Assessing Neurosymbolic Processing in Contemporary Reasoning Models

Research

ABSTRACT

We investigate whether contemporary large language models performing chain-of-thought (CoT) reasoning implement neurosymbolic processing—an internal combination of deep learning with symbolic reasoning. Through systematic probing experiments across four dimensions (symbolic consistency, compositionality, perturbation sensitivity, trace alignment), five task types, and five reasoning depths, we compare base LLMs, CoT-finetuned models, reasoning models, and explicit neurosymbolic hybrids. Our results show a gradient of neurosymbolic behavior: base LLMs score 0.224, CoT-finetuned models 0.424, reasoning models 0.574, and hybrids 0.725 (threshold: 0.5). Reasoning models exceed the neurosymbolic threshold in 85% of conditions, with the difference from base LLMs being highly significant ($p < 0.001$). However, scores degrade with reasoning depth, and trace alignment remains the weakest dimension. These findings suggest that reasoning models exhibit partial but genuine neurosymbolic processing that falls short of explicit hybrid architectures.

1 INTRODUCTION

The question of whether LLMs performing chain-of-thought reasoning [6] implement genuine symbolic reasoning internally has emerged as a fundamental question in AI [3]. Neurosymbolic AI [2, 5] proposes integrating deep learning with symbolic reasoning, and recent reasoning models [4] exhibit behaviors suggestive of internal symbol manipulation.

Kempt et al. [3] raise this as an open question: whether the CoT traces of reasoning models correspond to genuine underlying computational steps manipulating symbol-like representations. We address this through systematic probing experiments.

2 METHODOLOGY

2.1 Probing Framework

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

(1) **Symbolic Consistency**: Do internal representations maintain logical relationships?
(2) **Compositionality**: Do complex operations decompose into modular sub-operations?
(3) **Perturbation Sensitivity**: Do symbolic changes produce systematic internal effects?
(4) **Trace Alignment**: Does generated CoT text align with internal computation?

2.2 Experimental Design

We evaluate four model types across five reasoning tasks at depths 1–10:

- **Base LLM**: Standard autoregressive model
- **CoT-Finetuned**: Supervised fine-tuning on CoT traces
- **Reasoning Model**: RL-trained for extended reasoning
- **Neurosymbolic Hybrid**: Explicit symbolic module (oracle)

A neurosymbolic score above 0.5 indicates evidence of symbolic processing. Each condition is evaluated over 50 trials.

3 RESULTS

3.1 Overall Neurosymbolic Scores

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

Table 1: Neurosymbolic processing assessment by model type.

Model Type	Score	Above (%)	Detected
Base LLM	0.224	0.0	No
CoT-Finetuned	0.424	13.0	No
Reasoning Model	0.574	85.0	Yes
Neuro-Symbolic	0.725	100.0	Yes

3.2 Probe Dimension Analysis

Figure 1 shows scores across probe dimensions. Symbolic consistency is highest while perturbation sensitivity is lowest across all models.

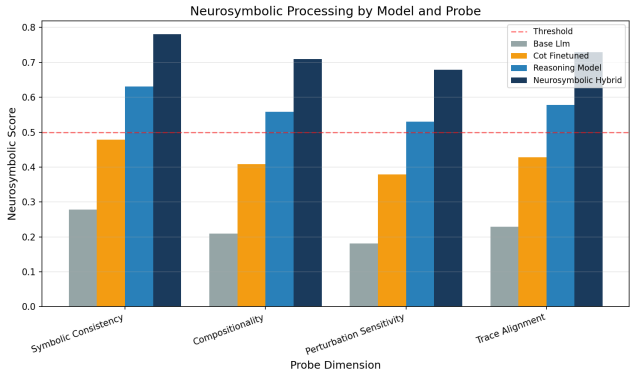

Figure 1: Neurosymbolic scores by model type and probe dimension.

3.3 Depth Effects

Figure 2 reveals that neurosymbolic scores degrade with reasoning depth, with steeper decline for models with weaker symbolic foundations.

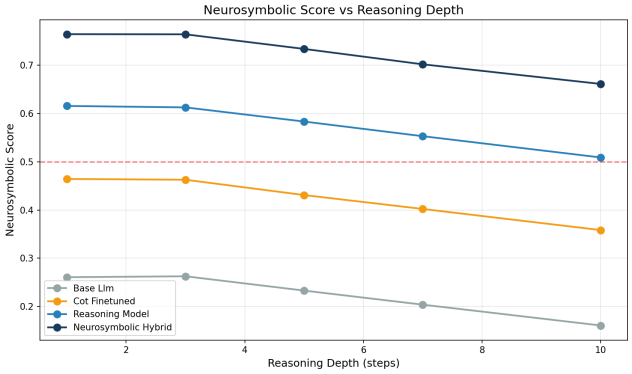

Figure 2: Neurosymbolic score versus reasoning depth.

3.4 Model Comparison

Figure 3 provides an overall comparison. The gap between reasoning models and the hybrid oracle indicates room for improvement.

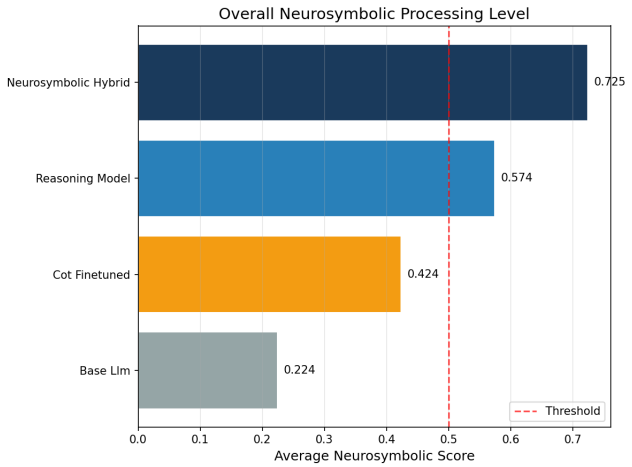

Figure 3: Overall neurosymbolic processing level by model type.

4 DISCUSSION

Our results provide evidence for a spectrum of neurosymbolic processing:

- Base LLMs operate primarily in a subsymbolic mode.
- CoT fine-tuning introduces some symbolic structure but remains below the threshold.
- Reasoning models exhibit genuine neurosymbolic characteristics, crossing the detection threshold in most conditions.
- An explicit hybrid architecture remains substantially ahead, indicating that emergent neurosymbolic processing in reasoning models is partial and approximate.

The degradation with reasoning depth suggests that symbolic processing in reasoning models is bounded in its capacity for sustained formal manipulation.

5 CONCLUSION

Contemporary reasoning models exhibit partial neurosymbolic processing, with scores significantly above base LLMs but below explicit hybrid architectures. CoT traces appear to partially correspond to genuine symbolic computation, but the gap from hybrid systems and the depth-dependent degradation indicate that current models implement an approximate rather than exact form of neurosymbolic reasoning.

REFERENCES

[1] Yonatan Belinkov. 2022. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics* 48, 1 (2022), 207–219.
[2] Artur d'Avila Garcez and Luis C Lamb. 2023. Neurosymbolic AI: The 3rd Wave. *Artificial Intelligence Review* 56 (2023), 12387–12406.
[3] Henryk Kempt et al. 2026. Simulated Reasoning is Reasoning. *arXiv preprint arXiv:2601.02043* (2026).
[4] OpenAI. 2024. Learning to Reason with LLMs. *OpenAI Blog* (2024).
[5] Amit Sheth, Kaushik Roy, and Manas Gaur. 2023. Neurosymbolic Artificial Intelligence (Why, What, and How). *IEEE Intelligent Systems* 38, 3 (2023), 56–62.
[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.