International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Beyond Prompt Engineering: The Evolution of Reasoning in Advanced Large Language Models

Nan Wu

Milpitas, CA, USA
Chris.wunan88@gmail.com

Abstract

This paper explores the evolving role of prompt engineering as large language models (LLMs) develop enhanced intrinsic reasoning capabilities. Initially essential for effective model performance, explicit prompting techniques are becoming less crucial with advanced models like GPT-4.5 and DeepSeek R1. Benchmark analyses indicate that intrinsic reasoning now solves most reasoning tasks efficiently, though explicit prompting still provides incremental benefits in specialized scenarios. Future directions emphasize intrinsic reasoning improvements, automated prompting strategies, and refined evaluation methods, marking a fundamental shift in leveraging LLMs.

Keywords: Prompt Engineering, Intrinsic Reasoning, Large Language Models, Chain-of-Thought, Benchmarks

Funding Declaration

The author declares that no funding was received for conducting this study or for the preparation of this manuscript.

I. INTRODUCTION

Prompt engineering, the practice of carefully crafting inputs (prompts) to elicit desired outputs from large language models (LLMs), emerged as a pivotal technique alongside early advances in models like GPT-3. Initially, large language models demonstrated limited ability to solve complex reasoning problems directly; thus, effective prompts became crucial in guiding models toward more accurate and detailed responses. Techniques such as zero-shot and few-shot prompting evolved, allowing users to instruct models through minimal examples, dramatically improving performance on diverse natural language processing (NLP) tasks.[1]

A significant advancement occurred with Chain-of-Thought (CoT) prompting, where models were explicitly guided to break down complex problems into intermediate reasoning steps[2]. This method substantially improved reasoning performance, making previously unsolvable tasks manageable for models like Google's PaLM and OpenAI's GPT series. Despite its effectiveness, prompt engineering introduced complexities, as subtle changes in prompt phrasing could dramatically affect outcomes, creating challenges around reliability, efficiency, and scalability.

Recently, state-of-the-art reasoning-focused LLMs such as GPT-4.5, GPT-O1/O1 Pro, DeepSeek R1, and Anthropic's Claude models have exhibited intrinsic reasoning capabilities due to advanced training

methods like reinforcement learning with human feedback (RLHF) and extensive instruction tuning. These models increasingly deliver accurate and efficient reasoning outcomes without needing explicitly detailed prompts.

This evolution prompts the critical question: Is traditional prompt engineering still essential, or is its importance declining as modern LLMs inherently possess more robust reasoning abilities?

This paper aims to address this question through a comprehensive analysis of the evolving role of prompt engineering in modern large language models, exploring empirical evidence from benchmarks and case studies, efficiency trade-offs, and future trends.

II. Evolution of Prompt Engineering Techniques

Prompt engineering began primarily with zero-shot and few-shot prompting techniques, significantly enhancing early language model performance. For example, GPT-3 demonstrated impressive results using these minimal prompting strategies, setting a foundation for more sophisticated methods[3].

The introduction of Chain-of-Thought (CoT) prompting marked a crucial evolution, explicitly instructing models to outline intermediate reasoning steps, thus significantly improving accuracy on complex tasks such as arithmetic and logic puzzles. Notably, Google's PaLM model increased its accuracy from 17.9% to 58.1% on the GSM8K math benchmark by employing CoTprompting[2].

Advanced frameworks, including ReAct prompting, further expanded the utility of prompt engineering by combining explicit reasoning with interactive tool-use actions. This hybrid approach allowed models to iteratively reason and act, enhancing performance in tasks requiring external information retrieval or calculations [4].

Despite their effectiveness, these sophisticated prompting methods, as shown in Figure 1[5] came with inherent challenges. Small alterations in prompt wording could result in drastically different outputs, introducing issues related to reproducibility and scalability. These limitations raised questions about the practicality of relying heavily on prompt engineering for consistent high-level performance across diverse tasks.

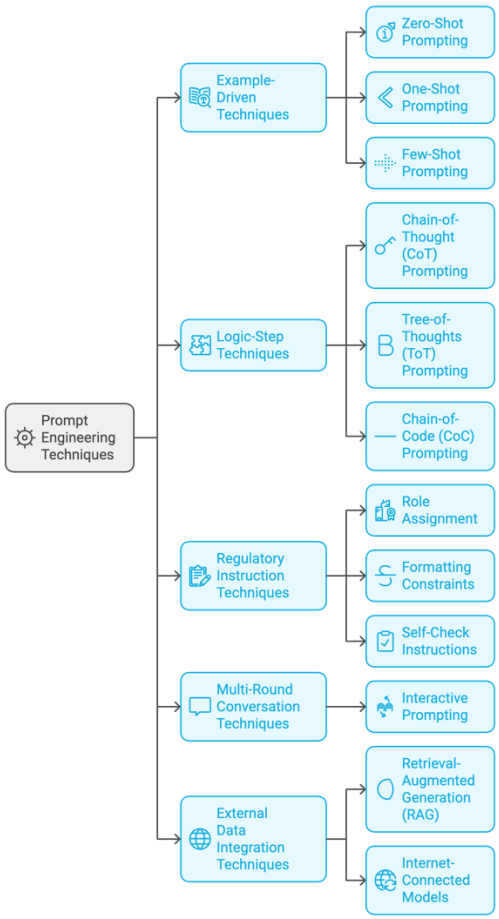

Fig. 1.Basic Prompt Technics

III. Explicit Prompting vs. Intrinsic Reasoning

1) *Modern Reasoning Models and Intrinsic Capabilities*

Recent large language models have evolved significantly in their inherent ability to reason effectively, thus potentially reducing their dependence on prompt engineering techniques. Modern examples include GPT-4.5, OpenAI's GPT-O1 and O1-Pro, and the open-source DeepSeek R1[6]. These models are explicitly trained or fine-tuned to enhance internal reasoning processes, either through supervised fine-tuning on reasoning exemplars or reinforcement learning techniques designed to encourage long-form reasoning outputs without explicit prompting.

Forinstance,DeepSeek R1 demonstrates advanced capabilities for step-by-step reasoning inherently, as it was explicitly optimized via reinforcement learning to produce coherent multi-step rationales internally without explicit prompting DeepSeek R1 Evaluation[7]. Similarly, GPT-O1 and its more advanced version, GPT-O1 Pro, have also incorporated advanced internal reasoning capabilities, achieving state-of-the-art performance on challenging reasoning benchmarks like GSM8K and MMLU[8].

2) *Benchmark-Based Empirical Comparisons*

To empirically assess the impact of explicit prompt engineering against intrinsic reasoning capabilities, several benchmark analyses were conducted. Key benchmarks include:

GSM8K (Grade School Math Problems)

MMLU (Massive Multitask Language Understanding)

Big-Bench Hard (BBH) (Difficult Reasoning Tasks)

GSM8K Benchmark Analysis

Historically, early models like GPT-3 (175B) struggled significantly on GSM8K tasks when relying solely on direct zero-shot prompting, achieving roughly 15%-20% accuracy. However, introducing Chain-of-Thought (CoT) prompting significantly boosted performance. For instance, Google's PaLM-540B improved from 17.9% to 58.1% accuracy solely by applying few-shot CoT examples.

Recent models demonstrate substantial intrinsic reasoning improvements. GPT-4, even without explicit CoT prompting, achieves approximately 90% accuracy on GSM8K, closely approaching human-level performance (OpenAI GPT-4 Technical Report). Similarly, GPT-O1 Pro reportedly scores at or above this level (OpenAI GPT-O1). DeepSeek R1 matches this high-level performance closely, with minimal or no prompting required to trigger stepwise reasoning (DeepSeek R1 Evaluation).

As an example GSM8K benchmarks rank are shown in Figure 2[9].

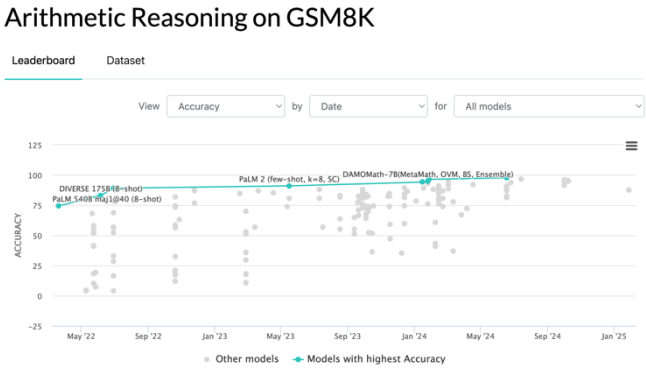

Fig. 2.Benchmark Example for GSM8K

MMLU Benchmark Analysis

On knowledge-intensive benchmarks like MMLU, earlier models saw modest improvements through prompt engineering. GPT-3.5 showed limited incremental accuracy gains through explicit CoT prompts[10]. By contrast, newer models trained explicitly on instruction-following datasets (e.g., GPT-4 series and Claude models) perform strongly with minimal or no prompt-engineering intervention, consistently surpassing 85% accuracy[11].

On harder variants like MMLU-Pro, however, explicit prompting still contributes notable accuracy improvements-up to +19%-demonstrating continued (though reduced) benefits from careful prompting[12]. This implies a situational dependence: intrinsic reasoning suffices for standard complexity tasks, while explicit prompts continue to assist models on extremely challenging variants.

Big-Bench Hard (BBH) Analysis

BBH tasks specifically designed to evaluate reasoning robustness showed that earlier models performed poorly without extensive CoT prompting. CoT prompts were essential to reach human-level or higher performance.

Recent reasoning-enhanced models inherently solve a large portion of BBH tasks even without explicit prompting. GPT-4 and DeepSeek R1 reportedly achieve competitive results directly. However, careful

CoT or ReAct prompting still incrementally enhances accuracy, underscoring that explicit prompting remains relevant for specific complex scenarios[13].

3) Efficiency and Practical Trade-offs

Explicit prompt engineering, particularly involving chain-of-thought or extended reasoning instructions, introduces computational overhead (additional tokens, longer generation times), thus increasing costs. Recent analyses indicate modern models (GPT-4.5, O1 Pro) achieve close-to-peak reasoning performance under simpler prompting, offering higher computational efficiency.

A recent study showed lengthy explicit reasoning chains might sometimes introduce errors or inefficiency ("overthinking") on simpler tasks.[14] This suggests a clear trade-off: intrinsic reasoning typically provides optimal balance for routine tasks, whereas explicit prompting yields maximal accuracy gains only on niche, complex tasks, as shown in Figure 3.

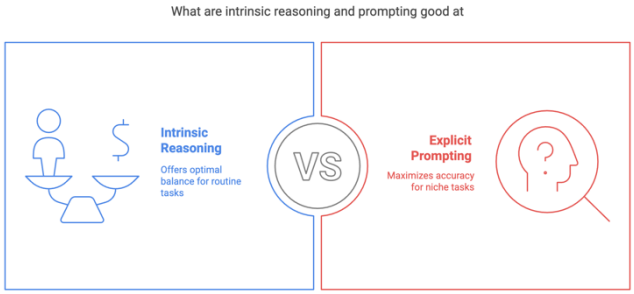

Fig. 3.Intrinsic reasoning and explicit prompting strong area

4) Summary of Findings and Implications

Overall, empirical comparisons reveal a critical insight: while explicit prompting still has value in highly complex reasoning scenarios, its marginal utility declines significantly as models' intrinsic reasoning abilities evolve. Thus, the trend clearly indicates a diminishing reliance on extensive prompt engineering strategies.

Modern developers and researchers should prioritize intrinsic model improvements, reserving explicit prompt engineering mainly for exceptionally challenging contexts requiring maximal accuracy and interoperability.

IV. Discussion and Future Directions

The empirical evidence presented highlights an evident transformation in the role of prompt engineering as large language models continue to evolve. Early reliance on carefully crafted explicit prompts has shifted considerably toward intrinsic reasoning capabilities. This shift has meaningful implications for future research, development, and practical application of large language models.

1) Interpretation of Findings

Our comparative analysis reveals a clear trend: explicit prompting remains beneficial primarily in highly complex and niche reasoning scenarios, whereas intrinsic reasoning capabilities suffice or even excel in most general and moderately challenging tasks. Particularly notable are the results from the GSM8K and MMLU benchmarks, where newer models achieve near-human-level performance without significant prompt engineering, showcasing substantial intrinsic reasoning capabilities.

However, explicit prompting continues to play a crucial role in specialized domains such as internal business cases, and highly interactive tasks involving external tools or iterative reasoning. In these scenarios, structured prompt frameworks (like ReAct) or detailed CoT instructions still meaningfully enhance model performance. Since it provides context that intrinsic reasoning can not.

2) Emerging Trends and Innovations

The ongoing development of language models emphasizes embedding reasoning skills directly into models during the training phase, making explicit prompts less necessary. The spontaneous emergence of reasoning abilities in the Deepseek R1 model has led to fine-tuning efficiency far exceeding previous expectations, significantly reducing the overall training costs. Consequently, future model architectures and training approaches will likely focus even more heavily on intrinsic reasoning optimization.

Moreover, there's growing interest in automated prompt optimization methods—such as prompt tuning, soft prompts, and prompt adaptation techniques—that help models internally adjust reasoning processes dynamically [15]. Such methods could reduce or eliminate the manual burden traditionally associated with crafting optimal prompts, making models both more flexible and easier to deploy across diverse tasks and contexts.

3) Practical Implications

For researchers, the diminishing returns from explicit prompt engineering suggest future studies should shift focus toward improving intrinsic reasoning mechanisms, understanding their limitations, and developing benchmarks that better capture models' inherent reasoning abilities. Benchmarks like MMLU-Pro and ARC-AGI provide promising directions for such research.

For developers, there will likely be a reduced necessity for extensive trial-and-error prompt optimization. The priority may shift to designing intuitive interfaces, integrated prompting systems, or adaptive prompting algorithms that leverage intrinsic reasoning capabilities while providing explicit prompts only when necessary.

End-users will experience easier interactions with LLMs, expecting effective reasoning without elaborate prompt structures. Ultimately, advanced reasoning models can offer more natural, intuitive interactions and reduce barriers to adoption, democratizing access to powerful AI-driven reasoning capabilities.

4) Recommended Future Research Directions

Several promising avenues warrant further exploration:

Advanced Intrinsic Reasoning Architectures: Explore new model architectures that inherently facilitate multi-step reasoning, potentially through memory augmentation, internal scratchpads, or recursive reasoning loops.

Automated Prompt Engineering: Investigate automated methods (e.g., learned prompts or dynamic prompt generators) that optimize reasoning capabilities internally, reducing manual interventions.

Interactive Reasoning Agents: Develop sophisticated conversational or interactive systems capable of self-clarification, iterative reasoning, and external tool usage without extensive manual prompting.

Evaluation of Intrinsic Reasoning Robustness: Refine existing or develop new benchmarks that specifically assess models' intrinsic reasoning capabilities, separating them clearly from prompt-dependent performance gains.

Figure 4 demos the key approach of potential future directions

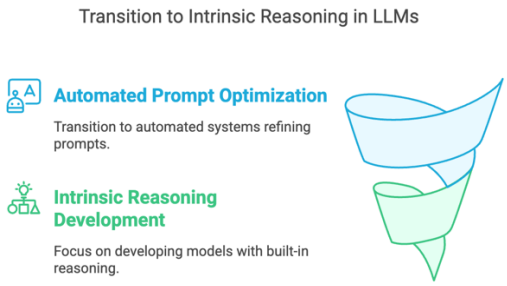

Fig. 4.key approach of potential future directions

V. CONCLUSIONS

This paper analyzed the evolving role of prompt engineering in large language models, particularly regarding their reasoning capabilities. Initially critical for guiding early models, explicit prompting techniques such as zero-shot, few-shot, and Chain-of-Thought (CoT) prompting significantly improved model performance. However, as newer, more sophisticated models have emerged—such as GPT-4.5, GPT-O1/O1 Pro, DeepSeek R1, and Anthropic's Claude—intrinsic reasoning capabilities have greatly diminished the need for explicit prompt engineering in most practical scenarios.

Our comparative analysis demonstrated that while advanced reasoning models inherently solve complex tasks efficiently, explicit prompting techniques continue to deliver incremental benefits primarily in extremely challenging or specialized contexts, such as advanced mathematics or tasks requiring external tools. Benchmarks including GSM8K, MMLU, and Big-Bench Hard (BBH) empirically supported this finding, showing high intrinsic reasoning performance with explicit prompting adding marginal but notable improvements in specific challenging cases.

Looking ahead, intrinsic reasoning capabilities are expected to further improve, driven by advancements in training techniques like Reinforcement Learning with Human Feedback (RLHF), instruction tuning, and automated prompt optimization strategies. Researchers, developers, and end-users should prepare for a shift toward models that inherently understand and reason effectively with minimal manual intervention.

Future research should emphasize developing advanced intrinsic reasoning architectures, creating automated and adaptive prompting methods, and refining evaluation benchmarks to better capture intrinsic reasoning capabilities. These efforts promise a future where explicit prompt engineering becomes a supplementary rather than foundational aspect of leveraging large language models.

REFERENCES

[1] P. Sahoo, "A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications." Feb. 2024. Accessed: Mar. 25, 2025. [Online]. Available: https://arxiv.org/html/2402.07927v1

[2] J. Lee et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," Jan. 01, 2022, Cornell University. doi: 10.48550/arxiv.2201.11903.

[3] T. B. Brown et al., "Language Models are Few-Shot Learners," Jan. 01, 2020, Cornell University. doi: 10.48550/arxiv.2005.14165.

[4] S. Yao et al., "ReAct: Synergizing Reasoning and Acting in Language Models," Jan. 01, 2022, Cornell University. doi: 10.48550/arxiv.2210.03629.

[5] N. Wu, "Expanding Horizons in Prompt Engineering: Techniques, Frameworks, and Challenges," Jan. 13, 2025. doi: 10.55041/ijsrem40635.

[6] B. Loki, "DeepSeek's Journey in Enhancing Reasoning Capabilities of Large Language Models." Feb. 2025. Accessed: Mar. 25, 2025. [Online]. Available: https://medium.com/@bernardloki/deepseeks-journey-in-enhancing-reasoning-capabilities-of-large-language-models-ff7217d957b3

[7] DeepSeek-AI et al., "DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning," Jan. 22, 2025, Cornell University. doi: 10.48550/arxiv.2501.12948.

[8] D. Cleary, "DeepSeek R-1 Model Overview and How it Ranks Against OpenAI's o1." Jan. 2025. Accessed: Mar. 25, 2025. [Online]. Available: https://www.prompthub.us/blog/deepseek-r-1-model-overview-and-how-it-ranks-against-openais-o1

[9] "Gsm8k Benchmark." Mar. 01, 2025. [Online]. Available: https://paperswithcode.com/sota/arithmetic-reasoning-on-gsm8k

[10] M. Süzgün et al., "Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them," Jan. 01, 2022, Cornell University. doi: 10.48550/arxiv.2210.09261.

[11] "OpenAI GPT4 Report." Mar. 14, 2023. [Online]. Available: https://openai.com/index/gpt-4-research/

[12] S. Bubeck et al., "Sparks of Artificial General Intelligence: Early experiments with GPT-4," Jan. 01, 2023, Cornell University. doi: 10.48550/arxiv.2303.12712.

[13] "Google Big-Bench." Mar. 01, 2023. [Online]. Available: https://github.com/google/BIG-bench

[14] B. Wang, X. Deng, and H. Sun, "Iteratively Prompt Pre-trained Language Models for Chain of Thought," Jan. 01, 2022. doi: 10.18653/v1/2022.emnlp-main.174.

[15] B. Lester, R. Al-Rfou, and N. Constant, "The Power of Scale for Parameter-Efficient Prompt Tuning," Jan. 01, 2021, Cornell University. doi: 10.48550/arxiv.2104.08691.