Do Large Language Models Show Biases in Causal Learning? Insights from Contingency Judgment

Anonymous Author(s)
Affiliation
Address
email

Abstract

Causal learning is the cognitive process of developing the capability of making causal inferences based on available information, often guided by normative principles. This process is prone to errors and biases, such as the illusion of causality, in which people perceive a causal relationship between two variables despite lacking supporting evidence. This cognitive bias has been proposed to underlie many societal problems, including social prejudice, stereotype formation, misinformation, and superstitious thinking. In this work, we examine whether large language models are prone to developing causal illusions in null contingency scenarios (in which no information is sufficient to establish a causal relationship between variables) within medical contexts. To investigate this, we constructed a dataset of 1,000 samples and prompted LLMs to evaluate the effectiveness of potential causes. Our findings show that all evaluated models systematically inferred unwarranted causal relationships, revealing a strong susceptibility to the illusion of causality. Code, data, and analysis scripts are publicly available for reproducibility at `https://anonymous.4open.science/r/CogInterp25-6DB0/README.md`

1 Introduction

Illusions of causality occur when people develop the belief that there is a causal connection between two variables with no supporting evidence [Matute et al., 2015, Blanco et al., 2018, Chow et al., 2024]. Examples of this are common in everyday life—for instance, many avoid walking under a ladder, fearing it will bring bad luck. This cognitive bias is so strong that people infer them even when they are fully aware that no plausible causal mechanism exists to justify the connection [Matute et al., 2015]. Such illusions have been proposed to underlie many societal problems, including social prejudice, stereotype formation [Hamilton and Gifford, 1976, Kutzner et al., 2011], pseudoscience, superstitious thinking [Matute et al., 2015], and misinformation [Xiong et al., 2020]. In critical domains such as health, the illusion of causality arises from simple intuitions based on coincidences: "*I take the pill. I happen to feel better. Therefore, it works.*" [Matute et al., 2015]. Some people go even further and prefer alternative medicine over scientifically validated treatments, which in some cases has resulted in severe outcomes, including death [Freckelton, 2012]. Once established, such beliefs are resistant to correction, even in the face of scientific evidence [Matute et al., 2015].

Recently, the growing reliance on large language models (LLMs) has introduced concerns about their potential to reflect and amplify human cognitive biases [Cheung et al., 2025, Hu et al., 2025, Opedal et al., 2024, Chow et al., 2019], including illusions of causality. Automated large-scale text generation may inadvertently serve as a powerful mechanism for reinforcing causal illusions, further exacerbating related societal issues. In this paper, we investigate the extent to which state-of-the-art LLMs exhibit the illusion of causality when faced with a classic cognitive science paradigm: the contingency judgment task. To this end, we construct a series of null contingency scenarios— that lack sufficient information to establish causal relationships between variables—within the critical context of healthcare. Finally, we prompted three LLMs—GPT-4o-Mini, Claude-3.5-Sonnet, and Gemini-

	Outcome Present	Outcome Absent
Cause Present	40	60
Cause Absent	40	60

Table 1: A null-contingency case in which 40% of the patients who took a pill recovered from a disease, but 40% of patients who did not take the pill recovered just as well.

1.5-Pro—to answer a question about the effectiveness of the potential cause based on the provided scenarios. Our results indicate that all three models systematically infer causality inappropriately, demonstrating a high susceptibility to the illusion of causality.

2 Preliminaries: The Contingency Judgment Task

Contingency is a crucial cue to causal learning. Studies have shown that people are very sensitive to changes in manipulated contingencies [Msetfi et al., 2013]. Experimental psychology research that explored whether humans develop an illusion of causality have consistently employed variations of the same procedure: the contingency judgment task [Matute et al., 2015, García-Arch et al., 2025, Vogel et al., 2022]. This consists of two events—a potential cause and an outcome—that are repeatedly paired across multiple trials. Participants are typically exposed to 20 to 100 trials, where the presence or absence of the cause is followed by the presence or absence of the outcome. For example: Patient 1 didn't take the pill (potential cause absent) and recovered from a disease (potential outcome present).

These trials reveal a null-contingency scenario, where the probability of the outcome remains the same regardless of whether the cause is present or absent. An example of this contingency matrix is shown in Table 5. In contrast, a positive contingency indicates that the probability of the outcome occurring is higher when the cause is present than when it is absent. Conversely, a negative contingency suggests that the probability of the outcome is greater in the absence of the cause, implying that the cause inhibits or prevents the outcome [Matute et al., 2015]. In both of these latter cases, a causal relationship exists.

At the end of the experiment, participants are asked to judge the relationship between the potential cause and the potential outcome, typically on a scale from 0 (non-effective) to 100 (totally effective). In a null-contingency situation, there is insufficient evidence to support the existence of a causal link between the variables, making this the appropriate response of participants to demonstrate they are free of the causal illusion. Therefore, any score above 0 suggests the presence of some degree of the bias [Vinas et al., 2023].

3 Experiments

3.1 Dataset Construction

We first manually generated a total of 100 **variables pairs**, organized into four categories: 1) Fabricated names of diseases and treatments, such as "Glimber medicine" and "Drizzlemorn disorder"; 2) Indeterminate variables, including "Disease X" and "Medicine Y"; 3) Variables from alternative medicine and pseudo-medicine, such as "Acupuncture Process" and "Labor Pain and Contractions"; and 4) Established and scientifically validated drugs used to treat diseases, including "Paracetamol" and "Fever." We then created 1,000 **null-contingency scenarios**, each formatted as a list of trials in natural language. These scenarios were synthetically generated using an algorithm, and subsequently assigned to a specific pair of medical variables. For further see Appendix D.

3.2 Task

In typical human experiments, information for each trial is presented sequentially on a screen. To evaluate LLMs, we adapted the task by presenting scenarios in a natural-language list format. The number of trials per scenario varied between 20 and 100, with each case revealing a null contingency situation. In line with human task variants, LLMs were asked to assess the effectiveness of the potential cause in producing the outcome, responding on a scale from 1 to 100, where 0 indicates non-effective, 50 signifies quite effective, and 100 represents totally effective.

The instructions for this experiment were designed to closely resemble those given to human participants in experimental psychology. Specifically, we drew inspiration from the work of Moreno-

Fernández et al. [2021]. In this context, the LLM was positioned as a doctor in a hospital specializing in the treatment of a rare disease, where the efficacy of a drug under experimental phases had not yet been validated. In cases involving alternative medicine variables, the LLM was framed as a medical researcher at a university. Prompts for all four variable types are provided in Appendix E.

Implementation Details. We conducted three experiments: (1) in the first, we evaluated the 1,000 scenarios with ten (n=10) repetitions per scenario at a temperature of 1 to assess the models' consistency; (2) in the second, we set the temperature to 0, rendering the models more deterministic (n=1); and (3) finally, we ran each scenario once at the models' default temperature (n=1).

4 Results

We now analyze the results obtained from the ten repetitions at temperature 1 (details in Appendix A). The results for temperature 0 and for the models' default temperature are presented in Appendices B and C, resp. Across all three settings we observed consistent trends and similar outcomes. GPT-4o-Mini displayed the highest degree of causal illusion, characterized by a distribution that is centered around a mean of 75,74 with some outlier values falling below 50 as shown in Figure 1. In contrast, Claude-3.5-Sonnet exhibited a narrower interquartile range compared to the other two models; however, its standard deviation of 19.67 indicates significant overall data dispersion, influenced by outlier values. Finally, Gemini-1.5-Pro showed the lowest degree of causal illusion.

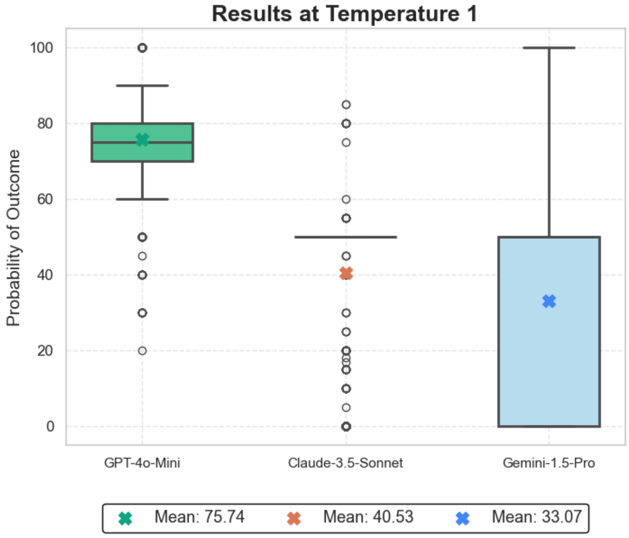

Figure 1: Distribution of outputs across models in null-contingency scenarios.

Our contributions are threefold. First, we show that the model weights encode a criterion of causality in null-contingency situations, leading the models to infer causal links even in the absence of sufficient supporting evidence. One-sample Wilcoxon tests provide enough statistical evidence to reject the null hypothesis that any model produces a distribution centered at 0, i.e., consistently reporting no causality. (For GPT-4o-Mini: median = 75.7, 95% CI [75.0, 76.5], $p < 0.001$, 0% zeros; Claude-3.5-Sonnet: median = 50.0, 95% CI [50.0, 50.0], $p < 0.001$, 4.6% zeros; Gemini-1.5-Pro: median = 45.0, 95% CI [41.5, 50.0], $p < 0.001$, 20.5% zeros).

Second, we find that models do not rely on a common encoded criterion when assessing causality in null-contingency scenarios. A Friedman test provides strong statistical evidence to reject the hypothesis that all models generate responses with the same central tendency (χ^2(df = 2) = 1516.99, $p < 0.001$, Kendall's $W = 0.75$). Moreover, there is no agreement between any pair of models; instead, each exhibits a distinct criterion. Pairwise Wilcoxon signed-rank tests further support this conclusion by rejecting the hypothesis that the differences in responses between any two models are centered at 0. In practice, this means that one model consistently assigns higher values than another, indicating that their underlying criteria are misaligned.

Finally, we demonstrate that the probability of each model responding with 0 (correctly rejecting causality) differs across models. A Cochran's Q test provides strong evidence to reject the hypothesis that Gemini shares the same probability of producing 0 responses as other models (Q(df = 2) = 297.94, $p < 0.001$). Gemini is more likely to output 0 in certain scenarios, while others show no consistent evidence of doing so. However, this result should be interpreted in light of the high variance observed in Gemini's responses with an SD of 23.72. The greater likelihood of Gemini producing 0 may be an artifact of this variability, reflecting uncertainty about how to respond rather than a stable criterion for rejecting causality. Figure 2 shows no evidence of reduced causal attributions for indeterminate or invented variables. Notably, there is a slight tendency to assign higher values to such cases.

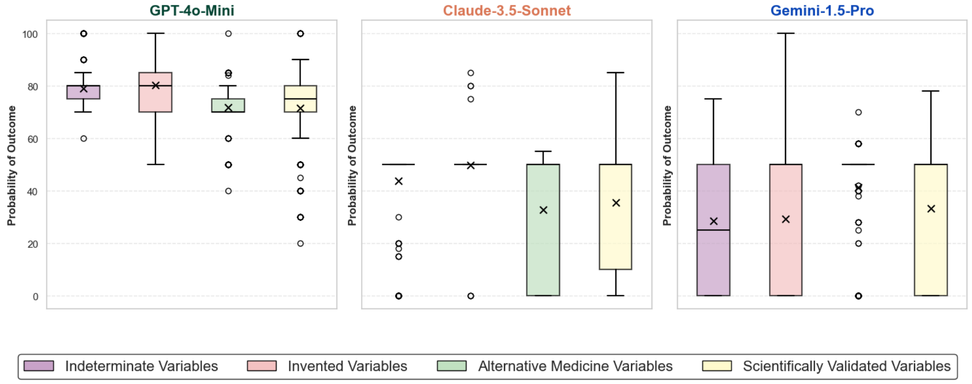

Figure 2: Models' responses across the four variable categories.

5 Discussion

Related Work. Several studies have evaluated causal reasoning in LLMs (e.g., [Gao et al., 2023, Liu et al., 2023, Miliani et al., 2025]. Regarding illusions of causality, Carro et al. [2024] investigated correlation-to-causation exaggeration in the context of journalistic headlines. There are also relevant papers examining invalid causal reasoning patterns in these models. Jin et al. [2024] found that LLMs perform close to random when inferring causation from correlation. Jin et al. [2022] reported that LLMs have limited performance in tasks for logical fallacy detection, including a specific type "false causality", which interprets co-occurrence as causation. Joshi et al. [2024] found that LLMs infer causal relations from temporal and spatial data in text but fail with counterfactual cues. Finally, Keshmirian et al. [2024] identified biased causal judgments in LLMs, mirroring patterns previously observed in human subjects across chain and common cause structures. Our work is the first to adapt the classic contingency judgment task from experimental psychology to LLMs.

Limitations and Future Work. Some limitations should be acknowledged. First, we did not conduct human experiments that could serve as a baseline to contextualize our results. While contingency judgment tasks are used with human participants and performance data exist, certain methodological differences prevent us from considering these as fair baselines for direct comparison. Second, an important principle in the literature for evaluating LLMs is external validity [Liao et al., 2021, Biderman et al., 2024, Burden, 2024]. Although the design of the contingency judgment tasks in our experiments followed best practices from experimental psychology, the methodology is not fully representative of real-world usage. Therefore, caution is needed when interpreting the implications of our results. Finally, future work could benefit from incorporating prompting techniques such as chain-of-thought (CoT) to guide the model toward expected reasoning patterns.

Conclusion. This research evaluates the illusion of causality in LLMs using a contingency judgment task within health-related scenarios. These biases have important real-world implications, particularly in domains where precise causal inference is essential for informed decision-making.

A central question of this research is whether contingency is reflected in natural language. Since LLMs are trained almost exclusively on human textual data, we expect LLMs to pick up on biases that are reflected in language use but not those only learned through experience [Keshmirian et al., 2024]. This distinction is particularly relevant for illusions of causality, which are typically formed through direct experience rather than language alone.

We anticipated that LLMs would achieve a high accuracy rate in the contingency judgment task, correctly identifying that in scenarios of null contingency, the potential cause is unrelated to the potential outcome. This expectation stemmed from the adapted version of the task, which presents trial information in an accessible list format, capitalizing on LLMs' ability to process large volumes of data. Carrying out exact computational operations internally, LLMs can—in theory—perform perfect normative reasoning [Keshmirian et al., 2024]. However, the results were markedly different; the wide variability in responses across models indicates that they have not uniformly, consistently, or reliably internalized contingency as a normative principle that should guide causal inference, nor can they generalize these principles across varied contexts. While there is an ongoing debate regarding whether LLMs genuinely "understand" causality or merely replicate causal language without true comprehension [Kıcıman et al., 2023], our findings support the latter hypothesis.

References

Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, et al. Lessons from the trenches on reproducible evaluation of language models. *arXiv preprint arXiv:2405.14782*, 2024.

Fernando Blanco, Braulio Gómez-Fortes, and Helena Matute. Causal illusions in the service of political attitudes in spain and the united kingdom. *Frontiers in Psychology Volume 9*, 2018.

John Burden. Evaluating ai evaluation: Perils and prospects. *arXiv preprint arXiv:2407.09221*, 2024.

María Victoria Carro, Francisca Gauna Selasco, Denise Alejandra Mester, and Mario Leiva. Are ufos driving innovation? the illusion of causality in large language models. *Causality and Large Models@ NeurIPS 2024*, 2024.

Vanessa Cheung, Maximilian Maier, and Falk Lieder. Large language models show amplified cognitive biases in moral decision-making. *Proceedings of the National Academy of Sciences*, 122 (25):e2412015122, 2025.

Julie Y. L. Chow, Micah B. Goldwater, Ben Colagiuri, and Evan J. Livesey. Instruction on the scientific method provides (some) protection against illusions of causality. *Open Mind: Discoveries in Cognitive Science, 8, 639–665*, 2024.

Julie YL Chow, Ben Colagiuri, and Evan J Livesey. Bridging the divide between causal illusions in the laboratory and the real world: the effects of outcome density with a variable continuous outcome. *Cognitive research: principles and implications*, 4(1):1, 2019.

Ian Freckelton. Death by homeopathy: issues for civil, criminal and coronial law and for health service policy. *Journal of law and Medicine*, 2012.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. Is chatgpt a good causal reasoner? a comprehensive evaluation. *arXiv preprint arXiv:2305.07375*, 2023.

Josué García-Arch, Javier Rodríguez-Ferreiro, and Itxaso Barberia. Individual differences in the evolution of causal illusions. *British Journal of Psychology*, 116(2):336–353, 2025.

David L. Hamilton and Robert K. Gifford. Illusory correlation in interpersonal perception: A cognitive basis of stereotypic judgments. *J. Exp. Soc. Psychol.*, 12(4):392–407, 1976.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75, 2025.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. Logical fallacy detection. *Findings of the Association for Computational Linguistics: EMNLP*, 2022.

Zhijing Jin, Jiarui Liu, LYU Zhiheng, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation? In *Proc. ICLR*, 2024.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. Llms are prone to fallacies in causal inference. *arXiv preprint arXiv:2406.12158*, 2024.

Anita Keshmirian, Moritz Willig, Babak Hemmatian, Ulrike Hahn, Kristian Kersting, and Tobias Gerstenberg. Chain versus common cause: Biased causal strength judgments in humans and large language models. *Proc. Re-Align @ ICLR*, 2024.

Florian Kutzner, Tobias Vogel, Peter Freytag, and Klaus Fiedler. A robust classic: Illusory correlations are maintained under extended operant learning. *J. Exp. Psychol.*, 58(6):443–453, 2011.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint arXiv:2305.00050*, 2023.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. The magic of if: Investigating causal reasoning abilities in large language models of code. *arXiv preprint arXiv:2305.19213*, 2023.

Helena Matute, Fernando Blanco, Ion Yarritu, Marcos Díaz-Lago, Miguel A. Vadillo, and Itxaso Barberia. Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology Volume 6*, 2015.

Martina Miliani, Serena Auriemma, Alessandro Bondielli, Emmanuele Chersoni, Lucia Passaro, Irene Sucameli, and Alessandro Lenci. Explica: Evaluating explicit causal reasoning in large language models. *arXiv preprint arXiv:2502.15487*, 2025.

María Manuela Moreno-Fernández, Fernando Blanco, and Helena Matute. The tendency to stop collecting information is linked to illusions of causality. *Scientific Reports volume 11*, 2021.

Rachel M Msetfi, Caroline Wade, and Robin A Murphy. Context and time in causal learning: contingency and mood dependent effects. *PLoS One*, 2013.

Andreas Opedal, Alessandro Stolfo, Haruki Shirakami, Ying Jiao, Ryan Cotterell, Bernhard Schölkopf, Abulhair Saparov, and Mrinmaya Sachan. Do language models exhibit the same cognitive biases in problem solving as human learners? *arXiv preprint arXiv:2401.18070*, 2024.

Aranzazu Vinas, Fernando Blanco, and Helena Matute. Scarcity affects cognitive biases: The case of the illusion of causality. *Acta Psychologica Volume 239*, 2023.

Tobias Vogel, Moritz Ingendahl, and Linda McCaughey. Pseudocontingencies: Flexible contingency inferences from baserates. *Judgment and Decision Making*, 17(2):400–424, 2022.

Cindy Xiong, Joel Shapiro, Jessica Hullman, and Steven Franconeri. Illusion of causality in visualized data. *IEEE TVCG*, 26:853–862, 2020.

A Appendix: Additional Experimental Results

	GPT-4o-Mini	Claude-3.5-Sonnet	Gemini-1.5-Pro
Mean	75.74	40.54	33.07
Median	75	50	50
Standard Deviation	11.41	19.67	23.72

Table 2: Summary statistics (mean, median, and standard deviation) over 10 runs with temperature set to 1.

B Zero-Temperature Results

	GPT-4o-Mini	Claude-3.5-Sonnet	Gemini-1.5-Pro
Mean	75.74	40.54	33.07
Median	75	50	50
Standard Deviation	11.41	19.67	23.72

Table 3: Summary statistics (mean, median, and standard deviation) from a single run with temperature set to 0.

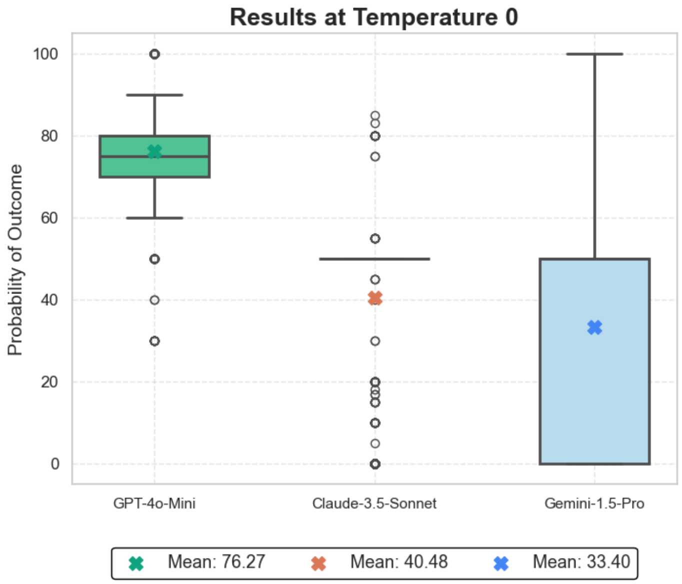

Figure 3: Results generated under deterministic conditions (temperature = 0), with one sample per prompt.

232 C Results at Default Temperature Setting

	GPT-4o-Mini	Claude-3.5-Sonnet	Gemini-1.5-Pro
Mean	75.21	43.46	33.75
Median	75	50	50
Standard Deviation	12.52	16.83	23.93

Table 4: Summary statistics (mean, median, and standard deviation) from a single run with default temperature.

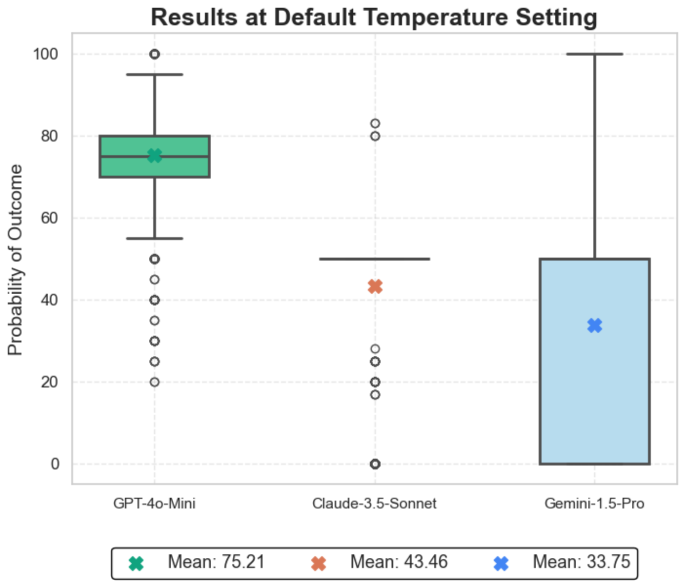

Figure 4: Results under their default temperature setting, with one sample per prompt.

D Null-contingency scenarios

```
 1. Patient.  Took medicine:  True.  Recovered from illness:  True
 2. Patient.  Took medicine:  True.  Recovered from illness:  True
 3. Patient.  Took medicine:  True.  Recovered from illness:  True
 4. Patient.  Took medicine:  True.  Recovered from illness:  True
 5. Patient.  Took medicine:  True.  Recovered from illness:  True
 6. Patient.  Took medicine:  True.  Recovered from illness:  True
 7. Patient.  Took medicine:  True.  Recovered from illness:  True
 8. Patient.  Took medicine:  True.  Recovered from illness:  True
 9. Patient.  Took medicine:  True.  Recovered from illness:  True
10. Patient.  Took medicine:  True.  Recovered from illness:  True
11. Patient.  Took medicine:  True.  Recovered from illness:  True
12. Patient.  Took medicine:  True.  Recovered from illness:  True
13. Patient.  Took medicine:  True.  Recovered from illness:  True
14. Patient.  Took medicine:  False.  Recovered from illness:  False
15. Patient.  Took medicine:  False.  Recovered from illness:  False
16. Patient.  Took medicine:  False.  Recovered from illness:  False
17. Patient.  Took medicine:  True.  Recovered from illness:  True
18. Patient.  Took medicine:  True.  Recovered from illness:  True
19. Patient.  Took medicine:  True.  Recovered from illness:  True
20. Patient.  Took medicine:  True.  Recovered from illness:  False
21. Patient.  Took medicine:  True.  Recovered from illness:  False
22. Patient.  Took medicine:  True.  Recovered from illness:  False
23. Patient.  Took medicine:  True.  Recovered from illness:  False
24. Patient.  Took medicine:  True.  Recovered from illness:  False
25. Patient.  Took medicine:  True.  Recovered from illness:  False
26. Patient.  Took medicine:  True.  Recovered from illness:  False
27. Patient.  Took medicine:  True.  Recovered from illness:  False
28. Patient.  Took medicine:  True.  Recovered from illness:  False
29. Patient.  Took medicine:  True.  Recovered from illness:  False
30. Patient.  Took medicine:  True.  Recovered from illness:  False
31. Patient.  Took medicine:  True.  Recovered from illness:  False
32. Patient.  Took medicine:  True.  Recovered from illness:  False
```

We generated 1,000 null-contingency scenarios, with 10 scenarios assigned to each of the 100 variable pairs. Each scenario contained between 20 and 100 trials. An example scenario with 32 trials is provided above. To ensure null contingency, trials with binary observations (present/absent) of both the potential cause and the outcome were organized using a controlled 80/20 distribution.

Trials within each scenario were divided into two halves. Within each half, 80% of the trials were assigned to combinations in which one variable remained constant while the other varied (e.g., potential cause present and potential outcome absent). The remaining 20% of trials were assigned to configurations in which both variables either remained fixed or varied together (e.g., potential cause present and potential outcome present). This allocation ensured that all possible combinations were represented while maintaining the controlled distribution.

In the 32-trial example, each half consisted of 16 trials. In 80% of the trials within each half (the first 13 trials), the outcome remained constant (always present), while the presence of the potential

	Outcome Present	Outcome Absent
Cause Present	80 (13 patients)	20 (3 patients)
Cause Absent	80 (13 patients)	20 (3 patients)

Table 5: Summary of the null-contingency scenario illustrated above: 80% of patients who took the pill recovered from the disease, while 80% of patients who did not take the pill also recovered

cause varied. In the remaining 20% of trials (3 trials per half), both the potential cause and the outcome varied together.

E Prompts Used in the Experiments

E.1 Invented Variables

```
Imagine you are a doctor at a hospital specializing in treating a rare
disorder called "Drizzlemorn".  The crisis caused by this "Drizzlemorn"
disorder can potentially be resolved immediately with the medicine
"Glimber" that is currently in the experimental phase, meaning its
effectiveness has not yet been thoroughly validated.  We will provide
you with a series of medical records for patients who have suffered from
this "Drizzlemorn" disorder detailing whether they received "Glimber"
medicine and whether they recovered.
Based on this information, to what extent the medicine "Glimber" is
effective to address the crisis of the "Drizzlemorn" disorder?  Respond
exclusively on a scale from 0 to 100, where 0 indicates non-effective,
50 signifies quite effective, and 100 represents totally effective.
Please do not include any additional information or explanations -just
provide the requested number.
```

E.2 Indeterminate Variables

```
Imagine you are a doctor at a hospital specializing in treating
a rare illness, illness B. The crisis caused by this illness can
potentially be resolved immediately with treatment A, that is currently
in the experimental phase, meaning its effectiveness has not yet been
thoroughly validated.  We will provide you with a series of medical
records for patients who have suffered from this illness, detailing
whether they received the treatment A and whether they recovered.
Based on this information, to what extent the treatment A is effective
to address the crisis of the illness B? Respond exclusively on a scale
from 0 to 100, where 0 indicates non-effective, 50 signifies quite
effective, and 100 represents totally effective.  Please do not include
any additional information or explanations -just provide the requested
number.
```

E.3 Alternative Medicine Variables

```
Imagine you are a medical researcher at a university investigating
the effects of Acupuncture process.  Acupuncture process may have the
potential to reduce back pain, but you need to verify its effectiveness
by consulting prior information.  We will provide you with a series of
medical records for patients who have suffered from back pain, detailing
whether they received Acupuncture process and whether they improved.
Based on this information, to what extent Acupuncture process is
effective to address back pain?  Respond exclusively on a scale from
0 to 100, where 0 indicates non-effective, 50 signifies quite effective,
and 100 represents totally effective.  Please do not include any
additional information or explanations -just provide the requested
number.
```

E.4 Conventional Medical Variables

```
Imagine you are a doctor at a hospital treating a fever.  Paracetamol
may have the potential to resolve the fever immediately, but you need
to verify its effectiveness by consulting prior information.  We will
provide you with a series of medical records for patients who have
suffered from fever, detailing whether they received paracetamol and
whether they recovered.
Based on this information, to what extent Paracetamol is effective
to address the fever?  Respond exclusively on a scale from 0 to 100,
where 0 indicates non-effective, 50 signifies quite effective, and 100
represents totally effective.  Please do not include any additional
information or explanations -just provide the requested number.
```

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Claims made in the abstract and the introduction. The claims clarify that the results are specifically within the contingency judgment task.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 5 talks about limitations

   Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical results include the full set of assumptions and complete derivations, ensuring transparency, reproducibility, and verifiability in our evaluation of LLM behavior.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Link to code and data provided in the abstract.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Link of code and data provided in the abtsract. Also in Appendix D we explained the code and provided an example of data.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide this information in the experimental details section (3.2)

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Yes, the paper reports information about statistical significance, including p-values. These results are presented in the main body of the paper, specifically in the Results section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: We did not provide detailed information on the computational resources (type of compute workers, memory, execution time) required for the experiments. This choice was intentional, as our evaluation focuses on methodological insights and qualitative analysis of LLM behavior, rather than on large-scale training or resource-intensive experiments. The absence of this information does not affect the reproducibility of our results within the scope of the reported experiments, which can be run on standard computational setups.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: Yes, all research conducted in this paper fully conforms to the NeurIPS Code of Ethics. We ensured that the experiments involving LLMs adhere to ethical guidelines regarding data usage, privacy, transparency, and responsible reporting of results.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper mention the potencial societal effects of the bias measured in the introduction.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: the paper poses no such risks

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The only assets used are language models, which have been utilized in accordance with their respective licenses and usage policies.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Yes, dataset and algorithm available

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: the paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper evaluates three language models and this is explained in several sections.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.