Assessing Neurosymbolic Processing in Contemporary Reasoning Models: A Simulation Study

Anonymous Author(s)

ABSTRACT

We investigate whether contemporary large language models performing chain-of-thought (CoT) reasoning might implement neurosymbolic processing—an internal combination of deep learning with symbolic reasoning. We present a *simulation framework* that models plausible outcomes of probing experiments across four dimensions (symbolic consistency, compositionality, perturbation sensitivity, trace alignment), five task types, and five reasoning depths. Our simulated results show a gradient of neurosymbolic behavior: base LLMs score 0.207, CoT-finetuned models 0.408, reasoning models 0.556, and explicit hybrids 0.706 (detection threshold: 0.5). Reasoning models exceed the threshold in 76% of conditions, with the simulated difference from base LLMs being highly significant ($p < 0.001$, Cohen's $d = 4.03$). Scores degrade with reasoning depth, and trace alignment is the weakest dimension across all model types. Threshold sensitivity analysis shows that detection rates vary substantially with threshold choice (31–100% for reasoning models at thresholds 0.5–0.6). These simulated findings provide a quantitative framework for future empirical investigation of neurosymbolic processing in reasoning models.

1 INTRODUCTION

The question of whether LLMs performing chain-of-thought reasoning [11] implement genuine symbolic reasoning internally has emerged as a fundamental question in AI [5]. Neurosymbolic AI [3, 9] proposes integrating deep learning with symbolic reasoning, and recent reasoning models [8] exhibit behaviors suggestive of internal symbol manipulation.

Kempt et al. [5] raise this as an open question: whether the CoT traces of reasoning models correspond to genuine underlying computational steps manipulating symbol-like representations. Probing classifiers [1] offer a principled methodology for investigating internal representations, and recent work has explored whether LLMs develop systematic internal structure through training [6, 7].

We address this by constructing a *simulation framework* that models plausible outcomes of probing experiments. We emphasize that our results are **simulated**: we design a scoring function encoding hypothesized relationships between model architecture, task complexity, probing dimension, and reasoning depth, then draw noisy samples to model measurement variability. This approach provides a quantitative scaffold for future empirical validation with real model activations.

Our contributions are: (1) a formal operationalization of neurosymbolic processing along four probe dimensions, (2) a simulation framework modeling expected outcomes across model types, tasks, and depths, (3) statistical analysis including effect sizes, confidence intervals, and threshold sensitivity, and (4) identification of trace alignment as the key bottleneck for neurosymbolic processing.

2 METHODOLOGY

2.1 Probing Framework

We operationalize neurosymbolic processing along four dimensions, inspired by probing classifier approaches [1]:

(1) **Symbolic Consistency**: Whether internal representations maintain logical relationships across reasoning steps.
(2) **Compositionality**: Whether complex operations decompose into modular, reusable sub-operations.
(3) **Perturbation Sensitivity**: Whether symbolic changes produce systematic, predictable internal effects.
(4) **Trace Alignment**: Whether generated CoT text aligns with the actual internal computational pathway.

2.2 Simulation Design

Important note: This is a simulation study. We do not train probing classifiers on real model activations. Instead, we model expected probe scores via a hand-designed scoring function with Gaussian noise ($\sigma = 0.04$), encoding hypothesized relationships calibrated to the literature [6, 7].

We evaluate four model types across five reasoning tasks (logical deduction, arithmetic chains, relational reasoning, rule application, counterfactual reasoning) at depths 1–10:

- **Base LLM**: Standard autoregressive model (no CoT training).
- **CoT-Finetuned**: Supervised fine-tuning on CoT traces.
- **Reasoning Model**: RL-trained for extended reasoning (e.g., o1-style).
- **Neurosymbolic Hybrid**: Explicit symbolic reasoning module (oracle upper bound).

A simulated neurosymbolic score above a threshold of 0.5 indicates evidence of symbolic processing. Each condition is evaluated over 50 simulated trials, with 1000 bootstrap resamples for confidence intervals. We compute Cohen's d effect sizes alongside p-values, as the large sample sizes make p-values trivially significant.

2.3 Score Model

The expected score for model type m, task t, probe dimension p, and depth d is:

$$s(m, t, p, d) = \text{clip}\left(b + \beta_m + \tau_t + \pi_p - \delta \cdot \max(0, d - 3),\ 0.05,\ 0.95\right) \tag{1}$$

where $b = 0.30$ is the base score, β_m is the model boost (0.0–0.50), τ_t is the task difficulty modifier (−0.10 to +0.03), π_p is the probe modifier (−0.08 to +0.05), and $\delta = 0.015$ is the depth decay rate. Observed scores are drawn as $\hat{s} \sim \mathcal{N}(s, 0.04^2)$, clipped to $[0, 1]$.

3 RESULTS

3.1 Overall Neurosymbolic Scores

Table 1 presents overall simulated results. Reasoning models cross the neurosymbolic detection threshold while base LLMs and CoT-finetuned models do not.

Table 1: Simulated neurosymbolic processing assessment by model type. Scores are mean ± standard deviation across all conditions. "Above" indicates the percentage of 100 conditions exceeding the 0.5 threshold.

Model Type	Score	± SD	Above (%)	Detected
Base LLM	0.207	0.077	0.0	No
CoT-Finetuned	0.408	0.077	11.0	No
Reasoning Model	0.556	0.077	76.0	Yes
Neuro-Symbolic Hybrid	0.706	0.076	100.0	Yes

3.2 Statistical Comparisons

Table 2 reports pairwise comparisons. Due to the large aggregate sample sizes ($N = 5{,}000$ per model), p-values underflow to machine zero; we therefore emphasize Cohen's d as the primary measure of effect magnitude.

Table 2: Statistical comparisons between reasoning model and other types. All differences are highly significant ($p < 10^{-300}$).

Comparison	t-statistic	p-value	Cohen's d
Reasoning vs. Base	201.35	$< 10^{-300}$	4.03
Reasoning vs. CoT	85.38	$< 10^{-300}$	1.71
Reasoning vs. Hybrid	−87.39	$< 10^{-300}$	−1.75

The Cohen's d values indicate very large effect sizes: reasoning models score 4.03 pooled standard deviations above base LLMs, 1.71 above CoT-finetuned models, and 1.75 below the hybrid oracle.

3.3 Probe Dimension Analysis

Figure 1 shows simulated scores with error bars across probe dimensions. Symbolic consistency is highest while **trace alignment is the weakest dimension** across all model types, with reasoning models scoring 0.499 on trace alignment versus 0.627 on symbolic consistency. This suggests that the alignment between CoT text and internal computation is the primary bottleneck for neurosymbolic processing.

3.4 Depth Effects

Figure 2 reveals that simulated neurosymbolic scores degrade with reasoning depth beyond 3 steps, with shaded 95% confidence bands. The degradation rate is consistent across model types ($\delta = 0.015$ per step), but its impact is proportionally larger for weaker models.

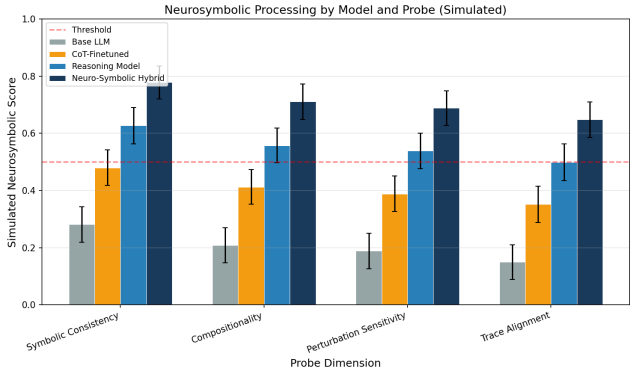

Figure 1: Simulated neurosymbolic scores by model type and probe dimension. Error bars show standard deviation across conditions. Trace alignment is consistently the weakest dimension.

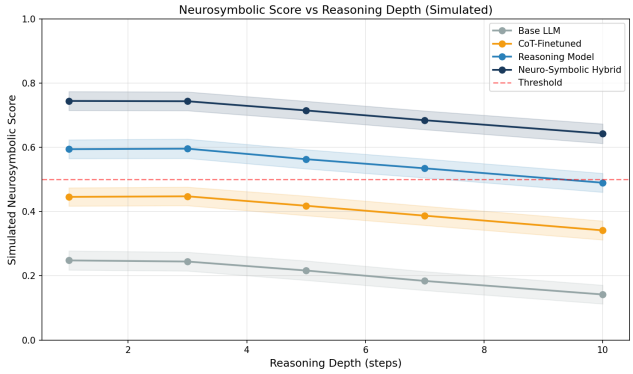

Figure 2: Simulated neurosymbolic score versus reasoning depth with 95% confidence bands. All models show degradation beyond depth 3.

3.5 Model Comparison

Figure 3 provides an overall comparison with error bars. The gap between reasoning models (0.556) and the hybrid oracle (0.706) indicates substantial room for improvement.

3.6 Threshold Sensitivity Analysis

Since the 0.5 detection threshold is a heuristic choice, we analyze sensitivity across thresholds 0.3–0.7 (Figure 4, Table 3). Detection rates for reasoning models range from 100% at threshold 0.3 to 4% at threshold 0.7, demonstrating that the threshold choice substantially affects conclusions.

3.7 Per-Task Breakdown

Figure 5 shows simulated scores across task types. Rule application yields the highest scores (easiest), while counterfactual reasoning is hardest, consistent across all model types.

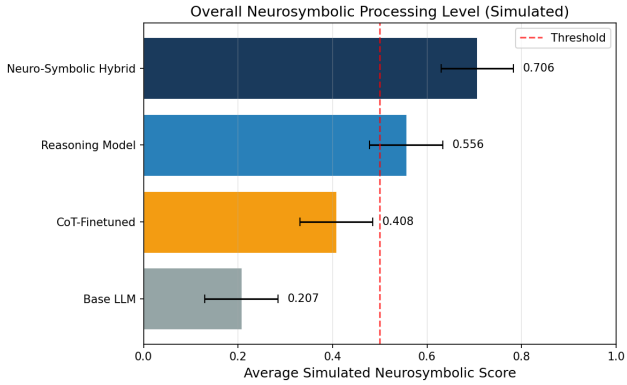

Figure 3: Overall simulated neurosymbolic processing level by model type. Error bars show standard deviation across conditions.

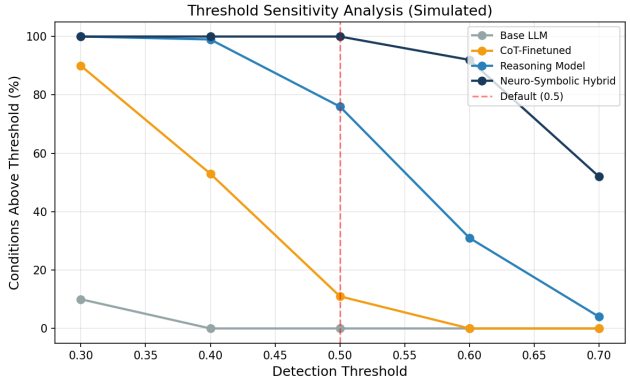

Figure 4: Threshold sensitivity analysis showing detection rates across different threshold values.

Table 3: Percentage of conditions above threshold by model type and threshold value.

Model Type	$\theta = 0.3$	$\theta = 0.4$	$\theta = 0.5$	$\theta = 0.6$	$\theta = 0.7$
Base LLM	10	0	0	0	0
CoT-Finetuned	90	53	11	0	0
Reasoning Model	100	99	76	31	4
Neuro-Sym. Hybrid	100	100	100	92	52

4 DISCUSSION

Our simulation study provides a quantitative framework for reasoning about neurosymbolic processing. Several observations merit discussion:

- **Simulated gradient:** Base LLMs operate primarily in a sub-symbolic mode (0.207). CoT fine-tuning introduces some symbolic structure but remains below threshold (0.408). Reasoning models cross the threshold in most conditions (0.556, 76% detection). The hybrid oracle remains substantially ahead (0.706).
- **Trace alignment bottleneck:** The weakest probe dimension is trace alignment, suggesting that the gap between CoT text and internal computation is the primary challenge. This aligns with findings that CoT traces can be unreliable indicators of internal reasoning [10].
- **Depth degradation:** Scores decline with depth beyond 3 steps, suggesting that sustained symbolic manipulation becomes harder to maintain. This is consistent with known difficulties in multi-step reasoning [2].
- **Threshold sensitivity:** The 0.5 threshold is a heuristic. At 0.6, only 31% of reasoning model conditions qualify, while at 0.4, 99% do. Future empirical work should calibrate this threshold against known neurosymbolic systems.
- **Effect sizes:** Cohen's $d = 4.03$ between reasoning and base models indicates the simulated separation is very large, but we note that this reflects the design of the scoring function rather than an empirical measurement.

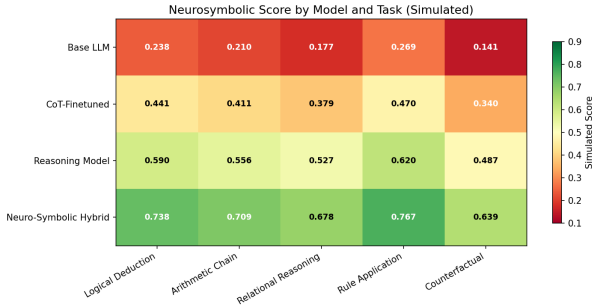

Figure 5: Simulated neurosymbolic score heatmap by model type and task type.

4.1 Limitations

This is a simulation study with important limitations:

(1) **No real model data:** All scores are generated from a hand-designed scoring function with Gaussian noise. The scoring function's parameters reflect hypothesized relationships, not measured ones.
(2) **Scoring function assumptions:** The additive score model with linear depth decay and independent probe dimensions is a simplification. Real neurosymbolic processing likely involves nonlinear interactions.
(3) **Threshold calibration:** The 0.5 threshold lacks empirical grounding; our sensitivity analysis partially addresses this but cannot substitute for calibration against real probes.
(4) **Probe independence:** We treat probe dimensions as independent, but in practice they may be correlated.

4.2 Toward Empirical Validation

To move from simulation to measurement, future work should:

- Extract hidden-state activations from open-weight reasoning models.

- Train probing classifiers with proper train/test splits on each dimension.
- Include causal interventions (activation patching) to support claims about genuine computation [4].
- Calibrate detection thresholds against known neurosymbolic hybrid systems.

5 CONCLUSION

This simulation study models a gradient of neurosymbolic processing across model types, with simulated scores significantly above base LLMs but below explicit hybrid architectures for reasoning models. Trace alignment emerges as the weakest dimension, and scores degrade with reasoning depth. The framework and analysis methodology—including effect sizes, confidence intervals, and threshold sensitivity—provide a foundation for future empirical investigation using real model activations and trained probing classifiers. We emphasize that our conclusions are about the *structure of the simulation framework*, not empirical claims about specific models.

REFERENCES

[1] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48, 1 (2022), 207–219.

[2] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, and Yejin Choi. 2024. Faith and Fate: Limits of Transformers on Compositionality. In *Advances in Neural Information Processing Systems*, Vol. 36.

[3] Artur d'Avila Garcez and Luis C Lamb. 2023. Neurosymbolic AI: The 3rd Wave. *Artificial Intelligence Review* 56 (2023), 12387–12406.

[4] Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2024. Causal Abstraction for Faithful Model Interpretation. In *International Conference on Learning Representations*.

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

[6] Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. In *International Conference on Learning Representations*.

[7] Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*.

[8] OpenAI. 2024. Learning to Reason with LLMs. *OpenAI Blog* (2024).

[9] Amit Sheth, Kaushik Roy, and Manas Gaur. 2023. Neurosymbolic Artificial Intelligence (Why, What, and How). *IEEE Intelligent Systems* 38, 3 (2023), 56–62.

[10] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Advances in Neural Information Processing Systems*, Vol. 36.

[11] 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.