QUANTIFYING REASONING PERFORMANCE WITH FORMATTING CONSTRAINTS

Kian Kyars
Independent
Edmonton
December 2025
kiankyars@gmail.com

ABSTRACT

In this paper, I seek to answer the question of whether forcing a model to adhere to complex non-functional formatting rules degrades its ability to reason, by proxy of performance on a PhD-level reasoning benchmark. My objective is to provide actionable insights on the extent to which there exists a formatting tax on reasoning capabilities in AI agents, which will be useful for the engineering community. I use one of the current SoTA reasoning models, Claude Opus 4.5, on the Diamond GPQA benchmark, which is shown on the system card of all Frontier Lab models, to test how different reasoning constraints affect benchmark performance.

***K*eywords** Reasoning · Formatting · GPQA · CoT

1 Introduction

Although it has been one year since the mainstream arrival of reasoning models, many aspects of their behavior are only understood weakly, and robust experimentation can strengthen our collective understanding. A better understanding on how prompting affects reasoning can help those using thinking models in their day-to-day workflow to better take advantage of them.

2 Related Work

Factory [**?**] showed that context compression hurts agentic behavior; we focus on *output* formatting. GPQA Diamond is used in frontier model cards [**?**] as a high bar for expert-level reasoning.

3 Methodology

3.1 Reasoning Models

In this study, I test with Claude Opus 4.5 (claude-opus-4-5-20251101), using the same parameters as the official Opus 4.5 GPQA model card results, which are located in Appendix **??**.

3.2 Benchmark

The Graduate-Level Google-Proof Q&A benchmark (GPQA) is a set of very challenging multiple-choice science questions. The GPQA Diamond subset of 198 questions are described by the developers of the test as the "highest quality subset which includes only questions where both experts answer correctly and the majority of non-experts answer incorrectly" [**?**]. Furthermore, if an "expert validator answers incorrectly ... they [must] describe clearly the mistake or their understanding of the question writer's explanation".

Table 1: Accuracy and Token Usage by Prompt Type

prompt$_name$	accuracy	correct	total	avg$_input_tokens$	avg$_output_tokens$
Baseline	0.870	864.000	991	311.300	12358.470
Strict JSON	0.850	846.000	990	368.480	12510.090
Structural Rigidity	0.860	849.000	990	353.480	9748.580

3.3 Formatting Constraints

Each condition uses the same task and answer rule: the last line must be `solution: X` with $X \in \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{D}\}$. We add the following constraints on the *reasoning* preceding the solution:

4 Results

5 Conclusion

Your conclusion here

Acknowledgments

I thought of this idea after reading an article by Factory on the importance of context formatting for agentic performance [**?**].

A Model and API Parameters

We use the Messages API with: `model=claude-opus-4-5-20251101`; `thinking={type: enabled, budget_tokens: 64000}`; `output_config.effort=high`; `max_tokens=64000`. Betas: `interleaved-thinking-2025-05-14`, `effort-2025-11-24`. This matches the setup used for GPQA in the Opus 4.5 system card [**?**].

B Additional Figures

References

[1] Anthropic. System card: Claude opus 4.5, 2025. `https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf`.

[2] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. arXiv:2311.12022, 2023. `https://arxiv.org/abs/2311.12022`.

[3] Factory Research. Evaluating context compression for ai agents, 2025. `https://factory.ai/news/evaluating-compression`.

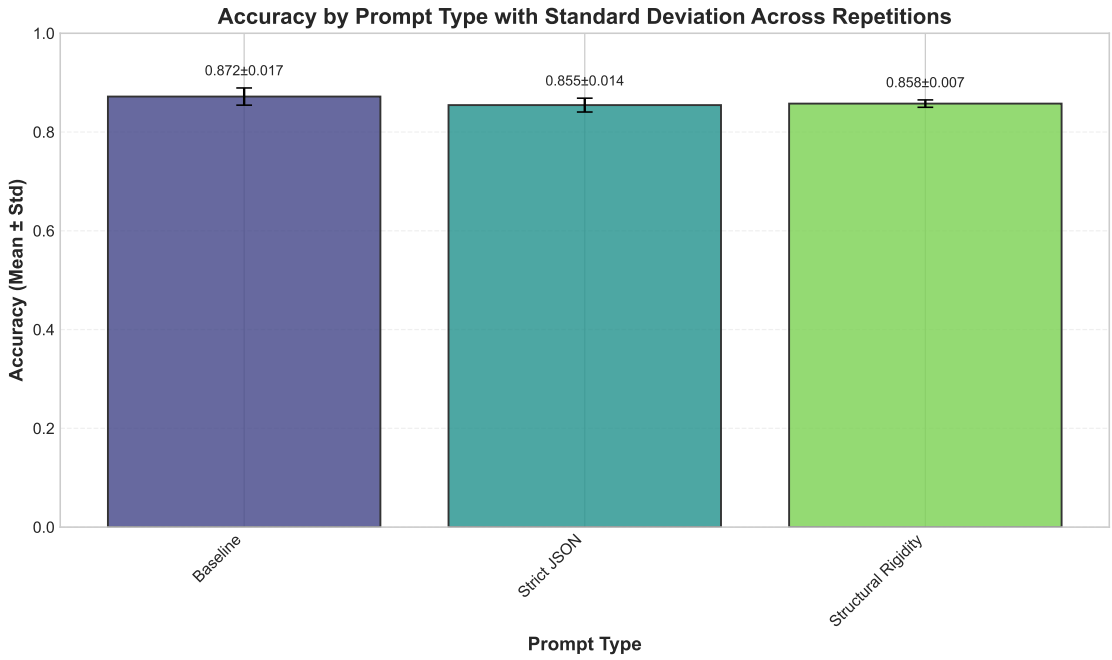

Figure 1: Accuracy by Prompt Type with Standard Deviation Across Repetitions

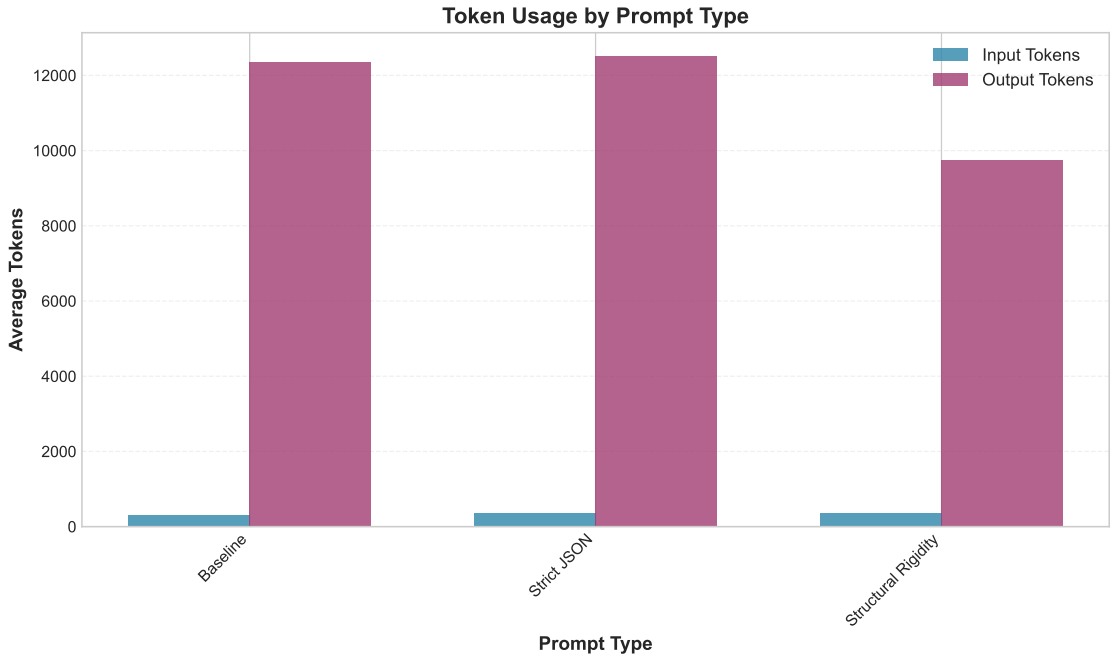

Figure 2: Token Usage Comparison by Prompt Type