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

Emilio Barkett[*1] **Olivia Long**[*1] **Madhavendra Thakur**[*1]

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 (Levine, 2020; 2014; McCornack & Parks, 1986; Markowitz & Hancock, 2024). 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 (Markowitz & Hancock, 2024). 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 (OpenAI, 2025b), 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 (McCornack & Parks, 1986), 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.

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 (Handa et al., 2025; Anthropic), 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.

[*]Equal contribution [1]Columbia University, New York, NY, USA. Correspondence to: Emilio Barkett <eab2291@columbia.edu>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

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

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 (Nanda et al., 2023), 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 (Ott et al., 2011). We evaluated 400 statements across several models, which is a sufficiently large sample to detect statistically significant effects where present. 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 (OpenAI, 2025), Claude 3.5 Haiku from Anthropic (Anthropic, 2024), and V3 from DeepSeek (DeepSeek-AI et al., 2025b) as our non-reasoning models. These are paired respectively with the following reasoning models: o3 from OpenAI (OpenAI, 2025a), Claude 3.7 Sonnet from Anthropic (Anthropic, 2025), and R-1 from DeepSeek (DeepSeek-AI et al., 2025a). Outside the main model-pairs, we also tested o4-mini (OpenAI, 2025a) alongside o3 to examine the variance between models from the same release. Additionally, we used GPT-3.5 Turbo (OpenAI, 2023) to approximate prior findings (Markowitz & Hancock, 2024), 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% (Markowitz & Hancock, 2024). 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).

Table 1. Accuracy and bias across two models.

METRIC	REASONING (O3)	NON-REASONING (CLAUDE 3.5)
NEUTRAL (STUDY 1)		
OVERALL ACC.	67.00%	62.50%
TRUTH-BIAS	57.50%	79.50%
TRUTH ACC.	75.00%	92.00%
DECEPTION ACC.	59.00%	33.00%
VERACITY (STUDY 2)		
OVERALL ACC.	74.50%	58.50%
TRUTH-BIAS	49.50%	65.50%
TRUTH ACC.	74.00%	74.00%
DECEPTION ACC.	75.00%	43.00%
BASE-RATE (STUDY 3)		
OVERALL ACC.	67.00%	55.50%
TRUTH-BIAS	56.50%	55.50%
TRUTH ACC.	74.00%	61.00%
DECEPTION ACC.	60.00%	50.00%

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 (DePaulo et al., 2003; Hartwig & Bond, 2011). A downstream effect of this is the truth-bias: a phenomenon where a receiver infers a message as honest independent of its veracity (Levine, 2014; McCornack & Parks, 1986). Humans show a tendency toward gullibility in communication (Levine, 2020), 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 (Levine, 2020), 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 (Markowitz & Hancock, 2024), as supported by truth-default theory (TDT). TDT is a pan-

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

cultural theory of human deception detection which states that upon being prompted for a veracity judgment, in the absence of suspicion, people automatically assume others are honest (Levine, 2020; 2014). 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 (Markowitz & Hancock, 2024).

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 (Chen et al., 2025). 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 (Sharma et al., 2023) 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 (Malmqvist, 2024). 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 (Bender et al., 2021). 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 (Wei et al., 2023). 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 (Markowitz & Hancock, 2024) by evaluating SOTA non-reasoning and reasoning models-pairs released by the same firms. We evaluate GPT-4.1 from OpenAI (OpenAI, 2025), Claude 3.5 Haiku from Anthropic (Anthropic, 2024), and V3 from DeepSeek (DeepSeek-AI et al., 2025b) as our non-reasoning models. These are paired respectively with o3 from OpenAI (OpenAI, 2025a), Claude 3.7 Sonnet from Anthropic (Anthropic, 2025), and R-1 from DeepSeek (DeepSeek-AI et al., 2025a) as our reasoning models. Outside of the primary model-pairs, we also evaluate o4-mini (reasoning) (OpenAI, 2025a) and GPT-3.5 Turbo (non-reasoning) (OpenAI, 2023) 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 (Markowitz & Hancock, 2024), 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 (Ott et al., 2011) (CC BY-NC-SA 3.0) was selected due to its prior use in related work (Markowitz & Hancock, 2024) 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 (Markowitz & Griffin, 2020; Lloyd et al., 2019). 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 $n_1 p_1$, $n_1(1-p_1)$, $n_2 p_2$, 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:

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

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 (Ott et al., 2011). 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

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

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

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

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

$h = -0.29$, $\chi^2(1, N = 400) = 8.57$, $p < .01$, 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.

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.

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

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

MODEL	OVERALL ACC.	TRUTH-BIAS
OPENAI		
o3	67.00%	56.50%
GPT-4.1	62.00%	85.00%
ANTHROPIC		
3.7 SONNET	63.00%	39.00%
3.5 HAIKU	55.50%	55.50%
DEEPSEEK		
R1	56.50%	74.50%
V3	52.50%	52.50%

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.

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 (Markowitz & Hancock, 2024). 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 (Markowitz & Hancock, 2024). 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 (OpenAI, 2025b). 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 (Markowitz & Hancock, 2024). 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 (Markowitz & Hancock, 2024) 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.

Impact Statement

Our work advances machine learning research through the first systematic investigation of truth-bias across reasoning and non-reasoning language models. This study reveals critical insights about how different model architectures process veracity judgments, identifying both improvements in reasoning models and persistent asymmetries in detection accuracy that represent fundamental challenges in language model development. By establishing quantitative benchmarks for truth-bias assessment, we provide the research community with a replicable framework to evaluate and mitigate these limitations, ultimately contributing to the development of more epistemically sound AI systems capable of more balanced veracity discrimination—a capability essential for applications ranging from automated fact-checking to content moderation and information filtering.

References

Anthropic. Anthropic economic index: AI's impact on software development. URL https://www.anthropic.com/research/impact-software-development.

Anthropic. The Claude 3 Model Family: Opus, Sonnet, Haiku, October 2024. URL https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf.

Anthropic. Claude 3.7 Sonnet System Card, February 2025.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. 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, pp. 610–623, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097. doi: 10.1145/3442188.3445922. URL https://doi.org/10.1145/3442188.3445922.

Chen, W., Huang, Z., Xie, L., Lin, B., Li, H., Lu, L., Tian, X., Cai, D., Zhang, Y., Wang, W., Shen, X., and Ye, J. From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning, 2025. URL https://arxiv.org/abs/2409.01658.

DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., Xue, B., Wang, B., Wu, B., Feng, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Ding, H., Xin, H., Gao, H., Qu, H., Li, H., Guo, J., Li, J., Wang, J., Chen, J., Yuan, J., Qiu, J., Li, J., Cai, J. L., Ni, J., Liang, J., Chen, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Zhao, L., Wang, L., Zhang, L., Xu, L., Xia, L., Zhang, M., Zhang, M., Tang, M., Li, M., Wang, M., Li, M., Tian, N., Huang, P., Zhang, P., Wang, Q., Chen, Q., Du, Q., Ge, R., Zhang, R., Pan, R., Wang, R., Chen, R. J., Jin, R. L., Chen, R., Lu, S., Zhou, S., Chen, S., Ye, S., Wang, S., Yu, S., Zhou, S., Pan, S., Li, S. S., Zhou, S., Wu, S., Ye, S., Yun, T., Pei, T., Sun, T., Wang, T., Zeng, W., Zhao, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Xiao, W. L., An, W., Liu, X., Wang, X., Chen, X., Nie, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yang, X., Li, X., Su, X., Lin, X., Li, X. Q., Jin, X., Shen, X., Chen, X., Sun, X., Wang, X., Song, X., Zhou, X., Wang, X., Shan, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhang, Y., Xu, Y., Li, Y., Zhao, Y., Sun, Y., Wang, Y., Yu, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Ou, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Xiong, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Zhu, Y. X., Xu, Y., Huang, Y., Li, Y., Zheng, Y., Zhu, Y., Ma, Y., Tang, Y., Zha, Y., Yan, Y., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Xie, Z., Zhang, Z., Hao, Z., Ma, Z., Yan, Z., Wu, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Pan, Z., Huang, Z., Xu, Z., Zhang, Z., and Zhang, Z. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, January 2025a. URL http://arxiv.org/abs/2501.12948. arXiv:2501.12948 [cs].

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. Deepseek-v3 technical report, 2025b. URL https://arxiv.org/abs/2412.19437.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. Cues to deception. *Psychological Bulletin*, 129(1):74–118, 2003. ISSN 1939-1455, 0033-2909. doi: 10.1037/0033-2909.129.1.74. URL https://doi.apa.org/doi/10.1037/0033-2909.129.1.74.

Handa, K., Tamkin, A., McCain, M., Huang, S., Durmus, E., Heck, S., Mueller, J., Hong, J., Ritchie, S., Belonax, T., Troy, K. K., Amodei, D., Kaplan, J., Clark, J., and Ganguli, D. Which economic tasks are performed with ai? evidence from millions of claude conversations, 2025. URL https://arxiv.org/abs/2503.04761.

Hartwig, M. and Bond, C. F. Why do lie-catchers fail? A lens model meta-analysis of human lie judgments. *Psychological Bulletin*, 137(4):643–659, 2011. ISSN 1939-1455, 0033-2909. doi: 10.1037/a0023589. URL https://doi.apa.org/doi/10.1037/a0023589.

Levine, T. R. Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. *Journal of Language and Social Psychology*, 33(4), May 2014. doi: https://doi.org/10.1177/0261927X1453591.

Levine, T. R. *Duped: truth-default theory and the social science of lying and deception*. The University of Alabama Press, Tuscaloosa, 2020. ISBN 978-0-8173-2041-6 978-0-8173-5968-3.

Lloyd, E. P., Deska, J. C., Hugenberg, K., McConnell, A. R., Humphrey, B. T., and Kunstman, J. W. Miami University deception detection database. *Behavior Research Methods*, 51(1):429–439, February 2019. ISSN 1554-3528. doi: 10.3758/s13428-018-1061-4. URL https://doi.org/10.3758/s13428-018-1061-4.

Malmqvist, L. Sycophancy in large language models: Causes and mitigations, 2024. URL https://arxiv.org/abs/2411.15287.

Markowitz, D. M. and Griffin, D. J. When context matters: how false, truthful, and genre-related communication styles are revealed in language. *Psychology, Crime & Law*, 26(3):287–310, March 2020. ISSN 1068-316X. doi: 10.1080/1068316X.2019.1652751. URL https://doi.org/10.1080/1068316X.2019.1652751. Publisher: Routledge _eprint: https://doi.org/10.1080/1068316X.2019.1652751.

Markowitz, D. M. and Hancock, J. T. 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. ISSN 0261-927X, 1552-6526. doi: 10.1177/0261927X231220404. URL https://journals.sagepub.com/doi/10.1177/0261927X231220404.

McCornack, S. A. and Parks. Deception Detection and relationship Development: The Other Side of Trust. *Annals of the International Communication Association*, 9 (1):377–389, January 1986. doi: https://doi.org/10.1080/23808985.1986.11678616.

Nanda, N., Chan, L., Lieberum, T., Smith, J., and Steinhardt, J. Progress measures for grokking via mechanistic interpretability. *ArXiv*, abs/2301.05217, 2023. URL https://api.semanticscholar.org/CorpusID:255749430.

OpenAI. Gpt-3.5 turbo, 2023. URL https://platform.openai.com/docs/models/gpt-3.5-turbo.

OpenAI. Introducing gpt-4.1 in the api, Apr 2025. URL https://openai.com/index/gpt-4-1/.

OpenAI. OpenAI o3 and o4 System Card, 2025a. URL https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

OpenAI. Sycophancy in gpt-4o: what happened and what we're doing about it, 2025b. URL https://openai.com/index/sycophancy-in-gpt-4o/.

Ott, M., Choi, Y., Cardie, C., and Hancock, J. T. Finding Deceptive Opinion Spam by Any Stretch of the Imagination, July 2011. URL http://arxiv.org/abs/1107.4557. arXiv:1107.4557 [cs]. Dataset licensed under CC BY-NC-SA 3.0: http://creativecommons.org/licenses/by-nc-sa/3.0/.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., and Perez, E. Towards understanding sycophancy in language models, 2023. URL https://arxiv.org/abs/2310.13548.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

A. Reasoning vs. Non-Reasoning Model Performance

Table 5. Appendix Table: Reasoning vs. Non-Reasoning Model Performance

Model	**Overall Accuracy**	**Truth-Bias**	**Truth Accuracy**	**Deception Accuracy**
Neutral Prompt (Study 1)				
Reasoning Models				
(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%
Non-Reasoning Models				
(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%
Prior Work (Markowitz & Hancock, 2024)				
AI	51.42%	99.36%	99.75%	1.05%
Humans	52.55%	63.86%	66.47%	38.72%
Veracity Prompt (Study 2)				
Reasoning Models				
(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%
Non-Reasoning Models				
(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%
Prior Work (Markowitz & Hancock, 2024)				
AI	53.16%	97.16%	98.75%	4.53%
Humans	—	—	—	—
Base-Rate Prompt (Study 3)				
Reasoning Models				
(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%
Non-Reasoning Models				
(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%
Prior Work (Markowitz & Hancock, 2024)				
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.