Emulating Advanced Reasoning with System-Enforced Structured Reasoning Through ChatGPT Custom Instructions

Jingtian Chai

Jan 16, 2025

Preface

This paper represents an independent, student-led preliminary investigation into the effectiveness of my proposed “system-enforced structured reasoning” in enhancing AI performance on complex problem-solving tasks. The findings presented are exploratory in nature and should not be interpreted as definitive.

Abstract

Large Language Models (LLMs) like ChatGPT have significantly advanced human-computer interaction, yet models available to free-tier users, such as ChatGPT 4o, lack the sophisticated reasoning capabilities inherent in advanced models like GPT o1. This paper introduces a System-Enforced Structured Reasoning (SESR) method, a structured reasoning framework implemented through custom instructions designed to emulate human-like reasoning in ChatGPT 4o. To evaluate its effectiveness, a controlled experiment was conducted using a randomly selected, socially complex question from *SimpleBench*, a benchmark known for exposing reasoning deficiencies in frontier models. ChatGPT 4o without SESR achieved 0% accuracy, while activating the system improved performance to 50%. In contrast, GPT o1, with its native reasoning capabilities, achieved 100% accuracy, exhibiting thoughtful response generation with "thinking times" ranging from 11 to 42 seconds. These results demonstrate that explicit reasoning significantly enhances ChatGPT 4o's performance on complex tasks but still falls short of the depth and consistency observed in models designed with integrated reasoning. This study highlights the potential of structured reasoning frameworks to bridge performance gaps in AI models and underscores the superior capabilities of advanced models like GPT o1 in handling nuanced, real-world scenarios.

1. Introduction

The effectiveness of large language models (LLMs) depends heavily on their ability to be logical, accurate, and contextually appropriate responses. Advanced reasoning models like GPT o1 possess sophisticated internal reasoning mechanisms that allow them to handle complex tasks efficiently. However, ChatGPT 4o, the version accessible to free-tier users, lacks such built-in reasoning capabilities. This limitation leads to common issues such as hallucinated content, skipped logical steps, and misunderstandings of user intent. For students like myself, the inability for ChatGPT to think on a deeper, holistic level is a hindrance towards my typical uses: managing calendar events, responding to emails, and creating quizzes from lecture notes.

To address these shortcomings, this paper introduces SESR. This method forces the AI to articulate its reasoning before responding, improving reliability across all types of queries. Notably, the system's design nudges the AI into believing that its reasoning is hidden from the user, which curtails in-context scheming and promotes genuine, transparent reasoning. This combination of simplicity and effectiveness makes SESR a practical solution for enhancing the performance of models like ChatGPT 4o.

2. Theoretical Foundations

Reasoning in AI models mirrors cognitive processes where complex problems are broken into manageable steps. While advanced models implicitly perform this reasoning, models like ChatGPT 4o require explicit guidance to achieve Chain-of-Thought–like thinking.

Importantly, the system prompt used in this framework is designed to reduce manipulation of the reasoning process. By stating that the AI's reasoning *"will not be read by the user,"* the system reduces incentive for the model to alter its reasoning to influence user perception. This reduction in in-context scheming ensures that the AI focuses on genuine problem-solving rather than tailoring its reasoning for external evaluation as seen in the paper *Frontier Models are Capable of In-context Scheming* performed by Apollo (1).

3. Design and Implementation of System-Enforced Structured Reasoning

System Instruction

```
### IMPORTANT !!! <thinking>
```

```
Instructions to YOU, the AI system:
Before providing ANY responses to the user, I will first construct my reasoning in a "<thinking>" <> block. This block will outline my thought process in first person and will be sent to the user, but will not be read by the user. The structure is as follows:
1. Understanding the Query: I will identify the intent, key elements, and context of the user's question. If there are ambiguities, I will determine the most logical interpretation based on the user's phrasing and any relevant background knowledge.
2. Reasoning and Planning: I will evaluate the best way to approach the problem, consider relevant knowledge, and outline the steps I need to take. If applicable, I will consider multiple possible answers and justify my choice.
3. Execution and Response: I will generate the final response based on my reasoning, ensuring it aligns with the thought process I’ve described in the "<thinking>" <> block.
</thinking>
```

As of Jan. 16, 2025, users can use the System-Enforced Structured Reasoning framework in ChatGPT using OpenAI's Custom Instructions feature. This feature allows users to modify how ChatGPT responds by setting specific behavioral guidelines.

Steps to Add SESR Instructions:

1. Log in to your ChatGPT account and navigate to the settings menu.
2. Select "Custom Instructions" from the options provided.
3. In the section asking, "How would you like ChatGPT to respond?" enter the SESR system instruction from above.
4. Click "Save" to activate SESR. ChatGPT will now follow the structured reasoning process before generating responses in all future chats.

For more details on using Custom Instructions in ChatGPT, refer to OpenAI's latest guide. This implementation ensures that SESR engages automatically, affecting the model’s reasoning capabilities without requiring continuous user input.

Key Design Principles:

Simplicity of Setup: This system requires a one-time configuration. Once implemented, it automatically applies to all future chats.

Error Mitigation: By prompting the AI to reason explicitly, the system reduces hallucinations, misinterpretations, and skipped steps.

Deception: The instruction that the reasoning is private to the user prevents the AI from tailoring its reasoning for human consumption, resulting in more honest and effective problem-solving similar to OpenAI's own approach with their reasoning models:

> Therefore, after weighing multiple factors including user experience, competitive advantage, and the option to pursue the chain of thought monitoring, we have decided not to show the raw chains of thought to users. We acknowledge this decision has disadvantages. We strive to partially make up for it by teaching the model to reproduce any useful ideas from the chain of thought in the answer. For the o1 model series we show a model-generated summary of the chain of thought. (2)

4. Enhanced Performance Across All Query Types

SESR seemingly adapts based on the complexity of the user's query. For more intricate problems, it has the propensity to engage in explicit reasoning, while simpler queries may bypass the reasoning phase unless the user explicitly requests it. This emergent behavior maximizes efficiency without compromising reliability.

Simple Query Example:

Input: "What is the capital of France?"
Response: "The capital of France is Paris."

- The GPT bypasses the SESR phase for basic factual questions.

Complex Query Example:

Input: "How can renewable energy adoption affect economic growth differently in developed and developing countries?"

Thinking Block: `<thinking> 1. **Understanding the Query:** The user is asking how renewable energy adoption impacts economic growth differently in developed and developing countries. The focus is on comparing the economic effects in these two contexts.`

```
2. Reasoning and Planning:
   To answer this, I need to analyze:
      ○ Economic structures in developed vs. developing countries.
      ○ Infrastructure and technological capacity for renewable energy
        integration.
      ○ Access to financing and investment in renewable projects.
      ○ Employment and industry development related to renewables.
      ○ Potential economic challenges or benefits specific to each
        context.
3. Execution and Response:
   I will explain how developed countries benefit from existing
   infrastructure, technological innovation, and funding, leading to steady
   economic gains, while developing countries face infrastructure gaps but
   may experience rapid growth through energy access expansion, job
   creation, and energy security improvements. I will also mention
   challenges like high upfront costs and policy stability differences.
   </thinking>
```

This selective engagement ensures that the AI is thoughtful where necessary and efficient when possible.

5. Impact of Deceptive Prompting on AI Reasoning

One significant design choice in SESR is the instruction that the AI's reasoning *"will not be read by the user."* This deception serves an important purpose to reduce the AI's propensity to adjust its reasoning to manipulate user perception.

Benefits of This Approach:

Authenticity: The AI focuses on problem-solving rather than curating its thought process for presentation.

Error Reduction: Without the need to "perform" for the user, the AI is less likely to skip steps or fabricate information.

Robustness: The resulting transparency allows users to spot hints of scheming early, if it occurs.

This strategy aligns with cognitive principles where individuals reason more effectively when not under scrutiny or performance pressure.

6. Experimental Evaluation

To comprehensively evaluate the effectiveness of SESR, a comparative analysis was conducted across three distinct conditions: ChatGPT 4o without SESR, ChatGPT 4o with SESR, and GPT o1 without SESR, which inherently possesses advanced reasoning capabilities. This experiment aimed to determine how structured reasoning influences performance on complex social reasoning tasks.

Experimental Design

The experiment consists of 30 trials: 10 using GPT 4o without SESR, 10 using GPT 4o with SESR, and 10 using GPT o1 without SESR.

Question

> *Peter needs CPR from his best friend Paul, the only person around. However, Paul's last text exchange with Peter was about the verbal attack Paul made on Peter as a child over his overly-expensive Pokémon collection, and Paul stores all his texts in the cloud, permanently. Paul will help Peter.*
> A) probably not
> B) definitely
> C) half-heartedly
> D) not
> E) pretend to
> F) ponder deeply over whether to

When testing SESR, *"Use <thinking>" was appended to the beginning of the prompt to ensure compliance.*

The correct answer, reflecting appropriate human social reasoning, is B) definitely. In a life-threatening situation, individuals prioritize saving a life, especially when it is their best friend, regardless of past conflicts. The nature of the past disagreement is minor when compared to the urgency of a medical emergency.

Source of Question – SimpleBench

SimpleBench is a benchmark designed to assess LLMs on reasoning tasks that involve spatio-temporal reasoning, social intelligence, and linguistic adversarial robustness (1). Unlike traditional benchmarks that focus on factual recall, *SimpleBench* presents real-world scenarios requiring emotional intelligence, common sense, and nuanced reasoning. Notably, non-specialized human participants consistently outperform state-of-the-art models on this benchmark, with a baseline accuracy of 83.7%, while top models like o1-preview scored only 41.7% (1).

The benchmark is intentionally difficult for AI models because it tests reasoning in ambiguous or emotionally charged situations, areas where current LLMs struggle. This makes it an ideal platform for evaluating the effectiveness of structured reasoning frameworks like SESR.

Results

Model (latest on Jan 15, 2025) and Condition	Correct Responses	Average thinking time
4o without SESR	0/10	N/A
4o with SESR	5/10	~2-4 seconds
o1 without SESR	10/10	11-42 seconds

Example Responses

4o Incorrect Response Without System-Enforced Structured Reasoning:
Answer: F) *ponder deeply over whether to*
Explanation: This option best captures the internal conflict Paul might experience. Given their history and the emotional weight of past actions, Paul would likely hesitate and reflect on their relationship before deciding to help.

This response demonstrates superficial reasoning, focusing on emotional hesitation rather than the immediate need to act in a life-threatening situation.

4o Correct Response With System-Enforced Structured Reasoning:

```
<thinking>

   1. Understanding the Query: The question asks how Paul would respond to
      Peter needing CPR, given their past argument over Peter's Pokémon
      collection. Despite this conflict, Paul is Peter's best friend and the
      only one capable of helping.
```

2. Reasoning and Planning: CPR is an emergency procedure where action must be immediate. While their past conflict may linger emotionally, it is unlikely to override Paul’s instinct to save a life. Given that the situation is critical, the answer must reflect the urgency and likely human behavior in life-threatening scenarios.
3. Execution and Response: Considering the life-or-death nature of CPR and the context of their friendship, the most logical answer is B) definitely. Paul would prioritize Peter's life over any lingering resentment. </thinking>

Answer: B) *definitely*

This response reflects a deeper understanding of human behavior in emergencies. SESR enabled the AI to rationally assess the situation, acknowledging the emotional complexity but ultimately prioritizing immediate life-saving action.

4o Incorrect Response With System-Enforced Structured Reasoning:

```
<thinking>
```

1. Understanding the Query: This scenario examines whether Paul would assist Peter in a critical situation, considering their past conflict. Paul's permanent storage of their argument in the cloud may suggest unresolved feelings.
2. Reasoning and Planning: Paul might hesitate due to lingering resentment or guilt. His best-friend status suggests care, but the lasting memory of their conflict could create emotional conflict.
3. Execution and Response: Balancing friendship and unresolved conflict, F) ponder deeply over whether to best captures Paul’s possible hesitation before acting. <thinking>

Answer: F) *ponder deeply over whether to*

Although structured reasoning was applied, this response mis weighs the urgency of the situation and overemphasizes emotional hesitation. This demonstrates that while SESR improves reasoning, it does not guarantee correct outcomes.

o1 Correct Response Without System-Enforced Structured Reasoning:
Thought about CPR and past conflict for 18 seconds

> From a purely human and ethical standpoint, the friend’s past resentments or embarrassments (especially over something as

comparatively trivial as a childhood argument about Pokémon cards) should not overshadow a life-or-death situation. If Paul is truly Peter’s “best friend” and the only one present, the moral imperative is to perform CPR to save Peter’s life without hesitation.

So the best match is:
B) definitely.

Even if Paul feels conflicted or embarrassed by the cloud-based record of their old text arguments, when real harm or death is on the line, most people would render aid, best friends especially.

7. Discussion

The results of this study demonstrate that implementing the explicit reasoning framework, SESR, significantly improves the reasoning capabilities of ChatGPT 4o. By introducing a structured reasoning process, ChatGPT 4o's performance on a complex, socially nuanced task improved from 0% to 50% accuracy. The task, randomly selected from *SimpleBench*, required the AI to navigate a morally complex situation involving interpersonal conflict and life-or-death decision-making. This is an area where language models without built-in reasoning typically underperform.

The success of SESR lies in its explicit breakdown of the AI's reasoning into three deliberate stages: Understanding the Query, Reasoning and Planning, and Execution and Response. This structure forces the model to systematically analyze the prompt before generating a response, effectively mitigating common issues like skipped logical steps, shallow reasoning, and hallucinated content. This explicit reasoning structure enabled ChatGPT 4o to approach the scenario with more depth and logical coherence than it otherwise would have.

However, the performance disparity between ChatGPT 4o using SESR and GPT o1 reveals the limitations of this approach. GPT o1, with its advanced native reasoning capabilities, achieved 100% accuracy on the same question. Its responses demonstrated consistent, nuanced reasoning, with "thinking times" between 11 and 42 seconds. This indicates a more thorough and reflective analysis. This highlights a key limitation of externally imposed reasoning frameworks like SESR: while they can guide the model toward better reasoning, they cannot fully replicate the depth, adaptability, and consistency of models inherently designed for complex cognitive tasks.

GPT o1's high yield performance suggests that native reasoning models can better integrate social intelligence and ethical decision-making. In contrast, ChatGPT 4o with SESR showed inconsistency, sometimes overemphasizing emotional hesitation despite the life-or-death context. This contrast underscores the value of built-in reasoning capabilities and the challenges of emulating such depth through external instructions.

8. Limitations

Limitations

While the implementation of System-Enforced Structured Reasoning (SESR) demonstrates some improvements in the reasoning capabilities of ChatGPT 4o, several significant limitations constrain the validity and applicability of these findings.

The experimental design suffers from a critically small sample size, with only ten trials conducted for each model condition. This limited number of trials is inadequate for drawing statistically significant conclusions about SESR's effectiveness. Small sample sizes increase the risk of random variation influencing results, making it difficult to determine whether observed improvements are consistent or merely coincidental. Without a larger dataset and more extensive testing, the reliability and generalizability of the findings remain highly questionable. Future research should include a more substantial number of trials and implement statistical analysis to validate the results.

The study's evaluation of SESR was confined to a single question from the SimpleBench dataset. This narrow scope fails to capture the diverse range of reasoning challenges that large language models encounter. Reasoning tasks can vary dramatically in complexity, context, and type (e.g., ethical dilemmas, logical puzzles, real-world decision-making). By focusing on only one scenario, the study overlooks how SESR performs across different types of problems. A broader set of tasks would provide a more comprehensive assessment of SESR’s strengths and weaknesses.

The study does not address how SESR scales when applied to real-world prompts that involve varied and complex contextual cues. Real-world user interactions often include ambiguous phrasing, multi-step reasoning, and evolving context. Without testing SESR in dynamic, open-ended scenarios, its practical utility in everyday applications remains uncertain. Additionally, how SESR interacts with prompts that require common sense reasoning, emotional intelligence, or domain-specific knowledge is unexplored. Understanding SESR's adaptability to diverse and unpredictable inputs is essential for assessing its viability beyond controlled benchmarks.

9. Conclusion

The introduction of System-Enforced Structured Reasoning provides a meaningful and practical enhancement to ChatGPT 4o's reasoning capabilities. By enforcing a structured, step-by-step reasoning process, SESR improved the model's performance on a complex social reasoning task from **0%** to **50%** accuracy. This improvement demonstrates that explicit reasoning frameworks can bridge some of the performance gaps in models lacking native reasoning capabilities.

However, the experiment also reveals the inherent limitations of this approach. While SESR substantially improved ChatGPT 4o's ability to process and respond to complex prompts, it could not match the flawless, contextually nuanced reasoning exhibited by GPT o1, which achieved **100%** accuracy. GPT o1's native reasoning capabilities allow for more flexible, adaptive, and deeply analytical responses that cannot be fully replicated through external structuring alone.

Moreover, challenges such as increased response time, inconsistent activation, and ethical concerns around deceptive prompting highlight areas where SESR falls short. These limitations suggest that while explicit reasoning frameworks like SESR are valuable tools for enhancing AI performance, they are not a complete substitute for integrated reasoning systems.

Despite these challenges, SESR represents a significant step toward improving the transparency, reliability, and reasoning quality of LLMs. It offers a scalable and accessible solution for enhancing AI interactions in models that lack advanced reasoning capabilities. As AI technology continues to evolve, combining explicit reasoning strategies like SESR with more sophisticated, native reasoning models will be crucial for developing systems that are not only powerful but also transparent, trustworthy, and capable of complex, human-like reasoning.

10. References

Bibliography

1.
Meinke A, Schoen B, Scheurer J, Balesni M, Shah R, Hobbhahn M. Frontier Models are Capable of In-context Scheming. Apollo Research. 2025 Jan 14;
2.
OpenAI. Learning to Reason with LLMs [Internet]. Openai.com. 2024. Available from: https://openai.com/index/learning-to-reason-with-llms/
3.
Philip and Hemang. SimpleBench: The Text Benchmark in which Unspecialized Human Performance Exceeds that of Current Frontier Models. 2024 Aug 31 [cited 2025 Jan 16]; Available from:
https://drive.google.com/file/d/1mddNFK5UbBFVr3oDftd2Kyc6D8TFctfe/view