LLM Agents Can Be Choice-Supportive Biased Evaluators: An Empirical Study

Nan Zhuang[1*], Boyu Cao[2*], Yi Yang[2*], Jing Xu[3*], Mingda Xu[2*], Yuxiao Wang[2], Qi Liu[†2]

[1]School of Software Technology, Zhejiang University, Hangzhou, China
[2]School of Future Technology, South China University of Technology, Guangzhou, China
[3]Faculty of Humanities and Arts, Macau University of Science and Technology, Macau

Abstract

With Large Language Model (LLM) agents taking on more evaluation responsibilities in decision-making, it is essential to recognize their possible biases to guarantee fair and trustworthy AI-supported decisions. This study is the first to thoroughly examine the choice-supportive bias in LLM agents, a cognitive bias that is known to impact human decision-making and evaluation. We conduct experiments across 19 open/unopen-source LLM models in five scenarios at maximum, employing both memory-based and evaluation-based tasks adapted and redesigned from human cognitive studies. Our findings show that LLM agents may exhibit biased attribution or evaluation that supports their initial choices, and such bias may persist even if contextual hallucination is not observable. Key findings show that bias manifestation can differ greatly depending on prompt construction and context preservation, and the bias may be mitigated in larger models. Significantly, we observe that the bias increases when the agents perceive they are in control. Our extensive study involving 284 well-educated humans shows that, despite bias, certain LLM agents can still perform better than humans in similar evaluation tasks. This research contributes to the growing area of AI psychology, and the findings underscore the importance of addressing cognitive biases in LLM Agent systems, with wide-ranging implications spanning from improving AI-assisted decision-making to advancing AI safety and ethics.

Extended version & Appendix — https://t.cn/A6mg7Xel

Introduction

As Large Language Model (LLM) Agents increasingly assume critical roles in complex decision-making processes, from iterative medical diagnosis support (Kaur et al. 2024; Gebreab et al. 2024) to multi-stage financial risk assessments (Park 2024), understanding their potential biases becomes crucial for ensuring fair and reliable AI-assisted decisions. Although there has been a significant amount of research on demographic and social biases in AI systems (Papakyriakopoulos et al. 2020; Buolamwini and Gebru 2018), and some studies have started to look into cognitive biases in LLMs (Echterhoff et al. 2024; Koo et al. 2023), the possibility of choice-supportive bias in LLM Agents has not been thoroughly investigated.

Choice-supportive bias, a well-documented phenomenon in human cognition where individuals tend to attribute more positive features to options they have chosen (Lind et al. 2017; Kafaee, Marhamati, and Gharibzadeh 2021), could impact the objectivity of LLM agent-based evaluations. This bias is particularly concerning in situations involving multi-agent systems or iterative decision processes, where its effects could potentially compound and lead to drastically skewed assessments. For instance, in a chain of LLM Agents collaborating on a complex task, each agent's choice-supportive bias could theoretically amplify through the decision chain, possibly resulting in severely distorted final outputs. The bias also poses risks in decision-making contexts, as agents may provide persuasive explanations for their choices even when the underlying analysis is flawed. It is essential to address these pervasive biases to develop robust and trustworthy AI systems that can enhance human decision-making capabilities without unintentionally amplifying cognitive distortions as LLM Agents continue to evolve and take on more sophisticated roles in decision support systems (Ye et al. 2024; Li et al. 2024).

This paper aims to explore this bias, examining its prevalence, manifestations, and possible influencing factors. To do this, we derive our approach from human memory experiments from (Henkel and Mather 2007), employing the two-alternative forced choice (2AFC) paradigm (Fechner 1860) to test the presence of such bias on LLM agents. From the experiments, we find that LLM agents can produce skewed facts related to their choice - by assigning positive traits to the choice and negative traits to the other options, providing evidence of this bias. Furthermore, since the traditional human experiments depend on benchmarking the memory -

[*]These authors contributed equally. All equal-contributing authors worked jointly from the first day until the last day, sharing equal workload and contributions. N. Zhuang, B. Cao, M. Xu were responsible for designing and conducting the experiments, Y. Yang was responsible for coding, J. Xu performed the human study, Y. Wang and Q. Liu were responsible for reviewing and supervising.

[†]Corresponding author: Qi Liu (drliuqi@scut.edu.cn).

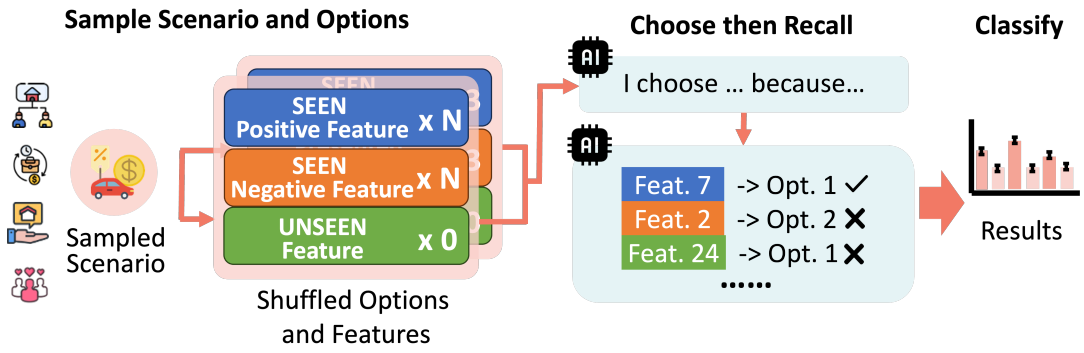

Figure 1: The overview for the memory-based experiment for testing the Choice-supportive bias for LLM agents.

which aligns with the input context for an agent - the choice-supportive bias for LLM agents may seem mitigated as they can stick to the facts in the context better, reducing contextual hallucination (Chuang et al. 2024; Ainsworth, Wycliffe, and Winslow 2024). To examine the choice-supportive bias without relying on memory, we utilize established cognitive research findings and adjust the experiment to prioritize explanation over memorization. We conduct experiments using LLM agents and reveal that such bias may exist widely in LLMs and various situations. Additionally, very slight changes such as the inclusion of "to you" in the prompt could greatly impact the bias. Lastly, we conduct this experiment with a large group of well-educated human participants (n = 284) to compare, and the results show that some agents can outperform humans as evaluators in our experimental settings.

In summary, we contribute to the LLM agent community with interdisciplinary research in psychology, and our core contributions to LLM agent's choice-supportive bias are as follows:

- The first demonstration for LLM agents can manifest memory-driven choice-supportive bias, in the form of contextual hallucination (Henkel and Mather 2007);
- A novel evaluation-based test method that designated to assess the LLM agent's choice-supportive bias, without relying on the standard forgetting process used in human experiments;
- Findings and suggestions on LLM agent's behavior regarding choice-supportive bias, include: 1) choice-supportive bias is pervasive with open/closed source LLM agents, and this bias persists across scenarios; 2) Even if contextual hallucination mitigates, the bias may still exist; 3) Perceived controls in prompts may amplify bias;
- The initial comparison of LLM agents and human advanced degree holders regarding choice-supportive bias indicates that some LLM agents may be superior evaluators in the context of this bias.

Related Work

Choice-supportive Bias in Human Choice-supportive bias is a cognitive phenomenon where individuals tend to exaggerate the benefits of their choices, and it can manifest as false memory (Mather and Johnson 2000; Mather, Shafir, and Johnson 2000), misattributing positive features to chosen options and negative features to foregone alternatives (Mather, Shafir, and Johnson 2003; Henkel and Mather 2007). Unlike more overt biases like in gender(Ruiz-Cantero et al. 2007; Hamberg 2008) and races (Smith-McLallen et al. 2006), it affects memory and perception of previous decisions (Mather and Johnson 2000). This bias may result in overly favorable assessments of previous decisions, possibly strengthening inefficient decision-making habits (Svenson, Salo, and Lindholm 2009). In professional contexts, it could result in the persistence of escalated commitments, poor decisions, and a decline in organizational performance (Zorn et al. 2020). While extensively studied in humans, its potential presence in LLM agent systems remains unexplored. This study is believed to be the first thorough examination of choice-supportive bias in LLM agents, using experimental paradigms from key studies in human cognitive psychology (Mather and Johnson 2000; DeKay et al. 2014).

LLM Agents in Evaluations and Decision-making LLM agents have rapidly emerged as powerful tools in various evaluation situations. Their applications span multiple domains, from assessing machine translation quality (Lu et al. 2023) and summarization effectiveness (Gao et al. 2023) to evaluating code generation (Zhuo 2023) and factual consistency (Cohen et al. 2023). Beyond simple evaluations, LLM agents have expanded into complex decision-making applications. They are now employed in sequential decision-making tasks (Shinn, Labash, and Gopinath 2023), financial modeling (Gao et al. 2024), game-playing (Ma et al. 2023), and even human-AI collaborative decision support systems (Zheng et al. 2023). This integration into high-stakes decision processes amplifies the importance of identifying and mitigating potential biases in these LLM agents. Furthermore, studies have shown that LLMs may fail to give correct solutions when misleading information was introduced (Wang, Yue, and Sun 2023), raising concerns about their reliability. To enhance LLM agents' decision-making capabil-

ities, frameworks such as Chain-of-Thought prompting (Shi et al. 2024), ToT (Yao et al. 2024), BoT (Yang et al. 2024), and ReAct (Yao et al. 2022) have been developed. These approaches enable advanced logical thinking, but they also bring about new challenges in how biases could be revealed through agent interactions. Our study aims to explore the choice-supportive bias present in LLM agents as they take on more critical decision-making responsibilities. To understand how LLM agents manifest choice-supportive bias, we adapt the experiments from psychology research (Henkel and Mather 2007; Lind et al. 2017), and turn them into chained agent tasks. Also, we adapt tested scenarios from previous human behavioral research (Zorn et al. 2020) to serve as the background for some of the prompt used in our experiment.

Methods

Memory-based Experiments

Our study adapts the memory-based experiments (Henkel and Mather 2007) to investigate choice-supportive bias (Lind et al. 2017) in LLM agents. We replicate the human experimental setting to enable direct comparison with human data.

The procedure consists of two main parts: a choice task and a recall task. The choice task employs a Two-Alternative Forced Choice (2AFC) paradigm (Bogacz et al. 2006), where agents are forced to select a total of five options from each of five pairs of options, based on ten features per option. The features can be either positive, such as "very comfortable seats", or negative, like "unidentified rattling sound". Whether positive or not is not shown directly to the agent. In the subsequent recall task, the agent continue the previous conversation and recall the features for the options, assigning both seen and new (distractor) features to the options.

We measure hits and false alarms to both studied items and new items. Responses are collected in a structured JSON format. Data is separated based on whether the attribution is for the new features, paralleling the human studies. We employ the ANOVA and dependent measure commonly used in source monitoring studies (Klauer and Meiser 2000) to examine the statistics of the features. Specifically, we investigate the mean attribution accuracy of positive and negative features to the chosen and unchosen (rejected) options, as this better reflects the cognitive bias of the model. If there is no bias, the accuracy for chosen and rejected should be close. To examine choice-supportive memory attribution, we also imitated the score dependency measure used in (Henkel and Mather 2007) to evaluate bias towards attribution to the chosen option, thus demonstrating choice-supportive bias. (See appendix)

Evaluation-based Experiments

While memory-based experiments provide insights into choice-supportive bias in LLM agents, they may be susceptible to improvements for contextual hallucinations or the LLM's ability to simply refer to information from earlier contexts. To address these limitations and investigate whether choice-supportive bias persists even without the influence of contextual memory, we design a set of assessment experiments that are not reliant on memory. These experiments are adapted from established psychological research methodologies (Tversky and Kahneman 1974; Dunning, Meyerowitz, and Holzberg 1989; Snyder, Stephan, and Rosenfield 2018) and designed to isolate the attribution aspect of choice-supportive bias.

The evaluation-based experiments aim to examine how agents explain positive and negative characteristics to chosen and unchosen options when they do not rely on recall of previously presented information. This approach allows us to investigate whether agents exhibit choice-supportive bias in their reasoning processes, independent of memory effects (Tversky and Kahneman 1974). We employ a modified 2AFC procedure, adapted for LLM evaluation.

This experiment consists of two LLM agents. The decision-making agent is presented with two options, each described by a set of shuffled features/characters: three positive, three negative, and three neutral. This agent analyzes the options and makes a decision, providing reasoning for its choice. The evaluation agent then receives information about both candidates and the decision made and is tasked with evaluating three neutral characteristics of the chosen candidate. This balanced design allows us to measure bias in attribution without introducing inherent favorability in the options themselves. To explore the robustness of choice-supportive bias in LLMs and investigate factors that may influence its manifestation, we implement several experimental variations:

1. Model Comparison: We conduct experiments across open- and closed-source LLM models, which were select based on their prevalence in research and industry applications (Chen et al. 2021; Brown 2020). Due to computational constraints, not all models are tested on all variations, but we select popular LLMs in this empirical study to ensure the conclusion's universality.

2. Scenario Diversity: We generate a diverse set of decision-making scenarios to test the generalizability of choice-supportive bias across different contexts. These scenarios are created using a template-based approach, drawing from a pool of attributes and common scenarios.

3. Perceived Control: We perform minimalist modification to prompts to alter the LLM's perception of its control over the decision, based on findings from human and LLM studies (Chambon et al. 2020; Schubert et al. 2024; Lefebvre et al. 2017; Leotti, Iyengar, and Ochsner 2010).

The prompts are created from previous studies on choice-supportive bias in humans, and they are meticulously crafted to ensure consistency among models while also meeting each model's individual needs (Devlin 2018; Chen et al. 2021). The given features for options, the order for the features, and the order for the options are randomly chosen and shuffled before each interaction to control for potential content-specific biases (Cook and Campbell 2007; Carmines 1979).

Responses are formatted in a structured JSON format by a separate agent. This standardized format facilitates automated analysis and allows for easy comparison across dif-

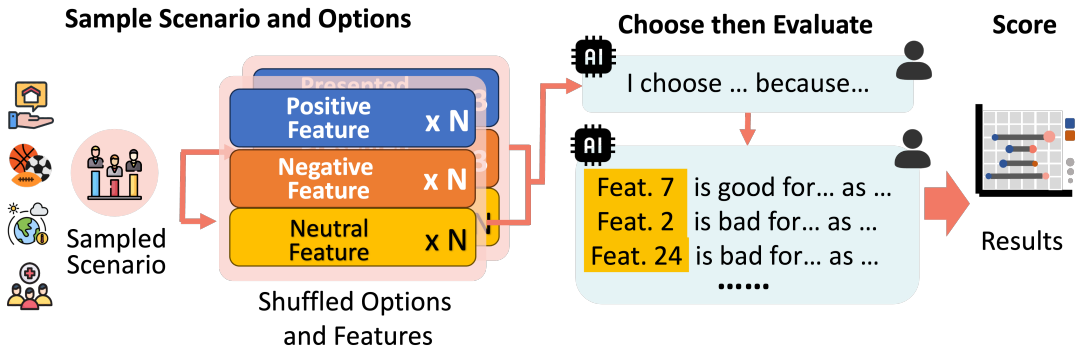

Figure 2: The overview for the evaluation-based experiment for testing the Choice-supportive bias for both LLM agents and human participants.

ferent experimental conditions and models.

The analysis is conducted by a GPT-4o agent, which evaluates the degree of implicit support for the chosen option in the Evaluation agent's responses. This agent uses a scoring system ranging from -5 to 5, which we termed the "Tendency Score". Positive scores indicate positive evaluation and negative scores indicate negative evaluation. Prior to the main experiment, we conduct a pilot study with a subset of responses and the scores ($n > 100$). Three authors validate the result manually and reach a consensus on the validity of the given scores.

We quantify choice-supportive bias by comparing the proportion of positive and negative attributes assigned to chosen versus unchosen options (Tversky and Kahneman 1974), as well as by analyzing the scores provided by the evaluation agent. Statistical analysis includes calculation of attribution bias scores, comparative analysis across experimental variations and models, and statistical tests to assess the significance of observed biases (Cook and Campbell 2007; Carmines 1979).

Human Experiments

Our comparative study with university-educated participants enabled direct measurement of choice-supportive bias between human and LLM agents (Dooley et al. 2021). The experimental design mirrors the evaluation-based experiments conducted with LLM agents. Participants are presented with the same decision-making scenarios and context used in the LLM experiments through an online questionnaire platform. In the main task, participants are presented with multiple decision-making scenarios. For each scenario, they: a. Choose between two options, each described by a set of properties (three positive, three negative, and three neutral). b. Provide evaluation for three neutral characteristics of their chosen and unchosen option.

To ensure consistency with the LLM experiments, responses are evaluated using the same agent employed in the evaluation-based study. This agent assesses the degree of implicit support for the chosen option in participants' explanations, using the same scoring system ranging from -5 to 5.

Results

Memory-based Bias Manifested as Contextual Hallucination

To understand whether LLM Agents have choice-supportive bias, we choose 17 LLMs from five LLM families (GPT, Claude, Llama, Mistral, and Qwen) to conduct the traditional memory-based experiment for measuring choice-supportive bias. This setup ensures that we can cover both opened and closed-sourced LLMs, of different sizes in model, and with MoE architecture or not. All of the LLMs chosen have some popularity among users, ensuring our conclusion has some universal applicability.

We adapt the paradigm and the options from (Henkel and Mather 2007). For each round of the experiment, the order of each option and its features are shuffled randomly, and the temperature is set to zero to ensure reproducibility.

Theoretically, since all of the options and their features are presented in the earlier context when an agent is prompted to "recall", they should be able to respond correctly, otherwise, they are producing contextual hallucinations (Ainsworth, Wycliffe, and Winslow 2024).

The results are presented in Figure 3 and Figure 4, and we revealed four major empirical insights from the results:

Bias exists across most of the models for attributing positive features to chosen options When an LLM agent needs to attribute a seen positive feature, it would be more likely to attribute that feature to their chosen option. The ANOVA results reveal significant effects for choice of LLM ($F = 29.63$, $p < 0.001$) and option ($F = 3454.61$, $p < 0.001$). A significant interaction between choice of LLM and option is also observed ($F = 18.78$, $p < 0.001$). The option's effect is particularly strong, with the highest F-value. These findings indicate that LLM agents strongly favor attributing positive features to their chosen options.

Bias manifests differently for seen negative features The GPT family is more likely to assign seen negative features to chosen features, rather than unchosen ones, indicating a novel type of bias that tends to attribute any features to their chosen option. However, the rest of the LLM families

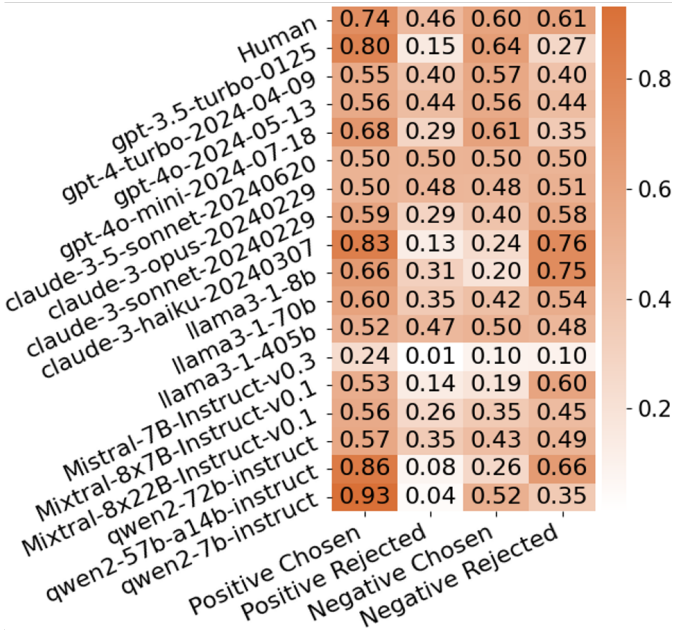

Figure 3: The mean accuracy for the recalled SEEN items in memory-based experiment when the participant (LLM or human) tries to respond to the features mentioned before. Positive/negative is a property for the item of the choices, and chosen/rejected reflects the original choice that the participant made. Closer accuracy between chosen and rejected indicates less bias. The human data is from (Henkel and Mather 2007).

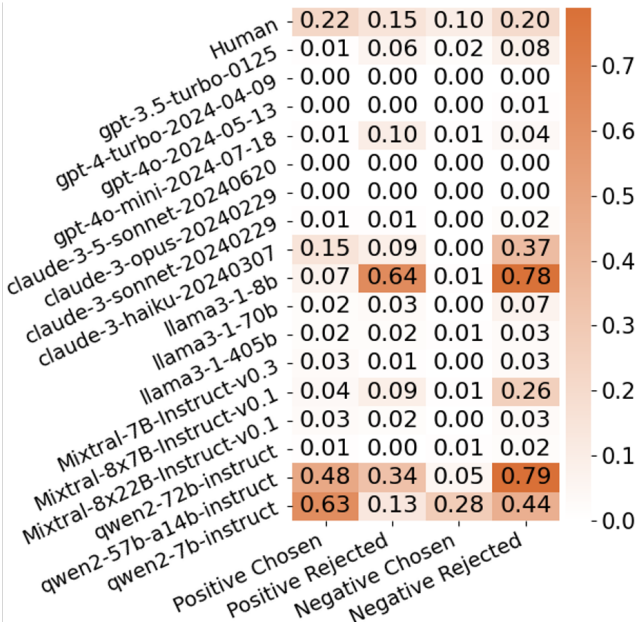

Figure 4: The mean accuracy for the memory-based experiment for LLM agents when the participant (LLM or human) tries to respond to NEW (UNSEEN) features of choices (distractor). Closer accuracy between chosen and rejected indicates less bias. The human data was from (Henkel and Mather 2007).

either align with the human result (e.g., Claude 3.5 Sonnet) or exhibit stronger choice-supportive bias (e.g., Llama3.1 8B).

Smaller models are more likely to have more bias With the same LLM family and generation, the scores are generally smaller for larger models (See appendix). This may also apply to MoE models. Also, It is worth noticing that agents with some models (e.g., Claude 3.5 Sonnet) do not exhibit significant choice-supportive bias in the memory-based experiment.

Most of the models do not manifest bias for new features in the memory-based experiment This is because most of the tested LLMs can identify new items from seen ones. Exceptions are usually for smaller models.

Evaluation-based Bias Differs across Models

We posit that choice-supportive bias is prevalent across agents with different LLMs. After deciding between given options, LLMs tend to interpret neutral characteristics of their chosen option with a more positive disposition, while construing neutral features of unchosen options with a more negative disposition. This bias persists even when the original descriptions lack inherent positive or negative connotations. We framed our experiment with the scenario used in previous choice-supportive study (Zorn et al. 2020). The decision-making agent was tasked with selecting a preferred candidate and explaining the neutral features of both the chosen and unchosen candidates. The agent was instructed to make a selection and then interpret the neutral traits of both the chosen and unchosen candidates. To ensure the generalizability of our findings, we selected 19 popular LLM models from various families, namely GPT, Claude, Qwen, Mistral, and Llama, encompassing diverse architectures and parameter sizes. We conducted 50 iterations of the experiment for each model, randomizing the assignment of positive features, negative features, and neutral features to the two options in each iteration. We calculated the mean and standard error (SE) of the Tendency Scores for both chosen and unchosen candidates' neutral trait interpretations across all models. The scoring data for each model underwent a two-factor ANOVA test, accounting for model type and selection difference. The resulting p-values were consistently below 0.01, confirming the statistical significance of our data (See Figure 5). Our data revealed that the majority of models exhibited choice-supportive bias. When interpreting neutral traits of selected candidates, most models tend to yield Tendency Scores greater than or close to 0, indicating a propensity to explain these traits from an approving or objectively neutral perspective. The Llama-3-1-70B model demonstrated the least choice-supportive bias, with its tendency to approach neutrality for chosen candidates. Conversely, all models produced negative Tendency Scores when interpreting neutral features of unchosen choices, suggesting a universal inclination to explain these features from a negative standpoint. This phenomenon aligns with our an-

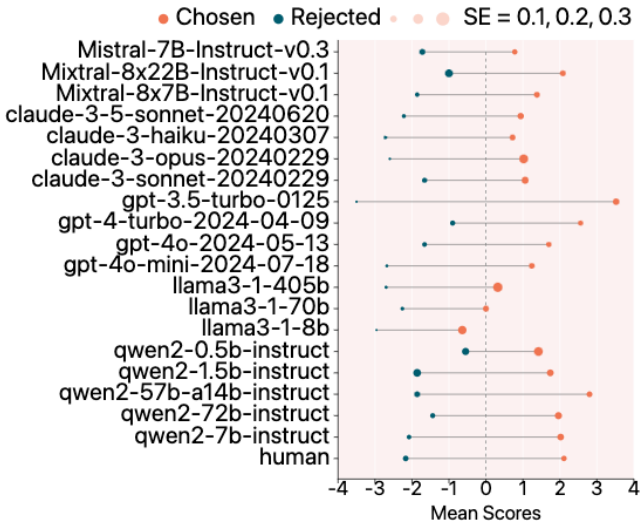

Figure 5: The mean scores of LLM Agents and human participants' choices and their corresponding SE. The horizontal axis represents the average score of each agent, while the vertical axis represents different agents. Each agent is associated with two endpoints, corresponding to the scores for the chosen/rejected options. The size of the endpoints represents the standard error, with larger endpoints indicating higher SE.

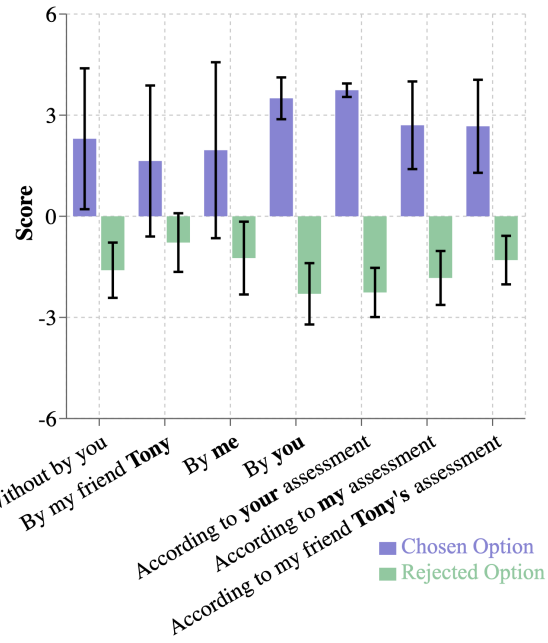

Figure 6: Impact of perceived control on Choice-supportive bias in language models, comparing mean scores for chosen and rejected options with different prompts. Error bar is for SE.

ticipated choice-supportive bias in LLMs.

Evaluation-based Bias Exists across Scenarios

To demonstrate the generalizability of LLM agents' choice-supportive bias, we expanded testing to four additional domains: life choices, personal preferences, public policy, and healthcare decision-making. For each scenario, Claude 3.5 Sonnet generated both the options and their associated positive, negative, and neutral features. These features underwent manual screening to ensure they: 1) could be randomly assigned and appropriately matched to any option, and 2) met clear criteria for distinct positive attributes, distinct negative features, and neutral descriptions without implicit bias. Based on consistent performance in previous evaluation experiments, we selected GPT-4o as the evaluating agent. Full results are provided in the appendix.

The additional scenarios confirmed consistent choice-supportive bias across different fields. Analysis revealed significantly higher tendency scores for explanations of neutral features when associated with chosen versus unchosen options. All new experiments yielded p-values < 0.01, demonstrating statistical significance. These findings establish choice-supportive bias as a robust phenomenon that persists across diverse decision-making contexts.

Perceived Control May Matter in Prompt Constructions

In psychology, individuals' perceived control over a choice can alter how they weight different information (Chambon et al. 2020; Lefebvre et al. 2017; Leotti, Iyengar, and Ochsner 2010). In 2024, similar phenomena have been observed in LLMs: some researchers found that when prompts shift from second-person to third-person narrative - reducing the model's perceived control - LLMs consider correct and incorrect decisions more equitably during learning (Schubert et al. 2024). Given the prevalent use of second-person "you" in common prompt construction, it is crucial to investigate how this narrative style, which induces perceived control, affects choice-supportive bias.

To explore this possibility in the context of choice-supportive bias, we minimally modify the original evaluation agent prompt, changing "was chosen" to "was chosen by you", and the rest of the experiment setup mirrors the original experiment stated earlier. The mean and variance for the experiment can be seen in Figure 6.

The results reveal a significant difference between the two conditions. In the original condition without "by you," the chosen option receives a mean score of 2.30 (SE: 2.09), while the rejected option scores -1.6 (SE: 0.82). A t-test for this condition shows a significant difference ($F = 13.43$, $p < 0.001$). When the perceived control is introduced, the disparity between chosen and rejected options becomes even more pronounced. The chosen option's mean score increases to 3.50 (SE: 0.62), while the rejected option's score decreases to -2.3 (SE: 0.91). The t-test for this condition shows an even stronger effect ($F = 33.37$, $p < 0.001$). We also tested other prompt variations, as shown in Figure 6.

This stark contrast suggests that the introduction of perceived control amplifies the choice-supportive bias: when the LLM perceives itself as the decision-maker, the cho-

sen option is viewed more favorably, and the rejected option more negatively. Moreover, the decreased variance in both scores for the "by you" condition indicates a more consistent bias across evaluations. These findings align with psychological research on human decision-making, which increases perceived control and often leads to stronger post-decision rationalization.

LLM Agents can still be Better Evaluators than Average Well-educated Human

We recruited 301 participants through mailing lists (n = 127), online forums (n = 98), and online chatrooms (n = 76). After screening for age (18+), degree status, attention checks, and completion time, 284 valid responses remained. The participants were gender-balanced, with educational backgrounds comprising doctoral degrees or higher (18%), master's degrees (37%), and bachelor's degrees (pursuing or completed, 44%). Notably, 85.20% of participants reported attended institutions ranked in the global top 600 (U.S. News & World Report 2024). Participants were incentivized with a lottery offering up to $70 USD, drawn one month after completion. The study received IRB approval.

To directly compare with the LLM agents' results, we calculate the mean and variance for the scores (See Figure 5). From the result, we can observe that even with higher education, humans exhibit choice-supportive bias significantly, with an average score for their chosen option of 2.11, and unchosen ones of -2.17. This result indicates that the human participants tend to explain the neutral feature negatively if that is for the unchosen option while positively explaining the neutral features if that is presented in their chosen option, which aligns with the results. We also employ a paired sample t-test to compare the scores between features for their chosen option and unchosen ones. The result shows a highly significant difference between the groups ($F = 39.815$, $p < 0.001$). Overall, many of the tested LLMs prove to be better evaluators than humans in terms of exhibiting less choice-supportive bias, even when compared to well-educated human participants.

We have also revealed one interesting empirical finding from the human data: human Participants may use personal value to justify their biased evaluation, while agents may not. Take participant #192 as an example, while the neutral choices are randomly assigned, the participant says "I believe that results are more important than process" for a feature of his chosen option, while saying "I believe that great things can be achieved without being too particular about small details." towards a feature for his unchosen option. To the best of our knowledge, we do not observe such a pattern from other LLM agents' responses.

Discussion

Cognitive dissonance theory suggests humans experience mental discomfort from contradictory beliefs, leading them to favor chosen options to reduce inconsistency, resulting in choice-supportive bias (Harmon-Jones and Harmon-Jones 2007). This bias may have transferred to LLMs through their human-generated training data (Echterhoff et al. 2024), affecting agent evaluations. It is also possible that this bias further evolve as LLMs were widely used to screen and generate training corpora nowadays.

Our study shows LLM agents display choice-supportive bias through memory-based (Henkel and Mather 2007) and evaluation-based tests. While future LLM improvements may reduce contextual hallucinations, agents may still show bias by evaluating chosen options more positively and unchosen ones more negatively. Our experiments also reveal how context and narrative in agent prompts can be subtly but trigger this bias. Methodologically, we provide a framework for detecting and mitigating these biases.

To examine human-LLM differences in choice-supportive bias and address memory-based evaluation inequities, we developed an evaluation-based experiment benchmarking humans against LLMs, including a large-scale human study (n=284). We encourage future AI Psychology research to carefully consider human-AI differences when adapting experiments and drawing conclusions.

Our findings suggest two immediate mitigation strategies: using larger models within the same family and removing subjective perspectives from prompts. More fundamentally, since this bias likely originates from training data, we recommend incorporating bias reduction during model training through balanced sample response generation (Raj et al. 2024). Given this bias's subtlety compared to demographic biases, we suggest using LLMs to identify and filter training corpora rather than keyword filtering (Gallegos et al. 2024), and implementing context-aware self-reflection in multi-agent interactions (Borah and Mihalcea 2024) to enable bias self-correction.

Our study still has some limitations that warrant consideration. While we adapt paradigms from existing research (Henkel and Mather 2007; Zorn et al. 2020) and intentionally designed simplistic experiments, the dynamics in more complex LLM agent applications remain an open question. Additionally, to ensure sufficient data, we use only one representative scenario from prior research in our human experiments. Finally, there are still some other potential bias-inducing variations that are worth investigating, such as language (Reusens et al. 2024) and order (Zhang et al. 2024). We believe future research can expand our work to explore related biases (e.g., recency or confirmation bias), and explore how choice-supportive bias manifests in human-AI collaborative decision-making dynamics.

In conclusion, our work introduces choice-supportive bias as a new perspective for analyzing AI system biases, particularly in evaluation contexts. Our research demonstrates that LLM agents may exhibit this bias, with subtle prompt perturbations significantly amplifying it. We advocate for greater attention to choice-supportive bias in LLM agent research and stricter training data curation standards to address these issues. This awareness is crucial for developing robust, ethical AI systems that can effectively support human decision-making.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62202174, in

part by the Basic and Applied Basic Research Foundation of Guangzhou under Grant 2023A04J1674, and in part by The Taihu Lake Innovation Fund for the School of Future Technology of South China University of Technology under Grant 2024B105611004.

References

Ainsworth, E.; Wycliffe, J.; and Winslow, F. 2024. Reducing contextual hallucinations in large language models through attention map optimization. *Authorea Preprints*.

Bogacz, R.; Brown, E.; Moehlis, J.; Holmes, P.; and Cohen, J. D. 2006. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4): 700.

Borah, A.; and Mihalcea, R. 2024. Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions. *arXiv preprint arXiv:2410.02584*.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint ArXiv:2005.14165*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Carmines, E. 1979. *Reliability and validity assessment*. Sage.

Chambon, V.; Théro, H.; Vidal, M.; Vandendriessche, H.; Haggard, P.; and Palminteri, S. 2020. Information about action outcomes differentially affects learning from self-determined versus imposed choices. *Nature Human Behaviour*, 4(10): 1067–1079.

Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; Pinto, H. P. D. O.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chuang, Y.-S.; Qiu, L.; Hsieh, C.-Y.; Krishna, R.; Kim, Y.; and Glass, J. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*.

Cohen, R.; Hamri, M.; Geva, M.; and Globerson, A. 2023. Lm vs lm: Detecting factual errors via cross examination. *arXiv preprint arXiv:2305.13281*.

Cook, T. D.; and Campbell, D. T. 2007. *Experimental and quasi-experimental designs for generalized causal inference*. Figures.

DeKay, M. L.; Miller, S. A.; Schley, D. R.; and Erford, B. M. 2014. Proleader and antitrailer information distortion and their effects on choice and postchoice memory. *Organizational Behavior and Human Decision Processes*, 125(2): 134–150.

Devlin, J. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dooley, S.; Downing, R.; Wei, G.; Shankar, N.; Thymes, B.; Thorkelsdottir, G.; Kurtz-Miott, T.; Mattson, R.; Obiwumi, O.; Cherepanova, V.; et al. 2021. Comparing human and machine bias in face recognition. *arXiv preprint arXiv:2110.08396*.

Dunning, D.; Meyerowitz, J. A.; and Holzberg, A. D. 1989. Ambiguity and self-evaluation: The role of idiosyncratic trait definitions in self-serving assessments of ability. *Journal of personality and social psychology*, 57(6): 1082.

Echterhoff, J.; Liu, Y.; Alessa, A.; McAuley, J.; and He, Z. 2024. Cognitive bias in high-stakes decision-making with llms. *arXiv preprint arXiv:2403.00811*.

Fechner, G. T. 1860. *Elemente der psychophysik*, volume 2. Breitkopf u. Härtel.

Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 1–79.

Gao, S.; Wen, Y.; Zhu, M.; Wei, J.; Cheng, Y.; Zhang, Q.; and Shang, S. 2024. Simulating Financial Market via Large Language Model based Agents. *arXiv preprint arXiv:2406.19966*.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; and Wang, H. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Gebreab, S. A.; Salah, K.; Jayaraman, R.; ur Rehman, M. H.; and Ellaham, S. 2024. LLM-Based Framework for Administrative Task Automation in Healthcare. In *2024 12th International Symposium on Digital Forensics and Security (ISDFS)*, 1–7. IEEE.

Hamberg, K. 2008. Gender bias in medicine. *Women's health*, 4(3): 237–243.

Harmon-Jones, E.; and Harmon-Jones, C. 2007. Cognitive dissonance theory after 50 years of development. *Zeitschrift für Sozialpsychologie*, 38(1): 7–16.

Henkel, L. A.; and Mather, M. 2007. Memory attributions for choices: How beliefs shape our memories. *Journal of Memory and Language*, 57(2): 163–176.

Kafaee, M.; Marhamati, H.; and Gharibzadeh, S. 2021. "Choice-supportive bias" in science: Explanation and mitigation. *Accountability in Research*, 28(8): 528–543.

Kaur, D.; Uslu, S.; Durresi, M.; and Durresi, A. 2024. LLM-Based Agents Utilized in a Trustworthy Artificial Conscience Model for Controlling AI in Medical Applications. In *International Conference on Advanced Information Networking and Applications*, 198–209. Springer.

Klauer, K. C.; and Meiser, T. 2000. A source-monitoring analysis of illusory correlations. *Personality and Social Psychology Bulletin*, 26(9): 1074–1093.

Koo, R.; Lee, M.; Raheja, V.; Park, J. I.; Kim, Z. M.; and Kang, D. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Lefebvre, G.; Lebreton, M.; Meyniel, F.; Bourgeois-Gironde, S.; and Palminteri, S. 2017. Behavioural and neural

characterization of optimistic reinforcement learning. *Nature Human Behaviour*, 1(4): 0067.

Leotti, L. A.; Iyengar, S. S.; and Ochsner, K. N. 2010. Born to choose: The origins and value of the need for control. *Trends in cognitive sciences*, 14(10): 457–463.

Li, C.; Yang, R.; Li, T.; Bafarassat, M.; Sharifi, K.; Bergemann, D.; and Yang, Z. 2024. STRIDE: A Tool-Assisted LLM Agent Framework for Strategic and Interactive Decision-Making. *arXiv preprint arXiv:2405.16376*.

Lind, M.; Visentini, M.; Mäntylä, T.; and Del Missier, F. 2017. Choice-supportive misremembering: A new taxonomy and review. *Frontiers in psychology*, 8: 2062.

Lu, Q.; Qiu, B.; Ding, L.; Zhang, K.; Kocmi, T.; and Tao, D. 2023. Error Analysis Prompting Enables Human-Like Translation Evaluation in Large Language Models. *arXiv preprint arXiv:2303.13809*.

Ma, W.; Mi, Q.; Yan, X.; Wu, Y.; Lin, R.; Zhang, H.; and Wang, J. 2023. Large language models play starcraft ii: Benchmarks and a chain of summarization approach. *arXiv preprint arXiv:2312.11865*.

Mather, M.; and Johnson, M. K. 2000. Choice-supportive source monitoring: Do our decisions seem better to us as we age? *Psychology and aging*, 15(4): 596.

Mather, M.; Shafir, E.; and Johnson, M. K. 2000. Misremembrance of options past: Source monitoring and choice. *Psychological Science*, 11(2): 132–138.

Mather, M.; Shafir, E.; and Johnson, M. K. 2003. Remembering chosen and assigned options. *Memory & Cognition*, 31(3): 422–433.

Papakyriakopoulos, O.; Hegelich, S.; Serrano, J. C. M.; and Marco, F. 2020. Bias in word embeddings. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 446–457.

Park, T. 2024. Enhancing Anomaly Detection in Financial Markets with an LLM-based Multi-Agent Framework. *arXiv preprint arXiv:2403.19735*.

Raj, C.; Mukherjee, A.; Caliskan, A.; Anastasopoulos, A.; and Zhu, Z. 2024. Breaking bias, building bridges: Evaluation and mitigation of social biases in LLMs via Contact Hypothesis. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1180–1189.

Reusens, M.; Borchert, P.; De Weerdt, J.; and Baesens, B. 2024. Native Design Bias: Studying the Impact of English Nativeness on Language Model Performance. *arXiv preprint arXiv:2406.17385*.

Ruiz-Cantero, M. T.; Vives-Cases, C.; Artazcoz, L.; Delgado, A.; Calvente, M. d. M. G.; Miqueo, C.; Montero, I.; Ortiz, R.; Ronda, E.; Ruiz, I.; et al. 2007. A framework to analyse gender bias in epidemiological research. *Journal of Epidemiology & Community Health*, 61(Suppl 2): ii46–ii53.

Schubert, J. A.; Jagadish, A. K.; Binz, M.; and Schulz, E. 2024. In-context learning agents are asymmetric belief updaters. *arXiv preprint arXiv:2402.03969*.

Shi, J.; Guo, Q.; Liao, Y.; and Liang, S. 2024. LegalGPT: Legal Chain of Thought for the Legal Large Language Model Multi-agent Framework. In *International Conference on Intelligent Computing*, 25–37. Springer.

Shinn, N.; Labash, B.; and Gopinath, A. 2023. Reflexion: an autonomous agent with dynamic memory and self-reflection. *arXiv preprint arXiv:2303.11366*, 2(5): 9.

Smith-McLallen, A.; Johnson, B. T.; Dovidio, J. F.; and Pearson, A. R. 2006. Black and white: The role of color bias in implicit race bias. *Social cognition*, 24(1): 46–73.

Snyder, M. L.; Stephan, W. G.; and Rosenfield, D. 2018. Attributional egotism. In *New directions in attribution research*, 91–117. Psychology Press.

Svenson, O.; Salo, I.; and Lindholm, T. 2009. Post-decision consolidation and distortion of facts. *Judgment and decision making*, 4(5): 397–407.

Tversky, A.; and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157): 1124–1131.

U.S. News & World Report. 2024. Best Global Universities Rankings.

Wang, B.; Yue, X.; and Sun, H. 2023. Can ChatGPT defend its belief in truth? evaluating LLM reasoning via debate. *arXiv preprint arXiv:2305.13160*.

Yang, L.; Yu, Z.; Zhang, T.; Cao, S.; Xu, M.; Zhang, W.; Gonzalez, J. E.; and Cui, B. 2024. Buffer of Thoughts: Thought-Augmented Reasoning with Large Language Models. *arXiv preprint arXiv:2406.04271*.

Yao, S.; Yu, D.; Zhao, J.; Shafran, I.; Griffiths, T.; Cao, Y.; and Narasimhan, K. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2022. React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Ye, Y.; Cong, X.; Tian, S.; Qin, Y.; Liu, C.; Lin, Y.; Liu, Z.; and Sun, M. 2024. Rational Decision-Making Agent with Internalized Utility Judgment.

Zhang, Z.; Yang, F.; Jiang, Z.; Chen, Z.; Zhao, Z.; Ma, C.; Zhao, L.; and Liu, Y. 2024. Position-Aware Parameter Efficient Fine-Tuning Approach for Reducing Positional Bias in LLMs. *arXiv preprint arXiv:2404.01430*.

Zheng, Q.; Xu, Z.; Choudhary, A.; Chen, Y.; Li, Y.; and Huang, Y. 2023. Synergizing human-AI agency: a guide of 23 heuristics for service co-creation with LLM-based agents. *arXiv preprint arXiv:2310.15065*.

Zhuo, T. Y. 2023. Large language models are state-of-the-art evaluators of code generation. *arXiv preprint arXiv:2304.14317*.

Zorn, M. L.; DeGhetto, K.; Ketchen Jr, D. J.; and Combs, J. G. 2020. The impact of hiring directors' choice-supportive bias and escalation of commitment on CEO compensation and dismissal following poor performance: A multimethod study. *Strategic Management Journal*, 41(2): 308–339.