SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Answer Positioning Biases in Large Language Model Responses to Medical Multiple-Choice Questions

Kartik Gupta, MSc, University of Western Ontario; Jaron Chong, MD

Introduction/Background

Large language models (LLMs) have demonstrated high performance on standardized medical examinations across various domains, typically employing a multiple-choice question (MCQ) format. Technical literature has reported that the precise order of answer choices may affect and bias LLM performance, leading to unreliable estimates of LLM performance. The objective of this study is to evaluate the accuracy of LLMs with forced re-positioning of multiple-choice answer options, utilizing MedQA, a widely recognized medical benchmarking dataset for LLMs.

Methods/Intervention

The comparative efficacy of GPT-3.5 and GPT-4 was assessed using three randomized subsets of the MedQA dataset, each comprising 1273 questions, representing 10% of the total dataset. For each subset, four permutations were generated by forced re-positioning of the correct answer into each of four possible answer positions. The models were evaluated utilizing two prompt templates: question-only format (QO) and chain-of-thought format (COT; "Think step-by-step."). Statistical analysis involved repeated measures ANOVA followed by post-hoc comparisons using Bonferonni's multiple comparison's test. The variance of performance was calculated by subtracting the accuracy of the least effective position from the most effective position, termed delta.

Results/Outcome

Using basic QO prompting, GPT-4 outperformed GPT-3.5 in accuracy (69.18% vs 57.53%; $p< 0.001$). COT outperformed QO prompting, with GPT-4 COT achieving a maximal performance of 80.36%, versus GPT-3.5 COT of 67.08% ($p< 0.001$). Across model-prompt interventions without COT, position A's performance was significantly greater than other positions. This positional bias is reduced by COT. Utilizing COT reduced Delta with GPT-3.5 (16.3% to 5.5%) and GPT-4 (15.6% to 2.9%).

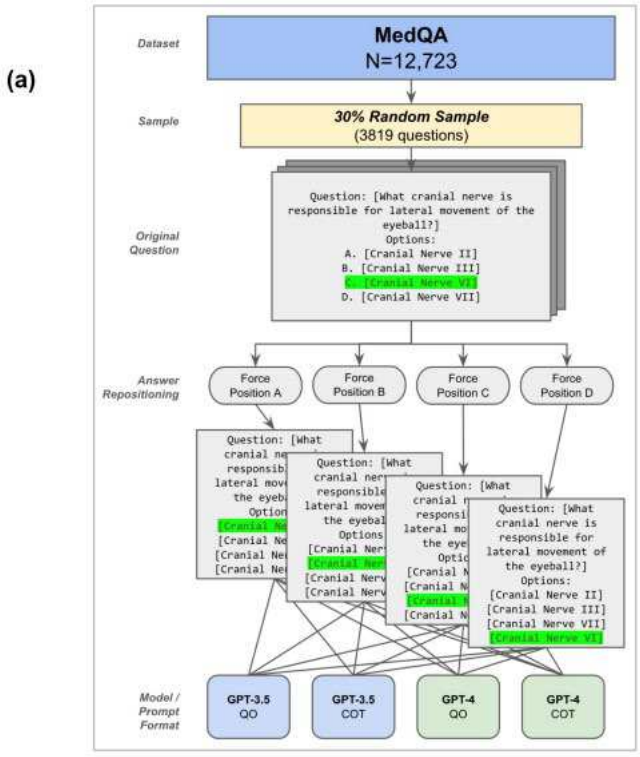
Conclusion

COT outperforms basic QO prompting, without which, there is strong LLM performance bias towards earlier answer positions. The distribution of answer choice positions in a MCQ evaluation may affect the apparent performance of an LLM.

Statement of Impact

Clinical LLM evaluation should carefully consider the effect of multiple-choice answer position, given systemic biases in performance based upon answer position. LLM evaluation should ideally incorporate randomization of answer position for evaluation.

FIGURE 1: (a) Flow chart of MedQA multiple-choice question sampling and forced answer re-positioning flow chart. Reformatted questions were submitted to Models GPT-3.5 & GPT-4 with both Question Only (QO) and Chain-of-Thought (COT) Prompt Formats (b) Jitter plot of LLM accuracy, stratified by model (GPT3.5 vs. GPT4) and Prompt Technique (QO-Question Only vs. COT-Chain of Thought). Answer position biases were strongly observed with GPT-3.5 and Question Only prompting techniques (Highlight: Red Dashed Boxes), significantly reduced with a combination of a stronger LLM model (GPT-4) and COT (Highlight: Green Dashed Box).



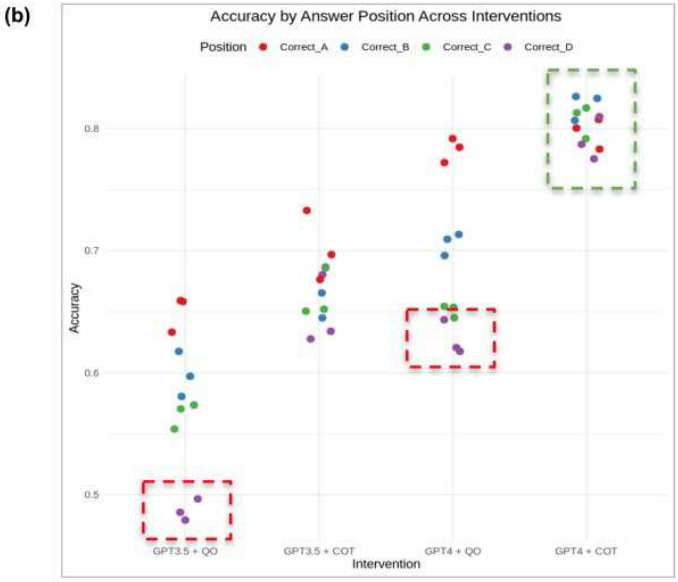


FIGURE 2: (a) Total correct answers for each model-prompt intervention along with resultant mean and delta for overall model-prompt performance. The denominator for each forced answer position set is 3819 questions total. QO refers to base prompting with only the question-and-answer choices. COT refers to Chain-of-Thought, where the model is additionally prompted to think step-by-step. The significance between the mean accuracy of each group is noted by a p-value <0.05.

(a)

Model	Prompt	Results (Correct Answers)	Mean	Delta*	P-values of Mean Performance
GPT-3.5	1: Base (QO)	Force A: 2483 Force B: 2285 Force C: 2161 Force D: 1860	0.575	0.163	---
	2: Chain-of-Thought (COT)	Force A: 2681 Force B: 2543 Force C: 2531 Force D: 2471	0.671	0.055	GPT 3.5 QO vs GPT 3.5 COT: **0.018**
GPT-4	1: Base (QO)	Force A: 2990 Force B: 2697 Force C: 2486 Force D: 2395	0.692	0.156	GPT 3.5 QO vs GPT 4 QO: **0.005**
	2: Chain-of-Thought (COT)	Force A: 3044 Force B: 3129 Force C: 3083 Force D: 3020	0.804	0.029	GPT 4 QO vs GPT 4 COT: **0.007** GPT 3.5 COT vs. GPT 4 COT: **0.002**

**Delta is defined as the difference in performance between the minimum and maximum performing MCQ answer positions.*

(b)

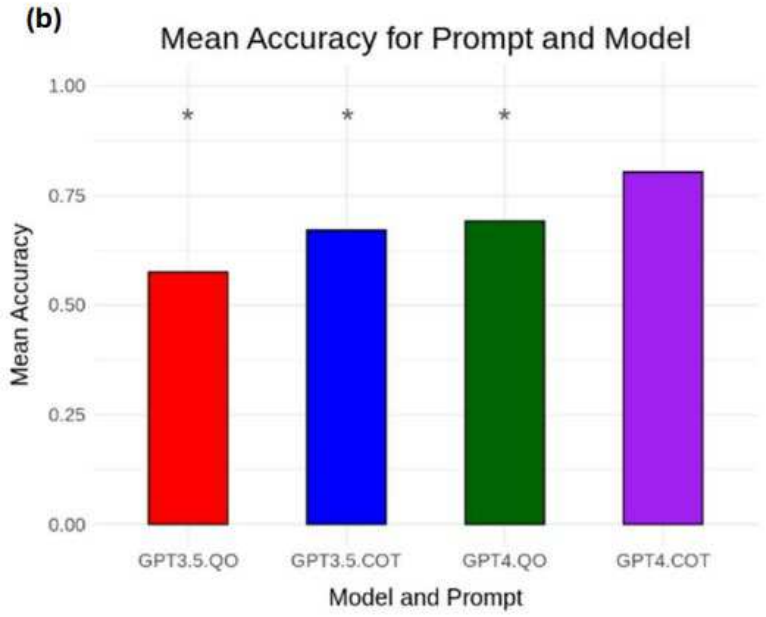
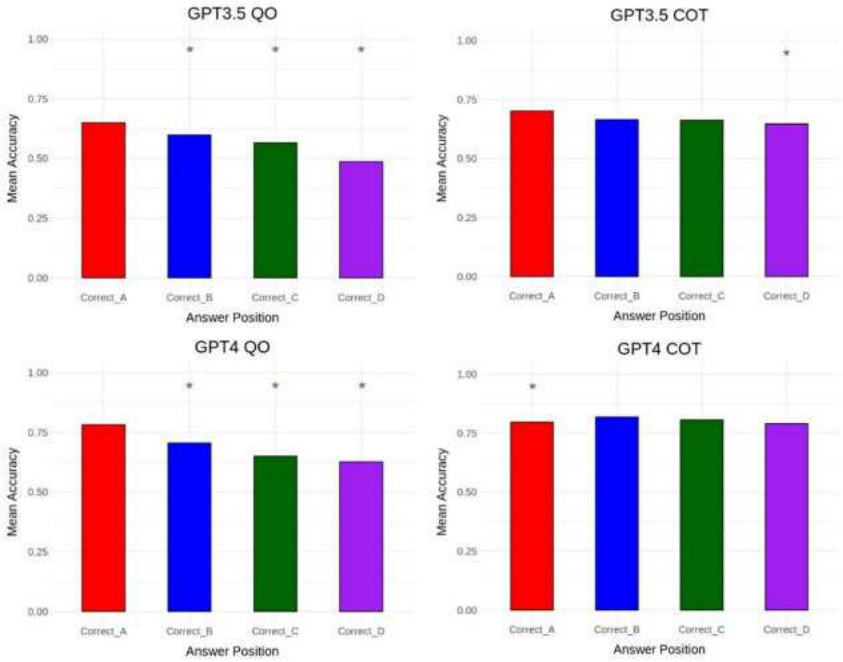


FIGURE 3: Mean Accuracy of GPT-3.5 and GPT-4 Models by Answer Position and Prompt Type. The graphs compare the mean accuracy of GPT-3.5 and GPT-4 models on different answer positions (Correct_A, Correct_B, Correct_C, Correct_D) using two methods: QO (Question Only) and COT (Chain of Thought). Position analysis showed that for GPT-3.5 QO, position A showed significantly higher performance compared to position D ($p=0.007$), and there were also significant differences between positions B and D ($p=0.032$), and positions C and D ($p=0.042$). For GPT-3.5 COT, there was a significant difference between position A and D ($p=0.04$). For GPT-4 QO, position A had significantly higher performance than all other positions (B, C, and D) with p-values < 0.05.



Keywords

Large Language Models (LLMs)**;** Medical Question Answering**;** Answering Bias**;** LLM Safety