

Prompt Repetition Improves Non-Reasoning LLMs

Yaniv Leviathan*
Google Research
leviathan@google.com

Matan Kalman*
Google Research
matank@google.com

Yossi Matias
Google Research
yossi@google.com

Abstract

When not using reasoning, repeating the input prompt improves performance for popular models (Gemini, GPT, Claude, and Deepseek) without increasing the number of generated tokens or latency.

1 Prompt Repetition

LLMs are often trained as *causal* language models, i.e. past tokens cannot attend to future tokens. Therefore, the order of the tokens in a user’s query can affect prediction performance. For example, a query of the form “<CONTEXT> <QUESTION>” often performs differently from a query of the form “<QUESTION> <CONTEXT>” (see *options-first* vs. *question-first* in Figure 1). We propose to *repeat the prompt*, i.e. transform the input from “<QUERY>” to “<QUERY><QUERY>”. This enables each prompt token to attend to every other prompt token, addressing the above. When not using reasoning, **prompt repetition improves the performance of LLMs (Figure 1) without increasing the lengths of the generated outputs or latency** (Figures 2 and 3).

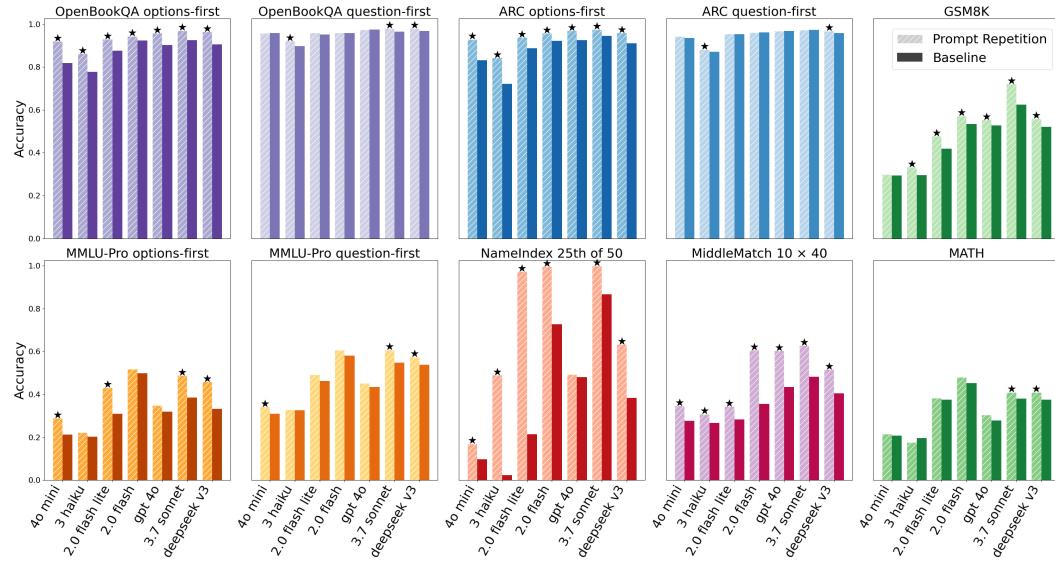


Figure 1: Prompt repetition vs. baseline accuracy for popular LLMs and various benchmarks when asking the models not to reason. A star indicates a statistically significant win ($p_{value} < 0.1$ according to the McNemar test [McNemar \[1947\]](#)). **Prompt repetition wins 47 out of 70 tests, with 0 losses.**

*Equal contribution.

As further motivation, we observe that reasoning models trained with RL often learn to repeat (parts of) the user’s request. Prompt repetition is *efficient*, moving the repetition to the parallelizable prefill stage. The number of generated tokens does not increase¹. Moreover, prompt repetition does not change the format of the generated outputs, keeping them interchangeable with those of the original prompts, enabling simple drop-in deployment in existing systems, and direct use by end-users. When reasoning is enabled, prompt repetition is neutral to slightly positive (Figure 4).

2 Experiments

We test *prompt repetition* on a range of 7 popular models from leading LLM providers of varying sizes: Gemini 2.0 Flash and Gemini 2.0 Flash Lite [Gemini Team Google, 2023], GPT-4o-mini and GPT-4o [OpenAI, 2024], Claude 3 Haiku and Claude 3.7 Sonnet [Anthropic, 2024], and Deepseek V3 [DeepSeek-AI, 2025]. We ran all tests using each provider’s official API in Feb and Mar 2025.

We test each of the models on a set of 7 benchmarks in several configurations: ARC (Challenge) [Clark et al., 2018], OpenBookQA [Mihaylov et al., 2018], GSM8K [Cobbe et al., 2021], MMLU-Pro [Wang et al., 2024], MATH [Hendrycks et al., 2021], and 2 custom benchmarks: NameIndex and MiddleMatch (Appendix A.3). For the multiple choice benchmarks (ARC, OpenBookQA, and MMLU-Pro) we report results both when placing the question first and the answer options later, as well as the other way around, the latter making the model process the options without seeing the question in context (unless using prompt repetition).

Accuracy. Without reasoning, prompt repetition improves the accuracy of all tested LLMs and benchmarks (Figure 1). We consider cases where one method is significantly better than the other according to the McNemar test [McNemar, 1947] with $p_{value} < 0.1$ as wins. With this criteria, **prompt repetition wins 47 out of 70 benchmark-model combinations, with 0 losses**. Notably, performance is improved for all tested models. As expected, we observe smaller improvements for the multiple-choice benchmarks with question-first, and larger improvements with options-first. On the custom tasks of NameIndex and MiddleMatch we observe strong gains with prompt repetition for all models (for example, prompt repetition improves the accuracy of Gemini 2.0 Flash-Lite on NameIndex from 21.33% to 97.33%). We also test a smaller set of benchmarks when encouraging thinking step-by-step (Figure 4), where the results are neutral to slightly positive (5 wins, 1 loss, and 22 neutral), as expected (see Appendix A.2).

Ablations and variations. We compare prompt repetition to 2 additional variants: *Prompt Repetition (Verbose)* and *Prompt Repetition × 3* (Appendix A.4). We observe that they perform similarly for most tasks and models (Figures 2 and 3), and sometimes outperform vanilla prompt repetition. Notably, *Prompt Repetition × 3* often substantially outperforms vanilla prompt repetition (which itself substantially outperforms the baseline) on NameIndex and MiddleMatch. It therefore seems worthwhile to further research variants. To demonstrate that the gains are indeed due to repeating the prompt and not to simply increasing the length of the input, we also evaluate the *Padding* method (Appendix A.4), which pads the inputs with periods (“.”) to the same length as prompt repetition, and, as expected, does not improve performance.

Efficiency. For each model, prompting method, and dataset, we measure the average and median of the lengths of the generated outputs, as well as the empirical latency². As expected, we observe similar latencies for all datasets and all tested models when reasoning is disabled. With reasoning enabled, all latencies (and the lengths of the generated outputs) are dramatically higher. Either way, in all cases **prompt repetition and its variants do not increase the lengths of the generated outputs or the measured latencies** (Figures 2 and 3), the only exception being the Anthropic models (Claude Haiku and Sonnet) for very long requests (from the NameIndex or MiddleMatch datasets or from the repeat × 3 variant) where the latencies increase (likely due to the prefill stage taking longer).

¹Unlike other prompting techniques, such as “Think step by step” [Kojima et al., 2023], see Figures 2 and 3.

²We measure end-to-end latencies via the official API for each of the providers. These might be affected by e.g., network delays or transient loads. For fairness, we ran all requests to the same provider in a round-robin fashion. As can be seen in Figures 2 and 3 the measured latencies are consistent with expectation, but given the above, should be taken with a grain of salt. Notably, the measured latencies for Deepseek are very high.

3 Related Work

Many prompting techniques for LLMs have been suggested, notably Chain of Thought (CoT) prompting [Wei et al., 2023] (which requires specific examples per task) and “Think step by step” [Kojima et al., 2023], which achieves substantial improvements, but increases the lengths of the generated outputs and thus the latency and compute requirements (we show that it can be used in tandem with prompt repetition, yielding mostly neutral results). More recently and independently, Shaier [2024] experimented with repeating just the question part of the prompt and found that it yields no gains, Springer et al. [2024] showed that repeating the input twice yields better text embeddings, and Xu et al. [2024] showed that asking the model to re-read the question improves reasoning.

4 Conclusion

We show that repeating the prompts consistently improves model performance for a range of models and benchmarks, when not using reasoning. In addition, latency is not impacted³, as only the parallelizable pre-fill stage is affected⁴. Prompt repetition does not change the lengths or formats of the generated outputs, and it might be a good default for many models and tasks, when reasoning is not used.

Future directions. (1) Fine tune the model with repeated prompts; (2) Train a reasoning model with prompt repetition to increase efficiency (the model might learn to avoid repeating the prompt); (3) Periodically repeat the last generated tokens during generation, as well as explore applicability to multi-turn scenarios; (4) Only keep the second repetition in the KV-cache (thus being completely performance neutral for the generation stage); (5) Repeat only parts of the prompt (especially for longer prompts); (6) Reorder the prompt, e.g. with a smaller model, instead of repeating everything; (7) Applicability to non-text modalities (e.g. images); (8) Further analyze different variants, e.g. when more than 2 repetitions might be advantageous; (9) Further analyze the attention patterns due to repetition (in the same vein as done in Xu et al. [2024]); (10) Use repetitions in tandem with techniques like selective attention [Leviathan et al., 2024]; (11) Explore interactions with techniques such as Prefix LM [Raffel et al., 2023]; (12) Investigate when repetition is helpful and how token representations vary between repetitions; (13) Explore promising variants (see Appendix A.1).

Acknowledgements

We’d like to extend a huge thank you to Raya Leviathan, Danny Lumen, Dani Valevski, Sergey Levi, the Theta Labs and Google Research teams, and our families for insightful feedback, ideas, suggestions, and support.

References

- Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024. URL https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL <https://arxiv.org/abs/1803.05457>.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL <https://arxiv.org/abs/2110.14168>.
- DeepSeek-AI. Deepseek-v3 technical report, 2025. URL <https://arxiv.org/abs/2412.19437>.

³Prompt repetition can affect latency for long prompts, and might be impossible for very long ones.

⁴This parallelization has a similar motivation to e.g., speculative decoding [Leviathan et al., 2022].

Gemini Team Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *NeurIPS*, 2021.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL <https://arxiv.org/abs/2205.11916>.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding, 2022. URL <https://arxiv.org/abs/2211.17192>.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Selective attention improves transformer, 2024. URL <https://arxiv.org/abs/2410.02703>.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL <https://arxiv.org/abs/1809.02789>.

OpenAI. Gpt-4o system card, 2024. URL <https://arxiv.org/abs/2410.21276>.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer, 2023. URL <https://arxiv.org/abs/1910.10683>.

Sagi Shaier. Asking again and again: Exploring llm robustness to repeated questions, 2024. URL <https://arxiv.org/abs/2412.07923>.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. Repetition improves language model embeddings, 2024. URL <https://arxiv.org/abs/2402.15449>.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark, 2024. URL <https://arxiv.org/abs/2406.01574>.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL <https://arxiv.org/abs/2201.11903>.

Xiaohan Xu, Chongyang Tao, Tao Shen, Can Xu, Hongbo Xu, Guodong Long, Jian guang Lou, and Shuai Ma. Re-reading improves reasoning in large language models, 2024. URL <https://arxiv.org/abs/2309.06275>.

A Appendix

A.1 Ablations and Variations

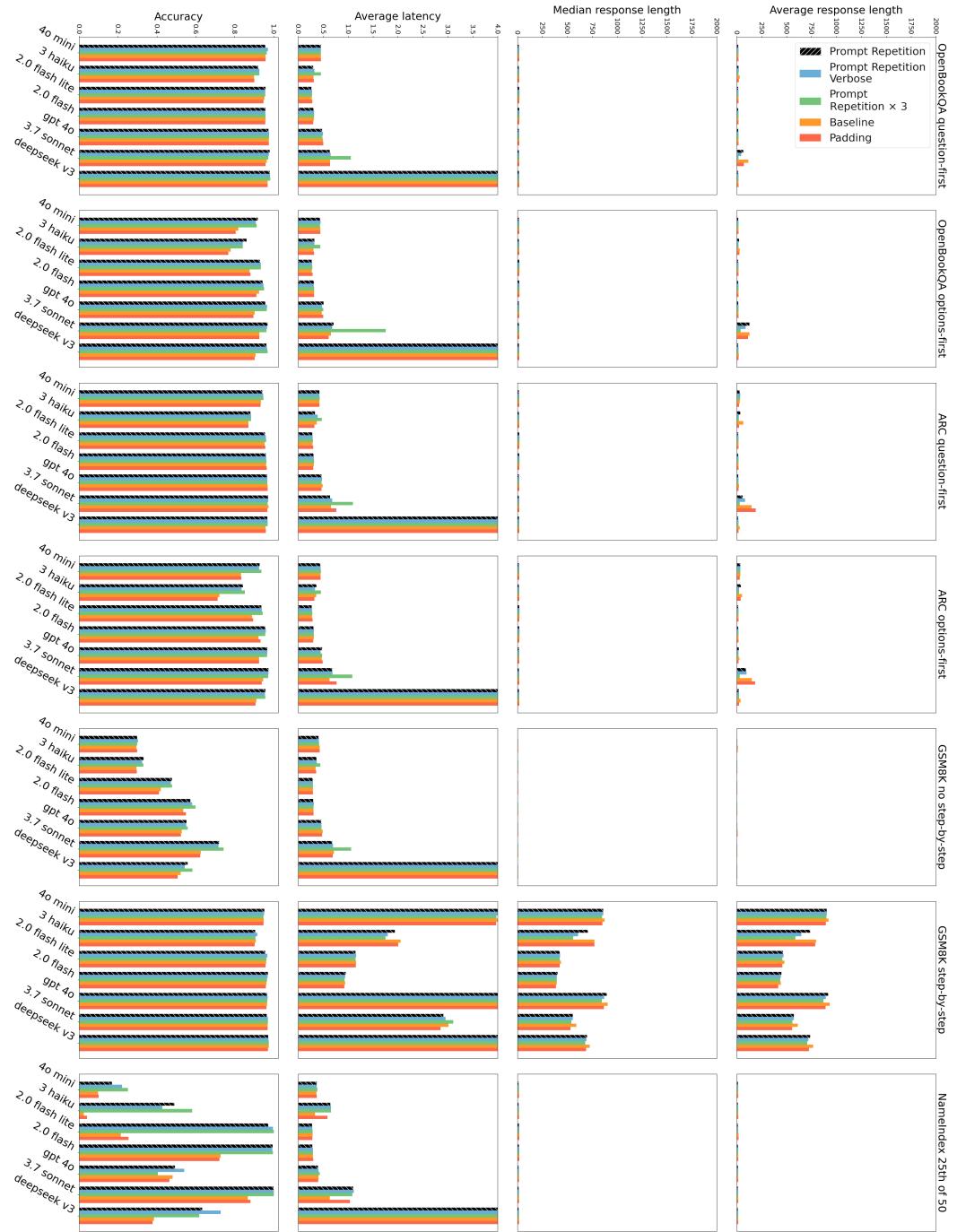


Figure 2: Comparison of accuracy, average, and median response length, as well as average latency, across methods and benchmarks (1).



Figure 3: Comparison of accuracy, average, and median response length, as well as average latency, across methods and benchmarks (2).

A.2 Prompt Repetition with Reasoning

When asking the models to think step by step, we observe that prompt repetition is neutral to slightly positive (5 wins, 1 loss, 22 ties), which is expected given that the reasoning often starts with repeating (parts-of) the prompt anyway.

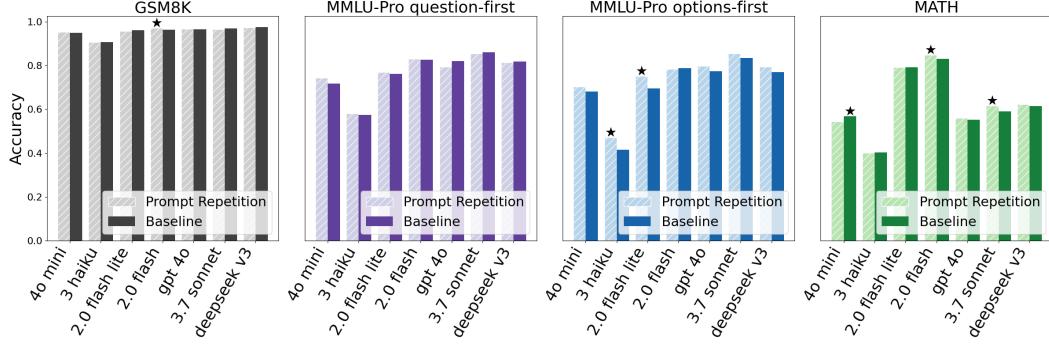


Figure 4: Prompt repetition vs. baseline accuracy for popular LLMs and various benchmarks when asking the model to think step by step. A star indicates a statistically significant win ($p_{value} < 0.1$ according to the McNemar test [McNemar \[1947\]](#)). Prompt repetition wins 5 out of 28 tests, with 1 loss.

A.3 Custom Tasks

In addition to the standard benchmarks, we also evaluate prompt repetition on two custom tasks specifically designed to demonstrate its usefulness.

NameIndex Here the model gets a list of N names and is asked to output the i th name on the list. We use $N = 50, i = 25$. Example:

Here's a list of names:

Dale Lopez, Peter Sanchez, Allen Harris, Scott Davis, Hudson Leviathan, Daphne Kalman, Dennis Davis, Henry King, Alfred Cooper, Bruce Usher, Travis Ramirez, Rafael Jennings, Richard Rogers, Walter Young, Caleb Harris, Ben Kalman, Donald Carter, Richard Sterling, Mark Nightingale, Steven Carter, Talia Kalman, Dennis Hanson, James Harris, Craig Chavez, Paul Sanchez, Samuel Curtis, Jacob James, Allen Thomas, Dale Evans, James Fox, Douglas Allen, Orion Johnson, Alexander Wright, Eugene Morrison, Nelson Lee, Alan Young, Caleb Ward, Alberto Robinson, Robert McCarthy, Mark Price, Kenneth Ramirez, Jeffrey White, Chad Cooper, Arthur Waters, Bruce Callahan, Liam Leviathan, Steven Robinson, Alberto Murphy, Leonard Johnson, Robert Murphy

What's the 25th name?

MiddleMatch Here the model gets a list of N names or numbers (out of a total possible set of K , i.e., $K < N$ means there will be repeating elements), and is asked to output the name/number located directly between two given ones. We use $N = 40, K = 10$. Example:

Here's a list (potentially with repetitions) of names:

Carlos Davis, Dale Sims, Carlos Davis, Dale Sims, Stephen Cruz, Dale Sims, Finnian Ross, Stephen Cruz, Stephen Cruz, Gregory Collins, Dale Sims, Stephen Cruz, Carlos Davis, Stephen Cruz, Dale Sims, Dale Sims, Stephen Cruz, Stephen Cruz, Leonard Kalman, Bruce Phillips, Raymond Roberts, Dale White, Leonard Kalman, Finnian Ross, James Wright, Finnian Ross, Raymond Roberts, Dale Sims, Dale Sims, Leonard Kalman, Dale Sims, Carlos Davis, Leonard Kalman, Bruce Phillips, Dale Sims, Raymond Roberts, Gregory Collins, Gregory Collins, Dale Sims, Finnian Ross

What is the single name that appears right between Carlos Davis and Bruce Phillips?

A.4 Query Examples

Method	Template	Example Query
Baseline	<QUERY>	<p>Which of the following combinations is a mixture rather than a compound?</p> <p>A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia</p> <p>Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>.</p>
Prompt Repetition	<QUERY> <QUERY>	<p>Which of the following combinations is a mixture rather than a compound?</p> <p>A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia</p> <p>Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>.</p> <p>Which of the following combinations is a mixture rather than a compound?</p> <p>A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia</p> <p>Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>.</p>

(Continued on next page)

(Continued from previous page)

Method	Template	Example Query
Prompt Repetition (Verbose)	<QUERY> Let me repeat that: <QUERY>	Which of the following combinations is a mixture rather than a compound? A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>. Let me repeat that: Which of the following combinations is a mixture rather than a compound? A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>.

(Continued on next page)

(Continued from previous page)

Method	Template	Example Query
Prompt Repetition x3		<p>Which of the following combinations is a mixture rather than a compound?</p> <p>A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia</p> <p>Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>.</p> <p>Let me repeat that:</p> <p><QUERY></p> <p>Let me repeat that:</p> <p><QUERY></p> <p>Let me repeat that one more time:</p> <p><QUERY></p> <p>Which of the following combinations is a mixture rather than a compound?</p> <p>A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia</p> <p>Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>.</p> <p>Let me repeat that one more time:</p> <p>Which of the following combinations is a mixture rather than a compound?</p> <p>A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia</p> <p>Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>.</p>

(Continued on next page)

(Continued from previous page)

Method	Template	Example Query
Padding	<QUERY> Ignore these periods (they are irrelevant) and answer the above question:<LEN(QUERY)>...	Which of the following combinations is a mixture rather than a compound? A. oxygen and nitrogen in air B. sodium and chlorine in salt C. hydrogen and oxygen in water D. nitrogen and hydrogen in ammonia Reply with one letter ('A', 'B', 'C', 'D') in the format: The answer is <ANSWER>.. Ignore these periods (they are irrelevant) and answer the above question: