
Reasoning Isn’t Enough: Examining Truth-Bias and Sycophancy in LLMs

Emilio Barkett
Columbia University
eab2291@columbia.edu

Olivia Long
Columbia University
o12256@columbia.edu

Madhavendra Thakur
Columbia University
mt3890@columbia.edu

Abstract

Despite their widespread use in fact-checking, moderation, and high-stakes decision-making, large language models (LLMs) remain poorly understood as judges of truth. This study presents the largest evaluation to date of LLMs’ veracity detection capabilities and the first analysis of these capabilities in reasoning models. We had eight LLMs make 4,800 veracity judgments across several prompts, comparing reasoning and non-reasoning models. We find that rates of truth-bias, or the likelihood to believe a statement is true, regardless of whether it is actually true, are lower in reasoning models than in non-reasoning models, but still higher than human benchmarks. Most concerning, we identify sycophantic tendencies in several advanced models (o4-mini and GPT-4.1 from OpenAI, R1 from DeepSeek), which displayed an asymmetry in detection accuracy, performing well in truth accuracy but poorly in deception accuracy. This suggests that capability advances alone do not resolve fundamental veracity detection challenges in LLMs.

1 Introduction

The *truth-bias*, or the perception that others are honest independent of message veracity, is one of the most replicated findings in deception research [12, 11, 17, 16]. Truth-bias is measured by calculating the proportion of messages judged to be truthful out of the total number of messages evaluated; a rate above 50% indicates a truth-bias. Its pervasiveness in humans has led to the investigation into whether truth-bias can be found in LLM judges, and if so, to what extent. Previous work showed the truth-bias in large language models (LLMs) at 67%-99%,¹ suggesting that AI judges most information to be true [16]. While prior work evaluated non-reasoning LLMs, models that predict the next token without engaging in structured reasoning, this study investigates whether reasoning LLMs, which are trained to perform step-by-step reasoning or “thinking,” exhibit a similar degree of truth-bias.

This study is relevant for several reasons. First, truth-bias is closely related to sycophancy, a phenomenon in which language models excessively agree with or flatter the user, often at the expense of factual accuracy. This became especially salient following OpenAI’s April 2025 rollback of GPT-4o [22], which was widely criticized for producing outputs that echoed user sentiments uncritically, even when those sentiments were factually incorrect or harmful. This underscores the practical consequences of sycophancy, highlighting how a model’s tendency toward agreement can amplify real-world risks, particularly in sensitive domains like health and well-being. Second, it tests the idea that reasoning models *should* outperform their non-reasoning counterparts on deception detection, a cognitively demanding task that requires deliberation [17], since reasoning models are capable of “thinking” rather than predicting the next token. Understanding whether reasoning models are more susceptible to truth-bias than non-reasoning models is crucial for evaluating their reliability and determining their suitability for epistemically demanding tasks.

¹This was aggregated across three models: GPT-3.5 from OpenAI; LaMDA from Google; GPT-4 from ChatSonic.

Third, it tests the assumption that state-of-the-art (SOTA) models will outperform previous models. While true in domains like generating code or images [9, 1], which do not necessarily involve judging statements, we test the model’s discernment of what is truthful and deceptive. Fourth, there is a concerning risk that LLMs could feed into preexisting user beliefs or nudge users toward accepting deceptive beliefs. Finally, it offers a replicable method that can serve as a model evaluation for tracking the developmental progress of LLMs.

While an experimental design to evaluate a model’s exhibited truth-bias is relatively simple, it remains substantially more difficult to understand *why* biases arise and *what* internal mechanisms are responsible. While answers to these questions fall outside the scope of the present study and into the domain of mechanistic interpretability [18], we aim to document a point-in-time evaluation of the truth-bias in non-reasoning and reasoning LLMs and leave room for future interpretability work.

This study uses a dataset of deceptive and truthful hotel reviews [23]. Across three studies, we evaluate truth-bias across *model-pairs* composed of SOTA non-reasoning and reasoning LLMs from the same firm. We evaluate GPT-4.1 from OpenAI [20], Claude 3.5 Haiku from Anthropic [2], and V3 from DeepSeek [7] as our non-reasoning models. These are paired respectively with the following reasoning models: o3 from OpenAI [21], Claude 3.7 Sonnet from Anthropic [3], and R-1 from DeepSeek [6]. Outside the main model-pairs, we also tested o4-mini [21] alongside o3 to examine the variance between models from the same release. Additionally, we used GPT-3.5 Turbo [19] to approximate prior findings [16], as the original model GPT-3.5 was no longer accessible.

On average, we find that reasoning models perform better, with a lower truth-bias (59.33%) than non-reasoning models (71.00%).² We show a marked improvement in model performance since previous work demonstrated a higher truth-bias in generative AI (aggregated across three models: GPT-3.5 from OpenAI; LaMDA from Google; GPT-4 from ChatSonic.) at 87.73% [16]. It is worth noting that the models we evaluate are over two years newer than the ones previously evaluated. Claude 3.7 Sonnet from Anthropic (44.83%), o3 from OpenAI (54.50%), and V3 from Deepseek (55.33%) exhibited the lowest truth-bias, whereas GPT-4.1 from OpenAI (90.83%), R1 from Deepseek (78.67%), and Claude 3.5 Haiku from Anthropic (66.83%) exhibited the highest truth-bias. The best performing reasoning model was o3 from OpenAI, and the best performing non-reasoning model was Claude 3.5 Haiku from Anthropic (Table 1).

In our most definitive study (Study 2), OpenAI’s SOTA reasoning model o3 demonstrated significantly lower truth-bias (49.50%) compared to the SOTA non-reasoning counterpart, GPT-4.1 (93.00%), and revealed a statistically significant difference: $z = -9.61$, $p < .001$, 95% CI = $[-52.4\%, -34.6\%]$, Cohen’s $h = -1.05$, $\chi^2(1, N = 400) = 92.38$, $p < .001$, indicating a large effect size.

We propose that reasoning models exhibit reduced truth-bias because reasoning processes emulate a form of reflective cognition that allows a model to evaluate statements more analytically. Unlike non-reasoning models, reasoning models are prompted to “slow down” via intermediate inference steps. This additional stepwise interrupts the tendency to truth-default completions, attenuating the truth-bias.

2 Background

People frequently encounter difficulty in accurately detecting deception. Instead of relying on false cues, people struggle as deception detectors because cues associated with deception are typically subtle, ambiguous, and unreliable [8, 10]. A downstream effect of this is the truth-bias: a phenomenon where a receiver infers a message as honest independent of its veracity [11, 17]. Humans show a tendency toward gullibility in communication [12], as they are likely to assume others are telling the truth when making judgments.

2.1 Truth-Bias in LLMs

The truth-bias is one of the most consistently replicated results in deception research [12], and has motivated investigations into its replicability among LLM judges. Prior work conducted the first empirical investigation into whether LLMs trained on human data had learned to be truth-biased like humans [16], as supported by truth-default theory (TDT). TDT is a pancultural theory of human

²These percentages only represent the main model-pairs and do not include o4-mini or GPT-3.5 Turbo.

Table 1: Results Across Best Performing Models

Measure	Reasoning (OpenAI o3)	Non-reasoning (Claude 3.5 Haiku)	Prior Work [16] (OpenAI GPT-3.5)
Neutral prompt (Study 1)			
Overall Accuracy	67.00%	62.50%	51.42%
Truth-Bias	57.50%	79.50%	99.36%
Truth Accuracy	75.00%	92.00%	99.75%
Deception Accuracy	59.00%	33.00%	1.05%
Veracity prompt (Study 2)			
Overall Accuracy	74.50%	58.50%	53.16%
Truth-Bias	49.50%	65.50%	97.16%
Truth Accuracy	74.00%	74.00%	98.75%
Deception Accuracy	75.00%	43.00%	4.53%
Base-rate prompt (Study 3)			
Overall Accuracy	67.00%	55.50%	61.33%
Truth-Bias	56.50%	55.50%	66.67%
Truth Accuracy	74.00%	61.00%	78.00%
Deception Accuracy	60.00%	50.00%	44.67%

Note: Overall Accuracy = (correct truths + correct lies) / total messages judged; Truth-Bias = messages judged as truthful / total messages judged; Truth Accuracy = correctly identified truths / total actual truths; Deception Accuracy = correctly identified lies / total actual lies.

deception detection which states that upon being prompted for a veracity judgment, in the absence of suspicion, people automatically assume others are honest [12, 11]. Through a replication of TDT principles across four studies, they demonstrated that non-reasoning LLMs, including GPT-3.5 from OpenAI, Bard LaMDA from Google, and GPT-4 from ChatSonic, are not only as accurate as humans in deception detection but consistently more truth-biased (67%-99%) across various prompting conditions [16].

This prior work offered worrying evidence about how LLMs detect deception relative to humans and how fundamental principles of human communication are extended to LLMs. The authors argue that the truth-bias likely emerged during pre-training on vast corpora of human language, and if so, this bias may be an emergent property of AI rather than a uniquely human trait. The authors suggest that future research should examine a broader range of LLMs to better document the prevalence of truth-bias, which they predict will persist as LLMs continue to advance.

Against this backdrop, the present study evaluates whether reasoning models—those that generate responses through structured, multi-step inference—exhibit the same truth-bias observed in non-reasoning LLMs. By extending analysis to these models, this work explores whether such structure mitigates truth-bias and enhances veracity judgments.

2.2 Sycophancy in LLMs

Sycophancy, in which models tend to agree with user beliefs instead of being truthful, has been extensively documented in LLMs [5]. A sycophantic model will excessively agree or otherwise flatter the user—this is especially dangerous, for instance, if the AI reinforces concerning, life-threatening behaviors. Prior research [24] found that sycophancy arises when human preference models prefer sycophantic responses over more truthful ones, and generally speaking, agreeableness and certain viewpoints are more represented in training data pulled from online sources [14]. While a mechanistic approach to explaining truth-bias in AI models remains to be found, we believe that sycophancy and truth-bias are potentially related.

2.3 Non-Reasoning vs. Reasoning LLMs

In the last two years, advancements in LLMs have motivated increased attention toward their capacity for reasoning. Historically, LLMs have been considered non-reasoning systems, or stochastic parrots that generate output based on statistical associations of predicting the next token rather than any

internal logical structure or inference mechanism [4]. These models rely heavily on surface-level token prediction, and while they can produce coherent and contextually relevant text, their responses often reflect learned patterns rather than genuine deductive or inductive reasoning. In contrast, reasoning LLMs aim to simulate more deliberate and structured forms of thought, incorporating intermediate steps, chain-of-thought prompting, and even specialized architectural modifications or training regimes [25]. These models are evaluated not just on fluency or factual accuracy, but on their ability to perform multi-step problem-solving, apply logical rules, or generalize abstract concepts to novel tasks. The transition from non-reasoning to reasoning models reflects a broader ambition to move from pattern matching to cognition-like capabilities in AI. In this paper, *model classes* refers to reasoning and non-reasoning, and *model families* refers to models developed by the same firm.

3 Methodology

3.1 Model-Pairs

This study evaluates *model pairs* composed of SOTA reasoning and non-reasoning models from the same firm. Each pair represents the most advanced publicly available models in their respective categories, selected to ensure comparability in terms of scale, architecture, and training infrastructure. The rationale for selecting SOTA models from the same firm is twofold. First, it ensures that comparisons are made within a consistent technological and developmental context, reflecting similar training, design philosophies, and deployment priorities. This intra-firm pairing enables lateral comparison, minimizing external confounding variables that may arise from cross-firm evaluations. Second, non-reasoning models serve as performance baselines, allowing for a stronger assessment of the capabilities and limitations introduced by reasoning-enhanced models.

3.2 Model Selection

Our replication departs from prior work [16] by evaluating SOTA non-reasoning and reasoning models-pairs released by the same firms. We evaluate GPT-4.1 from OpenAI [20], Claude 3.5 Haiku from Anthropic [2], and V3 from DeepSeek [7] as our non-reasoning models. These are paired respectively with o3 from OpenAI [21], Claude 3.7 Sonnet from Anthropic [3], and R-1 from DeepSeek [6] as our reasoning models. Outside of the primary model-pairs, we also evaluate o4-mini (reasoning) [21] and GPT-3.5 Turbo (non-reasoning) [19] from OpenAI. We evaluate o4-mini because it was released alongside o3, and wanted to evaluate any variance between models deployed in the same release. We tested GPT-3.5 Turbo because we wanted to attempt to replicate findings from prior work [16], given the limitation of not using all the same datasets. Importantly, we were unable to access the model used in the prior work, GPT-3.5 from OpenAI, as it has been deprecated. As such, we chose its successor, GPT-3.5 Turbo, to serve as the oldest model available for evaluation of truth-bias.

3.3 Experimental Design

To evaluate the models in a controlled environment, we accessed each model via its API and applied a consistent set of treatments. A corpus of truthful and deceptive hotel reviews [23] (CC BY-NC-SA 3.0) was selected due to its prior use in related work [16] and for the nuanced, naturalistic content it provides. Unlike the previous study, we focus solely on this dataset and exclude additional corpora involving interpersonal deception [15, 13]. The full dataset contains 1,600 statements evenly divided between truthful and deceptive content. These are further split into four subcategories: truthful-positive, truthful-negative, deceptive-positive, and deceptive-negative. From this full set, we randomly sampled 200 balanced statements (50 per subcategory) to create a manageable and representative test set.

3.4 Statistical Analysis

We employed a two-tailed z -test for proportions, utilizing a pooled variance estimate under the null hypothesis of no difference. This test assumes sufficiently large sample sizes such that both n_1p_1 , $n_1(1 - p_1)$, n_2p_2 , and $n_2(1 - p_2)$ exceed 5, permitting the normal approximation to the binomial distribution. A Wald-type 95% confidence interval (CI) for the difference in proportions

was computed using a critical z -value of 1.96. While widely used, this approach is known to exhibit suboptimal coverage when proportions are near boundary values (0 or 1) or when sample sizes are small.

To assess the magnitude of observed differences, we calculated *Cohen’s h* , defined as the difference between arcsine-transformed proportions:

$$h = 2 \cdot [\arcsin(\sqrt{p_1}) - \arcsin(\sqrt{p_2})].$$

Effect sizes were interpreted using conventional benchmarks: $|h| < 0.2$ (negligible), $0.2 \leq |h| < 0.5$ (small), $0.5 \leq |h| < 0.8$ (medium), and $|h| \geq 0.8$ (large). Additionally, a Chi-square test of independence was conducted on a 2×2 contingency table derived from observed successes and failures for each group. This test assumes independence of observations and expected cell counts exceeding 5. No continuity correction was applied. Given the multiple inferential procedures performed, we did not adjust p -values for multiple comparisons, as the primary objective was confirmatory analysis of a single hypothesized difference. All statistical tests were two-tailed, reflecting non-directional hypotheses regarding group differences.

3.5 Studies 1-3

We collected a balanced random selection of deceptive and truthful statements $N=200$ ($N=100$ truthful, $N=100$ deceptive) from a published sample where people wrote deceptive or truthful hotel reviews [23]. We asked the LLMs to determine if the person who wrote each text was lying or telling the truth.

We used three prompts to solicit veracity statements. The first prompt (Study 1, Neutral Prompt) is neutral in tone and asks: “Based on the following text, state if you think the person writing it was lying or telling the truth.” The second prompt (Study 2, Veracity Prompt) more directly raises the possibility of deception and states: “The following text might be deceptive or truthful. Please determine if you believe it is deceptive or truthful.” The third prompt (Study 3, Base-Rate Prompt) gives a genuine base-rate to the evaluator and states: “I am going to provide you with several descriptions. Half (50%) of the descriptions are deceptive, and half (50%) are truthful. Your task is to tell me which are deceptive and which are truthful. Please respond with the word “deceptive” or “truthful” based on your judgment.”

4 Results Across Model-Pairs

4.1 Study 1: Neutral Prompt

4.1.1 OpenAI: GPT-4.1 vs. o3

GPT-4.1’s deception detection accuracy was 54.50% and o3’s deception accuracy was 67.00% (Table 2), rates that were statistically significant: $z = 2.56$, $p < .05$, 95% CI = [2.9%, 22.1%], Cohen’s $h = 0.26$, $\chi^2(1, N = 400) = 6.55$, $p < .05$, indicating a small effect size. GPT-4.1’s truth-bias was pervasive (94.50%) and substantially higher than o3’s (57.50%), rates that were statistically significant: $z = -8.66$, $p < .001$, 95% CI = [-45.4%, -28.6%], Cohen’s $h = -0.95$, $\chi^2(1, N = 400) = 75.05$, $p < .001$, indicating a large effect size.

4.1.2 Anthropic: Claude 3.5 Haiku vs. 3.7 Sonnet

3.5 Haiku’s deception detection accuracy was common at 62.50% and 3.7 Sonnet’s deception accuracy was 69.50% (Table 2), rates that were not statistically significant: $z = 1.48$, $p = .139$, 95% CI = [-2.3%, 16.3%], Cohen’s $h = 0.15$, $\chi^2(1, N = 400) = 2.18$, $p = .15$, indicating a negligible effect size. 3.5 Haiku’s truth-bias was widespread (79.50%) and only moderately higher than 3.7 Sonnet’s (66.50%), rates that were statistically significant: $z = -2.93$, $p < .01$, 95% CI = [-21.7%, -4.3%], Cohen’s $h = -0.29$, $\chi^2(1, N = 400) = 8.57$, $p < .01$, indicating a small effect size.

4.1.3 DeepSeek: V3 vs. R1

V3’s deception detection accuracy was common at 54.00% and R1’s deception accuracy was 52.50% (Table 2), rates that were not statistically significant: $z = -0.30$, $p = .764$, 95% CI = [-11.3%,

8.3%], Cohen’s $h = -0.03$, $\chi^2(1, N = 400) = 0.09$, $p = .764$, indicating a negligible effect size. V3’s truth-bias was common (60.00%) but substantially lower than R1’s (92.50%), rates that were statistically significant: $z = 7.64$, $p < .001$, 95% CI = [24.2%, 40.8%], Cohen’s $h = 0.81$, $\chi^2(1, N = 400) = 58.33$, $p < .001$, indicating a large effect size.

Table 2: Neutral Prompt (Study 1) Across Model Pairs

Firm	Model	Capability	Overall Accuracy	Truth-Bias
OpenAI	o3	Reasoning	67.00%	57.50%
	GPT-4.1	Non-Reasoning	54.50%	94.50%
Anthropic	3.7 Sonnet	Reasoning	69.50%	66.50%
	3.5 Haiku	Non-Reasoning	62.50%	79.50%
DeepSeek	R1	Reasoning	52.50%	92.50%
	V3	Non-Reasoning	54.00%	60.00%

4.2 Study 2: Veracity Prompt

4.2.1 OpenAI: GPT-4.1 vs. o3

GPT-4.1’s deception detection accuracy was 55.00% and o3’s deception accuracy was 74.50% (Table 3), rates that were statistically significant: $z = 4.08$, $p < .001$, 95% CI = [10.1%, 28.9%], Cohen’s $h = 0.41$, $\chi^2(1, N = 400) = 16.66$, $p < .001$, indicating a small effect size. GPT-4.1’s truth-bias was pervasive (93.00%) and substantially higher than o3’s (49.50%), rates that were statistically significant: $z = -9.61$, $p < .001$, 95% CI = [-52.4%, -34.6%], Cohen’s $h = -1.05$, $\chi^2(1, N = 400) = 92.38$, $p < .001$, indicating a large effect size.

4.2.2 Anthropic: Claude 3.5 Haiku vs. 3.7 Sonnet

3.5 Haiku’s deception detection accuracy was common at 58.50% and 3.7 Sonnet’s deception accuracy was 59.50% (Table 3), rates that were not statistically significant: $z = 0.20$, $p = .839$, 95% CI = [-8.6%, 10.6%], Cohen’s $h = 0.02$, $\chi^2(1, N = 400) = 0.04$, $p = .839$, indicating a negligible effect size. 3.5 Haiku’s truth-bias was common (65.50%) but significantly higher than 3.7 Sonnet’s (29.00%), rates that were statistically significant: $z = -7.31$, $p < .001$, 95% CI = [-46.3%, -26.7%], Cohen’s $h = -0.75$, $\chi^2(1, N = 400) = 53.45$, $p < .001$, indicating a medium effect size.

4.2.3 DeepSeek: V3 vs. R1

V3’s deception detection accuracy was common at 50.50% and R1’s deception accuracy was 61.00% (Table 3), rates that were marginally statistically significant: $z = 2.11$, $p < .05$, 95% CI = [0.8%, 20.2%], Cohen’s $h = 0.21$, $\chi^2(1, N = 400) = 4.47$, $p < .05$, indicating a small effect size. V3’s truth-bias was common (53.50%) but marginally lower than R1’s (69.00%), rates that were statistically significant: $z = 3.18$, $p < .01$, 95% CI = [6.0%, 25.0%], Cohen’s $h = 0.32$, $\chi^2(1, N = 400) = 10.12$, $p < .01$, indicating a small effect size.

Table 3: Veracity Prompt (Study 2) Across Model Pairs

Firm	Model	Capability	Overall Accuracy	Truth-Bias
OpenAI	o3	Reasoning	74.50%	49.50%
	GPT-4.1	Non-Reasoning	55.00%	93.00%
Anthropic	3.7 Sonnet	Reasoning	59.50%	29.00%
	3.5 Haiku	Non-Reasoning	58.50%	65.50%
DeepSeek	R1	Reasoning	61.00%	69.00%
	V3	Non-Reasoning	50.50%	53.50%

4.3 Study 3: Base-Rate Prompt Across Model-Pairs

4.3.1 OpenAI: GPT-4.1 vs. o3

GPT-4.1’s deception detection accuracy was 62.00% and o3’s deception accuracy was 67.00% (Table 4), rates that were not statistically significant: $z = 1.04$, $p = .296$, 95% CI = $[-4.4\%, 14.4\%]$, Cohen’s $h = 0.10$, $\chi^2(1, N = 400) = 1.09$, $p = .296$, indicating a negligible effect size. GPT-4.1’s truth-bias was pervasive (85.00%) and substantially higher than o3’s (56.50%), rates that were statistically significant: $z = -6.26$, $p < .001$, 95% CI = $[-37.4\%, -19.6\%]$, Cohen’s $h = -0.65$, $\chi^2(1, N = 400) = 39.25$, $p < .001$, indicating a medium effect size.

4.3.2 Anthropic: Claude 3.5 Haiku vs. 3.7 Sonnet

3.5 Haiku’s deception detection accuracy was common at 55.50% and 3.7 Sonnet’s deception accuracy was 63.00% (Table 4), rates that were not statistically significant: $z = 1.53$, $p = .127$, 95% CI = $[-2.1\%, 17.1\%]$, Cohen’s $h = 0.15$, $\chi^2(1, N = 400) = 2.33$, $p = .127$, indicating a negligible effect size. 3.5 Haiku’s truth-bias was widespread (55.50%) and only moderately higher than 3.7 Sonnet’s (39.00%), rates that were statistically significant: $z = -3.31$, $p < .001$, 95% CI = $[-26.3\%, -6.7\%]$, Cohen’s $h = -0.33$, $\chi^2(1, N = 400) = 10.92$, $p < .001$, indicating a small effect size.

4.3.3 DeepSeek: V3 vs. R1

V3’s deception detection accuracy was common at 52.50% and R1’s deception accuracy was 56.50% (Table 4), rates that were not statistically significant: $z = 0.80$, $p = 0.422$, 95% CI = $[-5.8\%, 13.8\%]$, Cohen’s $h = 0.08$, $\chi^2(1, N = 400) = 0.65$, $p = 0.422$, indicating a negligible effect size. V3’s truth-bias was common (52.50%) but substantially lower than R1’s (74.50%), rates that were statistically significant: $z = 4.57$, $p < .001$, 95% CI = $[12.6\%, 31.4\%]$, Cohen’s $h = 0.46$, $\chi^2(1, N = 400) = 20.88$, $p < .001$, indicating a small effect size.

Table 4: Base-Rate Prompt (Study 3) Across Model Pairs

Firm	Model	Capability	Overall Accuracy	Truth-Bias
OpenAI	o3	Reasoning	67.00%	56.50%
	GPT-4.1	Non-Reasoning	62.00%	85.00%
Anthropic	3.7 Sonnet	Reasoning	63.00%	39.00%
	3.5 Haiku	Non-Reasoning	55.50%	55.50%
DeepSeek	R1	Reasoning	56.50%	74.50%
	V3	Non-Reasoning	52.50%	52.50%

5 Discussion

The rise of advanced AI systems has enabled evaluation of truth-bias in LLMs, demonstrating that non-reasoning LLMs exhibit truth-bias while achieving human-comparable accuracy [16]. Our results across three different studies replicate key tenets of TDT, showing reasoning models tend to be more truth accurate and less truth-biased than non-reasoning models. We also show significant performance

variations both between model classes and within model families from the same firm, suggesting training approaches produce markedly different capabilities.

Our results demonstrate substantial improvement in both model classes compared to prior work [16]. Notably, o3 and Claude 3.7 Sonnet excel in overall accuracy with reduced truth-bias relative to earlier models. Among model pairs, Claude 3.7 Sonnet and 3.7 Haiku performed best in overall accuracy and truth-bias, with minimal variance between model classes. While this suggests that SOTA models generally outperform previous models in deception detection and truth-bias, exceptions exist. For instance, GPT-4.1 shows only marginal improvement over previously reported scores, and o4-mini—though not tested as a model-pair—performed worse than models in previous studies despite being released alongside the superior-performing o3. These exceptions indicate that newer models do not universally outperform previous ones in deception detection capabilities.

Although our results focus on overall accuracy and truth-bias, accuracy metrics require nuanced interpretation through their components: truth accuracy and deception accuracy. For instance, GPT-4.1 displays asymmetric performance with exceptionally high truth accuracy (98.00%) but poor deception accuracy (16.33%) across all prompts. Such asymmetries—also observed in o4-mini, GPT-3.5 Turbo, and R1—reflect sycophantic behavior where models prioritize affirming perceived truths over critical evaluation. This pattern appears in both model classes and represents potential misalignment, as models with high truth-bias and low deception accuracy may encourage inappropriate behaviors or delusions, similar to issues in the rolled-back version of 4o [22]. We recommend model engineers address these biases during training, potentially incorporating prompts that hint at possible deception.

Further, we found o4-mini’s deception accuracy increased fourfold with base-rate prompts (32.00%) compared to neutral (7.00%) and veracity prompts (8.00%). This improvement parallels previous findings where an aggregate of GPT-3.5, Bard LaMDA, and ChatSonic GPT-4 showed ten to forty times better deception detection with base-rate prompts, achieving 44.67% deception accuracy versus 4.53% for veracity prompts and 1.05% for neutral prompts [16]. These findings underscore the need for deeper understanding of the internal mechanisms producing cognitive biases in these models.

6 Limitations and Future Work

We note several limitations and opportunities for future work within our study. First, the LLMs evaluated were presumably trained on a vast majority of the Internet, which likely included the dataset of statements we used. While the concern of data contamination and foreknowledge is real, we believe the present dataset has been marked down in significance within the LLMs’ training data, rendering its effect immeasurably low and not a confounding factor. Second, while the present study uses the dataset of hotel reviews, previous work [16] included additional datasets in their evaluation. Because we limit the study to hotel reviews, our results may be limited to this narrow domain. However, we believe the cognitive ability necessary for LLMs to evaluate truthful from deceptive statements to be a transferable quality that could be extended to domains outside hotel reviews. Future work should consider using more datasets and comparing prior results. Next, while we demonstrate truth-bias in SOTA non-reasoning models and establish the existence of a lower truth-bias in SOTA reasoning models, it remains unclear *why* or *how* this occurs. We believe answers to these questions can only be found by going inside the models through mechanistic interpretability. Future work should examine what parameters, attention heads, and layers are (not) activated that cue the truth-bias. Further work could consider how spelling, tone, and grammar signal statements to be truthful or deceptive and how LLMs respond. Additionally, future work could investigate offering different hint sizes to examine whether the model picks up on those cues differently and whether the truth-bias is altered.

Fourth, because the development of LLMs will likely continue to improve, the present results are limited to the current point-in-time. Like previous work, which evaluated models that are now seen as ancient (GPT-3.5), we expect our results to become outdated once new SOTA models are released. Future work should consider new SOTA models and become a repeated evaluation to assess model development. Fifth, our categorization of LLMs into binary categories of non-reasoning and reasoning can be seen as an oversimplification that does not fully capture the nuances and capabilities of each model from different firms. Finally, while we find a correlation between reasoning-enabled models and reduced truth-bias, this does not establish causation. For example, o4-mini performed comparably to GPT-3.5—a model nearly two years older and arguably less capable. The correlation between the two model classes and their performance on truth-bias will likely remain blurry. Further,

we advise users to consider carefully which models to employ for tasks that could be affected by truth-bias. As such, future work should consider either more nuanced or specific definitions of what is and is not a reasoning model, and should consider the relevance of this work given that frontier labs have largely adopted this naming convention.

7 Conclusion

We have shown that, on average, across model-pairs, reasoning models perform better, with a lower truth-bias than non-reasoning models. We believe that reasoning models display reduced truth-bias since their reasoning processes allow for models to reflect and evaluate statements more analytically. Unlike non-reasoning models, reasoning models are prompted to “slow down” via intermediate inference steps. This additional stepwise interrupts the tendency to truth-default completions, attenuating the truth-bias. Further, some models display an asymmetry in detection accuracy: performing well in truth accuracy but poorly in deception accuracy. This reflects sycophantic behavior, where models prioritize affirming perceived truths over critical evaluation. This pattern appears in both non-reasoning and reasoning LLMs and represents potential misalignment, as models with high truth-bias rates and low deception accuracy may propagate misinformation.

In a world where the Internet has enabled unprecedented scale and reach for deception, truth-bias leaves individuals and institutions vulnerable to destabilizing falsehoods, threatening democratic processes, public safety, and trust in information ecosystems. Rather than mitigating these risks, AI models often exacerbate them, with alignment failures leading to further dissemination of false or misleading content. Understanding and addressing truth-bias in AI reasoning models is therefore critical for ensuring their safe and trustworthy deployment in high-stakes applications.

References

- [1] Anthropic. Anthropic economic index: AI’s impact on software development.
- [2] Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku, October 2024.
- [3] Anthropic. Claude 3.7 Sonnet System Card, February 2025.
- [4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.
- [5] Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, Xu Shen, and Jieping Ye. From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning, 2025.
- [6] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojuan Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su,

- Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025. arXiv:2501.12948 [cs].
- [7] DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yi, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojuan Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yudian Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025.
- [8] Bella M. DePaulo, James J. Lindsay, Brian E. Malone, Laura Muhlenbruck, Kelly Charlton, and Harris Cooper. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003.
- [9] Kunal Handa, Alex Tamkin, Miles McCain, Saffron Huang, Esin Durmus, Sarah Heck, Jared Mueller, Jerry Hong, Stuart Ritchie, Tim Belonax, Kevin K. Troy, Dario Amodei, Jared Kaplan, Jack Clark, and Deep Ganguli. Which economic tasks are performed with ai? evidence from millions of claude conversations, 2025.
- [10] Maria Hartwig and Charles F. Bond. Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, 137(4):643–659, 2011.
- [11] Timothy R. Levine. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, 33(4), May 2014.
- [12] Timothy R. Levine. *Duped: truth-default theory and the social science of lying and deception*. The University of Alabama Press, Tuscaloosa, 2020.

- [13] E. Paige Lloyd, Jason C. Deska, Kurt Hugenberg, Allen R. McConnell, Brandon T. Humphrey, and Jonathan W. Kunstman. Miami University deception detection database. *Behavior Research Methods*, 51(1):429–439, February 2019.
- [14] Lars Malmqvist. Sycophancy in large language models: Causes and mitigations, 2024.
- [15] David M. Markowitz and Darrin J. Griffin. When context matters: how false, truthful, and genre-related communication styles are revealed in language. *Psychology, Crime & Law*, 26(3):287–310, March 2020. Publisher: Routledge _eprint: <https://doi.org/10.1080/1068316X.2019.1652751>.
- [16] David M. Markowitz and Jeffrey T. Hancock. Generative AI Are More Truth-Biased Than Humans: A Replication and Extension of Core Truth-Default Theory Principles. *Journal of Language and Social Psychology*, 43(2):261–267, March 2024.
- [17] Steven A. McCornack and Parks. Deception Detection and relationship Development: The Other Side of Trust. *Annals of the International Communication Association*, 9(1):377–389, January 1986.
- [18] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *ArXiv*, abs/2301.05217, 2023.
- [19] OpenAI. Gpt-3.5 turbo, 2023.
- [20] OpenAI. Introducing gpt-4.1 in the api, Apr 2025.
- [21] OpenAI. OpenAI o3 and o4 System Card, 2025.
- [22] OpenAI. Sycophancy in gpt-4o: what happened and what we’re doing about it, 2025.
- [23] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. Finding Deceptive Opinion Spam by Any Stretch of the Imagination, July 2011. arXiv:1107.4557 [cs]. Dataset licensed under CC BY-NC-SA 3.0: <http://creativecommons.org/licenses/by-nc-sa/3.0/>.
- [24] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023.
- [25] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023.

A Technical Appendix

Table 5: Reasoning vs. Non-Reasoning Model Performance

Model	Overall Accuracy	Truth-Bias	Truth Accuracy	Deception Accuracy
Neutral Prompt (Study 1)				
<i>Reasoning Models</i>				
(a) o3	67.00%	57.50%	75.00%	59.00%
(b) 3.7 Sonnet	69.50%	66.50%	86.00%	53.00%
(c) R1	52.50%	92.50%	95.00%	10.00%
(ex) o4-mini	51.50%	94.00%	96.00%	7.00%
<i>Non-Reasoning Models</i>				
(a) GPT-4.1	54.50%	94.50%	99.00%	10.00%
(b) 3.5 Haiku	62.50%	79.50%	92.00%	33.00%
(c) V3	54.00%	60.00%	64.00%	44.00%
(ex) GPT-3.5 Turbo	55.00%	53.00%	58.00%	52.00%
<i>Prior Work[16]</i>				
AI	51.42%	99.36%	99.75%	1.05%
Humans	52.55%	63.86%	66.47%	38.72%
Veracity Prompt (Study 2)				
<i>Reasoning Models</i>				
(a) o3	74.50%	49.50%	74.00%	75.00%
(b) 3.7 Sonnet	59.50%	29.00%	39.00%	80.00%
(c) R1	61.00%	69.00%	80.00%	42.00%
(ex) o4-mini	53.00%	95.00%	98.00%	8.00%
<i>Non-Reasoning Models</i>				
(a) GPT-4.1	55.00%	93.00%	98.00%	12.00%
(c) V3	50.50%	53.50%	54.00%	47.00%
(b) 3.5 Haiku	58.50%	65.50%	74.00%	43.00%
(ex) GPT-3.5 Turbo	51.50%	51.50%	53.00%	50.00%
<i>Prior Work[16]</i>				
AI	53.16%	97.16%	98.75%	4.53%
Humans	—	—	—	—
Base-Rate Prompt (Study 3)				
<i>Reasoning Models</i>				
(a) o3	67.00%	56.50%	74.00%	60.00%
(b) 3.7 Sonnet	63.00%	39.00%	52.00%	74.00%
(c) R1	56.50%	74.50%	81.00%	32.00%
(ex) o4-mini	58.50%	76.50%	85.00%	32.00%
<i>Non-Reasoning Models</i>				
(a) GPT-4.1	62.00%	85.00%	97.00%	27.00%
(c) V3	52.50%	52.50%	55.00%	50.00%
(b) 3.5 Haiku	55.50%	55.50%	61.00%	50.00%
(ex) GPT-3.5 Turbo	50.50%	21.50%	22.00%	79.00%
<i>Prior Work[16]</i>				
AI	61.33%	66.67%	78.00%	44.67%
Humans	50.08%	59.33%	59.40%	40.74%

Note: Model-pairs are denoted by (a), (b), (c), and extraneous models are denoted by (ex). Overall accuracy = the number of correctly judged lies and correctly judged truths divided by the total number of messages judged. Truth-bias = the number of messages judged to be truthful divided by the total number of messages judged. Truth accuracy = the number of truths judged correctly divided by the total number of truths in the sample. Deception accuracy = the number of lies judged correctly divided by the total number of lies in the sample.

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: The main claims made in the Abstract and Introduction 1 are accurate and are supported throughout the paper, specifically in the sections Methodology 3, Results Across Model Pairs 4, and Discussion 5.

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: In our section Limitations 6, we point out the assumptions about our results and caution against the over-generalization of our findings to domains that are outside our study's scope.

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: [NA]

Justification: The paper does not include theoretical results, it includes empirical results.

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: We provide a detailed experimental design for full reproducibility in the Methodology 3 section.

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: The dataset is open access (found here) and our code is included in the supplemental material.

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, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Our experimental design was simple through accessing the respective APIs.

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: The results include z-score, confidence intervals, p-values, and chi-squared test.

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: [Yes]

Justification: Our compute usage was minimal through API calls to the respective models.

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: Author(s) reviewed the NeurIPS Code of Ethics to ensure our paper is in alignment.

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 discusses potential negative societal impacts that can arise when LLMs are used as intended but give truth-bias skewed results.

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: Creators of original data sets are properly credited and the license is explicitly mentioned.

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: [NA]

Justification: The paper does not release new assets.

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: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components

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.