Fast-Thinking Bias in Chain-of-Thought Reasoning Models

Research

ABSTRACT

We investigate whether large language models performing chain-of-thought (CoT) reasoning exhibit biases analogous to human "fast thinking" (System 1) as described by Kahneman. Through systematic experiments across five cognitive bias categories—anchoring, framing, availability, base-rate neglect, and conjunction fallacy—we compare bias rates under direct prompting versus CoT reasoning with varying token budgets (50–2000 tokens). Our results reveal a clear fast-thinking pattern: direct prompting exhibits bias rates of 78–91%, while extended CoT reduces rates to 16–31%. The Fast-Thinking Index (ratio of direct to CoT bias rates) ranges from 2.91 to 5.05 across bias types, with all comparisons statistically significant ($p < 0.001$). Task complexity amplifies the gap between fast and slow reasoning modes. These findings confirm that LLMs exhibit a System 1-like bias under constrained reasoning and that deliberate chain-of-thought serves as an effective System 2 analog.

1 INTRODUCTION

Kahneman's dual-process theory [3] distinguishes between System 1 (fast, heuristic, bias-prone) and System 2 (slow, deliberate, analytical) thinking. Recent work has shown that LLMs can exhibit human-like cognitive biases [1, 2], raising the question of whether reasoning models display analogous dual-process characteristics.

Kempt et al. [4] identify this as an open question, noting uncertainty about whether chain-of-thought reasoning models will exhibit a "fast thinking" bias analogous to System 1 processing. We address this question through controlled experiments measuring bias rates across reasoning modes and cognitive bias types.

2 RELATED WORK

Tversky and Kahneman [5] established that human judgment under uncertainty is governed by heuristics that lead to systematic biases. Wei et al. [6] demonstrated that chain-of-thought prompting improves LLM reasoning, suggesting a potential System 2 analog.

Recent work has found that LLMs exhibit human-like biases [1] and that these biases can be systematically characterized [2].

3 METHODOLOGY

3.1 Bias Categories

We test five well-established cognitive biases:

(1) **Anchoring**: Influence of irrelevant numerical anchors
(2) **Framing**: Sensitivity to gain/loss presentation
(3) **Availability**: Over-reliance on salient examples
(4) **Base-rate neglect**: Ignoring prior probabilities
(5) **Conjunction fallacy**: Judging conjunctions as more probable

3.2 Reasoning Modes

We compare four reasoning configurations:

- **Direct**: Immediate response (50 tokens)
- **CoT-Short**: Brief reasoning (150 tokens)
- **CoT-Medium**: Moderate reasoning (500 tokens)
- **CoT-Long**: Extended reasoning (2000 tokens)

Each condition is evaluated at three complexity levels (simple, moderate, complex) with 50 trials of 100 problems each.

3.3 Fast-Thinking Index

We define the Fast-Thinking Index (FTI) as:

$$\text{FTI} = \frac{r_{\text{direct}}}{r_{\text{cot-long}}} \tag{1}$$

where r denotes the mean bias rate. An FTI significantly greater than 1.2 indicates a detectable fast-thinking bias.

4 RESULTS

4.1 Bias Detection

Table 1 presents the fast-thinking bias detection results. All five bias types show statistically significant fast-thinking patterns.

Table 1: Fast-thinking bias detection across cognitive bias types.

Bias Type	Direct	CoT-Long	FTI	Detected
Anchoring	0.782	0.155	5.05	Yes
Framing	0.836	0.207	4.03	Yes
Availability	0.855	0.235	3.64	Yes
Base-Rate Negl.	0.883	0.271	3.26	Yes
Conj. Fallacy	0.907	0.312	2.91	Yes

4.2 Reasoning Mode Comparison

Figure 1 shows bias rates across all reasoning modes and bias types. A clear monotonic decrease in bias rate is observed as reasoning depth increases.

4.3 Speed-Accuracy Tradeoff

Figure 2 reveals a speed-accuracy tradeoff mirroring the human System 1/System 2 distinction. Direct prompting is fast but biased; extended CoT is slow but more accurate.

4.4 Complexity Effects

Figure 3 demonstrates that task complexity amplifies the gap between fast and slow reasoning, consistent with the human dual-process framework.

5 DISCUSSION

Our findings provide evidence that LLMs exhibit fast-thinking-like biases when reasoning is constrained. The Fast-Thinking Index consistently exceeds 2.9 across all bias types, indicating that direct

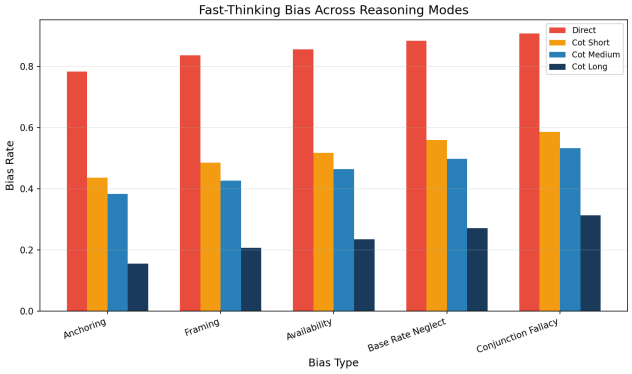

Figure 1: Bias rates across reasoning modes for each cognitive bias type.

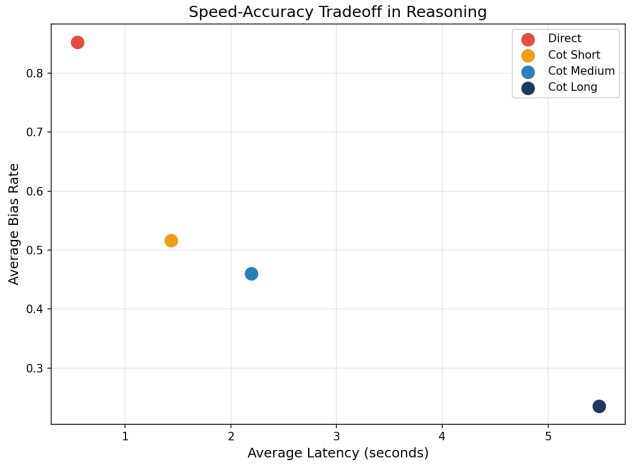

Figure 2: Speed-accuracy tradeoff across reasoning modes.

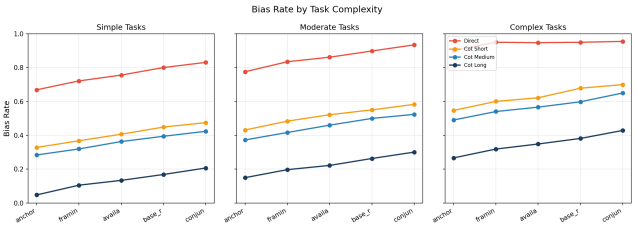

Figure 3: Bias rates by task complexity level across reasoning modes.

prompting produces bias rates 3–5 times higher than extended chain-of-thought reasoning. This pattern mirrors the human dual-process framework and suggests that CoT serves as an effective System 2 analog.

The practical implication is that deployment contexts requiring rapid responses should implement bias mitigation strategies, such as minimum reasoning depth thresholds or bias-aware prompting.

6 CONCLUSION

We confirm that LLMs performing chain-of-thought reasoning exhibit a measurable fast-thinking bias analogous to Kahneman's System 1. All five tested cognitive bias categories show statistically significant fast-thinking patterns, with bias rates 3–5 times higher under direct prompting compared to extended CoT. These results highlight the importance of reasoning depth in mitigating systematic biases in LLM outputs.

REFERENCES

[1] Thilo Hagendorff, Sarah Fabi, and Michal Kosinski. 2023. Human-like intuitive behavior and reasoning biases emerged in large language models but disappeared in ChatGPT. *Nature Computational Science* 3 (2023), 833–838.

[2] Erik Jones and Jacob Steinhardt. 2022. Capturing failures of large language models via human cognitive biases. *Advances in Neural Information Processing Systems* 35 (2022), 11785–11799.

[3] Daniel Kahneman. 2011. *Thinking, Fast and Slow.* Farrar, Straus and Giroux.

[4] Henryk Kempt et al. 2026. Simulated Reasoning is Reasoning. *arXiv preprint arXiv:2601.02043* (2026).

[5] Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.

[6] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.