be-made-of-fungi/>
- Thompson, V. A., Turner, J. A. P., Pennycook, G., Ball, L. J., Brack, H., Ophir, Y., & Ackerman, R. (2013). The role of answer fluency and perceptual fluency as metacognitive cues for initiating analytic thinking. *Cognition*, 128(2), 237–251. <https://doi.org/10.1016/j.cognition.2012.09.012>
- Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 74952–74965. <https://dl.acm.org/doi/10.5555/3666122.3669397>
- Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, 79(4), 281–299. <https://doi.org/10.1037/h0032955>

- Van Knippenberg, D., Dahlander, L., Haas, M. R., & George, G. (2015). Information, attention, and decision making. *Academy of Management Journal*, 58(3), 649–657. <https://doi.org/10.5465/amj.2015.4003>
- Vermersch, P. (2009). Describing the practice of introspection. *Journal of Consciousness Studies*, 16(10–12), 20–57.
- von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge University Press.
- Von Zahn, M., Liebich, L., Jussupow, E., Hinz, O., & Bauer, K. (2025). *Knowing (not) to know: XAI and human metacognition*. SSRN. <https://doi.org/10.2139/ssrn.5383106>
- Wang, B., Min, S., Deng, X., Shen, J., Wu, Y., Zettlemoyer, L., & Sun, H. (2022). *Towards understanding chain-of-thought prompting: An empirical study of what matters* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2212.10001>
- Wang, Y., Fang, X., Xu, Z., Li, J., & Wang, L. (2025). Exploring the impact of explainability in large language model (LLM) applications on user experience. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–8. <https://doi.org/10.1145/3706599.3719941>
- Wang, Z., Wang, W., Chen, Q., Wang, Q., & Nguyen, A. (2024). Generating valid and natural adversarial examples with large language models. *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, 1716–1721. <https://doi.org/10.1109/CSCWD61410.2024.10580402>
- Weisz, J. D., Kumar, S. V., Muller, M., Browne, K.-E., Goldberg, A., Heintze, K. E., & Bajpai, S. (2025). Examining the use and impact of an AI code assistant on developer productivity and experience in the enterprise. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–13. <https://doi.org/10.1145/3706599.3706670>
- Wester, J., De Jong, S., Pohl, H., & Van Berkel, N. (2024). Exploring people’s perceptions of LLM-generated advice. *Computers in Human Behavior: Artificial Humans*, 2(2), 100072. <https://doi.org/10.1016/j.chbah.2024.100072>
- Wilson, T. D., & Nisbett, R. E. (1978). The accuracy of verbal reports about the effects of stimuli on evaluations and behavior. *Social Psychology*, 41(2), 118. <https://doi.org/10.2307/3033572>
- Wu, J., Yang, S., Zhan, R., Yuan, Y., Chao, L. S., & Wong, D. F. (2025). A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1), 275–338. https://doi.org/10.1162/coli_a_00549

- Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5), 1122–1136. <https://doi.org/10.1109/JAS.2023.123618>
- Wu, X., Zhao, H., Zhu, Y., Shi, Y., Yang, F., Hu, L., Liu, T., Zhai, X., Yao, W., Li, J., Du, M., & Liu, N. (2024). *Usable XAI: 10 strategies towards exploiting explainability in the LLM era* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2403.08946>
- Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., & Hooi, B. (2023). *Can LLMs express their uncertainty? An empirical evaluation of confidence elicitation in LLMs* (Version 2). arXiv. <https://doi.org/10.48550/ARXIV.2306.13063>
- Yang, J. C., Dailisan, D., Korecki, M., Hausladen, C. I., & Helbing, D. (2024). LLM voting: Human choices and AI collective decision-making. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7, 1696–1708. <https://doi.org/10.1609/aies.v7i1.31758>
- Yeh, C., Chen, Y., Wu, A., Chen, C., Viégas, F., & Wattenberg, M. (2023). AttentionViz: A global view of transformer attention. *IEEE Transactions on Visualization and Computer Graphics*, 1–11. <https://doi.org/10.1109/TVCG.2023.3327163>
- Yeo, W. J., Satapathy, R., Goh, R., & Cambria, E. (2024). How interpretable are reasoning explanations from prompting large language models? *Findings of the Association for Computational Linguistics: NAACL 2024*, 2148–2164. <https://doi.org/10.18653/v1/2024.findings-naacl.138>
- You, Y., Chen, H. W., & Zhao, J. (2025). Enhancing AI explainability for non-technical users with LLM-driven narrative gamification. *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, 1–7. <https://doi.org/10.1145/3706599.3719795>
- Yu, Y., Wu, M., Lin, Y., & Lobczowski, N. G. (2025). *THiNK: Can large language models think-aloud?* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2505.20184>
- Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D., & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1–38. <https://doi.org/10.1145/3639372>