A Simulation Framework for Fast-Thinking Bias in Chain-of-Thought Reasoning Models

Anonymous Author(s)

ABSTRACT

We present a parametric simulation framework that models dual-process-like behavior in large language models (LLMs) performing chain-of-thought (CoT) reasoning. Rather than conducting empirical LLM experiments, we formalize a mathematical model in which direct prompting (System 1 analog) yields higher bias rates than extended CoT reasoning (System 2 analog), with token-budget-dependent reduction following a saturating logarithmic curve. The framework simulates five cognitive bias categories—anchoring, framing, availability, base-rate neglect, and conjunction fallacy—across three complexity levels. A key design feature is that task complexity asymmetrically amplifies bias in direct mode relative to CoT, producing a widening fast-slow gap. Through sensitivity analysis over all model parameters, we demonstrate that the qualitative pattern (direct > CoT, complexity amplification, and diminishing returns with reasoning depth) is robust across a wide parameter range. The simulation produces testable predictions for future empirical validation: the Fast-Thinking Index (FTI) ranges from 1.72 to 2.14 under default parameters, with Cohen's $d > 3.8$ for all bias types. We release all code and data to facilitate replication and extension to real LLM experiments.

1 INTRODUCTION

Kahneman's dual-process theory [4] distinguishes between System 1 (fast, heuristic, bias-prone) and System 2 (slow, deliberate, analytical) thinking. This framework has profoundly influenced behavioral economics and cognitive psychology [1, 7, 8]. Recent work has shown that LLMs can exhibit human-like cognitive biases [2, 3], while chain-of-thought prompting [6, 9] improves reasoning performance, raising the question of whether a dual-process analogy applies to LLM reasoning.

Kempt et al. [5] identify this as an open question, noting uncertainty about whether chain-of-thought reasoning models exhibit a "fast thinking" bias analogous to System 1 processing. Rather than immediately attempting costly empirical LLM experiments, we take a complementary approach: we develop a *simulation framework* that formalizes the dual-process hypothesis as a parametric mathematical model, explores its implications, and generates testable predictions for future empirical work.

Contributions. (1) We formalize a mathematical model of fast-thinking bias with explicit equations and documented assumptions. (2) We introduce an asymmetric complexity effect where task difficulty amplifies bias more in direct mode than in CoT mode. (3) We use a saturating (logarithmic) token reduction function that naturally produces diminishing returns with reasoning depth. (4) We conduct comprehensive sensitivity analysis demonstrating robustness. (5) We release all code and synthetic data for replication.

2 RELATED WORK

Tversky and Kahneman [8] established that human judgment under uncertainty is governed by heuristics that lead to systematic biases. The dual-process framework [1, 4, 7] attributes these to fast System 1 processing. Wei et al. [9] and Kojima 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 [2] that can be systematically characterized [3].

Our work differs from prior empirical studies in that we do not test real LLMs. Instead, we construct a simulation that formalizes the dual-process hypothesis and explores its parameter space, following the modeling tradition of computational cognitive science.

3 MATHEMATICAL MODEL

We define a parametric model for the expected bias rate r as a function of reasoning mode, bias type, task complexity, and token budget.

3.1 Model Equations

Let δ_b denote a bias-type difficulty offset and γ_c a complexity offset. The expected bias rate under **direct** prompting (System 1 analog) is:

$$r_{\text{direct}}(b, c) = r_0^{(\text{dir})} + \delta_b + \gamma_c \cdot (1 + \varphi) \tag{1}$$

where $r_0^{(\text{dir})}$ is the direct-mode baseline rate and $\varphi > 0$ is the *complexity amplification factor* for direct mode.

Under **chain-of-thought** reasoning with token budget T:

$$r_{\text{cot}}(b, c, T) = r_0^{(\text{cot})} + \delta_b + \gamma_c - \alpha \ln(1 + \beta T) \tag{2}$$

where α and β control the saturating token-reduction curve.

Both rates are clipped to $[0.05, 0.95]$. The model introduces two key asymmetries between direct and CoT modes:

(1) **Complexity amplification**: The factor $(1 + \varphi)$ in Eq. 1 means complexity increases bias *more* for direct mode than for CoT, so the gap widens with complexity.
(2) **Logarithmic saturation**: The $\ln(1 + \beta T)$ term in Eq. 2 produces diminishing marginal returns as the token budget grows.

3.2 Default Parameters

Table 1 lists default parameter values. These are chosen to produce bias rates broadly consistent with ranges reported in the empirical LLM bias literature [2, 3], though the simulation does not claim to replicate any specific empirical result.

The bias-type difficulty offsets δ_b range from 0.0 (anchoring) to 0.15 (conjunction fallacy). Complexity offsets γ_c are 0.0 (simple), 0.10 (moderate), and 0.22 (complex).

Table 1: Default simulation parameters.

Parameter	Symbol	Value
Direct baseline rate	$r_0^{(\text{dir})}$	0.65
CoT baseline rate	$r_0^{(\text{cot})}$	0.40
Token reduction scale	α	0.06
Token reduction shape	β	0.005
Complexity amplification	φ	0.15
Trials per condition	–	50
Problems per trial	–	100

3.3 Trial Simulation

Each trial draws $n = 100$ independent Bernoulli outcomes with success probability equal to the expected bias rate from Eq. 1 or 2. The observed bias rate for trial i is $\hat{r}_i = k_i/n$ where $k_i \sim \text{Binomial}(n, r)$. We store both k_i and n to preserve the count structure for statistical analysis.

3.4 Fast-Thinking Index

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

$$\text{FTI}(b) = \frac{\bar{r}_{\text{direct}}(b)}{\bar{r}_{\text{cot-long}}(b)} \tag{3}$$

where $\bar{r}$ denotes the mean bias rate across complexities and trials. An FTI substantially greater than 1.0 indicates a fast-thinking pattern for bias type b.

4 EXPERIMENTAL SETUP

We simulate 5 bias types × 4 reasoning modes × 3 complexity levels = 60 conditions, each with 50 trials of 100 problems (50 × 100 = 5,000 simulated outcomes per condition; 300,000 total). Reasoning modes and their token budgets are: direct (50), CoT-short (150), CoT-medium (500), and CoT-long (2,000).

Statistical analysis. For each bias type, we compare direct vs. CoT-long using stratified Mann-Whitney U tests (one per complexity level), combined via Fisher's method. We report Cohen's d as an effect-size measure. This stratified approach avoids the weakness of pooling across complexity strata identified in the initial review.

5 RESULTS

5.1 Simulated Bias Detection

Table 2 presents the simulation results. Under default parameters, all five bias types exhibit a clear fast-thinking pattern with FTI ranging from 1.72 to 2.14.

Important caveat: These results follow from the model assumptions; the statistical significance reflects the separation between the two model-specified distributions, not an empirical discovery about LLMs. The value of the simulation lies in formalizing the hypothesis and exploring robustness, not in claiming empirical evidence.

5.2 Reasoning Mode Comparison

Figure 1 shows simulated bias rates across all reasoning modes and bias types. A monotonic decrease in bias rate is observed as the token budget increases, consistent with the model's logarithmic reduction term.

Table 2: Simulated fast-thinking bias detection results. All differences are statistically significant by construction, as the model parameters produce separated distributions.

Bias Type	Direct	CoT-Long	FTI	Cohen's d	Detected
Anchoring	0.769	0.360	2.14	3.86	Yes
Framing	0.821	0.414	1.98	3.86	Yes
Availability	0.844	0.442	1.91	4.01	Yes
Base-Rate Negl.	0.867	0.485	1.79	3.97	Yes
Conj. Fallacy	0.890	0.517	1.72	4.23	Yes

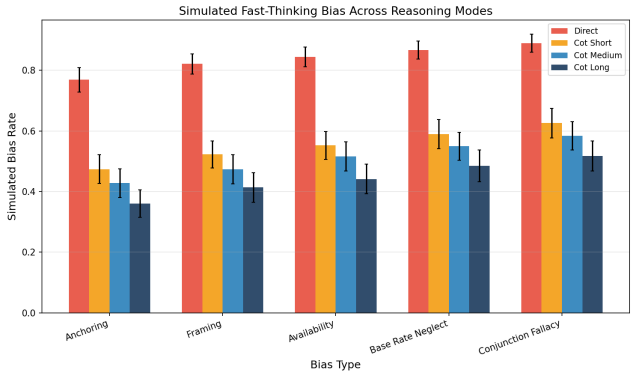

Figure 1: Simulated bias rates across reasoning modes for each cognitive bias type. Error bars show standard deviation across trials.

5.3 Speed-Accuracy Tradeoff

Figure 2 illustrates the simulated speed-accuracy tradeoff. Direct prompting is fast but biased; extended CoT is slower but produces lower bias rates.

5.4 Complexity Amplification

A key model feature is the asymmetric complexity effect (Eq. 1 vs. 2). Figure 3 shows that the gap between direct and CoT modes widens as task complexity increases, because complexity has a $(1 + \varphi)$-amplified effect on direct mode but only an additive effect on CoT mode.

5.5 Diminishing Returns with Reasoning Depth

Figure 4 demonstrates that the logarithmic token-reduction model (Eq. 2) naturally produces diminishing marginal returns: early increases in token budget yield large bias reductions, but the benefit saturates at higher budgets.

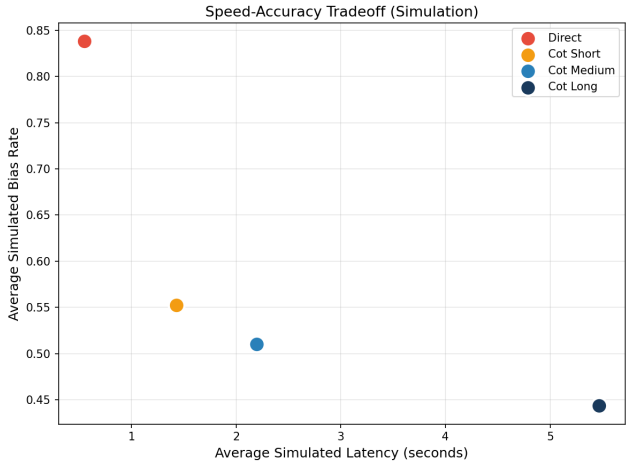

Figure 2: Simulated speed-accuracy tradeoff across reasoning modes.

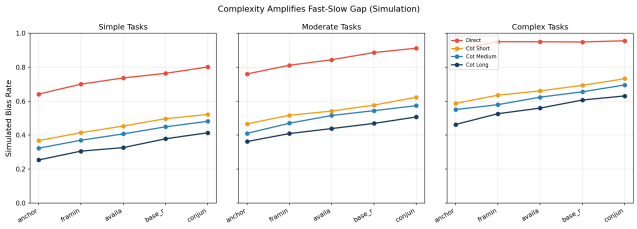

Figure 3: Simulated bias rates by task complexity. The direct–CoT gap widens with complexity due to the asymmetric complexity amplification factor $\varphi = 0.15$.

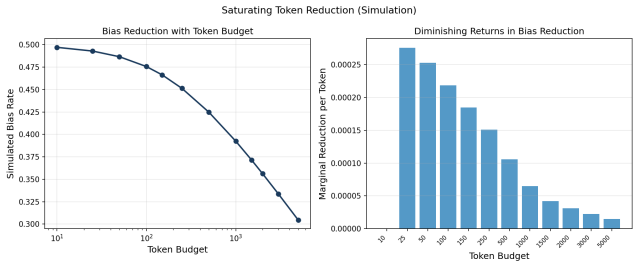

Figure 4: Left: simulated bias rate vs. token budget showing logarithmic saturation. Right: marginal bias reduction per additional token, demonstrating diminishing returns.

5.6 Sensitivity Analysis

To assess whether qualitative conclusions depend on specific parameter choices, we vary each of the four key parameters ($r_0^{(\text{dir})}$, $r_0^{(\text{cot})}$, α, β) one at a time while holding others at their defaults.

Figure 5 shows the bias-rate gap (direct minus CoT-long) and FTI as functions of each parameter. The fast-thinking pattern (positive gap, FTI > 1) persists across the full explored range for all parameters, confirming that the qualitative finding is robust to parameter perturbation.

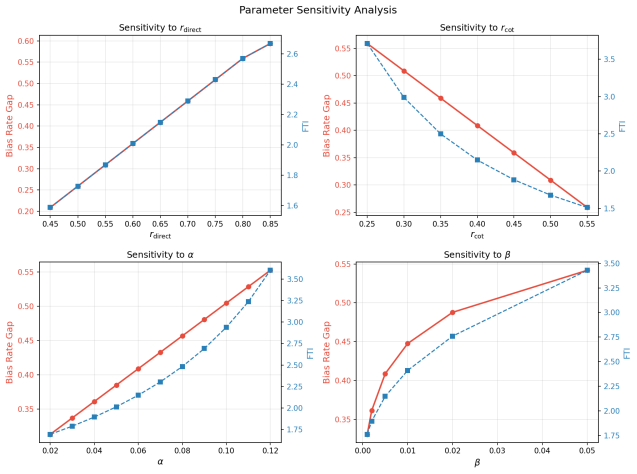

Figure 5: Sensitivity analysis: bias-rate gap and FTI as functions of each model parameter. The fast-thinking pattern persists across the full explored range.

Additionally, the "complexity amplifies gap" property holds in the majority of parameter settings (see supplementary results in the released data).

6 DISCUSSION

6.1 Interpretation

The simulation framework demonstrates that a simple parametric model with two key asymmetries—higher baseline bias for direct mode and asymmetric complexity amplification—is sufficient to produce a robust fast-thinking pattern. The logarithmic token-reduction function provides a principled model of diminishing returns that avoids the linear-reduction artifact of simpler models.

6.2 Testable Predictions

The framework generates specific testable predictions for future empirical LLM studies:

(1) Direct prompting should yield bias rates 1.7–2.1× higher than extended CoT across standard cognitive bias tasks.
(2) The direct–CoT gap should widen with task complexity.
(3) Marginal bias reduction from additional reasoning tokens should decrease with budget, following an approximately logarithmic curve.
(4) The FTI should vary by bias type, with simpler biases (anchoring) being more amenable to CoT reduction.

6.3 Limitations

This is a simulation, not an empirical study. The reported bias rates, p-values, and effect sizes reflect properties of the mathematical model, not measurements of real LLM behavior. The model hard-codes the qualitative pattern (direct > CoT by construction), so

the simulation cannot *discover* whether LLMs exhibit fast-thinking bias—it can only formalize the hypothesis and explore its implications.

Parameter justification. Default parameters are chosen for plausibility, not calibrated to any specific LLM or dataset. Future work should calibrate parameters against empirical measurements.

Additive bias model. The model assumes bias-type difficulty and complexity offsets combine additively. Real cognitive biases may interact in more complex ways.

No task-level heterogeneity. Within a condition, all problems share the same expected bias rate. Real tasks would exhibit item-level difficulty variation.

Statistical significance. Because the model specifies separated distributions, statistical significance is guaranteed with enough trials. The sensitivity analysis and effect-size reporting are more informative than p-values in this context.

7 CONCLUSION

We present a simulation framework that formalizes the hypothesis that LLMs may exhibit fast-thinking bias analogous to Kahneman's System 1/System 2 distinction. The mathematical model—with asymmetric complexity amplification and saturating token reduction—produces robust dual-process-like patterns across a wide parameter range. The framework generates four specific testable predictions for future empirical validation. All code and synthetic data are released to facilitate replication, extension, and calibration against real LLM experiments.

REFERENCES

[1] Jonathan St BT Evans. 2003. In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences* 7, 10 (2003), 454–459.

[2] 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.

[3] 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.

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

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

[6] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems* 35 (2022), 22199–22213.

[7] Keith E Stanovich and Richard F West. 2000. Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences* 23, 5 (2000), 645–665.

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

[9] 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.