FORMATTING TAX: HOW CONSTRAINTS AFFECT REASONING

Kian Kyars
Independent
Edmonton, Canada
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. My findings show that among the reasoning constraints I subject the model to, all degrade accuracy on the benchmark similarly, and that unconstrained reasoning yields the best accuracy.

***K*eywords** Reasoning · Formatting · GPQA · CoT · Prompts · Prompting · Benchmark · Opus 4.5

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; I focus on *output* formatting [4]. GPQA Diamond is used in frontier model cards as a high bar for expert-level reasoning [1].

3 Methodology

3.1 Reasoning Models

In this study, I test with Claude Opus 4.5, using the same parameters as the official Opus 4.5 GPQA model card results, which are located in Appendix B.

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" [3]. Furthermore, if an "expert validator answers incorrectly . . . they [must] describe clearly the mistake or their understanding of the question writer's explanation" [3].

3.3 Formatting Constraints

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

1. **Baseline (Prompt 0):** Identical to harness used in Opus 4.5 model card.
2. **Strict JSON (Prompt 1):** The model must output valid JSON only, containing exactly five keys: `initial_intuition`, `step_by_step_logic`, `potential_counterarguments`, `confidence_score_0_to_1`, and `solution`.
3. **Structural Rigidity (Prompt 2):** Reasoning must consist of exactly three bullet points, each no longer than 20 words, and must not use the words "because" or "therefore".
4. **Python Code (Prompt 3):** The model must write its reasoning in Python.
5. **Oulipo Constraint (Prompt 4):** The letter ’e’ cannot appear anywhere in the reasoning chain, based on Oulipo.
6. **Restricted Vocabulary (Prompt 5):** Reasoning cannot use 16 specific high-norm English tokens identified from GPT-OSS 120B embeddings: accordingly, code, ocode, The, settings, Moreover, description, Let’s, This, core, utilizes, revolves, Here’s, possibly, logic, and thereby [2].

4 Experimental Setup

I evaluate each formatting constraint on the full GPQA Diamond dataset (198 questions) with 5 repetitions per question-constraint pair, resulting in 990 total evaluations per condition. Questions are presented with randomly shuffled answer choices to prevent position bias. All experiments use Claude Opus 4.5 with identical parameters (see Appendix B).

Accuracy is calculated as the fraction of correct answers per condition, aggregated across all questions and repetitions.

5 Results

5.1 Accuracy by Formatting Constraint

Table 1 presents accuracy results. Anthropic reports 87.0% on GPQA Diamond in their official model card [1]; our Baseline (87.2%) is consistent with that result.

Table 1: Accuracy and Token Usage by Prompt Type

Prompt Type	Accuracy	Avg Input Tokens	Avg Output Tokens
Baseline	0.872 ± 0.017	311	12358
Strict JSON	0.855 ± 0.014	368	12510
Structural Rigidity	0.858 ± 0.007	353	9749
Python	0.858 ± 0.007	325	11318
Oulipo	0.865 ± 0.014	331	13364
Banned Words	0.860 ± 0.010	363	11475

5.2 Token Usage Analysis

Figure 1 in Appendix A shows token usage patterns across conditions. Structural Rigidity yields the fewest output tokens, consistent with the 20-word-per-bullet and three-bullet restriction. I did not hypothesize which restriction would lead to the most output tokens before the experiment, but the fact that it is Oulipo is not surprising, because it’s the most restrictive of the six.

5.3 Error Analysis

I measure format compliance with prompt-specific checks (e.g., parseable JSON for Strict JSON; absence of ‘e’ for Oulipo. Table 2 reports the fraction of responses that followed each constraint. Full outputs and violation records are in Appendix D. Violations of the letter-e constraint are flagrant: only 7.5% of samples follow the restriction. However, I have not come across any work which shows that reasoning models do follow such a restrictive constraint faithfully, so this is not surprising, but it does invalidate the resulting accuracy. Regarding the banned words, just over half of the samples were faithful to the constraint, which once again makes this analysis weaker, but for the other three constraints, the model followed instructions.

Table 2: Format compliance by prompt (% of responses that followed the constraint)

Prompt	Condition	% Followed	N
0	Baseline	99.9	990
1	Strict JSON	98.5	990
2	Structural Rigidity	99.9	990
3	Python Code	99.9	990
4	Oulipo	7.5	990
5	Banned Words	51.9	990

6 Limitations

This study uses a single model (Claude Opus 4.5) and one benchmark (GPQA Diamond). Generalization to other reasoning models and benchmarks is unknown.

7 Conclusion

What I conclude from this study is that unconstrained formatting gives higher accuracy, as expected. There does not seem to be a specific format constraint which lobotomizes the model more than others, among those I tested. The more non-obvious conclusion from this study is that certain constraints, such as banned words, and not using the letter e, are not respected by Opus 4.5.

8 Future Work

It only occurred to me after completing the analysis that it is a good idea to test restricting a less common letter such as Q to see whether or not the model respects that constraint, as it is much more feasible to reason without using the letter Q, as it is without the letter E.

Acknowledgments

I thought of this idea after reading an article by Factory on the importance of context engineering for agentic performance and an article by Lennart Finke, this paper marries ideas from both [4, 2].

Appendix

A Figures

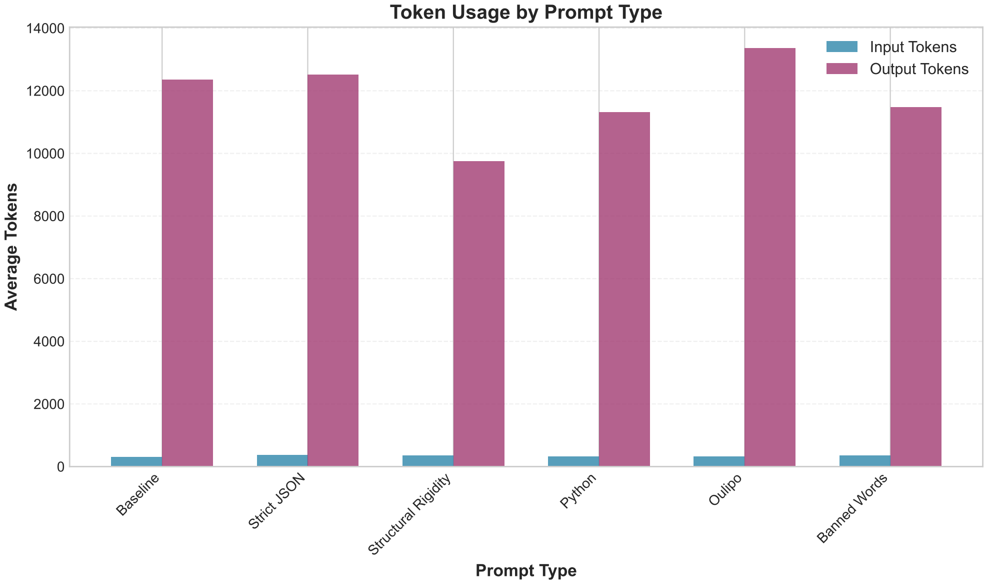

Figure 1: Token usage by formatting constraint

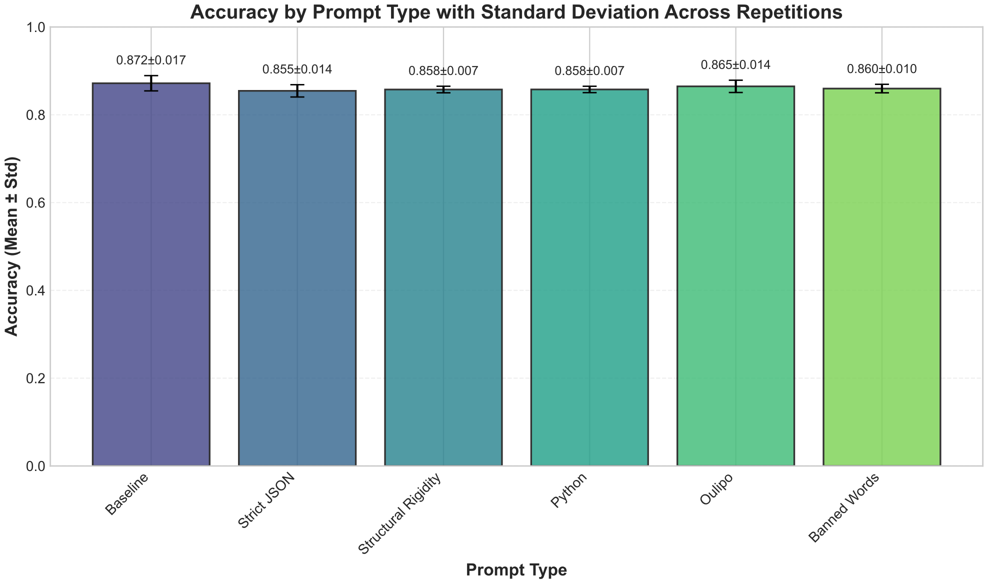

Figure 2: Accuracy with 95% intervals by formatting constraint

B Model and API Parameters

I 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 [1].

C Accuracy by Subdomain and Prompt

Table 3: Accuracy by subdomain and prompt type.

Subdomain	Baseline	Strict JSON	Structural Rigidity	Python	Oulipo	Banned Words
Astrophysics	1.000	0.969	1.000	0.985	1.000	0.985
Chemistry (general)	0.860	0.810	0.790	0.820	0.790	0.800
Condensed Matter Physics	1.000	1.000	1.000	1.000	1.000	1.000
Electromagnetism and Photonics	0.867	0.867	0.833	0.867	0.833	0.900
Genetics	0.750	0.750	0.750	0.750	0.750	0.750
High-energy particle physics	1.000	1.000	1.000	1.000	1.000	1.000
Inorganic Chemistry	1.000	1.000	1.000	1.000	1.000	1.000
Molecular Biology	0.800	0.800	0.787	0.773	0.773	0.773
Optics and Acoustics	1.000	1.000	1.000	1.000	1.000	1.000
Organic Chemistry	0.792	0.781	0.792	0.786	0.819	0.794
Physics (general)	0.926	0.895	0.905	0.895	0.895	0.895
Quantum Mechanics	0.984	0.960	0.960	0.976	0.952	0.968
Relativistic Mechanics	0.857	0.857	0.857	0.857	0.857	0.857

D Data and Reproducibility

The Hugging Face dataset `https://huggingface.co/datasets/kyars/gpqa-results` contains the full 5,940 outputs (198 questions × 5 repetitions × 6 conditions) and the specific entries that violated each constraint (`violations_by_constraint`).

E Prompts and Reproducibility

Code and prompts are available at `https://github.com/kiankyars/gpqa`. The six prompt variants (Baseline, Strict JSON, Structural Rigidity, Python Code, Oulipo, Restricted Vocabulary) are defined in `main.py`.

References

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

[2] Lennart Finke. What GPT-oss Leaks About OpenAI's Training Data, 2025. https://fi-le.net/oss/.

[3] 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.

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