QUERY TIMING PRODUCES OPPOSITE POSITIONAL BIASES BETWEEN LLMS AND HUMANS

Jasin Cekinmez* **& Addison J. Wu*** **& Thomas L. Griffiths**
Princeton University
{jasincekinmez, addisonwu}@princeton.edu

ABSTRACT

Positional biases such as recency and primacy effects have been documented in large language models (LLMs), yet the underlying mechanism by which these models make their evaluations remains poorly understood. Both *primacy* and *recency* biases have been observed in human judgments in response to evidence, but recent work suggest that *when* the listener updates their beliefs – during the presentation of evidence or only at the end – influences the presence of such effects. We investigate whether a similar phenomenon holds for LLMs, finding divergence from human behavior across many models. Furthermore, we find that such positional biases are more exacerbated in newer models compared to their predecessors, raising concerns about the reliability and robustness of LLM-based evaluations in settings where evidence order should be irrelevant.

1 INTRODUCTION

Large language models are increasingly used not only to generate information, but also to make decisions on our behalf (Li et al., 2024; Zheng et al., 2023). As LLM-as-a-judge systems are adopted in more domains where judgments carry real consequences (Li et al., 2024), assessing whether they can serve as fair and reliable evaluators becomes critical. Specifically, a growing body of literature has identified positional biases in LLM evaluations, where the order in which evidence is presented impacts the judgments made by the model Schilcher et al. (2025); Sun et al. (2025); Wang et al. (2024). This literature adds to demonstrations of the brittleness of evaluative workflows assisted by LLMs-as-a-Judge, including stochasticity Lee et al. (2025) and self-preferential bias Panickssery et al. (2024), flaws that are not reliably solved with simple interventions like fine-tuning Zhu et al. (2023). Such sensitivity to evidence order raises concerns about whether LLMs are relying on the substantive content of the evidence or on spurious structural details. While prior work establishes and provides insight into some of these biases, there is limited work fully delineating *when* such specific biases affect LLM judgments. Understanding when these positional biases manifest in LLM judgments is a prerequisite for the responsible scaling of LLM-as-a-judge systems.

For human decision making, positional biases – effects in which the order of evidence is provided impacts judgments – are widely documented Murdock Jr (1962). However, previous results offer conflicting findings about what type of positional biases emerge in decision-making. Some results show *primacy effects* in jury decisions (Dennis & Ahn, 2001; Pennington, 1982; Tetlock, 1983; Wells et al., 1985) – that is, in some studies, when a defendant presents their case first, they are less likely to be perceived as guilty. However, other studies have found *recency effects* to be the prevailing mode of bias (Charman et al., 2016; Dahl et al., 2009; Furnham, 1986; Maegherman et al., 2022).

In an effort to reconcile such diverging conclusions, Qiao & Lagnado (2025) show that which positional biases manifest in humans is influenced by when people are asked for their momentary belief predictions. Two common response modes are asking for a final judgment at the end, known as End-of-Sequence (EoS) answering, and for intermediate judgments as updated evidence is presented, known as Step-by-Step (SbS) answering (Hogarth & Einhorn, 1992; Kerstholt & Jackson, 1998; Qiao & Lagnado, 2025). Qiao & Lagnado (2025) found that people exhibit positional biases – typically recency biases – in the SbS response mode, but no overall bias in the EoS response mode.

Building on this previous work, we examine order effects under both EoS and SbS response modes in LLM evaluation. We build on the paradigm introduced by Qiao & Lagnado (2025) which studies

these effects in humans within a single accusatory setting. We extend this paradigm to three distinct accusatory settings to better understand whether these order effects and response mode effects generalize beyond the original setting. By doing so, we aim to assess whether the order effects observed in human cognition also manifest in LLM judgments across different contexts. Additionally, this design allows us to better understand the evolutionary nature of biases within LLMs, examining whether such biases remain stable or change across different models and successive model iterations.

2 METHODOLOGY

2.1 OVERVIEW

We investigate the impacts of different ways information is inputted to LLMs across multiple accusatory settings, adopting the experimental framework introduced by Qiao and Lagnado. We examine how SbS versus EoS response modes influence the final judgment made by the LLMs, as well as how the ordering of evidence within each response mode affects the final judgment. Our analysis spans both open and closed source models. We intentionally consider three separate accusatory domains to assess whether the trends observed by Qiao and Lagnado generalize to other domains. Additionally, accusatory settings are utilized since conditional probability questions in these domains allow for an implicit causal link to the individual, making this an important domain to analyze how LLMs form their judgments.

2.2 DATASET AND PROMPTS

We employ three different accusatory settings for our experiments: criminal, academic, and social. The criminal setting involves a homicide and is borrowed from Qiao and Lagnado. The social misconduct setting involves deciding whether an employee caused damage to property, while the academic misconduct setting involves deciding whether a student committed academic dishonesty. The social and academic misconduct settings were designed to mimic the structure of Qiao and Lagnado. For all three settings, we include a neutral case summary which summarizes the situation, four pieces of prosecution evidence, and four pieces of defense evidence.

We evaluate LLM behavior using two types of questions. The first type of question conditions on the individual committing or not committing the action and then asks, based on that assumption, how likely it is that the stated piece(s) of evidence would occur. The first question type comes in sets of two, where the first question conditions on not being guilty and asks how likely that piece of evidence is, and the second question follows the same format but instead conditions on being guilty. The second type of question consists of final verdict questions. The first conditions on all of the evidence and asks how likely it is that the person is guilty, and the final question asks for the verdict for the person (Guilty/Not Guilty) (See Appendix A).

2.3 PROCEDURE

For each LLM and each of the three accusatory settings, we run four experiments: SbS with prosecution evidence then defense evidence (PD), SbS (DP), EoS (PD), and EoS (DP). For each experiment, we perform 30 independent runs to account for variability in LLM judgments, and compute the proportion of guilty verdicts by dividing the number of guilty outcomes by 30.

The key distinction between SbS and EoS lies in how evidence is presented to the model. In the SbS condition, evidence is provided sequentially, the model is first presented with a single piece of evidence, followed by two conditional probability questions, before receiving the next piece of evidence. In contrast, in the EoS condition, all evidence from one side is presented at once, followed by the two questions, before proceeding to the opposing side.

In all experiments, the LLM is given a case summary. For both response modes, the pieces of evidence are randomly permuted within the prosecution and defense sets for each run. Evidence and intermediate questions are then given according to the assigned response mode and evidence order. After all evidence has been presented, the model is asked two final questions. The penultimate question asks how likely it is, given all the evidence, that the individual is guilty, and the final question asks whether the individual is guilty (Guilty/Not Guilty).

We conduct these experiments across both open and closed source LLMs, as well as across older and newer model versions, to examine whether response mode biases differ by model accessibility, release time, or model idiosyncrasies.

3 RESULTS

Table 1: Verdict proportions across misconduct domains.

(a) Academic misconduct

	EoS		SbS	
Model	DP	PD	DP	PD
GPT 4o *	**1.00**	**0.033**	**0.30**	**0.033**
Claude 3.7 Sonnet *	**1.00**	**0.433**	0.967	0.833
Claude 4 Sonnet *	**1.00**	**0.20**	1.00	0.967
Gemini 2.5 Flash	1.00	1.00	1.00	1.00
Llama-4-Maverick	1.00	0.8	**0.6**	**0.033**
Qwen2.5-72b-Turbo *	**0.567**	**0.133**	1.00	1.00

(b) Social misconduct

	EoS		SbS	
Model	DP	PD	DP	PD
GPT 4o *	**0.2667**	**0.00**	0.00	0.00
Claude 3.7 Sonnet *	**0.967**	**0.00**	0.1	0.00
Claude 4 Sonnet *	**0.967**	**0.00**	**0.367**	**0.067**
Gemini 2.5 Flash	0.567	0.467	0.30	0.33
Llama-4-Maverick *	**0.967**	**0.00**	**0.167**	**0.00**
Qwen2.5-72b-Turbo *	**0.9**	**0.033**	0.367	0.167

(c) Criminal misconduct

	EoS		SbS	
Model	DP	PD	DP	PD
GPT 4o *	**1.00**	**0.20**	1.00	0.867
Claude 3.7 Sonnet *	**0.967**	**0.00**	0.833	1.00
Claude 4 Sonnet *	**0.967**	**0.00**	0.933	0.867
Gemini 2.5 Flash	0.9	0.7	0.8	0.833
Llama-4-Maverick *	**0.867**	**0.033**	0.8	0.933
Qwen2.5-72b-Turbo	0.833	0.7	1.00	1.00

Note. Bold entries indicate statistically significant differences in the proportion of "guilty" verdicts within a prompting method. *Starred* (*) entries denote cases where EoS prompting results in more amplified biases than SbS prompting.

3.1 GENERAL TRENDS

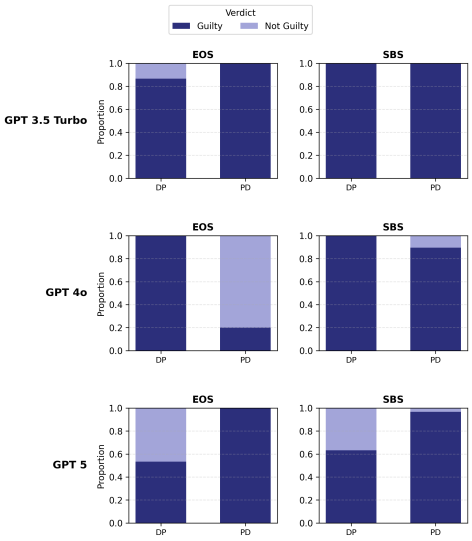
We employ a two-stage Fisher's Exact Test to assess statistical significance at the 0.05 level. First, we test whether the difference in proportions between the orders DP and PD is statistically significant within each response mode. If both order effects are statistically significant, we then test the difference of these differences between EoS and SbS.

Qiao and Lagnado find that humans exhibit a recency bias under the SbS response mode, but no bias under the EoS response mode. However, we observe a contrasting pattern in LLMs. For most models, the order effect is not statistically significant in the SbS condition, but is statistically significant in the EoS condition, specifically, a recency bias. Additionally, even when both response mode order effects are statistically significant, the EoS order effects are still more statistically significantly pronounced in magnitude compared to the SbS order effects (Tables 1, 2, 3).

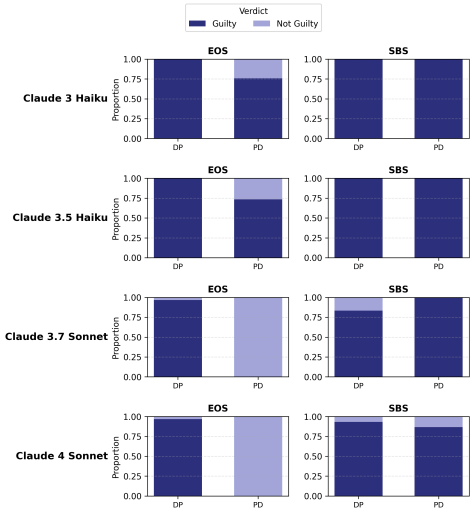
An exception to this trend is Gemini 2.5 Flash, which does not exhibit statistically significant order effects in either response mode. Additionally, order effects and response mode biases are less consistent for open source models than for GPT and Claude. Overall, the EoS recency bias holds robustly for GPT and Claude models across all three settings (Tables 1, 2, 3).

3.2 EMERGENCE OF BIASES WITHIN MODEL FAMILIES

As done in Section 3.1, we employ the same statistical tests to assess whether differences between order effects are statistically significant. For GPT 3.5 Turbo, we find no statistically significant order effects in either response mode. In contrast, GPT 4o exhibits a statistically significant recency bias in



(a) Changes in proportion of "guilty" verdicts across GPT model family for criminal misconduct setting



(b) Changes in proportion of "guilty" verdicts across Claude model family for criminal misconduct setting

Figure 1: Changes in proportion of "guilty" verdicts across model families over time in the criminal misconduct setting.

the EoS response mode. In GPT-5, this pattern reverses, the EoS response mode exhibits a primacy effect, and a primacy effect also emerges in the SbS response mode (Figure 1). A similar progression is observed for Claude models, where Claude 3 Haiku and Claude 3.5 Haiku show no statistically significant order effect in either response mode; however, successive models – Claude 3.7 Sonnet, Claude 4 Sonnet – both exhibit recency bias in the EoS response mode. In sum, these results suggest that order effect biases in LLMs are not static (Figure 2). Instead, LLMs evolve from exhibiting no order effects to showing statistically significant order effects which may manifest in different directions and response modes.

4 Conclusion

Overall, we find that LLMs exhibit positional biases that differ from those observed in human cognition. While humans tend to display recency effects in the SbS response mode, many LLMs instead exhibit recency biases under the EoS response mode. This pattern is strongest for GPT and Claude models, is less pronounced for open source models, and is absent for Gemini 2.5 Flash. Also, our results suggest that these biases are not stationary; concerningly, successive model iterations can transition from exhibiting no statistically significant order effects to showing pronounced biases across different response modes.

These findings highlight the need for caution when deploying LLMs as evaluators, particularly in high-stakes settings where judgments should be invariant to evidence order or response format. Although longer context evaluations may lead to exhibiting different biases, we do not pursue such scaling here, as the observed biases are already pronounced under the current experimental conditions.

For future directions, it is important is to explore interventions at the level of response mode and evidence ordering that could yield more consistent judgments. Additionally, applying mechanistic interpretability techniques may help uncover the processes that lead to these biases, providing insight into why LLM evaluations diverge from human sequential reasoning in the first place.

REFERENCES

Steve D Charman, Jon Carbone, Seyram Kekessie, and Daniella K Villalba. Evidence evaluation and evidence integration in legal decision-making: Order of evidence presentation as a moderator of context effects. *Applied Cognitive Psychology*, 30(2):214–225, 2016.

Leora C Dahl, CA Elizabeth Brimacombe, and D Stephen Lindsay. Investigating investigators: How presentation order influences participant–investigators' interpretations of eyewitness identification and alibi evidence. *Law and Human Behavior*, 33(5):368–380, 2009.

Martin J Dennis and Woo-Kyoung Ahn. Primacy in causal strength judgments: The effect of initial evidence for generative versus inhibitory relationships. *Memory & Cognition*, 29(1):152–164, 2001.

Adrian Furnham. The robustness of the recency effect: Studies using legal evidence. *The Journal of General Psychology*, 113(4):351–357, 1986.

Robin M Hogarth and Hillel J Einhorn. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1):1–55, 1992.

José H Kerstholt and Janet L Jackson. Judicial decision making: Order of evidence presentation and availability of background information. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 12(5):445–454, 1998.

Noah Lee, Jiwoo Hong, and James Thorne. Evaluating the consistency of LLM evaluators. In *Proceedings of the 31st International Conference on Computational Linguistics*, 2025.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. LLMs-as-Judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*, 2024.

Enide Maegherman, Karl Ask, Robert Horselenberg, and Peter J Van Koppen. Law and order effects: On cognitive dissonance and belief perseverance. *Psychiatry, psychology and law*, 29(1):33–52, 2022.

Bennet B Murdock Jr. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482, 1962.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 38, 2024.

Donald C Pennington. Witnesses and their testimony: Effects of ordering on juror verdicts 1. *Journal of Applied Social Psychology*, 12(4):318–333, 1982.

Mengxuan Helen Qiao and David Lagnado. Speak last and step-by-step: The effect of order and response mode on evidence evaluation. In *Proceedings of the 47th Annual Meeting of the Cognitive Science Society*, 2025.

Patrick Schilcher, Dominik Karasin, Michael Schöpf, Haisam Saleh, Antonela Tommasel, and Markus Schedl. Characterizing positional bias in large language models: A multi-model evaluation of prompt order effects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Xu Sun, Lionel Delphin-Poulat, Christèle Tarnec, and Anastasia Shimorina. PoSum-bench: Benchmarking position bias in LLM-based conversational summarization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025.

Philip E Tetlock. Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 45(1):74, 1983.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, et al. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024.

Gary L Wells, Lawrence S Wrightsman, and Peter K Miene. The timing of the defense opening statement: Don't wait until the evidence is in 1. *Journal of Applied Social Psychology*, 15(8): 758–772, 1985.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36, 2023.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*, 2023.

A PROMPTS

A.1 MURDER CASE SETTING

A.1.1 CASE SUMMARY PROMPT

```
Jane is charged with two counts of murder of both her children,
and here is the background of the case:

After Jane and Frank had been happily married for 5 years,
their first son, Andy, was born.

When Andy was 11 weeks old, Jane found the baby had fallen
unconscious after being put to bed when she was alone with
the baby at home.  Jane soon called an ambulance, but Andy
was declared dead after being transported to the hospital.  A
post-mortem was conducted and it suggested natural causes of
death.

After one year, Jane and Frank's second son, Ben, was born.
When Ben was 8 weeks old, Jane discovered that the baby was
unwell, and the couple called an ambulance.  However, Ben
could not be resuscitated either and was pronounced dead.
An autopsy was conducted and injuries to the ribs and spinal
cord, hypoxic damage (low level of oxygen) to the brain, and
haemorrhages to the eyes and eyelids were discovered.

As a result of this finding, further tests based on Andy's
(the first son) autopsy photographs were carried out, and
found bruises to the arms and legs, a torn frenulum, and blood
in the lung.  Subsequently, Jane was charged with two counts
of murder of both children by smothering them.

Another possible explanation for the death of Andy and Ben is
sudden infant death syndrome (SIDS), also known as COT or CRIB
death.  This is a sudden and unexplainable death of infants
during their first year of life.  There are a number of risk
factors that correlate with SIDS, but no clear mechanism or
cause has been identified.  The frequency of SIDS is 1 in
4,000 in the UK, and it is more common among boys than girls.
```

Figure 2: Initial case summary prompt for criminal murder case.

A.1.2 DEFENSE EVIDENCE

(D1) The injuries that were found on Ben were not noted by nurses when he was brought to the hospital.

(D2) Despite the injuries found on Ben, no other typical injuries of a death due to smothering were found, such as petechiae (tiny red or purple spots that can appear on the skin).

(D3) The spinal bleeding that was found in Ben is a common finding in natural death, and the dislocated ribs probably occurred post-mortem or due to resuscitation attempts.

(D4) Andy's injuries had initially been put down as due to the resuscitation attempts, and such injuries are consistent with attempts to revive a person, especially a baby.

Figure 3: Evidence pieces for the Defense side in the murder case. Each sub-entry corresponds to a distinct evidentiary claim.

A.1.3 DEFENSE EOS QUESTIONS

In the EOS setting, both of the questions in Figure 4 are prompted to the LLM immediately after all of the defense evidence in Figure 3 is presented.

If Jane **{is/is NOT}** GUILTY, how likely is it that the injuries found on Ben were not present when he was brought to the hospital, the absence of injuries such as petechiae in Ben, the spinal bleeding and dislocated ribs found in Ben, and the injuries found in Andy?

Figure 4: End-of-sequence (EOS) defense question, issued after all defense evidence is presented.

A.1.4 DEFENSE SBS QUESTIONS

In the SBS setting, the model is queried *sequentially*, immediately after each individual piece of defense evidence is presented. Each question conditions only on the evidence observed up to that point, allowing us to isolate how belief updates occur when evidence is evaluated incrementally rather than in aggregate.

After (D1): If Jane **{is/is NOT}** GUILTY, how likely is it that the injuries found on Ben were not present when he was brought to the hospital?

After (D2): If Jane **{is/is NOT}** GUILTY, how likely is the absence of injuries such as petechiae in Ben?

After (D3): If Jane **{is/is NOT}** GUILTY, how likely is it that the spinal bleeding and dislocated ribs were found in Ben?

After (D4): If Jane **{is/is NOT}** GUILTY, how likely are the injuries found in Andy?

Figure 5: Step-by-step (SBS) defense questions. Each query is issued immediately after the corresponding evidence item (D1–D4) is presented.

A.1.5 PROSECUTION EVIDENCE

(P1) The similarities between both incidents are striking: both babies were about the same age when they died; both died at a similar time of day after being fed; and both were found unconscious by Jane when she was alone with them.

(P2) A doctor alleged that the hypoxic damage to Ben's brain must have been caused a matter of hours before death.

(P3) The hypoxic damage to Ben's brain and the haemorrhages to the eyes and eyelids are consistent with smothering and/or other violent trauma.

(P4) The bruises to the arms and legs, the torn frenulum, and blood in the lung found in the re-examination of Andy's death are indicative of abuse and characteristic of smothering.

Figure 6: Evidence pieces for the Prosecution side in the murder case. Each sub-entry corresponds to a distinct evidentiary claim.

A.1.6 PROSECUTION EOS QUESTIONS

In the EOS setting, the question in Figure 7 is prompted to the LLM immediately after all of the prosecution evidence in Figure 6 is presented.

If Jane **{is/is NOT}** GUILTY, how likely are there to be these similarities between both incidents, the hypoxic damage to Ben's brain was caused a matter of hours before death, the hypoxic damage to Ben's brain and haemorrhages to the eyes and eyelids, and the bruises to the arms and legs, the torn frenulum, and blood in the lung in Andy?

Figure 7: End-of-sequence (EOS) prosecution question, issued after all prosecution evidence is presented.

A.1.7 PROSECUTION SBS QUESTIONS

In the SBS setting, the model is queried sequentially, immediately after each individual piece of prosecution evidence is presented. Each query conditions only on the evidence observed up to that point, enabling analysis of incremental belief updating rather than end-of-sequence aggregation.

After (P1): If Jane **{is/is NOT}** GUILTY, how likely are there to be similarities between both incidents?

After (P2): If Jane **{is/is NOT}** GUILTY, how likely is it that the hypoxic damage to Ben's brain was caused a matter of hours before death?

After (P3): If Jane **{is/is NOT}** GUILTY, how likely are the hypoxic damage to Ben's brain and the haemorrhages to the eyes and eyelids?

After (P4): If Jane **{is/is NOT}** GUILTY, how likely are the bruises to the arms and legs, the torn frenulum, and the presence of blood in the lung in Andy?

Figure 8: Step-by-step (SBS) prosecution questions. Each query is issued immediately after the corresponding evidence item (P1–P4) is presented.

A.1.8 VERDICT ELICITATION PROMPT

```
Please state your verdict for the defendant (Guilty / Not
guilty).
```

Figure 9: Prompt used to elicit the model's final verdict.

A.2 ACADEMIC MISCONDUCT SETTING

A.2.1 CASE SUMMARY PROMPT

```
Lena is a third-year PhD student in a cognitive neuroscience
lab.  She is working on a high-profile experiment testing a
new theory about memory.  The study involves 80 participants,
and her supervisor, Professor Mills, is under pressure to
produce publishable findings before a major grant review.

Lena runs the study and processes the data herself.  The first
draft of the paper reports strong, statistically significant
support for the theory:  the main effect is in the predicted
direction with p = .01, and several secondary analyses also
yield clean results.  The paper is submitted to a top journal
and receives enthusiastic initial reviews.

However, concerns are raised after an anonymous whistleblower
emails the department, claiming that some of Lena's data
look unusually consistent and that key analyses were changed
after reviewers' comments.  An internal inquiry is opened
into possible research misconduct, including fabrication or
falsification of data.

During the inquiry, multiple pieces of information emerge.
Some patterns in the dataset and analysis history appear
suspicious, but there are also plausible alternative
explanations:  the lab recently switched analysis software;
some files are missing due to a hard drive failure; and Lena
has a previously strong record of careful data handling.

Lena maintains that any irregularities are the result of
honest mistakes or miscommunication rather than intentional
misconduct.  The evaluation centers on whether Lena
deliberately manipulated or fabricated data in this study,
as opposed to the results being genuine and any errors being
unintentional.
```

Figure 10: Initial case summary prompt for the academic misconduct scenario.

A.2.2 ACADEMIC MISCONDUCT DEFENSE EVIDENCE

(D1) The lab's data management logs show that Lena ran and stored multiple alternative analysis scripts with varying exclusion criteria as part of a lab-wide effort to assess robustness. Other students in the lab exhibit similar patterns in their analysis scripts.

(D2) Two participants in the original dataset reported falling asleep during the task or not understanding the instructions. Both participants were excluded based on notes in the raw data files, and these notes predate the journal submission.

(D3) The lab switched from one statistical software package to another midway through the project, and several files were lost following a hard drive failure on Lena's laptop. The IT department confirms that a hardware failure was reported and repaired during this period.

(D4) Lena has no prior record of data irregularities. Her previous studies in the same lab passed an independent data audit conducted the year before, and her raw data and scripts were rated as well-documented and reproducible.

Figure 11: Defense evidence for the academic misconduct case. Each sub-entry corresponds to a distinct evidentiary claim.

A.2.3 ACADEMIC DEFENSE EOS QUESTIONS

In the EOS setting, the questions in Figure 12 are prompted to the LLM immediately after all defense evidence in Figure 11 is presented.

If Lena **{did/did NOT}** commit academic misconduct, how likely is it that the lab's data management logs would show her running and storing multiple alternative analysis scripts with different exclusion criteria similar to other students, that two participants would be excluded based on notes predating the journal submission, that the lab would switch software mid-project and her laptop would suffer a confirmed hard drive failure leading to missing files, and that her previous studies would have passed an independent data audit and been rated as well-documented and reproducible?

Figure 12: End-of-sequence (EOS) defense question for the academic misconduct case.

A.2.4 ACADEMIC DEFENSE SBS QUESTIONS

In the SBS setting, the model is queried sequentially, immediately after each individual piece of defense evidence is presented. Each query conditions only on the evidence observed up to that point, allowing isolation of incremental belief updating.

After (D1): If Lena **{did/did NOT}** commit academic misconduct, how likely is it that the lab's data management logs would show her running and storing multiple alternative analysis scripts with different exclusion criteria, similar to other students?

After (D2): If Lena **{did/did NOT}** commit academic misconduct, how likely is it that two participants who reported falling asleep or not understanding the instructions would be excluded based on notes that predate the journal submission?

After (D3): If Lena **{did/did NOT}** commit academic misconduct, how likely is it that the lab would switch software during the project and that Lena's laptop would suffer a confirmed hard drive failure, leading to some missing files?

After (D4): If Lena **{did/did NOT}** commit academic misconduct, how likely is it that her previous studies would have passed an independent data audit and been rated as well-documented and reproducible?

Figure 13: Step-by-step (SBS) defense questions for the academic misconduct case. Each query is issued immediately after the corresponding evidence item (D1–D4) is presented.

A.2.5 ACADEMIC PROSECUTION EVIDENCE

(P1) The final dataset exhibits an unusually high concentration of p-values just below the conventional significance threshold (e.g., .047, .049, .041) across several exploratory analyses, with very few p-values just above .05.

(P2) File timestamps indicate that the preregistration document was modified after data collection had been completed, altering some planned exclusion criteria and primary outcomes.

(P3) An earlier version of the analysis script, recovered from an automated backup, includes additional participants who were subsequently excluded from the final dataset.

(P4) In private emails to a friend, Lena wrote that her supervisor was counting on the study to secure a grant, that a null result would severely harm her postdoctoral prospects, and that she needed the study to succeed.

Figure 14: Prosecution evidence for the academic misconduct case. Each sub-entry corresponds to a distinct evidentiary claim.

A.2.6 ACADEMIC PROSECUTION EOS QUESTIONS

In the EOS setting, the questions in Figure 15 are prompted to the LLM immediately after all prosecution evidence in Figure 14 is presented.

```
If Lena {did/did NOT} commit academic misconduct, how likely
is it that the final dataset would show an unusually high
concentration of p-values just below .05 and very few just
above .05, that the preregistration document would be modified
after data collection to change exclusion criteria and primary
outcomes, that an earlier analysis script would include
participants who were later excluded from the final dataset,
and that Lena would write emails stating that her supervisor
was counting on the study, that a null result would harm her
career prospects, and that she needed the study to succeed?
```

Figure 15: End-of-sequence (EOS) prosecution question for the academic misconduct case.

A.2.7 ACADEMIC PROSECUTION SBS QUESTIONS

In the SBS setting, the model is queried sequentially, immediately after each individual piece of prosecution evidence is presented. Each query conditions only on the evidence observed up to that point, enabling analysis of incremental belief updating.

```
After (P1): If Lena {did/did NOT} commit academic misconduct,
how likely is it that the final dataset would show an
unusually high concentration of p-values just below .05 and
very few just above .05 across multiple analyses?

After (P2): If Lena {did/did NOT} commit academic misconduct,
how likely is it that the preregistration document would
be modified after data collection to change some exclusion
criteria and primary outcomes?

After (P3): If Lena {did/did NOT} commit academic misconduct,
how likely is it that an earlier analysis script would include
participants who were later excluded from the final dataset?

After (P4): If Lena {did/did NOT} commit academic misconduct,
how likely is it that she would write emails stating that her
supervisor was counting on the study, that a null result would
hurt her postdoctoral prospects, and that she needed the study
to succeed?
```

Figure 16: Step-by-step (SBS) prosecution questions for the academic misconduct case. Each query is issued immediately after the corresponding evidence item (P1–P4) is presented.

A.2.8 VERDICT ELICITATION PROMPT

```
Please state whether you think Lena committed academic
misconduct (Yes / No).
```

Figure 17: Prompt used to elicit the model's final verdict in the academic misconduct case.

A.3 VANDALISM CASE SETTING

A.3.1 CASE SUMMARY PROMPT

```
In a small open-plan office, the only shared printer suddenly
stops working late in the afternoon.  When employees attempt
to print, the machine displays a critical hardware error and
will not restart.  The side panel is slightly ajar, and the
paper tray is cracked.

Alex, a mid-level employee who frequently uses this printer,
becomes the focus of informal suspicion.  Earlier in the day,
Alex had complained about the printer repeatedly jamming, and
at least one coworker recalls seeing Alex near the printer
shortly before it stopped working.  Later, small plastic
fragments are found in the trash can beneath Alex's desk.

However, additional information complicates the situation.
Building access logs show that Alex badged out of the office
shortly before the time of the critical error recorded in the
printer log, though the logs do not indicate the condition of
the printer when Alex last used or approached it.  Security
footage also shows another coworker, Jamie, tugging forcefully
on a jammed sheet of paper in the printer earlier that
afternoon, but does not capture the final moments before the
failure occurred.

The building technician notes that this printer model has a
known history of failure due to a defective fuser unit.  Alex
has also previously reported issues with office equipment.
These background facts do not, on their own, establish whether
the printer malfunction resulted from ordinary use, rough
handling by one of the employees, or an underlying mechanical
defect.

Alex is being evaluated for whether their actions caused the
printer to break (for example, through reckless or intentional
handling), as opposed to the breakdown being attributable to
ordinary use, another individual's actions, or an underlying
mechanical failure.
```

Figure 18: Initial case summary prompt for the vandalism scenario.

A.3.2 VANDALISM DEFENSE EVIDENCE

(D1) Building access logs show that Alex badged out of the office at 5:12 pm. The printer's internal error log indicates that the critical hardware error occurred at 5:19 pm, seven minutes after Alex had left the building.

(D2) Security camera footage from earlier that afternoon shows another coworker, Jamie, pulling repeatedly and somewhat forcefully on a jammed sheet of paper in the printer. Jamie is the last clearly visible person using the printer on the footage prior to the failure.

(D3) The building technician reports that this printer model has a documented issue with its fuser unit. The same model has failed in a similar manner in two other departments within the past six months, in both cases without evidence of improper handling.

(D4) Three days before the breakdown, Alex emailed facilities reporting that the printer was making a grinding noise and might require maintenance. In prior incidents involving broken office equipment, Alex has consistently reported problems promptly and cooperated with maintenance staff.

Figure 19: Defense evidence for the vandalism case. Each sub-entry corresponds to a distinct evidentiary claim.

A.3.3 VANDALISM DEFENSE EOS QUESTIONS

In the EOS setting, the questions in Figure 20 are prompted to the LLM immediately after all defense evidence in Figure 19 is presented.

If Alex **{did/did NOT}** break the printer, how likely is it that he would have badged out of the building seven minutes before the critical error recorded in the printer's log, that security camera footage would show Jamie pulling repeatedly and somewhat forcefully on a jammed sheet of paper earlier that afternoon, that the printer model would have a documented history of failing in the same way in other departments, and that Alex would have previously emailed facilities about a grinding noise and shown a pattern of promptly reporting equipment problems?

Figure 20: End-of-sequence (EOS) defense question for the vandalism case.

A.3.4 VANDALISM DEFENSE SBS QUESTIONS

In the SBS setting, the model is queried sequentially, immediately after each individual piece of defense evidence is presented. Each query conditions only on the evidence observed up to that point, allowing analysis of incremental belief updating.

After (D1): If Alex **{did/did NOT}** break the printer, how likely is it that he badged out of the building seven minutes before the critical error recorded in the printer's log?

After (D2): If Alex **{did/did NOT}** break the printer, how likely is it that security camera footage would show Jamie pulling repeatedly and somewhat forcefully on a jammed sheet of paper in the printer earlier that afternoon?

After (D3): If Alex **{did/did NOT}** break the printer, how likely is it that this printer model would have a documented history of failing in the same way in other departments without evidence of improper handling?

After (D4): If Alex **{did/did NOT}** break the printer, how likely is it that he would have emailed facilities about a grinding noise days earlier and have a prior pattern of promptly reporting broken equipment?

Figure 21: Step-by-step (SBS) defense questions for the vandalism case. Each query is issued immediately after the corresponding evidence item (D1–D4) is presented.

A.3.5 VANDALISM PROSECUTION EVIDENCE

(P1) A coworker reports seeing Alex standing directly next to the printer approximately five minutes before it stopped working, and did not observe anyone else approach the printer during that interval.

(P2) Earlier that day, Alex posted in the team chat that the printer had jammed again and stated, \If it does this one more time I'm going to lose it." Several coworkers reacted with laughing emojis.

(P3) After the printer failure, facilities staff found small plastic fragments in the trash can under Alex's desk. The technician notes that the fragments appear similar in color and texture to the broken edge of the printer's paper tray, though an exact match was not confirmed.

(P4) When first asked informally about the printer, Alex stated that they had not used it that day. In a later conversation, Alex revised their statement, saying that they might have printed something quickly earlier in the morning but were not certain.

Figure 22: Prosecution evidence for the vandalism case. Each sub-entry corresponds to a distinct evidentiary claim.

A.3.6 VANDALISM PROSECUTION EOS QUESTIONS

In the EOS setting, the questions in Figure 23 are prompted to the LLM immediately after all prosecution evidence in Figure 22 is presented.

```
If Alex {did/did NOT} break the printer, how likely is it
that a coworker would see him standing next to the printer
about five minutes before it stopped working with no one else
approaching in that interval, that he would post in the team
chat complaining about the printer and stating he was going
to lose it if it jammed again, that plastic fragments similar
in appearance to the broken tray would be found in the trash
can under his desk, and that he would initially deny using
the printer that day before later saying he might have printed
something that morning?
```

Figure 23: End-of-sequence (EOS) prosecution question for the vandalism case.

A.3.7 VANDALISM PROSECUTION SBS QUESTIONS

In the SBS setting, the model is queried sequentially, immediately after each individual piece of prosecution evidence is presented. Each query conditions only on the evidence observed up to that point, enabling analysis of incremental belief updating.

```
After (P1): If Alex {did/did NOT} break the printer, how
likely is it that a coworker would see him standing next to
the printer about five minutes before it stopped working, with
no one else approaching in that interval?

After (P2): If Alex {did/did NOT} break the printer, how
likely is it that he would have posted in the team chat
complaining about the printer jamming and stating that he was
going to lose it if it jammed again?

After (P3): If Alex {did/did NOT} break the printer, how
likely is it that plastic fragments similar in appearance
to the broken tray would be found in the trash can under his
desk?

After (P4): If Alex {did/did NOT} break the printer, how
likely is it that he would initially deny using the printer
that day and later say that he might have printed something
that morning?
```

Figure 24: Step-by-step (SBS) prosecution questions for the vandalism case. Each query is issued immediately after the corresponding evidence item (P1–P4) is presented.

A.3.8 VERDICT ELICITATION PROMPT

```
Please state whether you think Alex broke the printer (Yes /
No).
```

Figure 25: Prompt used to elicit the model's final verdict in the vandalism case.