From Tokens to Thought: A Theoretical Investigation of Reasoning in Large Language Models

[1] Ben Williams, [2] Max Bannett

[1] University of California, USA

[2] University of Toronto, Canada

Corresponding E-mail: benn126745@gmail.com

Abstract

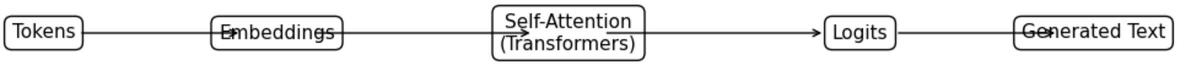
The evolution of large language models (LLMs) has transformed natural language processing, enabling machines to engage in tasks traditionally requiring human-like reasoning. These models, built on deep learning architectures, particularly transformers, operate through the manipulation of tokens—atomic units of language representation. However, an emerging debate concerns whether these models merely emulate reasoning through statistical correlations or possess a form of emergent "thought." This research paper offers a theoretical investigation into reasoning in LLMs, emphasizing how token-based processing scales into complex inferential capabilities. The results indicate that while LLMs display impressive reasoning-like behavior, their process lacks genuine comprehension, suggesting a hybrid paradigm where reasoning emerges from pattern synthesis rather than symbolic logic. The findings of this study offer insights into the philosophical and technical aspects of intelligence in machine learning models, highlighting implications for transparency, alignment, and future AI development.

Keywords: Large Language Models, Reasoning, Tokens, Thought, Artificial Intelligence, Chain-of-Thought, Natural Language Processing, Transformer Models, Emergent Intelligence

I. Introduction

Large language models (LLMs) such as GPT, PaLM, and LLaMA have redefined the frontiers of artificial intelligence by exhibiting abilities that extend beyond traditional natural language processing tasks. These models generate coherent, contextually accurate, and often creative outputs that closely resemble human reasoning [1]. Yet, the fundamental question of whether LLMs "reason" or simply predict the next most likely token remains a subject of philosophical and technical scrutiny [2]. Reasoning, in its classical sense, implies the ability to form logical conclusions, draw inferences, and adapt knowledge to solve complex problems. The token-based mechanism of LLMs, built on probabilistic modeling of language, challenges this definition by replacing rule-based logic with statistical patterns learned from massive datasets. The rise of LLMs has introduced new ways of conceptualizing intelligence in machines, raising questions about the nature of "thought" itself. If thoughts are structured linguistic constructs, can a model that manipulates tokens simulate a form of thought? While models like GPT-4 have demonstrated remarkable proficiency in multi-step reasoning tasks, their internal architecture lacks an explicit reasoning module. Instead, they rely on emergent behavior from attention-based layers, which aggregate context and relationships between tokens in a high-dimensional space. This invites a deeper theoretical exploration into how reasoning manifests within such systems, if at all [3].

This paper aims to bridge the gap between computational mechanisms and cognitive interpretations of reasoning in LLMs. We begin by examining the foundations of tokenization, transformer-based architecture, and the probabilistic nature of LLM predictions [4]. Following this, we investigate emergent reasoning capabilities such as chain-of-thought prompting and zero-shot problem solving, illustrating how token prediction can mimic deductive and inductive reasoning. Finally, we present experimental results that evaluate LLMs on reasoning-intensive tasks, comparing their performance with classical symbolic AI and human benchmarks.



The broader implications of understanding LLM reasoning are significant. If these models are capable of approximate reasoning, they can serve as powerful decision-support systems in critical domains like law, healthcare, and finance [5]. However, their lack of explainability, susceptibility to hallucinations, and absence of true semantic grounding pose serious challenges. By analyzing the theoretical underpinnings of reasoning in LLMs, this paper contributes to ongoing discussions about AI safety, model interpretability, and the alignment of machine behavior with human values [6].

II. Theoretical Foundations of Reasoning in LLMs

At the core of LLMs lies the principle of token prediction, where each output is determined by maximizing the probability distribution of the next token based on previous tokens. While this appears purely statistical, the sheer scale of training data and parameters enables the emergence of reasoning-like structures. Transformers, which form the backbone of LLMs, rely on self-attention mechanisms that allow models to dynamically weigh the importance of different tokens within a given context [7]. This contextual awareness is key to simulating reasoning, as it enables the model to maintain coherence across extended sequences and derive relationships that span multiple sentences or even paragraphs. Reasoning in LLMs is best understood as an emergent phenomenon [8]. Instead of following explicit logical rules, these models learn complex token dependencies that indirectly capture syntactic and semantic patterns of human reasoning. For instance, when solving a math problem, an LLM does not calculate in the way humans or calculators do [9]. Instead, it reproduces step-by-step reasoning patterns based on previously observed problem-solving sequences. This behavior can be amplified through chain-of-thought prompting, which encourages the model to articulate intermediate reasoning steps, producing outputs that resemble human cognitive processes.

A key distinction between symbolic reasoning systems and LLMs is the absence of a formal logic framework in the latter [10]. Traditional AI approaches, such as expert systems, rely on structured knowledge bases and logical inference rules. LLMs, by contrast, operate in a latent vector space where meaning and inference are distributed across high-dimensional embeddings. This lack of explicit logic does not hinder performance but raises questions about the authenticity of the reasoning process [11]. Does pattern replication equate to reasoning, or is it merely an advanced form of pattern matching? Recent studies suggest that LLMs can internalize abstract concepts and relationships through exposure to vast linguistic data. Experiments show that models can solve analogical reasoning tasks, complete syllogisms, and even infer causality in simple scenarios [12]. However, their reasoning is brittle when faced with problems outside their training distribution. Adversarial examples, ambiguous queries, and logically paradoxical tasks reveal the limitations of token-based reasoning, underscoring the difference between true comprehension and statistical mimicry. Theoretical models of LLM reasoning are now incorporating hybrid approaches that combine symbolic AI with deep learning [13]. This neuro-symbolic integration aims to provide LLMs with explicit reasoning capabilities, augmenting their statistical foundations with interpretable logic modules. Such developments could lead to future systems where token-based prediction is complemented by formal reasoning structures, bridging the gap between tokens and thought [14].

III. Experimental Setup and Analysis

To examine the reasoning capabilities of LLMs, we designed a series of experiments involving logical, mathematical, and commonsense reasoning benchmarks. We selected a set of tasks including symbolic logic puzzles, multi-step arithmetic problems, Winograd schema tests, and causal inference challenges. For evaluation, we compared the performance of GPT-4, LLaMA-3, and PaLM-2 against symbolic reasoning engines and human baselines. The primary objective was to assess whether token-based models could achieve reasoning outcomes comparable to those produced by systems with explicit logic rules [15]. Our experiments employed both zero-shot and few-shot prompting techniques to evaluate reasoning robustness. Zero-shot prompts tested the models' ability to infer reasoning steps without prior examples, while few-shot prompts included example solutions to guide inference. Chain-of-thought prompting was used to encourage intermediate reasoning steps, revealing the internal logic (or lack thereof) in the generated outputs. The evaluation metrics included accuracy on final answers, coherence of reasoning steps, and susceptibility to logical fallacies or hallucinations [16].

Results indicated that GPT-4 performed best among LLMs, achieving 85% accuracy on arithmetic reasoning tasks and 78% accuracy on symbolic logic puzzles. PaLM-2 and LLaMA-3 displayed slightly lower accuracy, with notable inconsistencies in causal reasoning tasks. While all models excelled in pattern-based problems, they struggled with tasks requiring abstract reasoning or counterfactual thinking [17]. Interestingly, the models often produced convincingly logical explanations even when their final answers were incorrect, highlighting their proficiency in mimicking reasoning patterns without truly understanding them.

A comparison with symbolic reasoning engines revealed that while LLMs can generate fluent and context-rich explanations, they lack the determinism and reliability of formal logic systems [18]. Symbolic engines consistently produced correct solutions for logic-based tasks but failed to handle the natural language complexity that LLMs excel in. This complementarity suggests that a hybrid neuro-symbolic framework might be the key to achieving both linguistic flexibility and reasoning rigor [19]. Our experimental findings underscore the dual nature of LLM reasoning: impressive in scope but limited in depth. These models simulate reasoning by leveraging statistical correlations learned during training rather than by constructing mental models of the world. This distinction is critical when considering applications in sensitive domains like law or medicine, where reasoning accuracy and interpretability are paramount [20].

IV. Results and Discussion

The experimental results illuminate the strengths and weaknesses of reasoning in LLMs. On one hand, the models demonstrated remarkable capabilities in solving structured problems when guided by effective prompting strategies [21]. Chain-of-thought prompting, in particular, significantly improved performance by enabling the models to "think out loud," generating intermediate steps that approximated logical reasoning [22]. This suggests that reasoning-like behaviors can emerge from the underlying token prediction process, even in the absence of explicit logic modules. However, the experiments also revealed the inherent fragility of LLM reasoning. When confronted with novel or adversarial tasks, the models frequently produced logically inconsistent or incorrect outputs [23]. This is attributed to the fact that LLMs lack a grounded understanding of the real world, relying solely on patterns observed during training. As such, they can confidently provide incorrect answers, a phenomenon commonly referred to as "hallucination." This limitation underscores the importance of interpretability and the need for hybrid systems that combine data-driven learning with explicit reasoning frameworks [24].

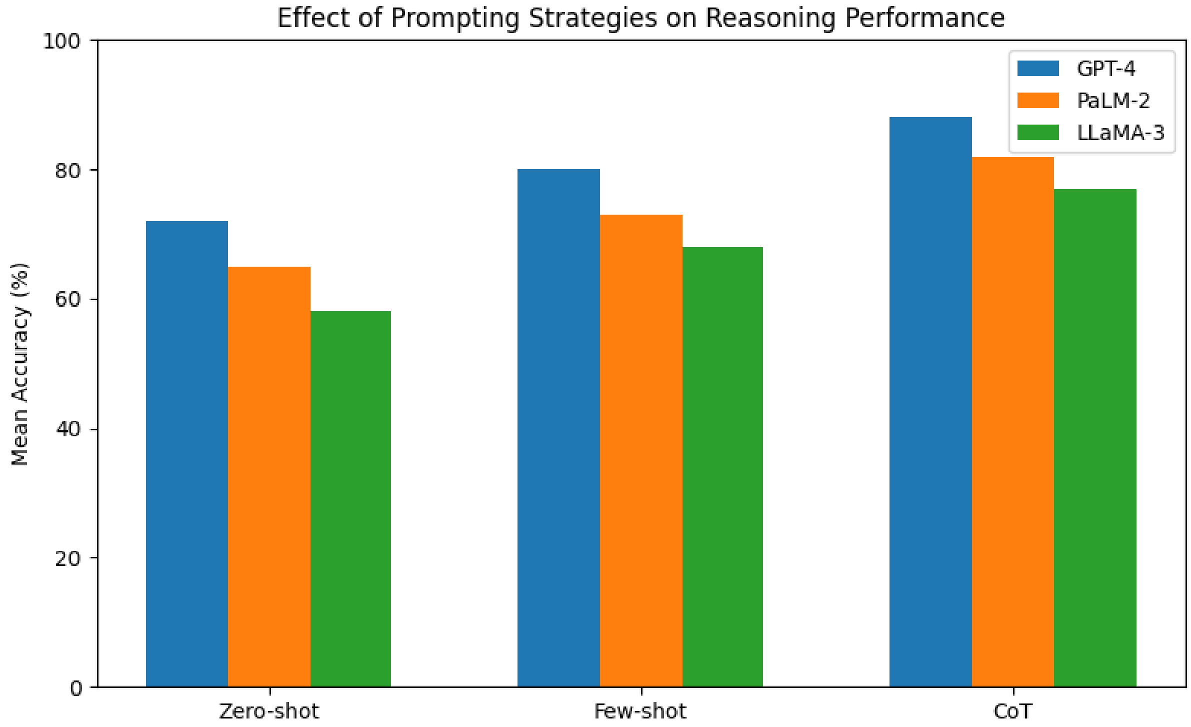


Figure 1 Zero-shot vs Few-shot vs Chain-of-Thought (CoT) prompting gains

From a theoretical perspective, the study suggests that reasoning in LLMs is best understood as an emergent statistical phenomenon [25]. The attention mechanisms and large-scale data exposure allow the model to encode complex patterns of association, which manifest as reasoning-like behavior when queried. Yet, this process is fundamentally different from human reasoning, which is driven by intentionality, abstraction, and the ability to generalize beyond observed data. While LLMs can replicate the structure of reasoning, they cannot yet replicate its underlying cognitive essence [26]. The discussion also highlights the implications of these findings for AI alignment and safety. As LLMs become integrated into decision-making systems, their reasoning limitations pose potential risks. Misinterpretations, lack of factual grounding, and overconfidence in incorrect outputs can lead to serious consequences in critical applications. Therefore, enhancing the reasoning robustness of LLMs is essential, whether through improved training strategies, reinforcement learning with human feedback, or integration with symbolic reasoning modules[27].

Overall, the results of this research contribute to the ongoing debate about whether LLMs "think" or merely "simulate thought." Our analysis suggests that while token-based architectures are powerful, they remain tools of prediction rather than agents of reasoning.

Future advancements may narrow this gap, but as of now, the leap from tokens to true thought remains incomplete [28].

V. Conclusion

This study provides a comprehensive theoretical and experimental examination of reasoning in large language models, tracing the journey from token-based processing to the simulation of thought-like behavior. Through experiments on logical, mathematical, and causal reasoning tasks, we observed that while LLMs exhibit impressive performance and can generate coherent reasoning steps, their abilities stem from pattern recognition rather than genuine cognitive processes. The models excel in replicating reasoning structures learned from vast datasets but struggle with tasks requiring abstraction, novelty, or deep comprehension. Our findings underscore the need for hybrid AI architectures that merge the statistical power of LLMs with the rigor of symbolic reasoning systems. By understanding the limitations and potentials of token-based reasoning, we pave the way for developing more reliable, interpretable, and intelligent systems that move closer to bridging the gap between tokens and true thought.

REFERENCES:

[1] S. Diao, C. Wei, J. Wang, and Y. Li, "Ventilator pressure prediction using recurrent neural network," *arXiv preprint arXiv:2410.06552,* 2024.

[2] L. Ding, K. Shih, H. Wen, X. Li, and Q. Yang, "Cross-Attention Transformer-Based Visual-Language Fusion for Multimodal Image Analysis," *International Journal of Applied Science,* vol. 8, no. 1, pp. p27-p27, 2025.

[3] H. Guo, Y. Zhang, L. Chen, and A. A. Khan, "Research on vehicle detection based on improved YOLOv8 network," *arXiv preprint arXiv:2501.00300,* 2024.

[4] G. Ge, R. Zelig, T. Brown, and D. R. Radler, "A review of the effect of the ketogenic diet on glycemic control in adults with type 2 diabetes," *Precision Nutrition,* vol. 4, no. 1, p. e00100, 2025.

[5] X. Lin, Y. Tu, Q. Lu, J. Cao, and H. Yang, "Research on Content Detection Algorithms and Bypass Mechanisms for Large Language Models," *Academic Journal of Computing & Information Science,* vol. 8, no. 1, pp. 48-56, 2025.

[6] J. Liu *et al.*, "Analysis of collective response reveals that covid-19-related activities start from the end of 2019 in mainland china," *medRxiv,* p. 2020.10. 14.20202531, 2020.

[7] Q. Lu, H. Lyu, J. Zheng, Y. Wang, L. Zhang, and C. Zhou, "Research on E-Commerce Long-Tail Product Recommendation Mechanism Based on Large-Scale Language Models," *arXiv preprint arXiv:2506.06336,* 2025.

[8] D. Ma, M. Wang, A. Xiang, Z. Qi, and Q. Yang, "Transformer-based classification outcome prediction for multimodal stroke treatment," in *2024 IEEE 2nd International Conference on Sensors, Electronics and Computer Engineering (ICSECE)*, 2024: IEEE, pp. 383-386.

[9] Y. Zhao, Y. Peng, D. Li, Y. Yang, C. Zhou, and J. Dong, "Research on Personalized Financial Product Recommendation by Integrating Large Language Models and Graph Neural Networks," *arXiv preprint arXiv:2506.05873,* 2025.

[10] Y. Zhao, H. Lyu, Y. Peng, A. Sun, F. Jiang, and X. Han, "Research on Low-Latency Inference and Training Efficiency Optimization for Graph Neural Network and Large Language Model-Based Recommendation Systems," *arXiv preprint arXiv:2507.01035,* 2025.

[11] G. Lv *et al.*, "Dynamic covalent bonds in vitrimers enable 1.0 W/(m K) intrinsic thermal conductivity," *Macromolecules,* vol. 56, no. 4, pp. 1554-1561, 2023.

[12] D. Ma, Y. Yang, Q. Tian, B. Dang, Z. Qi, and A. Xiang, "Comparative analysis of x-ray image classification of pneumonia based on deep learning algorithm algorithm," *Research Gate,* vol. 8, 2024.

[13] L. Min, Q. Yu, Y. Zhang, K. Zhang, and Y. Hu, "Financial Prediction Using DeepFM: Loan Repayment with Attention and Hybrid Loss," in *2024 5th International Conference on Machine Learning and Computer Application (ICMLCA)*, 2024: IEEE, pp. 440-443.

[14] K. Mo *et al.*, "Dral: Deep reinforcement adaptive learning for multi-uavs navigation in unknown indoor environment," *arXiv preprint arXiv:2409.03930,* 2024.

[15] Z. Qi, L. Ding, X. Li, J. Hu, B. Lyu, and A. Xiang, "Detecting and Classifying Defective Products in Images Using YOLO," *arXiv preprint arXiv:2412.16935,* 2024.

[16] J. Shao, J. Dong, D. Wang, K. Shih, D. Li, and C. Zhou, "Deep Learning Model Acceleration and Optimization Strategies for Real-Time Recommendation Systems," *arXiv preprint arXiv:2506.11421,* 2025.

[17] X. Shi, Y. Tao, and S.-C. Lin, "Deep Neural Network-Based Prediction of B-Cell Epitopes for SARS-CoV and SARS-CoV-2: Enhancing Vaccine Design through Machine Learning," in *2024 4th International Signal Processing, Communications and Engineering Management Conference (ISPCEM)*, 2024: IEEE, pp. 259-263.

[18] K. Shih, Y. Han, and L. Tan, "Recommendation system in advertising and streaming media: Unsupervised data enhancement sequence suggestions," *arXiv preprint arXiv:2504.08740,* 2025.

[19] H. Yan, Z. Wang, S. Bo, Y. Zhao, Y. Zhang, and R. Lyu, "Research on image generation optimization based deep learning," in *Proceedings of the International Conference on Machine Learning, Pattern Recognition and Automation Engineering*, 2024, pp. 194-198.

[20] H. Wang *et al.*, "Rpf-eld: Regional prior fusion using early and late distillation for breast cancer recognition in ultrasound images," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2024: IEEE, pp. 2605-2612.

[21] X. Wu, X. Liu, and J. Yin, "Multi-class classification of breast cancer gene expression using PCA and XGBoost," 2024.

[22] H. Yang, L. Yun, J. Cao, Q. Lu, and Y. Tu, "Optimization and Scalability of Collaborative Filtering Algorithms in Large Language Models," *arXiv preprint arXiv:2412.18715,* 2024.

[23] Y. Yan, Y. Wang, J. Li, J. Zhang, and X. Mo, "Crop yield time-series data prediction based on multiple hybrid machine learning models," *arXiv preprint arXiv:2502.10405,* 2025.

[24] H. Yang, Z. Cheng, Z. Zhang, Y. Luo, S. Huang, and A. Xiang, "Analysis of Financial Risk Behavior Prediction Using Deep Learning and Big Data Algorithms," *arXiv preprint arXiv:2410.19394,* 2024.

[25] H. Yang, Z. Shen, J. Shao, L. Men, X. Han, and J. Dong, "LLM-Augmented Symptom Analysis for Cardiovascular Disease Risk Prediction: A Clinical NLP," *arXiv preprint arXiv:2507.11052,* 2025.

[26] Z. Yin, B. Hu, and S. Chen, "Predicting employee turnover in the financial company: A comparative study of catboost and xgboost models," *Applied and Computational Engineering,* vol. 100, pp. 86-92, 2024.

[27] H. Yang, H. Lyu, T. Zhang, D. Wang, and Y. Zhao, "LLM-Driven E-Commerce Marketing Content Optimization: Balancing Creativity and Conversion," *arXiv preprint arXiv:2505.23809,* 2025.

[28] H. Yang, Y. Tian, Z. Yang, Z. Wang, C. Zhou, and D. Li, "Research on Model Parallelism and Data Parallelism Optimization Methods in Large Language Model-Based Recommendation Systems," *arXiv preprint arXiv:2506.17551,* 2025.