# Does Self-Consistency Improve the Recall of Encyclopedic Knowledge?

**Sho Hoshino, Ukyo Honda, Peinan Zhang**
CyberAgent
{hoshino_sho,honda_ukyo,zhang_peinan}@cyberagent.co.jp

## Abstract

While self-consistency is known to improve performance on symbolic reasoning, its effect on the recall of encyclopedic knowledge is unclear due to a lack of targeted evaluation grounds. To address this, we establish such a knowledge recall split for the popular MMLU benchmark by applying a data-driven heuristic from prior work. We validate this split by showing that the performance patterns on the symbolic reasoning and knowledge recall subsets mirror those of GSM8K and MedMCQA, respectively. Using this solid ground, we find that self-consistency consistently improves performance across both symbolic reasoning and knowledge recall, even though its underlying CoT prompting is primarily effective for symbolic reasoning. As a result, we achieve an 89% accuracy on MMLU, the best performance to date with the use of GPT-4o.

## 1 Introduction

The chain-of-thought prompting (CoT; Wei et al., 2022; Nye et al., 2022) has become the de facto standard for achieving the best performance on large language model (LLM) benchmarks (OpenAI et al., 2024; Gemini Team et al., 2024). Nevertheless, it is widely held that CoT is primarily effective for tasks requiring symbolic reasoning. Sprague et al. (2025) reported that on the Massive Multitask Language Understanding (MMLU) benchmark (Hendrycks et al., 2021), 95% of the performance gain from CoT is attributed to questions involving symbolic reasoning.

Building on top of CoT, self-consistency (SC; Wang et al., 2023) further improves performance by sampling multiple CoT reasoning paths and selecting the most consistent answer. However, as self-consistency is fundamentally an extension of CoT, it is unclear whether self-consistency also improves performance on non-math questions that involve the recall of encyclopedic knowledge (hereafter, *knowledge recall*[1]), such as *"How high is Mt. Fuji?"* shown in Figure 1. To investigate this gap, we address two primary research questions: Does knowledge recall benefit from multiple reasoning paths (**RQ1**)? If so, how does self-consistency improve knowledge recall with multiple samples (**RQ2**)?

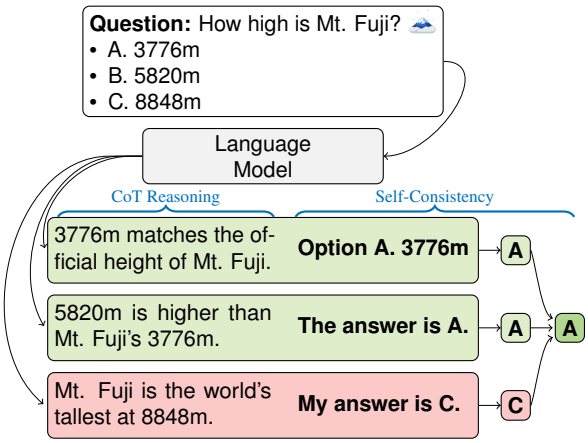


Figure 1: Illustrating how self-consistency can mitigate incorrect reasoning path for a knowledge recall question.

Our assessment reveals that self-consistency improves both the symbolic reasoning and knowledge recall performance over vanilla CoT, despite its premise. To enable this analysis, we apply a subject-level split to the MMLU dataset based on a prior heuristic. We validate this split by showing that the performance patterns on the symbolic reasoning and knowledge recall subsets mirror those of prototypical benchmarks, GSM8K (Cobbe et al., 2021) and MedMCQA (Pal et al., 2022), respectively. As a result, we achieve an 89% accuracy on MMLU, the best performance to date with the use of GPT-4o.

We further analyzed the mechanism behind this unexpected improvement. Our quantitative analysis found that the agreement among answers serves

[1]We followed the term coined by Chung et al. (2024).



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 9–17

as a reliable confidence score, providing evidence that self-consistency filters out incorrect paths by leveraging this signal. Our qualitative analysis further illustrates how self-consistency stabilizes unreliable CoT reasoning by filtering out these incorrect paths, making it effective for both symbolic reasoning and knowledge recall.

## 2 Data Splitting Methodology

Since the MMLU benchmark (Hendrycks et al., 2021) contains a wide range of 57 subjects, several categorization methods have been proposed. Indeed, Hendrycks et al. (2021) defined their own subject-level categorization with "supercategory", such as STEM and humanities. Their grouping was not designed to distinguish knowledge recall from symbolic reasoning. For example, the subject of econometrics, which requires symbolic reasoning, is categorized as humanities.

A more targeted approach by Sprague et al. (2025) used the "=" sign as an instance-level cue for symbolic reasoning. Their post-hoc analysis was model-dependent, as it relied on the cue's presence in either the question or the LLM's output. While this analysis was post-hoc, the heuristic originated from a classifier trained to differentiate subjects, providing a basis for subject-level split.

Building on these insights, we apply Sprague et al.'s (2025) heuristic across all 57 MMLU subjects to create a stable, *a priori* categorization independent of model outputs, as shown in Figure 2. We aggregated subjects guided solely by the presence of "=" in the questions and then propagated this classification within a discipline (e.g., from college math to elementary math). Our application disentangles knowledge recall from symbolic reasoning in a ratio of about 2:1.

To validate our split, we select two prototypical benchmarks, including GSM8K, a math dataset, and MedMCQA, a medical dataset where symbolic reasoning cues are almost entirely absent (16 out of 4,183 instances). We expect the performance patterns on our MMLU subsets to mirror those on these prototypical benchmarks. A data-driven analysis in Appendix E, which shows a high overlap (AUC of 0.96) between our split and one based on CoT performance gains, provides further validation.

**Symbolic Reasoning Subjects**

Abstract Algebra, Business Ethics, College Chemistry, College Computer Science, College Mathematics, College Physics, Conceptual Physics, Econometrics, Elementary Mathematics, Formal Logic, High School Chemistry, High School Computer Science, High School Macroeconomics, High School Mathematics, High School Microeconomics, High School Physics, Machine Learning, Professional Accounting

**Knowledge Recall Subjects**

Anatomy, Astronomy, Clinical Knowledge, College Biology, College Medicine, Computer Security, Electrical Engineering, Global Facts, High School Biology, High School European History, High School Geography, High School Government and Politics, High School Psychology, High School Statistics, High School US History, High School World History, Human Aging, Human Sexuality, International Law, Jurisprudence, Logical Fallacies, Management, Marketing, Medical Genetics, Miscellaneous, Moral Disputes, Moral Scenarios, Nutrition, Philosophy, Prehistory, Professional Law, Professional Medicine, Professional Psychology, Public Relations, Security Studies, Sociology, US Foreign Policy, Virology, World Religions

Figure 2: Listing symbolic reasoning and knowledge recall subjects defined for the MMLU dataset.

## 3 Experiments

We first validate our proposed MMLU split to ensure it effectively separates symbolic reasoning and knowledge recall. With the split validated, we then present our main results on the effectiveness of self-consistency, configured as follows.[2]

### 3.1 Setup

**Data.** We use the MMLU (Hendrycks et al., 2021) and MedMCQA (Pal et al., 2022) for multiple-choice question answering (MCQA; Balepur et al., 2025), and GSM8K (Cobbe et al., 2021) as an open-ended question answering prototype.

**Metrics.** To ensure a rigorous assessment, we evaluate using classification accuracy for the MCQA datasets, and a similar accuracy metric for GSM8K. We have performed experiments using the zero-shot setting, as opposed to the few-shot setting (Chung et al., 2024).[3]

**LLM and Prompt.** We use GPT-4o version 2024-08-06, GPT-4o-mini version 2024-07-18 (OpenAI et al., 2024), and Qwen2.5-32B-Instruct

[2]We provide full implementation details, including hyperparameters and data splitting, in Appendix A.

[3]We also performed experiments with a similar few-shot setting, as discussed in Appendix C.

| Prompt | Sampling | Accuracy (%) ↑ MMLU test All | Reasoning | Knowledge | GSM8K test | MedMCQA valid |
|---|---|---|---|---|---|---|
| DA | nucleus | 83.26 | 75.45 | 85.56 | 46.93 | 75.07 |
| CoT | nucleus | 87.86 (+4.60) | 90.38 (+14.93) | 87.12 (+1.56) | 84.23 (+37.30) | 76.76 (+1.69) |
| CoT | +SC ($n$=5) | **88.64**$^*$ (+5.38) | **91.32**$^*$ (+15.87) | **87.85**$^*$ (+2.29) | **84.31** (+37.38) | **77.67** (+2.60) |
| CoT | +SC ($n$=20) | **88.93**$^*$ (+5.67) | **91.94**$^*$ (+16.49) | **88.04**$^*$ (+2.48) | **84.46** (+37.53) | **77.41** (+2.34) |

Table 1: Performance of GPT-4o on MMLU, GSM8K, and MedMCQA. The top section validates our MMLU split, which consists of the full test set ("All"), the symbolic reasoning subset ("Reasoning"), and the knowledge recall subset ("Knowledge"). The bottom section shows that self-consistency (SC) consistently improves performance over the vanilla CoT baseline for both symbolic reasoning and knowledge recall, with these improvements highlighted in bold. The asterisk ($^*$) denotes statistical significance ($p < 0.05$, detailed in Appendix A).

(Qwen et al., 2025), without changing hyperparameters unless explicitly mentioned.[4] We use two types of prompts, including the zero-shot CoT (Kojima et al., 2022) and direct answer without CoT (abbreviated as DA).

**Sampling.** We use nucleus sampling (Holtzman et al., 2020) with top-$p$=0.9. For the vanilla CoT baseline, the first sample is selected deterministically. For the self-consistency, we varied the number of samples ($n$) from 3 to 20.

### 3.2 Validation of the MMLU Split

To validate our proposed split, we examine the performance difference between CoT and direct answer. We expect symbolic reasoning to benefit significantly from CoT, while knowledge recall should show little to no gain. The top section of Table 1 confirms this. The MMLU symbolic reasoning subset shows a substantial +14.93 point gain from CoT (75.45 to 90.38), while the knowledge recall subset shows only a modest +1.56 point gain (85.56 to 87.12).

This pattern is mirrored on our prototypical benchmarks, GSM8K (symbolic reasoning) and MedMCQA (knowledge recall). GSM8K sees a massive +37.3 point gain from CoT (46.93 to 84.23), whereas MedMCQA sees a small +1.69 point gain (75.07 to 76.76). The strong alignment between the MMLU splits and their corresponding prototypical benchmarks validates our split.

### 3.3 Self-Consistency on Knowledge Recall

Having validated our split, we now investigate whether knowledge recall benefits from multiple reasoning paths (**RQ1**). The bottom section of

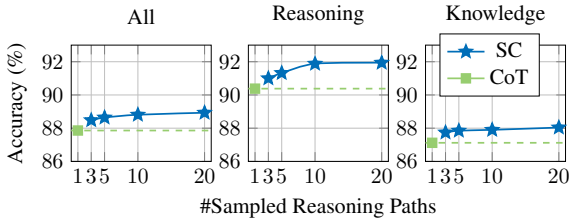


Figure 3: Comparing SC and vanilla CoT on MMLU using different numbers of samples from GPT-4o.

Table 1 shows that self-consistency consistently improves performance over vanilla CoT for both symbolic reasoning and, crucially, knowledge recall. As a result, we achieved an overall accuracy of 89% on the MMLU, the highest score to date using GPT-4o.[5]

This trend holds on our prototypical benchmarks as well. As shown in Table 1, self-consistency provides gains for both GSM8K (symbolic reasoning) and MedMCQA (knowledge recall). These results further validate the robustness of our findings.

As shown in Figure 3, self-consistency outperformed vanilla CoT with no degradation observed, even when using different numbers of samples.

## 4 Analysis

To further investigate how self-consistency improves knowledge recall with multiple samples (**RQ2**), we present quantitative evidence for its core mechanism and then provide a qualitative illustration of the underlying reasoning patterns.

### 4.1 Quantitative Evidence of the Mechanism

We quantitatively test self-consistency's core mechanism, which relies on agreement among answers as a strong signal for correctness. To

[4]Our main analysis focuses on GPT-4o for clarity. Full results for all models are in Appendix D.

[5]Similar scores are already reported for different LLMs.

| Prompt | Sampling | Pearson's $\rho$ [-1, 1] ↑ MMLU test | | |
|---|---|---|---|---|
| | | All | Reasoning | Knowledge |
| CoT | +SC ($n$=5) | 0.40 | 0.43 | 0.40 |
| CoT | +SC ($n$=20) | 0.42 | 0.46 | 0.42 |

Table 2: Correlation between prediction correctness and confidence based on answer agreement.

do so, we formalize a confidence score as $s = \frac{\text{the count of the majority answer}}{\text{the number of valid answers}}$. For example, from the three answers $\{\text{A}, \text{A}, \text{C}\}$ in Figure 1, our confidence in the final answer A is calculated as $s = \frac{2}{3}$. This principle aligns with findings from a separate line of work showing that confidence metrics based on answer consistency are more reliable than than methods based on prediction logits (Yang et al., 2025), especially since the first token may not indicate the final answer (Wang et al., 2024a). To validate that this score is a meaningful signal, we measure its Pearson correlation coefficient ($\rho$; Pearson, 1895) with the correctness of the predictions.

As summarized in Table 2, using the count of the majority answer as a confidence score showed a moderate correlation with the correctness of the LLMs' predictions. This finding provides quantitative evidence for self-consistency's core mechanism, confirming that majority voting effectively filters out incorrect reasoning paths by leveraging this signal.

### 4.2 Qualitative Illustration

To complement the quantitative finding, we analyzed typical CoT reasoning paths from the MMLU development set to understand why different samples produce different answers, as shown in Figure 4. Unlike symbolic reasoning that involves multi-hop reasoning, one would expect knowledge recall to be a single deduction that does not require multiple reasoning paths.

Unexpectedly, we found that multiple reasoning paths exist because the model generates different plausible justifications for the same question, even for knowledge recall. For example, both reasoning paths in Figure 4 contain seemingly plausible justifications (e.g., *"A situation analysis would typically be conducted here"*), despite leading to different answers. Our observations indicate that self-consistency does not alter how LLMs work, but rather stabilizes unreliable CoT reasoning by filtering out these incorrect paths. By selecting the paths that lead to the most consistent final answer, the technique makes CoT effective for both symbolic reasoning and knowledge recall.

Question: At which stage in the planning process would a situation analysis be carried out?
A. Defining the program
B. Planning the program
C. Taking action and implementing ideas

**Correct Reasoning Path**

A. Defining the program: This stage involves identifying the overall goals and objectives of the program. It sets the foundation for what the program aims to achieve. **Situation analysis is crucial here** because it helps understand the current environment, needs, and gaps, which are essential to defining the goals and objectives accurately.
Therefore, the correct answer is: A. Defining the program

**Incorrect Reasoning Path**

2. **Planning the program**: In this stage, specific plans are formulated based on the initial definition of the program. This is where detailed analysis and assessments are often conducted to inform the planning process. **A situation analysis would typically be conducted here** to understand the current state, identify opportunities and threats, and develop strategies based on this analysis.
Thus, the correct option is: B. Planning the program

Figure 4: An example of typical reasoning paths for a multiple-choice question involving knowledge recall. In this example, the option A is the correct answer.

### 4.3 Cost-Performance Trade-off

While self-consistency consistently improves knowledge recall, this gain comes with a practical trade-off because the inference cost scales linearly with the sample size $n$. As illustrated in Figure 3, increasing the number of samples yields diminishing returns in performance. To quantify this trade-off, achieving a +0.73 point gain on the MMLU knowledge recall subset ($n = 5$) requires approximately five times the computational cost compared to vanilla CoT ($n = 1$).

## 5 Related Work

Wang et al. (2023) proposed self-consistency and performed experiments on arithmetic and commonsense reasoning benchmarks. After that, Chung et al. (2024) conducted additional experiments including the MMLU benchmark, which is perhaps the most related work. However, their study did not provide a detailed breakdown of MMLU results by subject, leaving the performance on knowledge recall unclear.

Gema et al. (2025) reported that more than 9% of MMLU examples are considered incorrect due to errors in the dataset creation. To mitigate such

errors, several fixes to the MMLU dataset have been proposed, including MMLU-Redux (Gema et al., 2025), MMLU-Pro (Wang et al., 2024b), and MMLU-CF (Zhao et al., 2025). Nevertheless, the subject-level split we use is orthogonal to these instance-level fixes.

## 6 Conclusion

To investigate self-consistency's performance on the recall of encyclopedic knowledge, we applied a subject-level split to MMLU based on a prior heuristic. Our assessment on MMLU demonstrated that self-consistency yields consistent improvements over vanilla CoT for both symbolic reasoning and knowledge recall. As a result, we achieved 89% accuracy, the highest score to date on the MMLU with the use of GPT-4o.

## Limitations

**On the Granularity of the Data Split** One might critique our subject-level split between symbolic reasoning and knowledge recall as a coarse-grained approximation. We agree that any such partition is an imperfect proxy, and there is no perfect approach, as even instance-level classifications can be ambiguous. Our subject-level split is no exception, as some subjects contain a mix of question types. However, our goal was to move beyond treating MMLU as a monolithic benchmark. Therefore, we provide a stable, *a priori* split as a solid ground, offering a clear improvement over the status quo.

**On the Practical Costs of Self-Consistency** One could argue that the performance gain from self-consistency comes with additional costs. We admit these practical drawbacks, as self-consistency requires multiple samples of CoT reasoning paths, which means longer runtime and higher cost compared to vanilla CoT. In this study, however, our focus was to establish a targeted evaluation ground. By creating a principled *a priori* split, we established the solid ground needed to answer the fundamental question of whether self-consistency improves the recall of encyclopedic knowledge.

**On the Scope of Task Formats** While our experiments on MMLU involve multiple-choice question answering to ensure a rigorous assessment via classification accuracy, our scope is not limited to this task format. In fact, we included GSM8K as a open-ended question answering prototype. Furthermore, future work can address arbitrary tasks with universal self-consistency (Chen et al., 2024).

## References


Nishant Balepur, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2025. Which of these best describes multiple choice evaluation with LLMs? A) forced B) flawed C) fixable D) all of the above. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 3394–3418.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. 2024. Universal self-consistency for large language models. In *ICML 2024 Workshop on In-Context Learning*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, and 16 others. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

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. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. Are we done with MMLU? In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5069–5096.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *Proceedings of the Ninth International Conference on Learning Representations*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *Proceedings of the Eighth International Conference on Learning Representations*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395.

Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35:22199–22213.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. 2022. Show your work: Scratchpads for intermediate computation with language models. In *Proceedings of the first Deep Learning for Code Workshop*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the 5th AHLI Conference on Health, Inference, and Learning*, pages 248–260.

Karl Pearson. 1895. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(vii):240–242.

Qwen, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, and 24 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To CoT or not to CoT? chain-of-thought helps mainly on math and symbolic reasoning. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024a. "My Answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7407–7416.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proceedings of the Eleventh International Conference on Learning Representations*.

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 Wenhu Chen. 2024b. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Yongjin Yang, Haneul Yoo, and Hwaran Lee. 2025. MAQA: Evaluating uncertainty quantification in LLMs regarding data uncertainty. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5846–5863.

Qihao Zhao, Yangyu Huang, Tengchao Lv, Lei Cui, Qinzheng Sun, Shaoguang Mao, Xin Zhang, Ying Xin, Qiufeng Yin, Scarlett Li, and Furu Wei. 2025. MMLU-CF: A contamination-free multi-task language understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 13371–13391.

## A Setup Details

**Data Splitting.** For the MMLU dataset, we use dev split (285 instances) for development and test split (14,042 instances) for testing. For the GSM8K dataset, we use test split (1,319 instances) for testing. For the MedMCQA dataset, we use validation split (4,183 instances) for testing, instead of test split, because the latter's gold labels are not publicly available.

**Hyperparameters.** We set max tokens to 20 for DA and 1,000 for CoT, respectively, because the GPT-4o models tend to output longer, surpassing the 128 tokens previously used for the GPT-3 models (Wang et al., 2023).

**Prompts.** Figure 5 lists our prompt used as direct answer and chain-of-thought. In our preliminary study, we observed significant performance gains with the combinations of general and negative instructions proposed in Appendix B.

**Statistical Testing.** We assessed statistical significance using paired bootstrap resampling (Koehn, 2004). The improvements from self-consistency are statistically significant ($p < 0.05$) on the MMLU benchmark across the evaluated models, with the exception of the knowledge recall subset for Qwen2.5-32B-Instruct at $n = 5$. The performance gains on GSM8K and MedMCQA were consistent in direction, while they did not reach statistical significance in all configurations, likely due to the smaller size of these test sets.

**Licenses.** We performed experiments using the MMLU dataset (Hendrycks et al., 2021) released under the MIT license[6], the GSM8K dataset (Cobbe et al., 2021) released under the MIT license[7], the MedMCQA dataset (Pal et al., 2022) released under the Apache License 2.0[8], GPT-4o API (OpenAI et al., 2024) released under a proprietary license via Azure OpenAI Service[9], and Qwen2.5-32B-Instruct (Qwen et al., 2025) released under the Apache License 2.0[10].

[6] https://huggingface.co/datasets/cais/mmlu

[7] https://huggingface.co/datasets/openai/gsm8k

[8] https://huggingface.co/datasets/openlifescienceai/medmcqa

[9] https://azure.microsoft.com/en-us/products/ai-services/openai-service

[10] https://huggingface.co/Qwen/Qwen2.5-32B-Instruct

**Direct Answer Prompt**

```
Please read the multiple-choice question below carefully and select ONE of the listed options. Provide only the symbols and do not output anything else after that.

Question: {question}
Options:
{options}
Answer:
```

**Chain-of-Thought Prompt**

```
Please read the multiple-choice question below carefully, explain step-by-step, and select ONE of the listed options. Provide only the explanations and symbols. Do not output anything else after that.

Question: {question}
Options:
{options}
Answer:
```

Figure 5: Listing our prompts used as the direct answer (DA) and the zero-shot chain-of-thought (CoT). The placeholders "{question}" and "{options}" are replaced with an actual question and its options, as in Figure 4.

## B Implementation Details

We incorporated self-consistency with the best performing MCQA instructions and the zero-shot CoT prompt by turning them into generalized pre- and post-processing steps for LLMs, as follows.

**Pre-processing.** We generate a LLM prompt using three types of instructions used in relevant studies: (a) the general instruction, e.g. "*Please read the multiple-choice question*" (Wang et al., 2024a), (b) the negative instruction that prevents LLMs from producing unnecessary output, e.g. "*Provide only the symbols*" (Robinson et al., 2023), and (c) the zero-shot CoT prompt, e.g. "*explain step-by-step*" (Kojima et al., 2022), which is used only when necessary.

**Post-processing.** After obtaining multiple samples from LLMs, we parse each sample to obtain candidate answers, and then use self-consistency (Wang et al., 2023) to determine the final answer. This is done in three steps: (i) We parse the last line in the LLM output and extract the first valid answer. For example, from *"Option A. 3776m"* in Figure 1, we extract the option A. Unlike Wang et al. (2023) who used a string template "The answer is X", we assumed the alphabetical options [A-Z] only. In the case of GSM8K, we extract the numeric answers instead. (ii) We dismiss the samples from which we cannot extract a valid answer. (iii) We perform majority vote over the extracted answers.

| Prompt | Sampling | Accuracy (%) ↑ MMLU dev All | Reasoning | Knowledge |
|---|---|---|---|---|
| | | WITH 0-SHOT | | |
| CoT | nucleus | 80.35 | 84.44 | 78.46 |
| CoT | +SC ($n$=5) | **83.16** | **86.67** | **81.54** |
| CoT | +SC ($n$=20) | **82.81** | **88.89** | 80.00 |
| | | WITH 4-SHOT | | |
| CoT | nucleus | 80.35 | 80.00 | 80.51 |
| CoT | +SC ($n$=5) | 80.35 | 78.89 | **81.03** |
| CoT | +SC ($n$=20) | **82.46** | **85.56** | **81.03** |

Table 3: Comparisons of our zero-shot method with the conventional few-shot method using GPT-4o-mini. Bold text highlights improvements over the vanilla CoT baseline.

For instance, from the valid answers $\{\text{A}, \text{A}, \text{C}\}$ in Figure 1, we determine that the final answer is A.

**Tie-Breaking.** To ensure reproducibility, voting ties in our implementation (e.g. two votes for A versus two votes for B) are resolved deterministically by selecting the first option in alphabetical order (i.e. A). The quantitative analysis in §4.1 inherently accounts for such disagreement (e.g. $s = \frac{2}{2+2} = 0.5$), appropriately reflecting the reduced confidence of tied votes.

## C Zero-shot versus Few-shot

Unlike our zero-shot inference setting, Chung et al. (2024) performed experiments with a few-shot setting by using the MMLU development data that contains five examples for each subject as demonstration exemplars. We conducted similar experiments on the MMLU development data. To prevent data leakage for our assessment, we used four examples from each subject instead of five.

Table 3 summarizes our preliminary experiments comparing our zero-shot method with the conventional few-shot method on the MMLU development data. These results demonstrate that our zero-shot method outperformed the previously used few-shot method, indicating its flexibility to parse arbitrary answers over a string template. The only exception is a single instance ($n$=20), which can be attributed to the limited performance of GPT-4o-mini.

## D Full Results

This section presents the full results of our experiments. Consistent with our main findings, the results show that self-consistency generally improves performance over vanilla CoT. Table 4 details the performance of GPT-4o, GPT-4o-mini, and Qwen2.5-32B-Instruct on the MMLU test set and our proposed splits. Table 5 shows the corresponding results for GPT-4o and GPT-4o-mini on the GSM8K and MedMCQA benchmarks.

| Prompt | Sampling | Accuracy (%) ↑ MMLU test All | Reasoning | Knowledge |
|---|---|---|---|---|
| | | GPT-4O | | |
| DA | nucleus | 83.26 | 75.45 | 85.56 |
| CoT | nucleus | 87.86 | 90.38 | 87.12 |
| CoT | +SC ($n$=5) | **88.64**$^*$ | **91.32**$^*$ | **87.85**$^*$ |
| CoT | +SC ($n$=20) | **88.93**$^*$ | **91.94**$^*$ | **88.04**$^*$ |
| | | GPT-4O-MINI | | |
| DA | nucleus | 74.80 | 66.76 | 77.18 |
| CoT | nucleus | 81.43 | 84.41 | 80.56 |
| CoT | +SC ($n$=5) | **82.56**$^*$ | **85.44**$^*$ | **81.71**$^*$ |
| CoT | +SC ($n$=20) | **82.61**$^*$ | **85.57**$^*$ | **81.74**$^*$ |
| | | QWEN2.5-32B-INSTRUCT | | |
| DA | nucleus | 79.82 | 77.01 | 80.65 |
| CoT | nucleus | 80.06 | 82.66 | 79.29 |
| CoT | +SC ($n$=5) | **81.21**$^*$ | **85.94**$^*$ | **79.81** |
| CoT | +SC ($n$=20) | **82.90**$^*$ | **87.13**$^*$ | **81.65**$^*$ |

Table 4: Performance on the MMLU test set and our proposed splits. The asterisk ($^*$) denotes statistical significance ($p < 0.05$).

| Prompt | Sampling | Accuracy (%) ↑ GSM8K | MedMCQA |
|---|---|---|---|
| | GPT-4O | | |
| DA | nucleus | 46.93 | 75.07 |
| CoT | nucleus | 84.23 | 76.76 |
| CoT | +SC ($n$=5) | **84.31** | **77.67**$^*$ |
| CoT | +SC ($n$=20) | **84.46** | **77.41** |
| | GPT-4O-MINI | | |
| DA | nucleus | 29.95 | 65.86 |
| CoT | nucleus | 86.13 | 67.51 |
| CoT | +SC ($n$=5) | 86.13 | **68.23** |
| CoT | +SC ($n$=20) | **86.88** | **68.66**$^*$ |

Table 5: Performance on GSM8K (symbolic reasoning) and MedMCQA (knowledge recall). These results on prototypical benchmarks further validate our MMLU split.

## E Data-driven Data Splitting

As an alternative to the heuristic-based split used in our main analysis, we also explored a pure data-driven data split. Following an analysis by Sprague et al. (2025), we ranked all MMLU subjects by the performance gain from CoT over a direct answer baseline, using results from GPT-4o. This model-dependent ranking, shown in Table 6, contrasts with our primary split, which applies a heuristic in a model-independent manner.

To quantify the agreement between the two splits, we calculated the Area Under the Receiver Operating Characteristic Curve (AUC; Fawcett, 2006) and achieved a score of 0.96, indicating a strong overlap. The ranking confirms that math subjects such as High School Mathematics enjoyed the most benefit from CoT, while humanity subjects such as High School History resulted in modest improvements. The notable discrepancies with our heuristic-based split include subjects such as Medicine, Management, Nutrition, and Biology.

| Subject | $\Delta$ CoT $-$ DA |
|---|---|
| High School Mathematics | 40.37 |
| Abstract Algebra | 24.00 |
| College Physics | 23.53 |
| College Mathematics | 23.00 |
| Elementary Mathematics | 21.96 |
| Formal Logic | 19.84 |
| Professional Accounting | 19.51 |
| College Chemistry | 16.00 |
| Moral Scenarios | 14.30 |
| College Computer Science | 14.00 |
| High School Chemistry | 13.30 |
| Econometrics | 13.16 |
| High School Statistics | 12.50 |
| High School Physics | 10.59 |
| Machine Learning | 8.04 |
| High School Computer Science | 8.00 |
| College Medicine | 5.78 |
| Management | 3.88 |
| High School Microeconomics | 3.36 |
| High School Macroeconomics | 3.34 |
| Nutrition | 3.27 |
| Conceptual Physics | 2.98 |
| Electrical Engineering | 2.76 |
| Astronomy | 2.63 |
| College Biology | 2.09 |
| Business Ethics | 2.00 |
| Sociology | 1.49 |
| Anatomy | 1.48 |
| High School US History | $1.47^{\sim}$ |
| Miscellaneous | 1.40 |
| High School Government and Politics | 1.03 |
| Medical Genetics | 1.00 |
| Professional Law | 0.91 |
| Professional Psychology | 0.66 |
| Logical Fallacies | 0.62 |
| High School Psychology | 0.37 |
| High School Biology | 0.32 |
| Computer Security | 0.00 |
| Human Aging | 0.00 |
| Professional Medicine | 0.00 |
| Prehistory | -0.30 |
| Clinical Knowledge | -0.38 |
| Marketing | -0.42 |
| International Law | -0.82 |
| Jurisprudence | -0.92 |
| US Foreign Policy | -1.00 |
| World Religions | -1.17 |
| High School European History | -1.22 |
| Security Studies | -1.63 |
| Virology | -1.81 |
| Public Relations | -1.81 |
| Moral Disputes | -2.32 |
| Human Sexuality | -3.05 |
| High School Geography | -3.53 |
| Philosophy | -3.86 |
| High School World History | -4.22 |
| Global Facts | -5.00 |

Table 6: Data-driven splitting obtained for the MMLU test using GPT-4o. The tilde ($\sim$) denotes the median value, which corresponds to High School US History.