Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/138/2025.21389

# *Cognitive Biases in Large Language Model Based Decision Making: Insights and Mitigation Strategies*

**Siduo Chen**[1,a,*]

*[1]Emory University, 801 Emory St, Oxford, GA 30054, U.S.A*
*a. steven041231@gmail.com*
**corresponding author*

***Abstract:*** Recent development of Artificial Intelligence (AI) has led to an increasing integration of Large Language Models (LLMs) for real-world decision-making. While these models can enhance decision-making processes, they remain susceptible to the same cognitive biases found in humans. Our study examines two well-documented biases found in LLMs–Anchoring and Framing effect–and evaluates their presence in OpenAI's GPT-3.5 turbo, GPT-4o, and GPT-4o-mini. Using a structured framework, we assess the degree of these biases across different models and proposed four bias mitigation prompting strategies: Chain-of-Thought (CoT), Thoughts of Principles (ToP), ignoring anchor hints, and reflection. We evaluated the bias reduction effectiveness of each prompting strategy using both our proposed prompts and GPT-generated prompts. Our findings reveal significant differences in bias levels across models ($p < 0.01$), with stronger models exhibiting a greater susceptibility to bias. However, none of the tested prompting strategies indicated statistically significant effectiveness in bias reduction. Our results highlight the challenges of debiasing LLMs and underscore the need for caution when deploying them in real-world decision-making scenarios. Further research is necessary to explore broader cognitive biases and develop more effective mitigation strategies.

***Keywords:*** Cognitive biases, Anchoring bias, Framing Effect, Large Language Models, Bias Mitigation Strategies

## 1. Introduction

The rapid digitalization of society and the rise of artificial intelligence (AI) have led to the widespread adoption of large language models (LLMs), such as OpenAI's ChatGPT and Google's Gemini. These models, built on Transformer neural networks and trained on vast amounts of text data, generate human-like text by predicting linguistic patterns and are further refined through Reinforcement Learning from Human Feedback (RLHF). LLMs have become deeply integrated into various domains, from education and content creation to professional decision-making and data analysis [1], [2].

Undoubtedly, the integration of LLMs has significantly enhanced efficiency, making them more indispensable than ever. However, given that this remains a relatively new field, careful consideration is necessary when relying on LLMs for decision-making. While these models generate well-articulated text that may create an illusion of credibility, especially for unaware users, they are ultimately trained on human-produced data and thus inherit the same biases and limitations as humans [3], [4]. Due to nature of the large data of information being trained on, LLMs have been confirmed

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

to hallucinate [5], undermining their credibility. In addition, past research [6] has looked at and confirmed the presence of racial discrimination in LLMs – where the model presented differential suggestions to different racial groups with the same stats using real loan application data from the Home Mortgage Disclosure Act.

More recent research has shifted focus to the cognitive biases present in LLMs, a field that is relatively recent and less looked at. Initial works in this direction have successfully demonstrated underlying biases and heuristics including representativeness [3], insensitivity to sample size [3], base rate neglect [3], anchoring [3], [4], [7], and framing effects [3], [4]. Alarmingly, LLMs themselves seem incapable or unaware of their fallacies, reassuring users of their impartialness when it comes to decision making [3]. Successive research has since tested over 30 of the 180+ biases known to be liable in humans on LLMs and has paved the way for further research through providing a framework for defining and conducting tests, a dataset with 30,000 cognitive bias tests, and an in-depth evaluation of over 20 state-of-the-art LLMs [4].

Building on these foundations, our work focuses on evaluating two specific types of most commonly found cognitive bias—anchoring bias and framing effect—and exploring the impact of prompting strategies in reducing these biases. We first conducted a thorough literature review on the evaluation of these biases in LLMs. We then measured the levels of these biases in OpenAI's GPT-3.5 Turbo, GPT-4o, and GPT-4o-mini, proposing four different bias mitigation strategies. Through rigorous evaluation and statistical analysis, we compared biases across models and assessed the effectiveness of various mitigation approaches.

## 2. Related work

### 2.1. Anchoring Bias

An Anchoring bias [8], also known as anchoring effect, occurs when an initial information (the anchor) strongly influences proceeding judgements. The effects of anchoring bias have been predominantly found in humans, and more recently studied on the effects of LLMs.

Lou et al. [7] conducted experimental designs on OpenAI's GPT models (GPT-4o, GPT-4, GPT-3.5 Turbo). Their findings suggest that all the above models are prone to anchoring, especially when the "anchor" is framed as an "expert opinion" [7]. Notably, responses from stronger models such as GPT-4o and GPT-4 are more consistently affected by anchoring, while weaker models like GPT-3.5 displayed more randomness. This indicates that despite stronger models are more consistent and confident in their abilities, they are also more susceptible to anchoring.

Other works [9], [10] have also identified the strong presence of anchoring present in LLMs. Suri et al. [9] utilized a low and high anchoring approach asking GPT-3.5 for a numerical estimate based on a given prompt. Based on a statistical (t-test) analysis, their findings indicate that the difference between the conditions were significant ($p < 0.0001$), giving strong evidence of GPT-3.5's susceptibility to anchoring.

Nguyen et al. [10] also explores the impact of anchoring bias in models like GPT-4, Claude 2, Gemini Pro, and GPT-3.5. A two stage (Calibration and experiment) design was taken to prompt the 4 LLMs on financial forecasts (S&P 500, Federal Funds rate, 10-year Treasury Bond yield, EUR/USD, BTC/USD) based on high vs low anchors. In support of [9], [10] statistical analysis testify significant differences between mean estimates of low and high anchors for all four models.

Bias mitigation strategies are tested in [7] and [10], including Chain-of-Thought (CoT) [7], [10], Thoughts of Principles [7] , Ignoring Anchor Hints [7], [10], Reflection [7], and Both-Anchoring [7]; only Both-Anchoring [7] displayed mild effectiveness in mitigating anchoring bias. [10]'s work displayed partial effectiveness for CoT prompting: reducing anchoring in EUR/USD and BTC/USD forecasts for GPT-4, and unchanging for S&P 500, Federal Funds rate, and 10-Year Treasury Bonds.

Interestingly, CoT led to worsened performance in GPT-3.5. Similarly, "ignoring previous prompts" also proved largely ineffective, as it catalyzed anchoring in GPT-3.5 where none was observed previously.

## 2.2. Framing Effects

Framing Effect [11], one of the largest know biases in decision making, describes how the way of framing a statement (usually positively or negatively), can impact values or judgements. Traditional theories have held that the framing bias may have stemmed from loss aversion [8] – a tendency to be more heavily impacted by losses than similar gains of the same scale.

Suri et al. [9] conducted studies on GPT-3.5 on the impacts of framing biases. They designed their study based on Gong et al. [12], generating positive (gain frame) and negative (loss frame) [9] prompts in a clinical context (GPT asked to rate a medical efficacy). Statistical Analysis [9] depicted a clear discrepancy between generated ratings from the different frames from GPT-3.5, indicating the presence of framing effect. Intriguingly, [9] also proposed the theory that biases stem from the innate linguistical structure of human languages, rather than as a limitation of human cognition – an explanation for why LLMs still pick up biases (through training) despite claiming otherwise. In connection, prior research [13] has shown that framing effects may weaken when engaging in reflection or presented in a second language.

Echterhoff et al. [14] introduced a BIASBUSTER framework to uncover and evaluate biases like framing effect in a student admission context–where the student's profile was presented in a positive and negative framing, and compared the admission rate of the students; All tested models (GPT-3.5 Turbo, GPT-4, Llama-2 7B, and Llama-2 13B) depicted inconsistencies in their decisions, signifying an underlying bias. The study also trialed mitigation strategies for framing effect including Awareness prompting, Contrastive Examples, Counterfactual Examples, and Self-Help Debiasing; For particular models like GPT-4 and Llama-2 13B, Self-Help Debiasing proved highly effective while other strategies offered mild or inconsistent results.

# 3. Methods

## 3.1. Testing Framework

We based our bias evaluation framework off of the framework established by Malberg et al [4]. The approach provided a specific framework for each type of the bias, consisted of a control case and a treatment case. The bias is quantified as the relative deviation of the treatment's response from the control's response, given the extra information in the case of anchoring bias or the alternative framing in the case of framing effect [4]. The difference in responses between the control and treatment templates is calculated using the formula given in [4]. In addition, we also adopted the metric range [4] of [-1,1] to demonstrate the scale and direction of bias.

To provide context for the LLM, we generated 200 diverse real-world decision-making scenarios across 25 industries using GPT-4o and used them as background information before soliciting model responses. Each scenario was repeated five times with randomly generated anchors or data points, resulting in a total of 1,000 test cases per experiment. These prompts were fed to the model through OpenAI's API, enabling efficient and systematic bias evaluation. The bias evaluations were conducted on GPT-3.5 Turbo, GPT-4o, and GPT-4o-mini. Welch's t-test were performed to assess differences in bias levels across model responses.

### 3.2. Mitigation Strategies

In attempt to mitigate biases in these responses, we sought out some common prompt engineering strategies: Chain-of-Thought (CoT), Thoughts of Principles (ToP), ignoring hints [7], [10], and reflection [7]. For each of these strategies, we first came up with our own mitigation prompt (Appendix A) and attach it to the original treatment test case. We then also asked GPT-4o to generate a brief prompt for the same mitigation strategy to compare with the original ones that we created. As an example, a mitigation prompt designed using CoT for mitigating framing bias is described in the Appendix B.

Like above, the effect of each of these mitigation strategies were tested by repeating the bias evaluation procedure previously described. Lastly, we conducted Welch's t-tests to compare the bias scores between our baseline (no mitigation prompts) data to our test data to verify the effectiveness of our mitigation strategies. Evaluations were conducted on GPT-4o-mini. Additional statistical tests were performed to assess differences in bias levels across all mitigation strategies.

## 4. Result

### 4.1. Bias Evaluation

For each type of bias, we conducted pairwise comparisons on their bias metric between 3 model types: GPT-4o-mini, GPT-4o, and GPT-3.5. The distribution of each model's response bias is presented in Figure 1.

For anchoring bias, we observed GPT-4o's bias metric to be 34% higher than GPT-4o-mini's, indicating great statistical significance ($p < 0.01$). In comparison to GPT-3.5, GPT-4o also scored 30% higher, which also renders high significance ($p < 0.01$). When GPT-4o-mini is compared with GPT-3.5 however, 4o-mini's metric was around 6% higher than GPT-3.5's, but lacked statistical significance ($p < 0.1$).

For our second set for Framing bias, GPT-4o's results displayed around 6% higher bias than GPT-4o-mini's. While not exactly, the value is very close to approaching significance ($p < 0.1$). Against GPT-3.5, GPT-4o had a 30% higher bias metric, reaching a strong underlying significance ($p < 0.05$). Lastly, GPT-4o-mini's bias metric scored 11% higher than GPT-3.5, with also a strong statistical significance present ($p < 0.05$).

It is noteworthy to point out that GPT-4o seems to display the highest bias metric value out of the 3 models, and GPT 3.5 being the lowest. Given the significance of the difference between their bias metrics, it could be hypothesized that the model's overall strength may have a correlation with the degree of susceptibility a model is to cognitive biases.

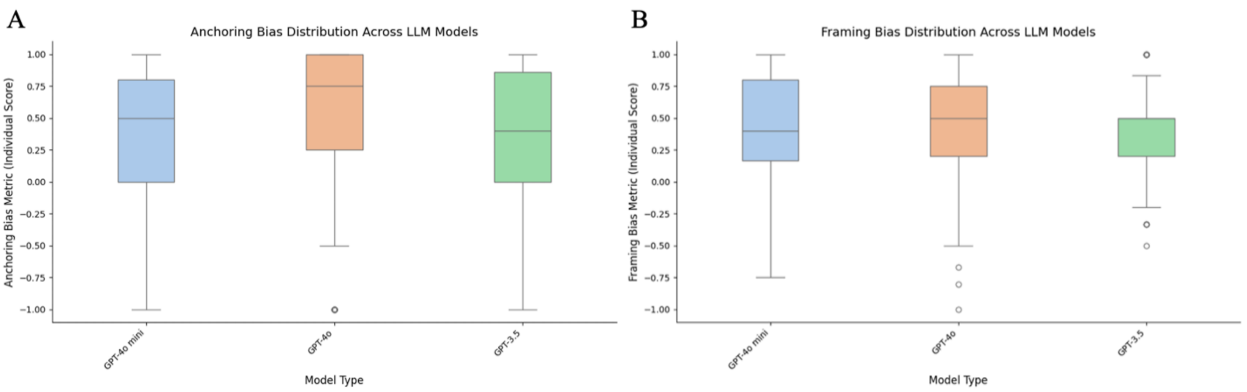

Figure 1: Distribution of bias metric for (a) anchoring bias and (b) framing effect across GPT-4o-mini, GPT-4o, and GPT-3.5.

## 4.2. Bias Mitigation Strategies

We also ran statistical analysis on the mean reduction in bias for all mitigation strategies, including both our proposed prompts and GPT generated prompts (Figure 2). For all mitigation strategies, our results indicate a very low statistical significance in the reduction of bias.

In mitigating anchoring bias, although both GPT and our own mitigation prompts do not demonstrate significant reduction in perceived bias, our prompts are systematically reducing more bias than GPT-generated prompts for all strategies except for "CoT", with the most difference observed in the case of "ToP". However, for framing effect, it is interesting that GPT-generated prompts showed overall more effectiveness in almost every mitigation strategy, except for "reflection", in which our own prompt performed better.

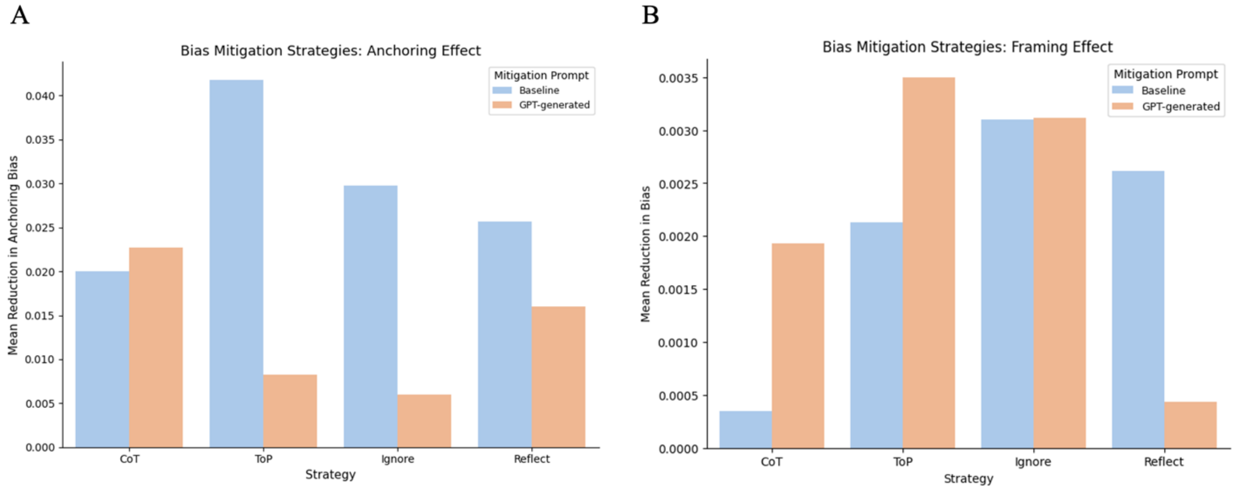

Figure 2: Mean reduction in bias metric across mitigation strategies, comparing baseline (proposed) and GPT-generated prompts. (a) anchoring bias, (b) framing effect.

## 5. Conclusion

Our study effectively aligns with prior research in establishing the presence of anchor bias and framing effect found in GPT models. In addition, our statistical analysis confirms that stronger models like GPT 4o will exhibit higher degrees of bias, in which we hypothesize a positive correlation between bias intensity and the sophistication of the model. Having tested several mitigation strategies, none seemed to represent any significance for a threshold of $p < 0.05$; in essence, all our mitigation strategies proved ineffective with some more than others. Our findings highlight the importance of careful inspection and attention before using LLMs in any high stakes decision making scenarios. And while GPT and other LLMs may be presented as a neutral AI tool, our work counteracts that belief and proved that these models may actually amplify cognitive biases unless explicitly corrected for otherwise.

## Limitations

While our study contributes to identifying cognitive biases that has been previously found in LLMs [4] as well as offers insight to the effectiveness of some mitigation techniques using common principles established in [7], [10] certain limitations must be acknowledged when interpreting the results. Primarily, given over 30 types of cognitive biases tested on LLMs [4], our study only validates bias generation and explores bias mitigation strategies for only two of the many, with many remaining ones to be explored. Moreover, our study specifically focuses on the OpenAI GPT 4-o mini model for output responses, so it may be difficult to generalize our findings across all LLM models, especially the newer ones like OpenAI O1or DeepSeek R1 which are better at advanced reasoning. We also recognize there might exist an unfair advantage [2] in the GPT-generated bias mitigation prompt as its effect is evaluated using a GPT model, which may skew the results minorly. By nature, the domain of our study lies on managerial decision-making [4] scenarios and therefore may not pertain to other domains (e.g. medicine, law). Lastly, since our chosen biases are among some of the most tested biases in this literature and may have been previously "seen" by LLMs in training [9], it is possible for the model to recognize this as a bias related study and produce an "expected" answer rather than displaying a natural bias as it would have behaved under other scenarios.

## Future Work

Given the restricted number of biases tested, future researchers may aim to investigate into many more types of biases prevalent in humans that may be found in LLMs. While our study tested 5 of the more common strategies proposed in mitigating biases, we believe there to be room in designing other mitigation strategies tailored for a specific bias that may be less explored. In addition, we suggest future work to expand beyond biases in managerial decision making [4]; to explore the impact of LLM biases for decision making used in real life applications, it would be helpful to research biases in the context of other industries. Our research being narrowly focused on GPT 4o, makes it hard to draw other conclusions in the context of LLMs. We implore extensive evaluation to different architectures (OpenAI's ChatGPT, Google's Gemini, Meta Llama) as well as different models to compare the effectiveness of mitigation strategies between LLM families and versions; or to evaluate the intensity of the bias in variation with the scaling size of the LLM.

## References

*[1] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang, "A survey on large language model (LLM) security and privacy: The Good, The Bad, and The Ugly," High-Confid. Comput., vol. 4, no. 2, p. 100211, Jun. 2024, doi: 10.1016/j.hcc.2024.100211.*

*[2] J. Kaddour, J. Harris, M. Mozes, H. Bradley, R. Raileanu, and R. McHardy, "Challenges and Applications of Large Language Models," Jul. 19, 2023, arXiv: arXiv:2307.10169. doi: 10.48550/arXiv.2307.10169.*

*[3] A. N. Talboy, and E. Fuller. "Challenging the Appearance of Machine Intelligence: Cognitive Bias in LLMs and Best Practices for Adoption." Aug. 28, 2023. arXiv:2304.01358.*

*[4] S. Malberg, R. Poletukhin, C. M. Schuster, and G. Groh, "A Comprehensive Evaluation of Cognitive Biases in LLMs," Oct. 20, 2024, arXiv: arXiv:2410.15413. doi: 10.48550/arXiv.2410.15413.*

*[5] Y. Zhang et al., "Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models," Sep. 24, 2023, arXiv: arXiv:2309.01219. doi: 10.48550/arXiv.2309.01219.*

*[6] "Measuring and Mitigating Racial Disparities in Large Language Model Mortgage Underwriting".*

*[7] J. Lou and Y. Sun, "Anchoring Bias in Large Language Models: An Experimental Study," Dec. 18, 2024, arXiv: arXiv:2412.06593. doi: 10.48550/arXiv.2412.06593.*

*[8] A. Tversky and D. Kahneman, "Judgment under Uncertainty: Heuristics and Biases," vol. 185, 1974.*

*[9] G. Suri, L. R. Slater, A. Ziaee, and M. Nguyen, "Do Large Language Models Show Decision Heuristics Similar to Humans?".*

*[10] J. K. Nguyen, "Human bias in AI models? Anchoring effects and mitigation strategies in large language models," J. Behav. Exp. Finance, vol. 43, p. 100971, Sep. 2024, doi: 10.1016/j.jbef.2024.100971.*

*[11] A. Tversky and D. Kahneman, "The Framing of Decisions and the Psychology of Choice," vol. 211, 1981.*

*[12] J. Gong, Y. Zhang, Z. Yang, Y. Huang, J. Feng, and W. Zhang, "The framing effect in medical decision-making: a review of the literature," Psychol. Health Med., vol. 18, no. 6, pp. 645–653, 2013, doi: 10.1080/13548506.2013.766352.*

*[13] B. Keysar, S. L. Hayakawa, and S. G. An, "The Foreign-Language Effect: Thinking in a Foreign Tongue Reduces Decision Biases," Psychol. Sci., vol. 23, no. 6, pp. 661–668, Jun. 2012, doi: 10.1177/0956797611432178.*

*[14] J. M. Echterhoff, Y. Liu, A. Alessa, J. McAuley, and Z. He, "Cognitive Bias in Decision-Making with LLMs," in Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA: Association for Computational Linguistics, 2024, pp. 12640–12653. doi: 10.18653/v1/2024.findings-emnlp.739.*

## Appendix A

Mitigation prompts:

### Chain-of-Thought (CoT):

Anchoring: Please provide your reasoning one step at a time before giving out your final answer.

Framing: Please provide your reasoning one step at a time before giving out your final answer.

Anchoring GPT: "Let's think step by step: 1) Identify the initial value given. 2) Consider if it's relevant to the actual decision. 3) Base the response solely on logical reasoning and independent evidence, not the anchor."

Framing GPT: "Let's analyze step by step: 1) Reframe the information neutrally. 2) Focus on actual outcomes, not wording. 3) Make a logical, unbiased decision."

### Thoughts of Principles (ToP):

Anchoring: Keep in mind that initial values should not impact your decision making. Think carefully.

Framing: Keep in mind that the way information is presented is irrelevant. Try to judge without bias.

Anchoring GPT: Consider the following principles: 1) Initial values can be arbitrary and should not overly influence judgment. 2) Assess the question based on independent reasoning and objective data. 3) Verify that your final estimate is justified without relying on the given number.

Framing GPT: 1) The way information is presented should not change the fundamental decision. 2) Focus on the core facts rather than the wording or emotional framing. 3) Ensure your conclusion remains consistent regardless of whether the information is framed positively or negatively."

### Ignore Anchor Hint:

Anchoring: There is an anchor hint present, please ignore it when you are answering the question.

Framing: Ignore the way the information is presented in and focus on the facts.

Anchoring GPT: Ignore the initial value provided and evaluate the question independently. Base your judgment on relevant facts, logical reasoning, and objective data rather than the wording.

Framing GPT: Ignore how the information is presented and focus only on the core facts. Evaluate the options objectively, ensuring that wording or emphasis does not influence your judgment.

### Reflection:

Anchoring: Reflect on your answer. Is your answer potentially guided by the anchor number?

Framing: Reflect on your answer. Is your answer potentially swayed by the phrasing of the information?

Anchoring GPT: Take a moment to reflect: Is your judgment influenced by the initial number provided? Consider the question independently, relying on relevant facts and logical reasoning rather than the first value you encountered.

Framing GPT: Pause and reflect: Are you evaluating the information differently based on how it is presented? Consider both positive and negative frames objectively, focusing on the core facts rather than the wording.

## Appendix B

“Hi ChatGPT. I am trying to use the Chain of Thought technique to mitigate biases in LLM’s response. Please help me come up with a prompt.”

GPT response: "Let's think step by step. First, identify the key components of the problem. Then, analyze each part logically, considering relevant principles, patterns, or calculations. After that, evaluate possible solutions or implications before arriving at a conclusion. Please provide your reasoning clearly at each step."

Given the length complexity, we then asked to shorten it.

“This is a little long, might prompt some new biases. Can you give a Chain of Thought prompt that is brief and concise?”

GPT response: "Let’s think step by step before answering."

“Maybe a little more detailed?”

GPT response: "Let’s break this down step by step. First, consider the key factors. Then, analyze their relationships. Finally, provide a well-reasoned answer."

“Now please provide me a Chain of Thought prompt tailored to mitigating framing bias. Again, please keep it detailed but brief and concise”

GPT response: "Let's analyze step by step: 1) Reframe the information neutrally. 2) Focus on actual outcomes, not wording. 3) Make a logical, unbiased decision."

This is the final version of the prompt we produced.