When Chain-of-Thought Backfires: Evaluating Prompt Sensitivity in Medical Language Models

Binesh Sadanandan[1] and Vahid Behzadan[1]

SAIL Lab, University of New Haven, West Haven, CT 06516, USA
bsada1@unh.newhaven.edu, vbehzadan@newhaven.edu

Abstract. Large Language Models (LLMs) are increasingly deployed in medical settings, yet their sensitivity to prompt formatting remains poorly characterized. We evaluate MedGemma (4B and 27B parameters) on MedMCQA (4,183 questions) and PubMedQA (1,000 questions) across a broad suite of robustness tests. Our experiments reveal several concerning findings. Chain-of-Thought (CoT) prompting *decreases* accuracy by 5.7% compared to direct answering. Few-shot examples degrade performance by 11.9% while increasing position bias from 0.14 to 0.47. Shuffling answer options causes the model to change predictions 59.1% of the time, with accuracy dropping up to 27.4 percentage points. Front-truncating context to 50% causes accuracy to plummet below the no-context baseline, yet back-truncation preserves 97% of full-context accuracy. We further show that cloze scoring (selecting the highest log-probability option token) achieves 51.8% (4B) and 64.5% (27B), surpassing all prompting strategies and revealing that models "know" more than their generated text shows. Permutation voting recovers 4 percentage points over single-ordering inference. These results demonstrate that prompt engineering techniques validated on general-purpose models do not transfer to domain-specific medical LLMs, and that reliable alternatives exist.

Keywords: Medical question answering · Prompt sensitivity · Language model robustness · Chain-of-thought · Position bias · Cloze scoring

1 Introduction

LLMs have achieved impressive performance on medical licensing exams, with GPT-4 exceeding passing thresholds by over 20 points [9] and Med-PaLM 2 reaching 86.5% on MedQA [13]. These results have fueled enthusiasm for deploying LLMs in clinical decision support. But benchmark accuracy tells only part of the story. How models respond to variations in prompt format, question ordering, and context presentation remains poorly understood, despite being critical for real-world deployment where inputs are rarely formatted identically to benchmark conditions.

We focus on MedGemma [3], Google's medical-specialist LLM built on the Gemma architecture. A widely held belief in the LLM community is that certain prompting strategies reliably improve performance. CoT prompting, which instructs models to reason step-by-step before answering, has shown consistent gains on mathematical and logical reasoning tasks [17]. Few-shot learning, where examples are provided in-context, helps models understand desired output formats [2]. These techniques are often treated as "best practices" that should transfer across domains and model families.

But should they? Domain-specific models may have internalized different patterns during training. A model trained extensively on medical literature might already encode structured clinical reasoning, making explicit CoT prompts redundant or even counterproductive. Similarly, few-shot examples drawn from one medical specialty might prime the model with concepts that are irrelevant or misleading for questions in other specialties.

We present a systematic evaluation of MedGemma's sensitivity to prompt variations across four experimental axes. First, we conduct a prompt ablation study comparing ten prompting strategies on 4,183 MedMCQA questions, measuring both accuracy and position bias. Second, we test option order sensitivity by shuffling answer choices across multiple random seeds and measuring how often the model changes its prediction. Third, we evaluate evidence conditioning on 1,000 PubMedQA questions across eleven context conditions, including five truncation strategies that reveal which parts of the abstract carry diagnostic value. Fourth, we benchmark robustness-oriented alternatives: cloze scoring, permutation voting, and CoT self-consistency. Our findings challenge conventional assumptions about prompt engineering in medical AI and identify practical mitigations for deployment.

2 Related Work

Medical Language Models and Benchmarks. Medical language models often demonstrate impressive benchmark scores on exam-style question answering. GPT-4 performs strongly on medical challenge sets [9], and Med-PaLM and Med-PaLM 2 report high scores on MedQA [12, 13]. Open, domain-tuned models have also emerged, including MedGemma [3] and BioMistral [5]. However, most reporting still emphasizes a single headline accuracy, while deployment inputs vary in formatting, context quality, and answer presentation.

Prompting for Reasoning. CoT prompting can improve performance on general reasoning tasks by eliciting intermediate steps [17]. Follow-up work such as self-consistency explores sampling multiple reasoning paths and aggregating answers [16]. These techniques are now common defaults, despite their added token cost and their sensitivity to output parsing.

When CoT Backfires. Recent work challenges the idea that CoT helps everywhere. Sprague et al. [15] identify settings where step-by-step prompting reduces

accuracy, connecting these failures to cases where deliberate reasoning hurts humans as well. Meincke et al. argue that the gains from CoT have diminished for newer models and can even reverse on some tasks [8]. In medical question answering, Omar et al. compare multiple CoT-style prompts and find that improvements depend on the specific method and dataset, rather than following a simple monotonic trend [10].

Prompt Sensitivity and Few-shot Formatting. Even without CoT, small prompt changes can shift performance. Lu et al. [7] show that few-shot example order can materially affect accuracy, and Zhao et al. [20] propose calibration methods that reduce sensitivity to label and prompt priors. ProSA provides a more systematic view by measuring how model outputs vary across prompt templates [22]. Together, this work suggests that comparisons between models can be misleading if prompt choices are not controlled.

Multiple-choice Artifacts. Multiple-choice evaluation introduces its own failure modes. Zheng et al. [21] show that LLMs exhibit selection bias, preferring certain option identifiers even when content is balanced. Our option reordering experiments build on this line of work by separating changes in answer position from changes in distractor content.

Retrieval and Context Quality. Retrieval-Augmented Generation (RAG) augments a model with external documents at inference time. In medicine, benchmarking work finds large variation in RAG performance across retrievers and corpora, with a pronounced "lost-in-the-middle" effect in biomedical settings [19, 6]. Retrieval can also introduce new failure modes: ClashEval shows that models can be led astray by incorrect retrieved context [18], and recent medical RAG methods aim to improve reliability under imperfect retrieval [14, 1].

3 Methods

3.1 Models and Datasets

We evaluate MedGemma-4B, the 4-billion parameter instruction-tuned variant at bfloat16 precision, and MedGemma-27B, the 27-billion parameter model at full bfloat16 precision on 80 GB A100 GPUs.

We use two standard medical QA benchmarks. MedMCQA [11] contains questions from Indian medical entrance examinations across 21 subjects; we use the 4,183-question validation split. PubMedQA [4] contains research questions derived from PubMed titles that must be answered using abstracts; we use the 1,000-question labeled subset.

3.2 Inference Protocol

All experiments use a frozen inference protocol for reproducibility. Table 1 summarizes the two configurations.

Table 1. Frozen inference parameters for all experiments.

Parameter	Determ.	Self-Cons.	Rationale
Temperature	0.0	0.7	Greedy for reproducibility; 0.7 for diverse reasoning paths
Top-p	1.0	0.95	No nucleus filtering for greedy; mild filtering for sampling
Top-k	0	50	Disabled for greedy; limits sampling to top 50 tokens
`do_sample`	False	True	Greedy decoding vs. stochastic sampling
`max_new_tokens`	256	256	Sufficient for CoT reasoning while bounding cost

Table 2. Overview of the four experiments in our evaluation.

Exp.	Name	Dataset	Conds.	Core Question
1	Prompt Ablation	MedMCQA (4,183)	10	Do CoT and few-shot help or hurt?
2	Option Order	MedMCQA (4,183)	5×3 seeds	Does shuffling options change the answer?
3	Evidence Cond.	PubMedQA (1,000)	11	How does context truncation affect accuracy?
4	Robustness Basel.	MedMCQA (4,183)	3 methods	Can alternative scoring mitigate fragility?

All runs use the default system prompt and constrain outputs to the valid answer set (A through D for MedMCQA; yes/no/maybe for PubMedQA). We extract answers via regex; unparseable outputs are scored as incorrect. Where randomized perturbations are involved, we report results over multiple seeds (42, 123, 456) with mean and 95% confidence intervals.

3.3 Experimental Overview

Table 2 summarizes the four experiments, their datasets, condition counts, and the core question each one addresses. Detailed descriptions follow below.

3.4 Experimental Conditions

Experiment 1: Prompt Ablation. We test ten prompting strategies on MedMCQA. Five *core* conditions vary the prompting approach: (1) zero-shot direct; (2) zero-shot CoT; (3) few-shot direct with three curated examples; (4) few-shot CoT with three curated examples; and (5) answer-only (minimal prompt).

Five additional conditions probe few-shot sensitivity: (6) random selection from the training set; (7) label-balanced selection (equal A/B/C/D representation); (8) subject-matched selection (same medical specialty as the test item); and (9–10) two alternative orderings of the curated examples to test order effects.

Experiment 2: Option Order Sensitivity. We apply four transformations to each MedMCQA question: random shuffle, rotate-1 (cyclic shift by one), rotate-2 (shift by two), and distractor swap (exchange incorrect options while preserving correct answer position). We measure the flip rate, defined as the proportion of questions where the model changes its answer when options are reordered. To quantify variance, we repeat the experiment across three random seeds and report mean accuracy and flip rate with 95% confidence intervals.

Experiment 3: Evidence Conditioning. On PubMedQA, we vary context presentation across eleven conditions. Six *core* conditions test context completeness: question-only (no context), full abstract, front-truncated 50%, front-truncated 25%, background-only (introduction sentences), and results-only (conclusion sentences). Five additional conditions test truncation strategy: back-truncated 50% (keeping the final half), middle-truncated 50% (keeping the central portion), sentence-boundary truncation at 50%, and salient-sentence extraction retaining the top-5 and top-3 sentences ranked by non-stopword overlap with the question.

Experiment 4: Robustness Baselines. We evaluate three complementary scoring and aggregation methods on MedMCQA. *Cloze scoring* bypasses free-text generation entirely: for each question we compute log-probabilities of the four option tokens (A through D) at the next-token position and select the highest. *Permutation voting* runs greedy inference on multiple option orderings and takes a majority vote over inverse-mapped predictions. *CoT self-consistency* samples N CoT responses at temperature 0.7 and takes a majority vote, testing whether sampling diversity can overcome single-sample fragility.

3.5 Metrics

We report accuracy with 95% Wilson score confidence intervals. Position bias is computed as a prediction frequency bias: Bias $= \frac{1}{4}\sum_{i\in\{A,B,C,D\}} |P_{\text{pred}}(i) - P_{\text{truth}}(i)|$, where $P_{\text{pred}}(i)$ is the fraction of model predictions for position i and $P_{\text{truth}}(i)$ is the fraction of correct answers at position i. A score of 0 indicates the model's answer distribution perfectly matches the ground truth; higher values indicate systematic over- or under-selection of certain positions. For option-order experiments, we compute the flip rate: the proportion of questions where the model's prediction changes under reordering. Table 3 describes our additional metrics.

Table 3. Additional metrics used in our evaluation.

Metric	Definition
CKLD	$\sum_i (O_i - E_i)^2 / E_i$, where O_i and E_i are observed and expected counts per position; quantifies departure from uniform answering
RStd	σ/μ of per-position accuracies; captures how unevenly performance is distributed across answer positions
Per-position acc.	Accuracy conditioned on the correct answer being A, B, C, or D; reveals position-dependent competence
McNemar's test	Paired significance testing between conditions, with Bonferroni correction for multiple comparisons

Table 4. Prompt ablation results on MedMCQA (n=4,183). Random baseline is 25%.

Condition	Accuracy	95% CI	Pos. Bias
Zero-shot direct	47.6%	[46.1, 49.1]	0.137
Zero-shot CoT	41.9%	[40.4, 43.3]	0.275
Few-shot direct (curated)	35.7%	[34.3, 37.0]	0.472
Few-shot CoT (curated)	40.8%	[39.4, 42.3]	0.413
Answer-only	43.0%	[41.5, 44.6]	0.096

4 Results

4.1 Prompt Ablation

Table 4 shows accuracy across prompting strategies for the five core conditions. Zero-shot direct achieves the highest accuracy at 47.6%, while CoT *reduces* accuracy by 5.7 percentage points (McNemar's test, $p < 0.001$). Few-shot examples cause an even larger degradation of 11.9% ($p < 0.001$), while simultaneously increasing position bias from 0.137 to 0.472 (measured as prediction frequency bias, i.e., the mean absolute deviation between predicted and ground-truth answer distributions; see Section 3.5). All pairwise differences between zero-shot direct and other conditions remain statistically significant after Bonferroni correction ($p < 0.001$, McNemar's test). The model picks up spurious patterns from in-context examples rather than useful output formats.

4.2 Qualitative Examples

Table 5 shows representative model outputs that illustrate our key findings. The first example shows CoT backfiring: direct prompting answers "D" (Candidiasis) correctly, but CoT prompting starts analyzing each option, notes that tuberculosis is "rare in children," and yet ultimately selects A (Tuberculosis). The second example shows an option-order flip: the model answers "B" (Anterior ethmoidal

artery) under the original ordering, correctly identifying a non-branch of the external carotid. When options are shuffled, the content previously at position B moves to position A, but the model again answers "B," now pointing to a different option entirely. It follows position, not content.

4.3 Option Order Sensitivity

Table 6 reveals extreme sensitivity to option ordering. The mean flip rate is 59.1%; the model changes its answer more often than not when options are shuffled. Rotation perturbations cause the largest accuracy drops (up to 27.4%), while distractor swaps show smaller impact (−8.9%). This confirms that position rather than distractor content drives the fragility.

4.4 Evidence Conditioning

Table 7 shows that context presentation substantially affects PubMedQA performance. Most critically, front-truncated context performs *worse* than no context: front-truncation at 50% yields 13.8% (4B) and 23.4% (27B), far below question-only baselines of 34.5% and 31.0%. Naively truncating from the front actively misleads the model.

Our expanded truncation analysis reveals that *how* context is truncated matters as much as *how much* is removed. Back-truncation at 50% preserves 97% of full-context accuracy (44.5% vs. 45.8%), while front-truncation at the same ratio destroys it (13.8%). The model relies on opening context for orientation; removing the beginning causes catastrophic failure, while removing the end is largely harmless.

Salient-sentence extraction offers a practical middle ground: selecting the top-5 sentences by question overlap achieves 40.1%, retaining 88% of full-context accuracy with roughly half the tokens.

4.5 Robustness Baselines

Table 8 compares three robustness-oriented methods against zero-shot direct prompting (47.6%).

Cloze scoring achieves 51.8% [50.3, 53.3] on 4B and 64.5% [61.4, 67.1] on 27B, the highest accuracy of any method we tested for both model sizes. The 27B result is striking: cloze scoring recovers 26.3 points over 27B's full-context evidence conditioning (38.2%) and 17.0 points over 4B's best prompting result (47.6%). Position bias scores are 0.013 (4B) and 0.054 (27B), an order of magnitude lower than zero-shot direct (0.137).

Permutation voting (4 orderings, n=1,000) achieves 49.0% [46.0, 52.1] aggregated accuracy, a 4-point improvement over the mean per-permutation accuracy of 45.1%. The agreement rate across orderings is 70.0%, with disagreement identifying items where position bias drives predictions.

The cloze result deserves attention. By bypassing generation entirely and reading the model's internal token preferences, we recover 4.2 points of accuracy

that are lost to the generation and parsing pipeline. This suggests that a significant fraction of MedGemma's apparent "errors" under standard prompting are not errors of knowledge but of output formatting.

5 Discussion

5.1 Why Chain-of-Thought Hurts

The 5.7% accuracy drop from CoT prompting aligns with recent findings that deliberation can reduce performance on certain tasks [15]. MedGemma was trained extensively on medical text and may have internalized domain reasoning patterns. Forcing explicit step-by-step logic may override these learned patterns with less reliable deliberation.

Case-level analysis reveals the mechanism. CoT prompting changed answers on 1,262 of 4,183 questions, hurting 750 (direct correct, CoT wrong) while helping only 512, a net loss of 238 questions. The predominant failure pattern involves verbose reasoning (90.7% exceeded 500 characters) where longer chains create opportunities for errors to compound. We also observed self-contradiction in 25.6% of cases and confident wrong conclusions in 11.1%.

The characteristic failure mode works like this: the model correctly identifies relevant medical concepts early in its reasoning, considers alternatives, then talks itself into the wrong answer. In one case involving organophosphate poisoning, CoT correctly identified the condition and atropine's role as antidote, then continued deliberating and selected neostigmine, which would worsen the condition.

5.2 The 59% Flip Rate Problem

MedGemma changes its answer 59.1% of the time when options are shuffled, far exceeding random noise. The maximum flip rate of 72.9% for rotation perturbations means that for nearly three-quarters of questions, answers depend more on option position than content. This magnitude exceeds typical findings [21], suggesting medical-specialist training may not mitigate position bias and could even exacerbate it.

For clinical applications, this fragility is unacceptable. A diagnostic support system that changes recommendations based on option ordering provides no reliable signal to clinicians.

5.3 Truncation Direction Matters More Than Amount

Front-truncating context to 50% yields 13.8% accuracy while no context achieves 34.5%, a 20.7-point gap showing that partial context actively misleads. But our expanded analysis reveals this is specifically a *front-truncation* problem: back-truncation at the same 50% ratio preserves 44.5% accuracy, just 1.3 points below full context. The model relies on opening context for orientation; once it has read the beginning of an abstract, losing the tail is relatively harmless.

This asymmetry has direct implications for RAG systems in medicine. Prior work shows that models can be led astray by incorrect retrieved evidence [18, 1]. Our results add specificity: the *position* of missing information matters as much as its quantity. RAG systems that truncate passages to fit context windows should truncate from the end, not the beginning. Better still, salient-sentence extraction (40.1% with top-5 sentences) provides a principled compression that retains 88% of full-context accuracy at roughly half the token cost.

Results-only context (41.9% for 4B, 40.0% for 27B) nearly matches or exceeds full context, while background-only achieves just 28.9%/19.8%. Both models benefit from conclusions rather than methodological background, suggesting RAG systems should prioritize high-information-density content over exhaustive retrieval.

5.4 Scale and Robustness

MedGemma-27B underperforms 4B on generative evidence conditioning (38.2% vs 45.8% with full context), apparently demonstrating that medical benchmark performance does not scale with model size. But our cloze scoring results invert this conclusion: 27B achieves 64.5% via log-probabilities, far exceeding 4B's 51.8%. The 27B model *does* encode substantially more medical knowledge than 4B; its generation pipeline simply fails to express it. This finding reframes the "scale doesn't help" narrative. Scale helps knowledge acquisition, but the generation-parsing pipeline becomes the bottleneck at larger sizes.

On evidence conditioning, 27B shows a "less is more" pattern: results-only context (40.0%) exceeds full-context accuracy (38.2%), suggesting larger models may be more susceptible to distraction from verbose context but respond well to concentrated information.

5.5 Cloze Scoring as a Diagnostic

The advantage of cloze scoring over zero-shot direct prompting, 4.2 points for 4B (51.8% vs. 47.6%), is striking because both methods use the same model and the same questions; the only difference is *how* the answer is extracted. The effect is even more dramatic for 27B: cloze scoring achieves 64.5%, making 27B the best-performing configuration in our entire study when answers are extracted via log-probabilities rather than parsed from generated text.

Two mechanisms likely contribute to this gap: (1) the generation pipeline adds noise through token-by-token autoregressive decoding and output formatting constraints; and (2) the answer parser fails on some responses that contain correct reasoning but not a cleanly extractable letter. The near-zero position bias (0.013 for 4B, 0.054 for 27B vs. 0.137 for zero-shot direct) supports this interpretation; the models' internal preferences are far more uniform across positions than their generated text.

This has practical implications. For applications where only the answer matters (not the reasoning), cloze scoring offers a strictly better extraction method.

For applications requiring explainability, the cloze-generation gap can serve as a diagnostic of how much accuracy is lost to the generate-parse pipeline.

5.6 Permutation Voting Partially Mitigates Position Bias

Permutation voting improves aggregated accuracy by 4 points over the mean single-ordering accuracy (49.0% vs. 45.1%), confirming that majority-vote aggregation across orderings can partially cancel position-dependent errors. The 70% agreement rate identifies a natural partition: questions with high agreement are reliably answered regardless of ordering, while low-agreement questions flag items where the model's answer depends on superficial cues. In a clinical deployment, the agreement rate could serve as a calibrated abstention signal.

5.7 Generalizability and Clinical Safety

Our experiments focus on the MedGemma family. Whether these "backfire" effects generalize to other domain-specific models, such as BioGPT, BioMistral [5], or clinical adaptations of Llama, remains an open question. Prior work has documented position bias and CoT sensitivity in general-purpose LLMs [21, 15], so these failure modes are likely not unique to MedGemma. But the magnitudes we observe (e.g., 59.1% flip rate, 5.7% CoT degradation) may well differ across architectures and training regimes. Future multi-model evaluations should test whether domain-specific fine-tuning amplifies or mitigates these effects.

Our evaluation also covers only multiple-choice and yes/no/maybe formats. Real clinical workflows involve free-text reasoning, open-ended differential diagnosis, and structured decision support. The mechanisms we identify, including position bias, verbose reasoning errors, and context truncation sensitivity, are likely to persist in those settings, though they may manifest differently.

Prompt-induced prediction changes are not uniformly benign. Among the 750 questions where CoT flipped a correct answer to an incorrect one, a subset involves high-risk clinical scenarios. In the organophosphate poisoning example, CoT selected a contraindicated drug. We observed multiple cases where prompt choice alone led to selection of contraindicated medications (e.g., neostigmine for organophosphate poisoning, allopurinol during acute gout; see Appendix B). While a comprehensive clinical severity taxonomy is beyond our scope, these examples demonstrate that prompt-induced errors can be clinically dangerous—not merely statistically inconvenient—underscoring the need for robustness testing before any safety-critical deployment.

6 Conclusion

Our evaluation demonstrates that standard prompt engineering techniques do not reliably transfer to domain-specific medical LLMs. On MedGemma, CoT decreases accuracy by 5.7% ($p < 0.001$), few-shot examples degrade accuracy by 11.9% while tripling position bias, option shuffling causes 59.1% prediction

flips, and front-truncated context performs worse than no context at all. Practical mitigations exist: cloze scoring achieves the highest accuracy (51.8%/64.5%) with near-zero position bias, back-truncation preserves 97% of full-context accuracy, and permutation voting recovers 4 points over single-ordering inference.

For deployment, we recommend: (1) default to zero-shot direct prompting or cloze scoring; (2) test option order sensitivity and use permutation voting with agreement-based abstention; (3) truncate retrieved context from the end, not the beginning; and (4) prefer selective high-density retrieval over exhaustive context. The cloze–generation gap suggests that benchmark accuracies reflect the prompt–generate–parse pipeline rather than underlying model knowledge, making rigorous per-use-case validation essential before deployment.

Acknowledgements. We thank the MedGemma team at Google for releasing open model weights. Experiments were conducted on NVIDIA A100 GPUs.

References

1. Barnett, S., Kurniawan, S., Thudumu, S., Brannelly, Z., Abdelrazek, M.: Seven failure points when engineering a retrieval augmented generation system. arXiv preprint arXiv:2401.05856 (2024). https://doi.org/10.48550/arXiv.2401.05856
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Google DeepMind: Medgemma: Medical language models. Google AI Blog (May 2025), technical report
4. Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: Pubmedqa: A dataset for biomedical research question answering. Proceedings of EMNLP-IJCNLP pp. 2567–2577 (2019)
5. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.A., Rouvier, M., Dufour, R.: Biomistral: A collection of open-source pretrained large language models for medical domains. arXiv preprint arXiv:2402.10373 (2024)
6. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics **12**, 157–173 (2024)
7. Lu, Y., Bartolo, M., Moore, A., Riedel, S., Stenetorp, P.: Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. Proceedings of ACL pp. 8086–8098 (2022)
8. Meincke, L., Mollick, E.R., Mollick, L., Shapiro, D.: The decreasing value of chain of thought in prompting. Tech. rep., Wharton Generative AI Labs (2024), sSRN 5285532
9. Nori, H., King, N., McKinney, S.M., Carignan, D., Horvitz, E.: Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:2303.13375 (2023)
10. Omar, R., et al.: A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. Computers in Biology and Medicine (2024)
11. Pal, A., Umapathi, L.K., Sankarasubbu, M.: Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. Proceedings of the Conference on Health, Inference, and Learning pp. 248–260 (2022)

12. Singhal, K., Azizi, S., Tu, T., Mahdavi, S.S., Wei, J., Chung, H.W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al.: Large language models encode clinical knowledge. Nature **620**(7972), 172–180 (2023)
13. Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., et al.: Towards expert-level medical question answering with large language models. arXiv preprint arXiv:2305.09617 (2023)
14. Sohn, K., Hong, S., Guevara, C., Amrollahi, R., Gholami, A., Niu, H., Mitra, B.: RAG^2: A full-stack framework for reliable medical RAG. arXiv preprint arXiv:2411.00300 (2024). https://doi.org/10.48550/arXiv.2411.00300
15. Sprague, Z., Pei, J., Chaturvedi, A., Lee, Z., Gao, N., Chen, Y., Zhang, R.: Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse. arXiv preprint arXiv:2410.21333 (2024)
16. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. Proceedings of ICLR (2023)
17. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems **35**, 24824–24837 (2022)
18. Wu, K., Wu, E., Zou, J.: Clasheval: Quantifying the tug-of-war between an LLM's prior knowledge and external evidence. arXiv preprint arXiv:2404.10198 (2024). https://doi.org/10.48550/arXiv.2404.10198
19. Xiong, G., Jin, Q., Lu, Z., Zhang, A.: Benchmarking retrieval-augmented generation for medicine. In: Findings of the Association for Computational Linguistics: ACL 2024. pp. 6233–6251. Association for Computational Linguistics, Bangkok, Thailand (2024). https://doi.org/10.18653/v1/2024.findings-acl.372, https://aclanthology.org/2024.findings-acl.372/
20. Zhao, Z., Wallace, E., Feng, S., Klein, D., Singh, S.: Calibrate before use: Improving few-shot performance of language models. Proceedings of ICML pp. 12697–12706 (2021)
21. Zheng, C., Zhou, H., Meng, F., Zhou, J., Huang, M.: Large language models are not robust multiple choice selectors. In: Proceedings of the International Conference on Learning Representations (ICLR) (2024), spotlight paper
22. Zhuo, J., Xing, S., Hu, Z., Wang, Z., Zhai, G., Zhang, X.P.: Prosa: Assessing and understanding the prompt sensitivity of llms. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 25372–25410 (2024)

A Prompt Templates

This section provides the exact prompt templates used in our experiments.

Zero-shot Direct.

```
Answer the following medical question by selecting
the correct option.

Question: {question}

A. {option_a}  B. {option_b}
```

```
C. {option_c}  D. {option_d}

Answer with just the letter (A, B, C, or D):
```

Zero-shot Chain-of-Thought.

```
Answer the following medical question.
Think step by step.

Question: {question}

A. {option_a}  B. {option_b}
C. {option_c}  D. {option_d}

Let's think step by step, then provide the
answer as a single letter (A, B, C, or D):
```

Few-shot Direct. Three example questions with correct answers were prepended to each test question. Examples were randomly sampled from different medical subjects to avoid information leakage.

Answer-only.

```
{question}
A. {option_a}  B. {option_b}
C. {option_c}  D. {option_d}
```

PubMedQA.

```
Based on the following context, answer the
research question.

Context: {context}
Question: {question}

Answer with one word: yes, no, or maybe.
```

B Detailed Chain-of-Thought Analysis

Case-level Breakdown. Of 4,183 MedMCQA questions evaluated with both zero-shot direct and zero-shot CoT:

Failure Patterns. We analyzed 200 cases where CoT hurt performance (Table 10). The predominant pattern is verbose reasoning (90.7%) where longer chains create opportunities for errors to compound. Categories are not mutually exclusive.

Overthinking Examples. Two additional cases illustrate the "Overthinking" failure pattern, where CoT introduces unnecessary complexity into straightforward questions.

Example A (Pharmacology, correct answer: B). *Question:* "Which drug is used for the treatment of acute gout?" Options: A) Allopurinol, B) Colchicine, C) Probenecid, D) Febuxostat. Direct prompting correctly answers "B." CoT begins: "Allopurinol is a xanthine oxidase inhibitor used for chronic gout prophylaxis... Colchicine is effective for acute flares... However, NSAIDs are often first-line for acute gout, and colchicine has a narrow therapeutic index... Probenecid is a uricosuric agent... Given the considerations about side effects..." and selects "A" (Allopurinol), a drug contraindicated during acute attacks. The model correctly identified colchicine as the acute treatment but then over-deliberated on side effect considerations irrelevant to the question.

Example B (Anatomy, correct answer: C). *Question:* "The longest cranial nerve is:" Options: A) Trigeminal, B) Facial, C) Vagus, D) Hypoglossal. Direct prompting correctly answers "C." CoT responds: "The trigeminal nerve has three divisions and covers a large facial area... The vagus nerve extends to the abdomen, making it anatomically the longest... However, if we consider the total fiber length including all branches, the trigeminal's three divisions might collectively exceed..." and selects "A" (Trigeminal). The model introduced a specious "total branch length" reasoning that turned a factual recall question into an unnecessary deliberation.

C Position Bias Analysis

The model shows a consistent preference for option A across all conditions, but this bias is dramatically amplified under few-shot prompting. With few-shot direct, the model predicts A for 76% of questions despite A being correct only 32.2% of the time, an overweight of 43.8 percentage points.

D Option Order Details

Only 23.4% of questions received consistent predictions across all five ordering conditions. For over three-quarters of questions, the model's answer depends on option ordering.

E Truncation Strategy Comparison

The 30.7 percentage point gap between back-truncation (44.5%) and front-truncation (13.8%) at the same reduction ratio is the single largest effect in our study.

F Robustness Baselines Details

Cloze Scoring. For each MedMCQA question, we extract the log-probabilities of the four option tokens (A, B, C, D) at the final position. The option with the highest log-probability is selected. This requires only a single forward pass with no autoregressive generation.

Permutation Voting. For each question, $K=4$ distinct orderings receive independent greedy decoding. Predictions are inverse-mapped to canonical labels; majority vote determines the final answer. The agreement rate serves as a calibration signal.

G Threats to Validity

Answer Parsing. Our evaluation relies on extracting answer letters via regex. Parse error rates are below 2.5% across all conditions (1.4% for direct, 2.1% for CoT). The 0.7-point difference cannot explain the 5.7% accuracy gap. Moreover, CoT being harder to parse makes our finding more conservative.

Dataset Imbalance. The model's overweighting of position A (45.9% predicted vs. 32.2% true) far exceeds what ground truth would justify. The 59.1% flip rate from option shuffling directly demonstrates position-based predictions that cannot be explained by dataset imbalance.

Dataset Contamination. Both benchmarks are publicly available and may have appeared in pretraining data. However, our analysis focuses on *relative* robustness across conditions. Memorized answers should not change when options are shuffled or prompts are reformatted. The large relative degradations we observe indicate genuine sensitivity independent of potential contamination.

H Reproducibility

All randomized experiments use seeds $\{42, 123, 456\}$. MedMCQA outputs are constrained to letters A through D; PubMedQA to yes/no/maybe. Answers are extracted via regex matching the first valid token. Unparseable outputs are scored as incorrect.

All experiment code, prompt templates, configuration files, and analysis scripts are available at: *https://github.com/UNHSAILLab/MedMCQA-Robustness-Study.*

Table 5. Sample model responses illustrating key failure modes. All examples from MedGemma-4B on MedMCQA. Prompts and outputs are reproduced verbatim.

Example 1: CoT backfires (Dental, correct answer: D)	

Prompt (direct)	`Question: In a 6-month-old child, thick curd like white patch appears on the buccal mucosa. On rubbing it leaves an erythematous patch. Most likely diagnosis is:`
	`Options: A) Tuberculosis B) Lichen planus C) Lupus erythematous D) Candidiasis`
	`Answer with a single letter (A, B, C, or D).`

Direct output	`D` ✓
Prompt (CoT)	`Question: [same as above]` `Think through this step-by-step, analyzing each option. Then provide your final answer as a single letter (A, B, C, or D).`
CoT output	`A) Tuberculosis: Oral tuberculosis is rare in children. While it can present with lesions, they are less likely to be the classic "curd-like" appearance.` `B) Lichen planus: less common in infants. It typically presents with flat-topped`... *(truncated at 256 tokens; selects* `A`*)* ×
Example 2: Option order flip (Anatomy, correct answer: B → A after shuffle)	

Prompt (orig.)	`Question: Which of the following are not a branch of external carotid artery in Kiesselbach's plexus.`
	`Options: A) Sphenopalatine artery B) Anterior ethmoidal artery C) Greater palatine artery D) Septal branch of superior labial artery`
	`Answer with a single letter (A, B, C, or D).`

Output (orig.)	`B` ✓ (Anterior ethmoidal is correct; it is not an external carotid branch)
Prompt (shuf.)	`Question: [same as above]` `Options: A) Anterior ethmoidal artery B) Septal branch of superior labial artery C) Sphenopalatine artery D) Greater palatine artery`
Output (shuf.)	`B` × (The correct content moved to position A, but the model still picks B)

Table 6. Option order sensitivity on MedMCQA (n=4,183). All intervals are 95% Wilson score CIs. Seed 42.

Perturbation	Accuracy	95% CI	Drop	Flip Rate
Original	47.6%	[46.1, 49.1]	–	–
Random shuffle	27.3%	[26.0, 28.7]	−20.3%	57.8%
Rotate-1	20.2%	[19.0, 21.5]	−27.4%	72.9%
Rotate-2	21.9%	[20.7, 23.2]	−25.7%	69.7%
Distractor swap	47.6%	[46.1, 49.1]	0.0%	36.1%
Mean	29.2%	–	−18.4%	59.1%

Table 7. Evidence conditioning on PubMedQA (n=1,000). Random baseline: 33.3%. 95% Wilson score CIs are ±2.0–3.1 pp for all conditions; see Appendix E for detailed intervals.

Condition	4B	27B	Description
Core conditions:			
Question only	34.5%	31.0%	No context
Full context	45.8%	38.2%	Complete abstract
Front-trunc. 50%	13.8%	23.4%	First half of words
Front-trunc. 25%	13.7%	18.6%	First quarter
Background only	28.9%	19.8%	Intro. sentences
Results only	41.9%	**40.0%**	Concl. sentences
Expanded truncation strategies (4B):			
Back-trunc. 50%	**44.5%**	–	Last half of words
Middle-trunc. 50%	33.9%	–	Central portion
Sent.-trunc. 50%	30.3%	–	Sentence boundary
Salient top-5	40.1%	–	By question overlap
Salient top-3	36.4%	–	By question overlap

Table 8. Robustness baselines on MedMCQA. Cloze scoring uses n=4,183 (4B) or n=1,000 (27B); permutation vote uses n=1,000.

Method	MedGemma-4B		MedGemma-27B	
	Acc.	Pos. Bias	Acc.	Pos. Bias
Zero-shot direct (ref.)	47.6%	0.137	–	–
Cloze scoring	51.8%	0.013	**64.5%**	0.054
Permutation vote (K=4)	49.0%	–	–	–

Table 9. Case-level comparison of direct vs. CoT prompting.

Outcome	**Count**	**Pct.**
Both correct	1,511	36.1%
Both wrong	1,410	33.7%
Direct correct, CoT wrong	750	17.9%
Direct wrong, CoT correct	512	12.2%
Net effect of CoT	−238	−5.7%

Table 10. Failure patterns in CoT responses (n=200 sample).

Pattern	**%**	**Description**
Verbose reasoning	90.7	>500 chars
Correct concept, wrong applic.	34.5	Misapplied knowledge
Distractor confusion	28.0	Rejected correct ans.
Self-contradiction	25.6	Conflicting logic
Overthinking	18.5	Over-complicated
Confident wrong concl.	11.1	“Therefore” + wrong

Table 11. Answer distribution: ground truth vs. predictions.

Condition	**A**	**B**	**C**	**D**
Ground truth	32.2%	25.1%	21.4%	21.3%
Zero-shot direct	45.9%	22.1%	17.8%	14.2%
Zero-shot CoT	52.3%	19.4%	15.7%	12.6%
Few-shot direct	76.0%	12.0%	7.0%	5.0%
Few-shot CoT	68.4%	15.2%	9.8%	6.6%
Answer-only	41.2%	24.3%	19.1%	15.4%

Table 12. Prediction consistency across perturbations.

Consistency Level	**Pct.**
All 5 conditions	23.4%
4 of 5 conditions	18.7%
3 of 5 conditions	21.2%
2 of 5 conditions	19.8%
All different	16.9%

Table 13. Truncation strategies at 50% reduction (MedGemma-4B, PubMedQA n=1,000).

Strategy	**Acc.**	**95% CI**	**Δ Full**
Full context (ref.)	45.8%	[42.7, 48.9]	–
Back-trunc. 50%	44.5%	[41.4, 47.6]	−1.3
Salient top-5	40.1%	[37.1, 43.2]	−5.7
Middle-trunc. 50%	33.9%	[31.0, 36.9]	−11.9
Sent.-trunc. 50%	30.3%	[27.5, 33.2]	−15.5
Front-trunc. 50%	13.8%	[11.8, 16.1]	−32.0
Question only	34.5%	[31.6, 37.5]	−11.3

Table 14. Cloze scoring detailed results.

Metric	**4B** (n=4,183)	**27B** (n=1,000)
Accuracy	51.8%	64.5%
95% CI	[50.3, 53.3]	[61.4, 67.1]
Mean logprob margin	4.07	3.14
Position bias	0.013	0.054

Table 15. Permutation voting (MedGemma-4B, n=1,000, K=4).

Metric	**Value**
Aggregated accuracy	49.0%
Per-perm. accuracy (mean ± std)	45.1% ± 1.7%
Mean agreement rate	70.0%
Agreement rate std	22.1%