

Does More Deliberation Help? A Comparative Evaluation of Chain-of-Thought Prompting Strategies for Grade-School Math Reasoning in Open-Source Instruction-Tuned Language Models

Arun Kumar

aarunkumar3333@gmail.com

Indian Institute of Technology Kharagpur <https://orcid.org/0009-0005-4775-1068>

Research Article

Keywords: chain-of-thought prompting, large language models, GSM8K, self-consistency, tree-of-thought, prompt engineering, grade-school math reasoning

Posted Date: October 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11327774/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Chain-of-thought (CoT) prompting has become a standard technique for eliciting multi-step reasoning from large language models (LLMs), and several refinements — zero-shot CoT, few-shot CoT, self-consistency decoding, and tree-of-thought (ToT) search — have each been reported to improve reasoning accuracy on frontier, closed-weight models. Comparatively little controlled evidence exists on how these strategies perform, relative to one another, on the small (3.8–8B parameter) open-weight instruction-tuned models that practitioners increasingly deploy for cost and privacy reasons. This paper reports a controlled empirical comparison of four CoT-based prompting strategies (zero-shot CoT, few-shot CoT, self-consistency, and tree-of-thought) across three open-weight instruction-tuned models — Mistral-7B-Instruct-v0.2, Phi-3-mini-4k-instruct, and Llama-3-8B-Instruct — on a 100-item subset of the GSM8K grade-school arithmetic benchmark. We report per-condition accuracy with 95% confidence intervals and test pairwise differences using McNemar's exact test on paired, per-item outcomes. Across all three models, few-shot CoT and self-consistency significantly outperform zero-shot CoT (accuracy gains of 12–40 percentage points, $p < 0.05$ in every model), while self-consistency with three sampled reasoning paths ($k = 3$) offers no statistically significant improvement over single-sample few-shot CoT. Tree-of-thought scores below few-shot CoT in every model (significantly so in two of three) and is statistically indistinguishable from the much cheaper zero-shot baseline in all three, despite requiring multiple sampled branches per answer. These results suggest that, for short, single-path arithmetic reasoning at this model scale, the presence of worked exemplars is the dominant driver of accuracy gains, while search-based deliberation (ToT) and low-sample self-consistency add inference cost without a corresponding, statistically reliable accuracy benefit.

I. Introduction

Large language models (LLMs) are known to struggle with multi-step arithmetic and logical reasoning when prompted to answer directly [5]. Chain-of-thought (CoT) prompting — instructing or demonstrating that a model should produce intermediate reasoning steps before a final answer — has emerged as the dominant remedy, substantially improving performance on arithmetic, commonsense, and symbolic reasoning benchmarks [1]. Building on this idea, subsequent work has proposed several refinements. Kojima et al. [2] showed that a single trigger phrase ("Let's think step by step") elicits useful reasoning chains even without worked examples (zero-shot CoT). Wang et al. [3] proposed self-consistency, which samples multiple independent reasoning paths at non-zero temperature and takes a majority vote over the extracted final answers, reporting large gains over greedy few-shot CoT decoding on GSM8K and related benchmarks. Yao et al. [4] proposed tree-of-thought (ToT) prompting, which frames generation as a search over a tree of intermediate "thoughts," allowing the model to explore, evaluate, and backtrack among partial solutions; ToT produced large gains on tasks that inherently require search and lookahead, such as the Game of 24 and Mini Crosswords.

Most of the empirical evidence behind these techniques comes from very large, often closed-weight models (GPT-3/3.5/4, PaLM-540B). It remains comparatively under-studied whether the same ranking of

strategies — and the same magnitude of gains — holds for the smaller (3.8–8B parameter) open-weight instruction-tuned models that are now widely used in cost-, latency-, and privacy-sensitive deployments. Because such models have weaker in-context reasoning capacity than frontier models, it is not obvious a priori whether more elaborate strategies such as self-consistency and tree-of-thought retain their advantage, or whether the benefits collapse to what a handful of worked examples already provide.

This paper addresses this gap with a controlled, head-to-head comparison of four CoT-based prompting strategies — zero-shot CoT, few-shot CoT, self-consistency, and tree-of-thought — evaluated on the same 100-item subset of the GSM8K benchmark [5] across three widely used open-weight instruction-tuned models. Beyond reporting point-estimate accuracy, we compute 95% confidence intervals for every condition and test pairwise differences on matched, per-item outcomes using McNemar's exact test [9], which is appropriate for comparing two prompting strategies evaluated on the identical set of questions. Specifically, we ask:

RQ1: Does exemplar-based prompting (few-shot CoT) significantly outperform zero-shot CoT in small, open-weight instruction-tuned models?

RQ2: Does self-consistency sampling provide an additional, statistically significant gain over single-sample few-shot CoT at a modest sample size ($k = 3$)?

RQ3: Does tree-of-thought's structured search improve accuracy over simpler CoT strategies on short, single-path elementary arithmetic problems?

The remainder of the paper is organized as follows. Section II reviews related work on CoT prompting and its variants. Section III describes the models, dataset, prompting strategies, and evaluation protocol. Section IV reports accuracy, confidence intervals, and significance tests. Section V discusses the findings and their implications for prompt-engineering practice. Section VI outlines limitations and threats to validity, and Section VII concludes.

II. Related Work

Wei et al. [1] first demonstrated that prepending a small number of worked reasoning examples to a prompt (few-shot CoT) substantially improves performance on arithmetic and symbolic reasoning tasks for sufficiently large language models, an effect they characterized as an emergent capability of scale. Kojima et al. [2] subsequently showed that a single instruction — "Let's think step by step" — elicits similar step-by-step reasoning without any worked examples (zero-shot CoT), though at generally lower accuracy than the few-shot variant. Wang et al. [3] introduced self-consistency, replacing greedy decoding with the sampling of several diverse reasoning paths and a majority vote over their final answers; the original study reported gains of up to 17.9 percentage points on GSM8K when combined with few-shot CoT on PaLM-540B, using considerably larger sample sizes (commonly 10–40 paths) than the $k = 3$ used in the present study. Yao et al. [4] generalized CoT into tree-of-thought (ToT) prompting, in which the model generates and evaluates multiple candidate "thoughts" at each reasoning step and can

search the resulting tree using breadth- or depth-first strategies; ToT produced dramatic gains on tasks that are difficult for single-pass CoT because they require lookahead or backtracking (e.g., the Game of 24), but at substantially higher inference cost — later analyses estimate ToT can consume well over 100× the compute of a single CoT completion for comparable tasks.

The GSM8K benchmark used in this study was introduced by Cobbe et al. [5] as a set of 8.5K linguistically diverse grade-school arithmetic word problems designed to be conceptually simple for humans but challenging for language models absent explicit reasoning steps. It has since become a standard benchmark for evaluating CoT-style prompting strategies. Most published comparisons among zero-shot CoT, few-shot CoT, self-consistency, and ToT, however, are conducted on a small number of frontier proprietary models; comparatively few studies report matched, statistically tested comparisons of all four strategies on multiple small (< 10B parameter) open-weight instruction-tuned models of the kind evaluated here (Mistral-7B [6], Phi-3-mini [7], and Llama-3-8B [8]). This paper aims to fill that gap with a compact, statistically grounded comparison.

III. Methodology

A. Models

Three publicly available, instruction-tuned, open-weight LLMs in a broadly comparable size class were evaluated: Mistral-7B-Instruct-v0.2 (7B parameters) [6], Phi-3-mini-4k-instruct (3.8B parameters, 4K context) [7], and Meta-Llama-3-8B-Instruct (8B parameters) [8]. All three models were used as distributed on Hugging Face, without any additional fine-tuning.

B. Dataset

Evaluation was performed on a fixed, randomly drawn subset of 100 questions from the GSM8K test split (1,319 items total) [5]. GSM8K problems are short, natural-language grade-school arithmetic word problems (e.g., the widely cited "Janet's ducks" example) that require between two and eight elementary arithmetic operations to solve, with a single unique numeric final answer per item. The same 100 questions were used for every model–strategy combination, giving 1,200 model outputs in total (3 models × 4 strategies × 100 questions).

C. Prompting Strategies

Zero-shot CoT [2]

The question is presented with no worked examples, followed by a step-by-step reasoning trigger; a single greedy completion is generated and the final numeric answer is extracted from it.

Few-shot CoT [1]

A fixed set of manually written exemplar questions, each paired with a worked step-by-step solution and a final answer, is prepended to the target question; a single completion is generated.

Self-consistency [3]

Using the same few-shot prompt, k independent reasoning chains are sampled at a non-zero decoding temperature; the final answer is decided by majority vote across the extracted per-chain answers. Inspection of the logged generations indicates that $k = 3$ chains were sampled per question for the great majority of items (a small number of items show 2 or 4 detected segments, attributable to the text-based delimiter occasionally appearing within a chain's own reasoning content rather than to a change in the intended sample count).

Tree-of-thought [4]

The target question is decomposed into an explicit sequence of intermediate steps, and $k = 3$ candidate reasoning branches are generated and aggregated (via majority vote over final answers) to produce a single output, mirroring the branch-and-aggregate structure of self-consistency but with an explicit multi-step decomposition scaffold rather than a single unstructured chain.

The exact decoding temperature and the ToT branch-evaluation/search procedure used at generation time were governed by the accompanying evaluation script and are not re-derived from the logs here beyond the sampled-path counts noted above; readers seeking to reproduce these results exactly should consult that script for the precise temperature, top- p , and branch-selection settings used.

D. Answer Extraction and Scoring

For each generated response, a final numeric answer was extracted programmatically from the raw model output and compared against the gold GSM8K answer for that question; a response was scored as correct only on an exact numeric match. For self-consistency and tree-of-thought, the majority answer across the k sampled/branched completions was scored against the gold answer.

E. Statistical Analysis

For each of the 12 model–strategy combinations, we report overall accuracy (percentage of the 100 items scored correct) together with a 95% confidence interval computed using the normal (Wald) approximation to the binomial proportion, appropriate given the per-cell sample size of $n = 100$. Because each of the four prompting strategies was evaluated on the identical set of 100 questions within a given model, per-item outcomes are paired rather than independent across strategies. We therefore test pairwise differences in accuracy within each model using McNemar's exact test [9] on the corresponding 2×2 table of discordant outcomes, rather than a between-subjects test such as an unpaired chi-square or t -test, which would understate statistical power by ignoring the item-level pairing. Five pre-specified pairwise comparisons, targeting the three research questions above, were tested within each model: few-shot CoT vs. zero-shot CoT; self-consistency vs. zero-shot CoT; self-consistency vs. few-shot CoT;

tree-of-thought vs. few-shot CoT; and tree-of-thought vs. zero-shot CoT. A significance threshold of $\alpha = 0.05$ was used without correction for multiple comparisons; given the modest number of pre-specified, non-exploratory comparisons, we report exact p-values throughout so that readers may apply their own correction if desired.

IV. Results

Table I reports accuracy and 95% confidence intervals for each of the three models under each of the four prompting strategies. Figure 1 presents the same data graphically.

TABLE I

ACCURACY AND 95% CONFIDENCE INTERVALS BY MODEL AND PROMPTING STRATEGY ($n = 100$ PER CELL)

Model	Prompting Strategy	Accuracy (%)	95% CI
Mistral-7B-Instruct-v0.2	Zero-Shot CoT	38.0	[28, 47]
	Few-Shot CoT	50.0	[40, 60]
	Self-Consistency (k = 3)	52.0	[42, 62]
	Tree-of-Thought (k = 3)	47.0	[37, 56]
Phi-3-mini-4k-instruct	Zero-Shot CoT	46.0	[37, 56]
	Few-Shot CoT	85.0	[78, 92]
	Self-Consistency (k = 3)	86.0	[79, 92]
	Tree-of-Thought (k = 3)	54.0	[44, 64]
Llama-3-8B-Instruct	Zero-Shot CoT	58.0	[49, 67]
	Few-Shot CoT	79.0	[70, 86]
	Self-Consistency (k = 3)	78.0	[69, 86]
	Tree-of-Thought (k = 3)	52.0	[42, 63]

Zero-shot CoT is the weakest strategy in two of the three models (Mistral-7B-Instruct-v0.2 and Phi-3-mini-4k-instruct); in Llama-3-8B-Instruct, tree-of-thought is lower (52.0% vs. 58.0%). Overall, zero-shot accuracy ranged from 38.0% to 58.0%, while few-shot CoT and self-consistency are consistently the strongest and are closely matched to one another (accuracy 50.0–86.0%). Tree-of-thought (accuracy 47.0–54.0%) is closer to zero-shot CoT than to few-shot CoT in Phi-3-mini and Llama-3-8B, and lies just below few-shot CoT in Mistral-7B (47.0% vs. 50.0%). Averaged across the three models, mean accuracy was 47.3% for zero-shot CoT, 71.3% for few-shot CoT, 72.0% for self-consistency, and 51.0% for tree-of-thought; notably, tree-of-thought showed the least cross-model variability (SD = 3.6 points), while few-

shot CoT and self-consistency showed the most (SD $\approx$ 18 points), driven mainly by Phi-3-mini's disproportionately large gain from exemplars relative to the other two models.

TABLE II

MCNEMAR EXACT TEST P-VALUES FOR PAIRWISE STRATEGY COMPARISONS, BY MODEL ($p < 0.05$)*

Pairwise Comparison	Mistral-7B	Phi-3-mini	Llama-3-8B
Few-Shot CoT vs. Zero-Shot CoT	0.043*	< 0.0001*	0.0001*
Self-Consistency vs. Zero-Shot CoT	0.024*	< 0.0001*	0.0012*
Self-Consistency vs. Few-Shot CoT	0.832	1.000	1.000
Tree-of-Thought vs. Few-Shot CoT	0.720	< 0.0001*	< 0.0001*
Tree-of-Thought vs. Zero-Shot CoT	0.078	0.256	0.418

The significance tests in Table II support three consistent patterns. First, few-shot CoT and self-consistency each significantly outperform zero-shot CoT in every model ($p \leq 0.043$ in all six tests, and $p < 0.002$ in four of the six). Second, self-consistency with $k = 3$ sampled paths produces no statistically significant improvement over single-sample few-shot CoT in any model (all $p \geq 0.83$), despite requiring roughly three times the inference cost. Third, tree-of-thought is significantly worse than few-shot CoT in two of three models (Phi-3-mini and Llama-3-8B, both $p < 0.0001$) and is statistically indistinguishable from the much cheaper zero-shot CoT baseline in all three models (all $p \geq 0.078$).

V. Discussion

RQ1 – worked exemplars matter. The consistent, statistically significant gap between zero-shot and few-shot CoT (12.0–39.0 percentage points, significant in every model) confirms that, at this model scale, seeing a small number of correctly formatted, step-by-step worked examples is far more valuable than an instruction to "think step by step" alone. This is consistent with the original zero-shot CoT study's own finding that Zero-shot-CoT underperforms Few-shot-CoT [2], and suggests that for practitioners deploying 4–8B parameter open models, investing effort in a handful of well-constructed exemplars is likely to yield a larger accuracy return than more elaborate decoding-time machinery.

RQ2 – self-consistency did not pay off at $k = 3$. Despite self-consistency's strong reported gains over greedy CoT decoding at large sample sizes on frontier models [3], we found no statistically significant improvement over single-sample few-shot CoT at $k = 3$ in any of the three models tested here (all $p \geq 0.83$). A plausible explanation is that at low k , a majority vote over only three sampled chains rarely overturns the single most likely (greedy-like) answer unless the model's chain-level error rate is unusually high and its errors are unusually diverse; the self-consistency literature itself typically employs considerably larger sample sizes (10–40 paths) to realize its reported gains. Our results are consistent with self-consistency's benefit being real but sample-size-dependent, and suggest that a $k = 3$ budget

may not be sufficient to justify its added inference cost over few-shot CoT alone for this task and model class.

RQ3 – tree-of-thought did not help, and sometimes hurt, on this task. Tree-of-thought was designed for, and shows its largest reported gains on, tasks that inherently require search, lookahead, or backtracking among alternative partial solutions, such as the Game of 24 or multi-step planning puzzles [4]. GSM8K problems, by contrast, typically admit a single, short, linear chain of arithmetic operations with little benefit to exploring alternative branches. Our results – ToT significantly worse than few-shot CoT in two of three models, and statistically indistinguishable from zero-shot CoT in all three – are consistent with the view that ToT's structured decomposition and branch-aggregation machinery adds inference-time cost (multiple sampled branches per answer) without a matching accuracy benefit when the underlying task does not require search. This mirrors prior observations that ToT's overhead can be substantial – on the order of 100× the compute of a single CoT completion in some settings – an overhead that is difficult to justify without a corresponding gain.

Model-level observations. Phi-3-mini-4k-instruct, despite having the fewest parameters (3.8B) of the three models, achieved the highest accuracy under both few-shot CoT (85.0%) and self-consistency (86.0%), exceeding both Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.2. Phi-3-mini's own technical report attributes its performance largely to training-data curation and scale of instruction tuning rather than to parameter count alone [7], and our results are consistent with the interpretation that data quality and instruction-tuning recipe can matter more than raw parameter count for how well a model exploits few-shot exemplars, at least within this narrow size band. Mistral-7B-Instruct-v0.2 was the weakest model under every strategy and the only model in which no pairwise comparison reached the $p < 0.01$ level, likely reflecting both a genuinely smaller effect size and reduced statistical power at $n = 100$ for a model with more variable per-item outcomes.

VI. Limitations and Threats to Validity

Sample size and statistical power. Each model–strategy cell is based on only 100 of GSM8K's 1,319 test items. The resulting confidence intervals are wide (often ± 10 points), and several comparisons – particularly within Mistral-7B-Instruct-v0.2 – may be under-powered to detect real but smaller effects. Replication on the full test set, or a larger stratified subset, would tighten these estimates.

Sample count for self-consistency and tree-of-thought. Both stochastic strategies used $k = 3$ sampled paths/branches, well below the 10–40 typically used in the self-consistency literature. The null result for self-consistency vs. few-shot CoT (RQ2) should be read as evidence about self-consistency at this specific, modest sampling budget, not as evidence against the technique in general; a systematic sweep over k would be needed to characterize the accuracy–cost trade-off more fully.

Single domain. All evaluation was conducted on GSM8K grade-school arithmetic word problems. Because tree-of-thought is designed for tasks requiring search or backtracking, its comparatively poor

showing here may not generalize to reasoning domains with a genuinely branching solution space (e.g., planning, combinatorial puzzles, or multi-hop question answering).

Answer-extraction errors. Final answers were extracted from free-form generations using an automated procedure; correct reasoning presented in an unexpected format could in principle be mis-scored as incorrect. We did not perform manual auditing of a sample of extraction failures, which would strengthen confidence in the reported accuracies.

Deterministic vs. stochastic conditions. Zero-shot and few-shot CoT were each evaluated with a single generation per item, while self-consistency and tree-of-thought aggregate over three sampled generations; this asymmetry means the point estimates for the single-sample conditions do not reflect any run-to-run variance from decoding stochasticity, which could modestly narrow or widen the gaps reported here under repeated runs.

No cost or latency accounting. This study reports accuracy only. Self-consistency and tree-of-thought both require multiple generations per answer and are therefore several times more expensive, in tokens and latency, than a single-sample strategy; a full cost–benefit comparison would weigh these against the (here, statistically non-significant) accuracy differences observed.

Uncorrected multiple comparisons. Five pairwise tests were run per model (15 in total) at an uncorrected $\alpha = 0.05$. All findings reported as significant remain significant under a conservative Bonferroni correction ($\alpha = 0.05/15 \approx 0.0033$) except the two comparisons at $p = 0.043$ and $p = 0.024$ for Mistral-7B-Instruct-v0.2, which should therefore be treated as suggestive rather than conclusive.

VII. Conclusion and Future Work

This study compared four chain-of-thought prompting strategies — zero-shot CoT, few-shot CoT, self-consistency, and tree-of-thought — across three open-weight instruction-tuned language models on a subset of the GSM8K arithmetic reasoning benchmark, using paired significance testing rather than accuracy alone. The evidence supports three conclusions. First, worked exemplars are the primary driver of accuracy gains at this model scale: few-shot CoT significantly outperforms zero-shot CoT in every model tested. Second, self-consistency at a modest sampling budget ($k = 3$) does not provide a statistically detectable additional benefit over few-shot CoT alone, suggesting that its cost–benefit trade-off is sensitive to sample size in ways that matter for practical deployment. Third, tree-of-thought's search-based deliberation, while effective for tasks that inherently require exploration, does not improve — and in several cases significantly hurts — accuracy on short, single-path arithmetic word problems, despite its higher inference cost. For practitioners working with small open-weight models on straightforward, single-path reasoning tasks, these results suggest that investing in well-constructed few-shot exemplars offers a substantially better accuracy-per-unit-cost trade-off than either low-sample self-consistency or tree-of-thought.

Future work should (i) replicate this comparison on the full GSM8K test set to narrow the confidence intervals reported here; (ii) systematically vary the self-consistency sample size k to characterize the point at which its accuracy gains become statistically detectable for small open-weight models; (iii) extend the comparison to reasoning domains that inherently require search or backtracking, where tree-of-thought is expected to have a genuine advantage; and (iv) report token-level cost and latency alongside accuracy so that strategy choice can be grounded in an explicit accuracy-per-cost trade-off rather than accuracy alone.

Declarations

Declaration of AI Assistance

Portions of the drafting, statistical analysis narrative, and manuscript structuring for this paper were produced with the assistance of Claude (Anthropic). The author reviewed, verified, and takes full responsibility for the accuracy of all data, statistical results, citations, and conclusions presented in this manuscript.

References

1. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models, in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 24824–24837
2. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y (2022) Large language models are zero-shot reasoners, in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 22199–22213
3. Wang X, Wei J, Schuurmans D, Le Q, Chi EH, Narang S, Chowdhery A, Zhou D (2023) Self-consistency improves chain of thought reasoning in language models, in Proc. International Conference on Learning Representations (ICLR)
4. Yao S, Yu D, Zhao J, Shafran I, Griffiths TL, Cao Y, Narasimhan K (2023) Tree of thoughts: Deliberate problem solving with large language models, in Advances in Neural Information Processing Systems (NeurIPS), vol. 36, pp. 11809–11822
5. Cobbe K, Kosaraju V, Bavarian M, Chen M, Jun H, Kaiser L, Plappert M, Tworek J, Hilton J, Nakano R, Hesse C, Schulman J Training verifiers to solve math word problems, arXiv preprint arXiv:2110.14168, 2021
6. Jiang AQ, Sablayrolles A, Mensch A, Bamford C, Chaplot DS, de las Casas D, Bressand F, Lengyel G, Lample G, Saulnier L, Lavaud LR, Lachaux M-A, Stock P, Le Scao T, Lavril T, Wang T, Lacroix T and W. El Sayed, Mistral 7B, arXiv preprint arXiv:2310.06825, 2023.
7. Abdin M, Jacobs SA, Awan AA, Aneja J, Awadallah A, Awadalla H, Bach N, Bahree A, Bakhtiari A, Behl H et al Phi-3 technical report: A highly capable language model locally on your phone, arXiv preprint arXiv:2404.14219, 2024

8. AI@Meta (2024) Llama 3 model card, Meta AI, [Online]. Available: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

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

Figures

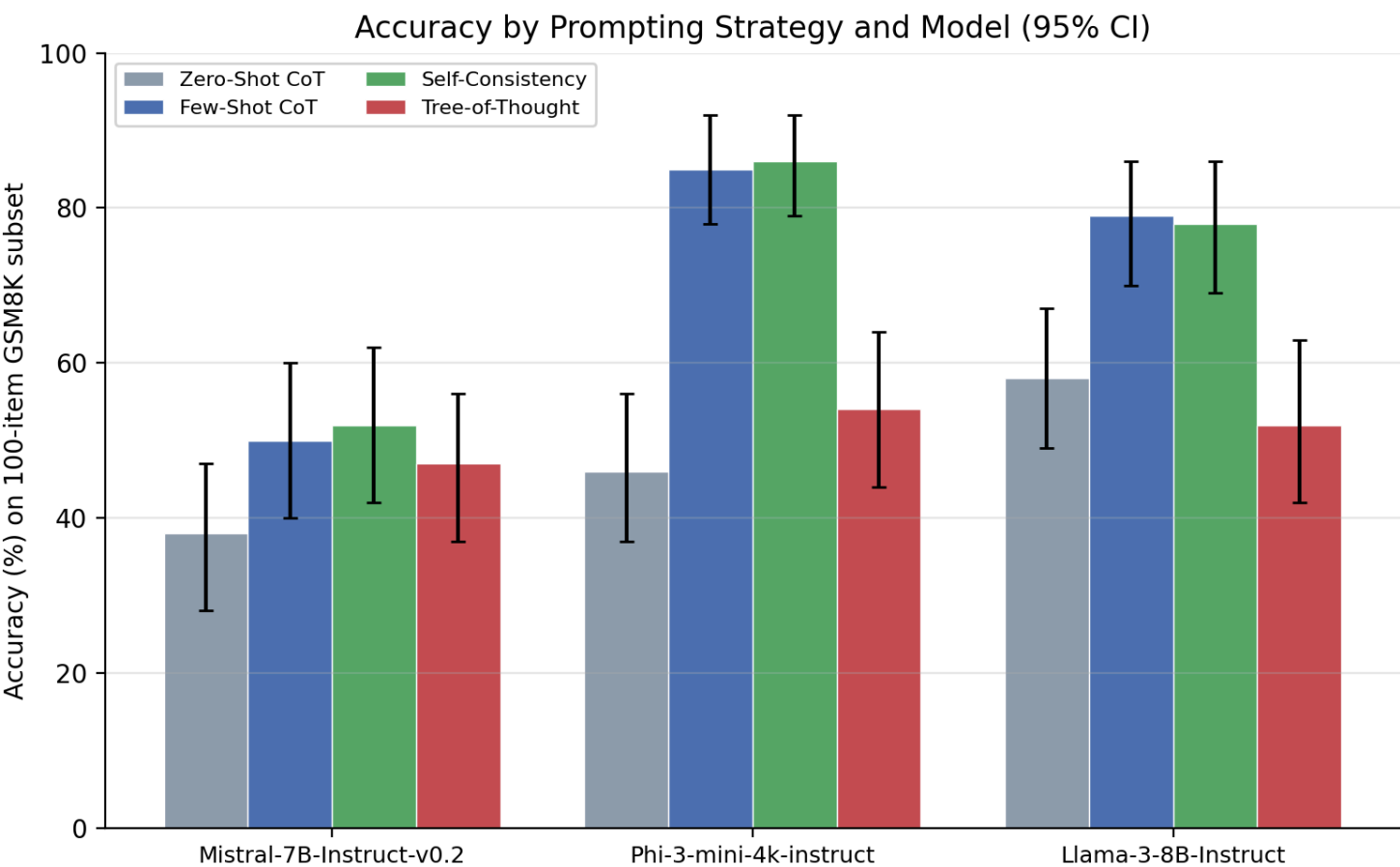


Figure 1

Accuracy (%) by prompting strategy and model, with 95% confidence intervals.