SMALL MODEL SYNDROME: TREATABLE WITH PROMPTING?

Anonymous authors
Paper under double-blind review

ABSTRACT

MedPrompt showed that compositional prompting (combining Chain-of-Thought (CoT), dynamic example retrieval, and answer-choice ensembling) can substantially improve medical question answering (MedQA) performance for large language models such as GPT-4. Whether these gains extend to smaller, open models remains unclear. We instantiate MedPrompt's core components on LLaMA-3.1-8B and observe a striking reversal: Zero-Shot CoT achieves the highest accuracy (64.03%), outperforming all few-shot and retrieval-based variants. Example-based prompting consistently degrades performance, yielding a U-shaped accuracy curve with 0-shot $>$ 5-shot $>$ 1-shot. These findings suggest that small models benefit from explicit reasoning instructions but are highly sensitive to noise from in-context examples. Our study offers a small-model perspective on prompt engineering and clarifies which MedPrompt techniques transfer, and which may backfire, when large models are infeasible.

1 INTRODUCTION

Large language models (LLMs) can reach strong performance on specialized domains such as medicine using only prompt engineering, without task-specific fine-tuning. The MedPrompt framework of Nori et al. (2023) showed that compositional prompting (combining Chain-of-Thought (CoT) reasoning, dynamic retrieval of in-context examples, and answer-choice ensembling) can steer generalist models like GPT-4 to state-of-the-art accuracy on medical benchmarks such as MedQA. This suggests that *compositional prompting* may serve as a powerful alternative to domain-specific training in high-stakes medical question answering (QA).

MedPrompt, however, is evaluated only on very large proprietary models (GPT-3.5 and GPT-4). At the same time, prior work on in-context learning (ICL) and CoT shows that their benefits are strongly scale-dependent: Brown et al. Brown et al. (2020) observe that few-shot performance improves dramatically as model size increases, while Wei et al. Wei et al. (2023) find that CoT yields large gains only for $\sim$100B-parameter models and can even hurt smaller ones. This raises a natural question that MedPrompt leaves open: *do these prompting methods transfer to small, open-source models that fit realistic compute budgets?*

This question matters in practice. Many deployments (e.g., teaching hospitals or low-resource educational settings) cannot afford GPT-4-scale inference and instead rely on 7–8B-parameter open models that fit on a single GPU. Understanding which MedPrompt components help, and which may backfire, is crucial for building reliable systems under such constraints.

We revisit MedPrompt from the perspective of an 8B-parameter model, LLaMA-3.1-8B, on the MedQA benchmark. We construct a "tiny reproduction" by instantiating MedPrompt's core components (few-shot prompting, CoT, dynamic example retrieval, and answer-choice manipulation) and evaluating their impact on MedQA accuracy. We find a striking reversal: Zero-Shot CoT is the best configuration, achieving 64.03% accuracy and outperforming all few-shot and retrieval-based variants. Naive five-shot prompting without CoT underperforms even a zero-shot baseline, while five-shot CoT (random or retrieved examples) improves over vanilla five-shot but never reaches Zero-Shot CoT. On MedQA, Nori et al. (2023) report that a prompt-engineered GPT-3.5 model reaches 60.2% accuracy, while the fine-tuned specialist model Med-PaLM reaches 67.2%; our best configuration, Zero-Shot CoT with LLaMA-3.1-8B, lies between these two systems.

Our experiments reveal a U-shaped pattern in accuracy as examples are added: zero-shot $>$ five-shot $>$ one-shot. These observations suggest that small models benefit from explicit reasoning instructions but are highly sensitive to noise introduced by in-context examples, which they cannot robustly ignore.

Contributions.

- We present a systematic reproduction of MedPrompt-style compositional prompting on a small (8B-parameter) open-source model for MedQA.
- We show that Zero-Shot CoT achieves the highest accuracy (64.03%) among all tested configurations, and that adding in-context examples induces a U-shaped accuracy curve in which example-based prompting often *hurts* performance relative to zero-shot prompting.

2 Background and Related Work

MedPrompt and medical QA. Prompt engineering has become a standard way to adapt large foundation models to new domains without gradient-based fine-tuning. The GPT-4 technical report documents broad capabilities across reasoning, coding, and domain-specific QA. Building on this, Nori et al. Nori et al. (2023) propose MedPrompt, a compositional prompting framework combining (1) CoT prompting, (2) dynamic retrieval of in-context examples, and (3) answer-choice ensembling. On MedQA, they report that a prompt-engineered GPT-3.5 model reaches 60.2% accuracy, a specialist fine-tuned model (Med-PaLM) reaches 67.2%, and GPT-4 with MedPrompt surpasses these baselines.

Scale dependence of in-context learning and CoT. MedPrompt builds on in-context learning and few-shot prompting, which are known to be strongly scale-dependent. Brown et al. Brown et al. (2020) show that few-shot performance improves dramatically as models scale to GPT-3 (175B parameters), and that larger models make more effective use of in-context examples. Wei et al. Wei et al. (2023) demonstrate that CoT prompting yields large gains on reasoning benchmarks only for larger models (like with $\sim$100B parameters), and can hurt smaller models, which often produce fluent but logically inconsistent chains of thought. These results motivate MedPrompt's focus on large models but leave open whether similar gains can be realized for smaller open models in the 5 to 20B parameter range. Our work targets precisely this gap.

3 Methodology

3.1 Task and dataset

We evaluate all prompting strategies on the MedQA benchmark, a multiple-choice question answering dataset constructed from United States medical licensing exam-style questions. Each question provides a stem and several candidate answers, with exactly one correct option. Following the MedPrompt setup, we treat MedQA as a pure inference task: the underlying language model is not fine-tuned on MedQA; instead, we only modify the prompts used at inference time.

We report accuracy as the primary metric: the proportion of questions for which the model's selected choice matches the ground-truth answer. For conditions involving ensembling (e.g., answer-choice shuffling), we compute the majority vote over multiple sampled model outputs and treat the ensembled prediction as the answer.

3.2 Base model

Our base model is LLaMA-3.1-8B, an 8B-parameter open-source LLM. We access the model via Purdue's RCAI API. To ensure fair comparison across prompting strategies, we only vary the input prompts across experiments. We apply a simple answer-extraction heuristic that parses the model's final answer choice from its response, ignoring any intermediate reasoning text when CoT is used. However, we save all reasoning responses in full for manual analysis.

Table 1: Accuracy under different prompting conditions on MedQA with LLaMA-3.1-8B.

Condition	Accuracy (%)	Notes
Zero-Shot	47.39	No reasoning, no examples
5-Shot Random	41.66	Standard few-shot, no CoT
Zero-Shot CoT	**64.03**	CoT, no examples
5-Shot Random + CoT	60.73	CoT with random examples
5-Shot Retrieval + CoT	61.30	CoT with retrieved examples
1-Shot Retrieval + CoT	57.22	CoT with single retrieved example
Zero-Shot CoT + Choice Shuffle	64.80	5× ensemble, 500-question subset on which Zero-Shot CoT had 62.80%

3.3 Prompting Conditions

We evaluate seven conditions mirroring MedPrompt's ablation structure:

- **Zero-Shot.** Question and options only, with instructions to select the best answer. No examples or reasoning prompts.
- **5-Shot Random.** Five randomly sampled question–answer pairs (without reasoning) precede the test question.
- **Zero-Shot CoT.** Zero-shot setting with Chain-of-Thought instruction: "Let's think step by step. First, reason about the question, then state the single best answer choice."
- **5-Shot Random + CoT.** Five random examples, each with reasoning chains explaining the correct answer, followed by CoT instruction for the test question.
- **5-Shot Retrieval + CoT.** Five nearest-neighbor examples (selected via embedding-based retrieval) with reasoning chains, followed by CoT instruction. Mirrors MedPrompt's dynamic retrieval.
- **1-Shot Retrieval + CoT.** Single nearest-neighbor example with reasoning, followed by CoT instruction. Isolates the effect of one highly similar example.
- **Zero-Shot CoT + Choice Shuffle.** On 500 questions, Zero-Shot CoT is run five times with permuted answer orders; predictions aggregated via majority vote.

3.4 Implementation details

All conditions share the same MedQA test split and the same API-based inference setup. For retrieval-based conditions, we embed questions into a shared vector space and use nearest-neighbor search to select examples; we do not use ground-truth answers or future information in the retrieval process. To reduce variance, we use a consistent random seed for sampling random examples and, where applicable, sample a single response per query per prompt configuration (except for the choice-shuffle ensemble, which explicitly uses multiple runs).

In the choice shuffle experiment, to mitigate position bias, we applied choice shuffle ensembling by generating five random permutations of the answer options for each question and querying the model using zero-shot Chain-of-Thought prompting. This is how they did it in the MedPrompt paper as well. The final prediction was determined via majority voting, where the generated answer labels were mapped back to their original semantic indices to ensure consistency across permutations.

4 Experiments

4.1 Overview of prompting conditions

Table 1 summarizes MedQA accuracy for each prompting condition.

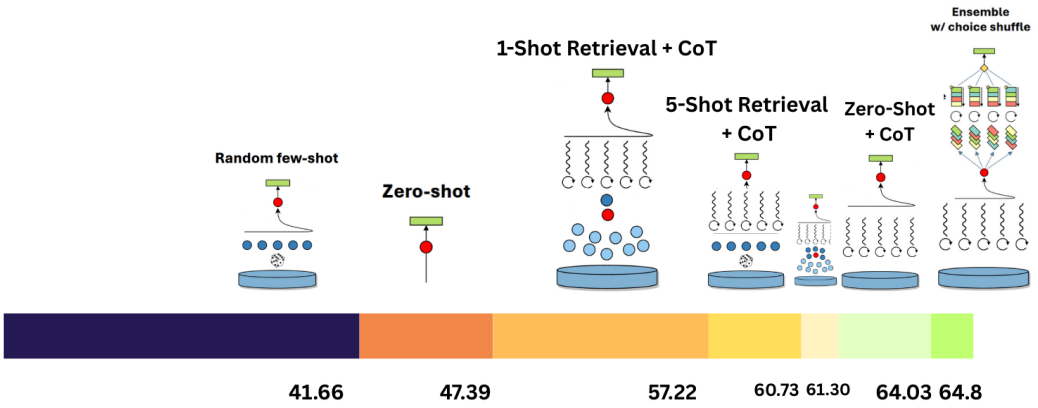

Figure 1: Ablation of MedPrompt-style components on MedQA with LLaMA-3.1-8B, visualized component-wise graphically like done in the MedPrompt paper.

4.2 Ablation Results

Zero-Shot CoT is the strongest configuration, outpacing all few-shot and retrieval-based variants. Naive five-shot prompting without CoT significantly *hurts* performance, dropping from 47.39% (Zero-Shot) to 41.66%. Adding CoT on top of five-shot prompts recovers much of this loss (60.73–61.30%) but still fails to match the simpler Zero-Shot CoT baseline. Retrieval-based five-shot CoT offers only a small improvement over random five-shot CoT (61.30% vs. 60.73%). The choice-shuffle ensemble on top of Zero-Shot CoT yields a modest gain (64.80% vs. 62.8% on the same 500-question subset), indicating some sensitivity to answer ordering but limited headroom from ensembling.

4.3 Effect of CoT and in-context examples

Comparing Zero-Shot to Zero-Shot CoT, adding a reasoning instruction improves accuracy by 16.64 percentage points (47.39% to 64.03%). This shows that even for an 8B-parameter model, instruction-level CoT can substantially improve performance on a complex domain-specific exam, likely by encouraging more structured reasoning and systematic elimination of distractors.

In contrast, in-context examples alone are harmful: 5-Shot Random underperforms Zero-Shot by 5.73 points. This supports the view that small models may not reliably perform meta-learning over a handful of examples and can be distracted by irrelevant context. When CoT is added, five-shot performance improves substantially but still lags behind Zero-Shot CoT, suggesting that examples provide, at best, marginal value once explicit reasoning instructions are present. The 1-Shot Retrieval + CoT condition sits between 5-Shot CoT and the non-CoT few-shot condition, again failing to match Zero-Shot CoT and indicating that even semantically similar examples can be misleading for small models.

4.4 Effect of answer-choice ensembling

The Zero-Shot CoT + Choice Shuffle ensemble improves accuracy from 62.8% to 64.80% on the 500-question subset, showing that some errors are sensitive to the ordering of answer choices. However, the improvement is small, and the additional inference cost may be difficult to justify in resource-constrained settings. MedPrompt also showed about a 2% gain from this so our results seem to follow the same pattern here.

4.5 WHY DOES ZERO-SHOT COT DOMINATE?

Zero-Shot CoT provides a clean reasoning scaffold that focuses the model on the current question. In contrast, in-context examples force the model to integrate additional context that may confuse its reasoning. If the model's meta-learning is limited, examples can act as noise. The fact that five-shot CoT configurations approach but do not surpass Zero-Shot CoT is consistent with a slightly negative net effect of examples: CoT helps, but not enough to offset exemplar noise.

4.6 A U-SHAPED RESPONSE TO EXAMPLES

We hypothesize that the U-shape stems from the model's ability to balance between clean question-only inputs (Zero-Shot) and the complexity introduced by examples. One-shot retrieval places too much weight on a single, potentially misleading example. With five-shot, the model sees more diverse patterns, but still struggles with the inconsistent quality of examples. Zero-Shot CoT is the best-performing configuration because it minimizes distractions and focuses on reasoning alone.

4.7 SMALL-MODEL PERSPECTIVE ON MEDPROMPT

Our findings suggest that, for small models like LLaMA-3.1-8B, CoT instructions provide the bulk of the performance gains. In contrast, examples and ensembling provide little additional value and can sometimes hurt performance. This implies that the runtime strategies that work well for very large proprietary models may not transfer straightforwardly to smaller, resource-constrained models. For deployment in such settings, practitioners should prioritize a strong Zero-Shot CoT baseline and only add complexity with additional components when they show consistent empirical improvements.

4.8 BRITTLE REASONING AND EXAMPLE NOISE

Smaller models are more sensitive to noisy or irrelevant examples. Zero-Shot CoT minimizes this risk by limiting input to the current question and answer choices, allowing the model to rely on its internalized reasoning abilities. This simpler setup appears to be more reliable than complex examples, especially for models with limited capacity.

5 LIMITATIONS AND FUTURE WORK

Our study has several limitations. Firstly, our choice-shuffle ensembling analysis is done on a much smaller subset (500 questions only[1]) which may not be fully representative of the effect of ensembling.

Secondly, we evaluate only a single small model (LLaMA-3.1-8B) on a single benchmark (MedQA). While MedQA is a challenging and widely used exam-style dataset, it represents only one facet of medical reasoning. Future work should test whether our observations generalize across other datasets in the MultiMedQA suite and to other specialized domains where MedPrompt has been applied (e.g., law, engineering).

Finally, our implementation relies on a specific set of prompt templates and reasoning instructions. While we attempted to keep these templates reasonable and stable across experiments, different prompt designs might lead to different conclusions. A more exhaustive search over prompt space, akin to the prompt optimization conducted in MedPrompt itself, could provide a more complete picture of what is achievable with an 8B model.

6 CONCLUSION

We set out to test whether MedPrompt-style compositional prompting, which dramatically improves GPT-4's performance on MedQA, could also benefit a smaller open-source model, LLaMA-3.1-8B. Our findings succeed in reproducing MedPrompt's core idea: prompting strategies can improve generalist model performance on specialist datasets without any fine-tuning.

[1] All other results have been derived from the full dataset of more than 14000 questions

We find that zero-Shot CoT is the best-performing configuration (64.03%), outperforming all few-shot and retrieval-based variants. Notably, our small generalist model with effective prompting not only outperforms GPT-3.5, a much larger model, but also gets surprisingly close to Med-PaLM, a specialized and fine-tuned system. This demonstrates that with the right prompting, smaller models can achieve significant performance gains, even on specialized, domain-specific tasks.

These results provide a new perspective on prompt engineering for small models, showing that explicit reasoning instructions (Zero-Shot CoT) are highly beneficial, while in-context examples and answer-choice ensembling add little value and can sometimes hurt performance. This suggests that while MedPrompt's techniques work well for very large models, they may not transfer as effectively to smaller, resource-constrained models.

For practitioners, the key takeaway is clear: when working with small open models in domains like medical QA, it's crucial to start with a strong Zero-Shot CoT baseline. Additional compositional prompting components, such as in-context examples and ensembling, should only be adopted if they consistently improve performance. More broadly, our work underscores the importance of evaluating prompting strategies across a range of model scales, highlighting that methods validated for GPT-4 may not automatically generalize to smaller models.

REFERENCES

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Renqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. Can generalist foundation models outcompete special-purpose tuning? case study in medicine, 2023. URL https://arxiv.org/abs/2311.16452.

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.