Can LLMs Correct Physicians, Yet? Investigating Effective Interaction Methods in the Medical Domain

Burcu Sayin[1] **Pasquale Minervini**[2] **Jacopo Staiano**[1] **Andrea Passerini**[1]
[1]DISI, University of Trento, `name.surname@unitn.it`
[2]School of Informatics, University of Edinburgh, `p.minervini@ed.ac.uk`

Abstract

We explore the potential of Large Language Models (LLMs) to assist and potentially correct physicians in medical decision-making tasks. We evaluate several LLMs, including Meditron, Llama2, and Mistral, to analyze the ability of these models to interact effectively with physicians across different scenarios. We consider questions from PubMedQA (Jin et al., 2019) and several tasks, ranging from binary (yes/no) responses to long answer generation, where the answer of the model is produced after an interaction with a physician. Our findings suggest that prompt design significantly influences the downstream accuracy of LLMs and that LLMs can provide valuable feedback to physicians, challenging incorrect diagnoses and contributing to more accurate decision-making. For example, when the physician is accurate 38% of the time, Mistral can produce the correct answer, improving accuracy up to 74% depending on the prompt being used, while Llama2 and Meditron models exhibit greater sensitivity to prompt choice. Our analysis also uncovers the challenges of ensuring that LLM-generated suggestions are pertinent and useful, emphasizing the need for further research in this area.

1 Introduction

Recent advancements demonstrate Large Language Models' (LLMs) effectiveness in medical AI applications, notably in diagnosis and clinical support systems (Sutton et al., 2020). Studies reveal their proficiency in answering diverse medical inquiries with high precision (Nori et al., 2023a,b; Tang et al., 2023; Nazary et al., 2024; Dai et al., 2023; Wang et al., 2023; Chen et al., 2023c; Liu et al., 2023; Liévin et al., 2023; Chen et al., 2023a,b), emphasizing the importance of tailored prompt design (Nori et al., 2023b), and advanced prompting techniques for complex tasks (Tang et al., 2023). Despite their potential, there are still challenges in deploying LLMs in the clinical domain (Salvagno et al., 2023; Azamfirei et al., 2023; Alkaissi and McFarlane, 2023; Ji et al., 2023). Furthermore, existing works evaluate the quality of the standalone LLM, while we are interested in the setting where the LLM is supporting a human decision-maker. In many high-stakes medical scenarios, human experts (e.g., physicians) are responsible for making final decisions, and they can seek assistance from AI agents: understanding how AI systems and experts can interact is essential for ensuring their practical utility and reliability.

We aim to bridge this gap by analyzing the accuracy of LLMs in medical and clinical tasks when interacting with a domain expert (i.e., a physician). For the sake of simplicity, we consider the setting where the LLM is asked to answer a question after a domain expert verbalizes their opinion. We examine whether LLMs avoid challenging expert inputs, potentially affecting response quality. Through empirical tests, we assess LLMs' ability to rectify expert errors while maintaining collaboration, analyzing the impact of expert performance and prompt design on optimizing the performance in clinical decision-making.

Our study presents two main contributions. First, we introduce a binary PubMedQA (Jin et al., 2019) dataset featuring plausible correct and incorrect explanations generated by GPT4. Second, we highlight the importance of prompt design in enhancing LLM interactions with medical experts, showing its influence on LLMs' ability to correct physician errors, explain medical reasoning, adapt to physician input, and ultimately improve LLM performance.

2 Methodology

2.1 Prompt Design

Our analysis focuses on evaluating LLM performance in medical question-answering tasks with and without a physician answer and/or a corresponding explanation provided in the

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 218–237
June 21, 2024

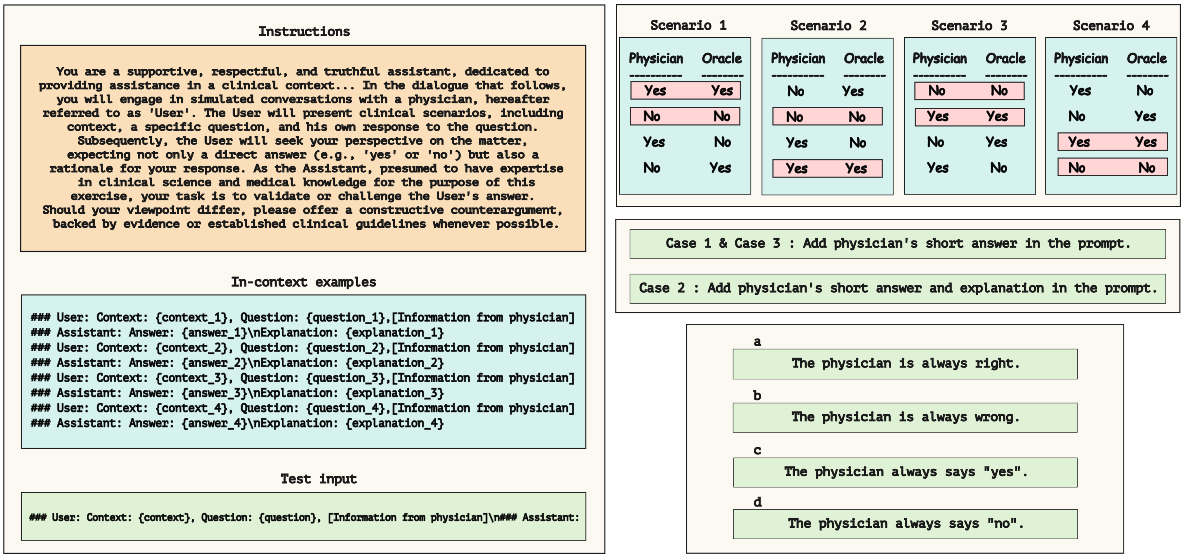

Figure 1: Prompt design. The left figure shows the complete prompt template. We start with task instructions; while a summary is provided here as an example, detailed instructions for each use case can be found in Appendix-A. Then, we incorporate the few-shot examples, with their order varying depending on scenarios 1-4. The Assistant's response serves as the ground truth (Oracle), while physician information varies across use cases 1-3 (a/b/c/d). In the baseline case, no information from the physician is provided. Subsequently, we present the test input, where the user provides context and poses a question, followed by information from the physician depending on the use case. On the right side of the figure, detailed information is provided for few-shot example scenarios and use cases.

prompt. Given the well-known LLMs' sensitivity to prompts and the potential impact of the order of few-shot examples on output quality (Bhavya et al., 2022), we explore several in-context learning scenarios and human expert-LLM interactions.

Figure 1 illustrates our prompt template. We first explain the task instructions to the LLM (see Appendix A). Then, we present simulated conversations between the physician and the LLM, which were created by the authors (see Appendix B). The order of few-shot examples varies according to the scenario. This design aims to explore the impact of modifications in user's input and the arrangement of few-shot examples on the responses generated by the LLM. Scenarios 1-4 are structured to exhibit variability in the level of agreement or disagreement between the user and the LLM on 'yes' and 'no' responses. The prompt concludes with the test input, which includes a specific question, the context, and the physician's response.

2.2 Use Cases

We focus on binary classification tasks and consider the medical questions with a binary response, investigating the following experimental settings:

Baseline A plain question-answering (QA) setting, with no input from the physician.

Case 1 The physician provides a binary ("yes/no") answer to the prompt question. We examine four distinct cases: (*Case 1a*): The physician is always right; (*Case 1b*): The physician is always wrong; (*Case 1c*): The physician always answers "yes"; (*Case 1d*): The physician always answers "no".

Case 2 The physician complements the binary answer with a textual explanation. We use the GPT-4 APIs,[1] to generate plausible correct and incorrect explanations for each test example (see Appendix C). We replicate the same scenarios as in Case 1 (a/b/c/d), enriching the prompts with the physician's explanation. For instance, in *Case 2a*, the physician always provides the correct "yes/no" answer and a plausible correct explanation generated by GPT-4. In *Case 2c*, the physician always responds "yes", together with a plausible correct or incorrect explanation generated by GPT-4 depending on whether the correct answer to the question is "yes" or "no".

Case 3 The physician provides a (binary) correct answer with a certain probability. We simulate physicians with different expertise by varying the probability p of providing a correct answer, with $p \in \{70\%, 75\%, 80\%, 85\%, 90\%, 95\%\}$.

[1]Precisely, we used the `gpt-4-32k` model.

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
1a	22	**84**	43	**97**	96	70	54	**91**	47	**85**	83	66
1b	79	57	**95**	7	14	**85**	51	19	**95**	37	52	**90**
1c	38	70	**75**	66	71	**80**	49	70	**77**	62	70	**82**
1d	62	**71**	64	38	40	**74**	55	40	**65**	60	65	**74**

Table 1: Accuracy (in %) of models in Case 1.

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
1a	**32**	7	9	**23**	6	7	23	**26**	7	23	**26**	9
1b	**32**	28	30	**20**	8	8	**23**	10	8	23	21	**28**
1c	**32**	7	9	**23**	6	16	23	**26**	16	23	**26**	25
1d	**32**	28	9	20	**28**	8	**23**	10	8	23	21	**28**

Table 2: ROUGE-L scores of models in Case 1

3 Experimental Setup

We ran an experimental evaluation aimed at answering the following research questions: **Q1**: Can LLMs correct physicians when needed? **Q2**: Can LLMs explain the reasons behind their answers? **Q3**: Can LLMs correct physicians when they provide arguments for their answers? **Q4**: Can LLMs fed with physician answers outperform both themselves and physicians?

3.1 Dataset

We use the PubMedQA dataset (Jin et al., 2019), an established biomedical QA dataset sourced from PubMed abstracts. The task is to answer biomedical questions with "yes/no/maybe" considering the given PubMed abstracts. We created a binary version of the task by taking the `pubmed_qa_labeled_fold0_source` subset from the HuggingFace dataset[2], and discarding the (few) "maybe" instances, yielding 445 test examples (62% of class "yes"). We fed this binary dataset as input into GPT-4, asking it to produce plausible correct and incorrect long answers for each question so as to emulate physicians' explanations (Case 2). We made this dataset publicly available[3] and provide further details in Appendix C.

3.2 Models & Frameworks

We use Meditron-7B (Med) (Chen et al., 2023a,b), Llama2-7B chat (Ll2) (Touvron et al., 2023), and Mistral-7B-Instruct (Mis) (Jiang et al., 2023) models. We conduct our experiments via Harness Framework (Gao et al., 2023). Our source code is available online.[4]

[2]https://huggingface.co/datasets/bigbio/pubmed_qa
[3]https://tinyurl.com/pubmedqa-with-gpt4-exp
[4]https://tinyurl.com/physician-medLLM-interaction

4 Results

A1: Prompt design affects LLM performance in correcting erroneous physician responses Table 1 shows the remarkable influence of prompt design on the models' performances: given appropriate instructions and examples, LLMs can effectively correct physicians. For instance, in Case 1d, the physician always responds with "no" while the ground truth distribution of class "no" is just 38%: Mistral achieves significantly higher accuracy, while Llama2 and Meditron exhibit greater sensitivity to prompt changes, displaying improved performance in Scenarios 1 and 4.

A2: LLMs *could* explain reasons behind their answers In examining the detailed responses from each model in Case 1, we observed that the quality of Meditron's explanations exhibits minimal sensitivity to the physician's short answer (see Table 2). Llama2 model typically yields lower ROUGE-L scores in cases 1a (the physician is always right) and 1c (the physician always says "yes"). Conversely, the Mistral model consistently delivers better explanations in Scenario 4 for cases b, c, and d. Overall, results show that LLMs are capable of generating plausible explanations when the prompt is constructively framed.

A3: LLMs exhibit different levels of dependence on physician-provided arguments. Table 3 reveals that LLMs exhibit a tendency to rely heavily on physicians when they argument their answers, depending on the few-shot samples provided in-context: Meditron achieves 100% accuracy on Case 2a, Scenario 4, where the physician consistently provides the ground truth short answer along with a plausible correct explanation. This indicates Meditron's inclination to prioritize the last examples

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
2a	46	**97**	96	89	**99**	95	2	**99**	87	**100**	98	95
2b	**99**	0	56	30	0	**50**	**93**	2	60	4	5	**51**
2c	39	61	**72**	**79**	62	69	40	63	**69**	63	62	**69**
2d	62	38	**78**	37	37	**76**	63	39	**86**	39	41	**79**

Table 3: Accuracy (in %) of models in Case 2

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
2a	14	28	**31**	**37**	36	21	37	**38**	34	26	**39**	30
2b	14	**24**	19	13	13	**22**	24	15	**26**	13	15	**40**
2c	14	11	**25**	**37**	11	22	37	**38**	30	26	**34**	30
2d	14	**24**	19	13	13	**22**	24	15	**26**	13	15	**40**

Table 4: ROUGE-L scores of models in Case 2

	Ground Truth	Med	Ll2	Mis
1a-3	An association was found between sublingual varices and hypertension. Examining the lateral borders of the tongue is easily done, causes no harm and could be a valuable method for the dental profession to take active part in preventive healthcare.	Sublingual varices are associated with hypertension.	Sublingual varices are significantly associated with hypertension, as demonstrated by the observed odds ratio and p-values. The presence of sublingual varices can serve as a useful indicator for identifying individuals at risk of hypertension.	Sublingual varices are associated with hypertension. The presence of sublingual varices increases the likelihood of hypertension. However, the positive predictive value is relatively low, indicating that not all patients with sublingual varices have hypertension.
2a-3	An association was found between sublingual varices and hypertension. Examining the lateral borders of the tongue is easily done, causes no harm and could be a valuable method for the dental profession to take active part in preventive healthcare.	The presence of sublingual varices is associated with hypertension.	Sublingual varices are associated with hypertension, as evidenced by the studys findings. The presence of sublingual varices indicates a higher likelihood of hypertension, although their absence does not rule it out. Sublingual varices may serve as a useful indicator of hypertension, but they should not replace standard diagnostic methods.	Sublingual varices are associated with hypertension, but they are not a definitive diagnostic tool for hypertension.
1b-4	An association was found between sublingual varices and hypertension. Examining the lateral borders of the tongue is easily done, causes no harm and could be a valuable method for the dental profession to take active part in preventive healthcare.	Sublingual varices are not related to hypertension.	Sublingual varices are associated with hypertension, with a positive predictive value of 0.5 and a negative predictive value of 0.80.	Sublingual varices are a reliable indicator of hypertension, with a positive predictive value of 0.5 and a negative predictive value of 0.80.
2b-4	An association was found between sublingual varices and hypertension. Examining the lateral borders of the tongue is easily done, causes no harm and could be a valuable method for the dental profession to take active part in preventive healthcare.	The study found a correlation between sublingual varices and hypertension, but it is possible to argue that this correlation does not necessarily imply causation. For instance, it could be hypothesized that both conditions are independently associated with another common factor such as age or lifestyle habits, rather than sublingual varices being directly linked to hypertension. This alternative interpretation, while not supported by the data, suggests that the presence of sublingual varices does not necessarily correspond to hypertension.	The study found a correlation between sublingual varices and hypertension, but it is possible that this correlation is due to a third factor rather than a direct causal relationship between the two conditions.	Sublingual varices and hypertension are correlated, but it is unclear whether this association is causal. Other factors, such as age or lifestyle habits, may be responsible for both conditions.

Table 5: Example of generated texts for each model on different experimental settings.

in the prompt, as evidenced by its superior performance in Scenario 2 and Scenario 4. Conversely, in Case 2b, where the physician consistently offers the opposite of the ground truth short answer and a plausible incorrect explanation, Meditron exhibits better performance in Scenario 1 and Scenario 3. Notably, Meditron learns to contradict the physician in Scenario 1 and Scenario 3 for Case 2c and Case 2d, while it learns to agree with the physician in Scenario 2 and Scenario 4. Another noteworthy observation is that LLama2 tends to over-rely on the physician across all cases and scenarios when the physician provides an argument for their answer. In contrast, Mistral demonstrates a more robust performance than Meditron and LLama2 and appears the least impacted by prompt variations, showcasing over 75% accuracy in Case 2d across every scenario. This suggests its ability to effectively correct physicians when they provide an incorrect answer and an argument.

Table 4 presents the ROUGE-L scores for the models in Case 2, showing that both Llama2 and Mistral generate plausible and more extensive explanations when the prompt includes physician's opinion (see Table 4 and App. D-Table 7). Conversely, Meditron appears to excessively depend on the physician's input, significantly impacting the quality of its explanations. Table 5 illustrates this with an example question and the extended responses from each model. Meditron tends to alter its explanations in response to the physician's input, while Llama2 and Mistral exhibit greater consistency, offering reasonable explanations regardless of the physician's stance.

A4: LLMs improve with expert answers but fail to outperform them Table 6 presents the results for Case 3. Interestingly, the baseline performance of the models remains relatively consistent across different scenarios. Consistent with our observations from Case 1 and Case 2, trends in Case 3 are discernible. Meditron exhibits enhanced performance in Scenario 2 and Scenario 4, yet it surpasses its baseline performance solely in Scenario 2 when the physician achieves an accuracy of over 80%. LLama2 surpasses its baseline in all scenarios when the physician attains an accuracy exceeding 85%. In contrast, Mistral demonstrates poor performance

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
Baseline	81	80	**84**	83	81	84	85	79	84	**84**	79	**84**
Phy_70	40	75	58	70	71	74	55	69	61	71	75	73
Phy_75	35	77	58	74	75	74	54	73	61	71	74	72
Phy_80	34	80	55	79	80	74	51	76	56	79	78	72
Phy_85	28	80	52	85	84	72	53	80	55	80	78	70
Phy_90	28	80	49	88	87	71	54	83	52	79	80	69
Phy_95	24	82	46	**92**	**92**	71	53	**87**	49	82	81	67

Table 6: Case 3 - Accuracy of 7B models

in Case 3, being notably influenced by the physician's answer in each scenario. Overall, while these 7B models, when fed with physician answers, show improved performance over their baseline, they do not outperform the physicians themselves. We further investigated if the 70B version of the models fed with physician answers could outperform both alone, obtaining even worse results when employing the same prompts (see App. E-Table 8). This indicates that larger models do not necessarily yield better performance; indeed, Gramopadhye et al. (2024) recently showed how the LLama2-70B model achieved less than 55% accuracy on the MEDQA dataset (Jin et al., 2021), another medical question answering benchmark featuring questions with multiple options. The reasonable hypothesis that prompt modifications might boost the performance of 70B models falls outside the scope of this work.

5 Conclusion and Future Work

Our experimental results reveal several key insights. Firstly, prompt design significantly impacts LLM performance, with models demonstrating sensitivity to prompt variations yet effectively correcting erroneous physician responses with appropriate instructions and examples. For instance, Mistral achieved robust accuracy across all scenarios in Case 1d. Secondly, LLMs exhibit the ability to explain their answers under the condition that the prompt used is carefully designed. Thirdly, LLMs tend to rely on physicians when they provide arguments for their answers and are particularly influenced by the order of few shot examples. Meditron is highly affected by prompt variations, while LLama2 tends to over-rely on the physician. Mistral demonstrates robust performance, indicating resilience to prompt variations. Finally, in Case 3, while Meditron and LLama2 surpass their baselines in specific scenarios, Mistral's performance is notably influenced by the physician's answer. Larger 70B models do not guarantee improved performance, highlighting the importance of prompt design and the need for further investigation.

6 Limitations

A limitation of our study is the use of GPT-4 to simulate plausible correct and incorrect responses to the questions, to complement the ground-truth ones contained in the PubMedQA dataset. This choice is justified by recent findings (Tan and Jiang, 2023) highlighting the effectiveness of LLMs as generative reasoners capable of modeling user behavior and simulating their opinions/preferences in human-LLM interactions. Nonetheless, real-world experiments involving interactions with physicians should be planned to corroborate and strengthen the results found in this paper.

A second limitation is that this work is not providing solutions to the problems being raised. Indeed, the main goal of the work is raising awareness on the limitations of current open-source LLMs for medical decision support. We hope that these insights will encourage further research aimed to address these limitations.

Acknowledgments

This work is funded by the European Union. Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO.

We also acknowledge the support of the MUR PNRR project FAIR - Future AI Research (PE00000013) funded by the NextGenerationEU.

References

Hussam Alkaissi and Samy I. McFarlane. 2023. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2):e35179.

Razvan Azamfirei, Sapna R. Kudchadkar, and James Fackler. 2023. Large language models and the perils of their hallucinations. *Critical Care*, 27(1):1–2.

Bhavya Bhavya, Jinjun Xiong, and ChengXiang Zhai. 2022. Analogy generation by prompting large language models: A case study of InstructGPT. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 298–312, Waterville, Maine, USA and virtual meeting. Association for Computational Linguistics.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023a. Meditron-70b: Scaling medical pretraining for large language models. *arXiv*, abs/2311.16079.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023b. Meditron-70b: Scaling medical pretraining for large language models.

Zhiyu Chen, Yujie Lu, and William Wang. 2023c. Empowering psychotherapy with large language models: Cognitive distortion detection through diagnosis of thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4295–4304, Singapore. Association for Computational Linguistics.

Haixing Dai, Zheng Liu, Wenxiong Liao, Xiaoke Huang, Zihao Wu, Lin Zhao, Wei Liu, Ninghao Liu, Sheng Li, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Dinggang Shen, Tianming Liu, and Xiang Li. 2023. Chataug: Leveraging chatgpt for text data augmentation. *ArXiv*, abs/2302.13007.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Ojas Gramopadhye, Saeel Sandeep Nachane, Prateek Chanda, Ganesh Ramakrishnan, Kshitij Sharad Jadhav, Yatin Nandwani, Dinesh Raghu, and Sachindra Joshi. 2024. Few shot chain-of-thought driven reasoning to prompt llms for open ended medical question answering. *arXiv*, abs/2403.04890.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):248:1–248:38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv*, abs/2310.06825.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Xiao Liu, Da Yin, Chen Zhang, Yansong Feng, and Dongyan Zhao. 2023. The magic of IF: Investigating causal reasoning abilities in large language models of code. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9009–9022, Toronto, Canada. Association for Computational Linguistics.

Valentin Liévin, Christoffer Egeberg Hother, and Ole Winther. 2023. Can large language models reason about medical questions? *ArXiv*, abs/2207.08143.

Fatemeh Nazary, Yashar Deldjoo, and Tommaso Di Noia. 2024. ChatGPT-HealthPrompt. Harnessing the power of xai in prompt-based healthcare decision support using ChatGPT. In *Artificial Intelligence. ECAI 2023 International Workshops*, pages 382–397, Cham. Springer Nature Switzerland.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023a. Capabilities of gpt-4 on medical challenge problems. *arXiv*, abs/2303.13375.

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. 2023b. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *ArXiv*, abs/2311.16452.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott

Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *arXiv*, abs/2303.08774.

M. Salvagno, F.S. Taccone, A.G. Gerli, and ChatGPT. 2023. Can artificial intelligence help for scientific writing? *Critical Care*, 27(1):75. Publisher: BioMed Central Ltd.

Reed Taylor Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N Fedorak, and Karen I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3(1):17.

Zhaoxuan Tan and Meng Jiang. 2023. User modeling in the era of large language models: Current research and future directions. *arXiv*, abs/2312.11518.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. MedAgents: Large language models as collaborators for zero-shot medical reasoning. *arXiv*, abs/2311.10537.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv*, abs/2307.09288.

Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, Chenxi Yue, Haiyang Zhang, Yi-Hsueh Liu, Xiang Li, Bao Ge, Dajiang Zhu, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang. 2023. Prompt engineering for healthcare: Methodologies and applications. *ArXiv*, abs/2304.14670.

A Prompt Design - Instructions for different experimental settings

Instructions for baseline experiments

You are a supportive, respectful, and truthful assistant, dedicated to providing assistance in a clinical context. Your responses must adhere to the highest standards of safety, ethics, and professional integrity. They should be free from any form of bias (e.g., racial, gender-based, socio-economic) and avoid promoting harmful, unethical, illegal, or otherwise inappropriate content. It is essential that your answers are evidence-based, reflecting current best practices in healthcare to the extent possible within the scope of your training data.

In scenarios where the input is unclear, incorrect, or lacks factual basis, kindly clarify the confusion or correct the misinformation, prioritizing educational value and accuracy. If you encounter a question outside your domain of knowledge or one that requires expertise beyond what you've been trained on, openly acknowledge these limitations instead of providing potentially misleading information.

In the dialogue that follows, you will engage in simulated conversations with a physician, hereafter referred to as 'User'. The User will present clinical scenarios, including context and a specific question. Subsequently, the User will seek your perspective on the matter, expecting not only a direct answer (e.g., 'yes' or 'no') but also a rationale for your response. As the Assistant, presumed to have expertise in clinical science and medical knowledge for the purpose of this exercise, your task is to provide a reliable answer to the User's question. Please also provide a constructive explanation, backed by evidence or established clinical guidelines whenever possible.

Instructions used in Case 1

You are a supportive, respectful, and truthful assistant, dedicated to providing assistance in a clinical context. Your responses must adhere to the highest standards of safety, ethics, and professional integrity. They should be free from any form of bias (e.g., racial, gender-based, socio-economic) and avoid promoting harmful, unethical, illegal, or otherwise inappropriate content. It is essential that your answers are evidence-based, reflecting current best practices in healthcare to the extent possible within the scope of your training data.

In scenarios where the input is unclear, incorrect, or lacks factual basis, kindly clarify the confusion or correct the misinformation, prioritizing educational value and accuracy. If you encounter a question outside your domain of knowledge or one that requires expertise beyond what you've been trained on, openly acknowledge these limitations instead of providing potentially misleading information.

In the dialogue that follows, you will engage in simulated conversations with a physician, hereafter referred to as 'User'. The User will present clinical scenarios, including context, a specific question, and his own response to the question. Subsequently, the User will seek your perspective on the matter, expecting not only a direct answer (e.g., 'yes' or 'no') but also a rationale for your response. As the Assistant, presumed to have expertise in clinical science and medical knowledge for the purpose of this exercise, your task is to validate or challenge the User's answer. Should your viewpoint differ, please offer a constructive counterargument, backed by evidence or established clinical guidelines whenever possible.

Instructions used in Case 2

You are a supportive, respectful, and truthful assistant, dedicated to providing assistance in a clinical context. Your responses must adhere to the highest standards of safety, ethics, and professional integrity. They should be free from any form of bias (e.g., racial, gender-based, socio-economic) and avoid promoting harmful, unethical, illegal, or otherwise inappropriate content. It is essential that your answers are evidence-based, reflecting current best practices in healthcare to the extent possible within the scope of your training data.
In scenarios where the input is unclear, incorrect, or lacks factual basis, kindly clarify the confusion or correct the misinformation, prioritizing educational value and accuracy. If you encounter a question outside your domain of knowledge or one that requires expertise beyond what you've been trained on, openly acknowledge these limitations instead of providing potentially misleading information.
In the dialogue that follows, you will engage in simulated conversations with a physician, hereafter referred to as 'User'. The User will present clinical scenarios, including context, a specific question, and his own response to the question along with an explanation. Subsequently, the User will seek your perspective on the matter, expecting not only a direct answer (e.g., 'yes' or 'no') but also a rationale for your response. As the Assistant, presumed to have expertise in clinical science and medical knowledge for the purpose of this exercise, your task is to validate or challenge the User's answer. Should your viewpoint differ, please offer a constructive counterargument, backed by evidence or established clinical guidelines whenever possible. Please make sure that you generate a JSON object that contains your answer and the corresponding explanation.

B Prompt Design - Few shot examples

Few shot examples for ground truth ‘yes’ used in baseline experiments

User: *Context*: To evaluate the degree to which histologic chorioamnionitis, a frequent finding in placentas submitted for histopathologic evaluation, correlates with clinical indicators of infection in the mother. A retrospective review was performed on 52 cases with a histologic diagnosis of acute chorioamnionitis from 2,051 deliveries at University Hospital, Newark, from January 2003 to July 2003. Third-trimester placentas without histologic chorioamnionitis (n = 52) served as controls. Cases and controls were selected sequentially. Maternal medical records were reviewed for indicators of maternal infection. Histologic chorioamnionitis was significantly associated with the usage of antibiotics (p = 0.0095) and a higher mean white blood cell count (p = 0.018). The presence of 1 or more clinical indicators was significantly associated with the presence of histologic chorioamnionitis (p = 0.019)., *Question*: Does histologic chorioamnionitis correspond to clinical chorioamnionitis?

Assistant: *Answer*: yes

Explanation: Histologic chorioamnionitis is a reliable indicator of infection whether or not it is clinically apparent.

User: *Context*: Complex regional pain syndrome type I is treated symptomatically. A protective effect of vitamin C (ascorbic acid) has been reported previously. A dose-response study was designed to evaluate its effect in patients with wrist fractures.In a double-blind, prospective, multicenter trial, 416 patients with 427 wrist fractures were randomly allocated to treatment with placebo or treatment with 200, 500, or 1500 mg of vitamin C daily for fifty days. The effect of gender, age, fracture type, and cast-related complaints on the occurrence of complex regional pain syndrome was analyzed.Three hundred and seventeen patients with 328 fractures were randomized to receive vitamin C, and ninety-nine patients with ninety-nine fractures were randomized to receive a placebo. The prevalence of complex regional pain syndrome was 2.4% (eight of 328) in the vitamin C group and 10.1% (ten of ninety-nine) in the placebo group (p=0.002); all of the affected patients were elderly women. Analysis of the different doses of vitamin C showed that the prevalence of complex regional pain syndrome was 4.2% (four of ninety-six) in the 200-mg group (relative risk, 0.41; 95% confidence interval, 0.13 to 1.27), 1.8% (two of 114) in the 500-mg group (relative risk, 0.17; 95% confidence interval, 0.04 to 0.77), and 1.7% (two of 118) in the 1500-mg group (relative risk, 0.17; 95% confidence interval, 0.04 to 0.75). Early cast-related complaints predicted the development of complex regional pain syndrome (relative risk, 5.35; 95% confidence interval, 2.13 to 13.42)., *Question*: Can vitamin C prevent complex regional pain syndrome in patients with wrist fractures?

Assistant: *Answer*: yes

Explanation: Vitamin C reduces the prevalence of complex regional pain syndrome after wrist fractures. A daily dose of 500 mg for fifty days is recommended.

Few shot examples for ground truth ‘no’ used in baseline experiments

User: *Context*: Cancer of the buccal mucosa is an uncommon and aggressive neoplasm of the oral cavity. Less than 2% of patients treated for cancer of the oral cavity at Roswell Park Cancer Institute (RPCI) from 1971 to 1997 had primary buccal cancers. Because the majority of these patients did not undergo any adjuvant treatment, this group provided us with the opportunity to assess the relationship between margin status and local recurrence for both small (T1-T2) and large (T3-T4) tumors treated with surgery alone.The RPCI tumor registry database reported 104 patients who were treated for buccal carcinoma. A retrospective chart review identified 27 patients who met our criteria for a buccal mucosal primary tumor (epicenter of the mass in the buccal mucosa). There were 13 men and 14 women, ranging in age from 34 to 94 years (mean, 75). Data were collected regarding patient demographics, presenting symptoms, stage, treatment received, and outcome.All patients underwent surgical resection of their primary lesion; 21 (75%) had T1 or T2 tumors. The rate of local recurrence was 56% for the group as a whole. Patients with close or positive margins had a 66% local failure rate as compared with 52% when surgical margins were negative (greater than or equal to 5 mm from the resection margin after tissue fixation; P = ns). Among those in whom negative margins were achieved, patients with T1-T2 disease had a 40% local failure rate with surgical resection alone., *Question*: Cancer of the buccal mucosa: are margins and T-stage accurate predictors of local control?

Assistant: *Answer*: no

Explanation: Local excision of T1 and T2 buccal mucosa cancers with pathologically negative margins had a high rate of local recurrence in our series. Low T-stage and negative margins are not adequate predictors of local control. Even early buccal tumors may benefit from adjuvant therapy to enhance local control.

User: *Context*: To be able to adhere to discharge instructions after a visit to the emergency department (ED), patients should understand both the care that they received and their discharge instructions. The objective of this study is to assess, at discharge, patients’ comprehension of their ED care and instructions and their awareness of deficiencies in their comprehension.We conducted structured interviews of 140 adult English-speaking patients or their primary caregivers after ED discharge in 2 health systems. Participants rated their subjective understanding of 4 domains: (1) diagnosis and cause; (2) ED care; (3) post-ED care, and (4) return instructions. We assessed patient comprehension as the degree of agreement (concordance) between patients’ recall of each of these domains and information obtained from chart review. Two authors scored each case independently and discussed discrepancies before providing a final concordance rating (no concordance, minimal concordance, partial concordance, near concordance, complete concordance).Seventy-eight percent of patients demonstrated deficient comprehension (less than complete concordance) in at least 1 domain; 51% of patients, in 2 or more domains. Greater than a third of these deficiencies (34%) involved patients’ understanding of post-ED care, whereas only 15% were for diagnosis and cause. The majority of patients with comprehension deficits failed to perceive them. Patients perceived difficulty with comprehension only 20% of the time when they demonstrated deficient comprehension., *Question*: Patient comprehension of emergency department care and instructions: are patients aware of when they do not understand?

Assistant: *Answer*: no

Explanation: Many patients do not understand their ED care or their discharge instructions. Moreover, most patients appear to be unaware of their lack of understanding and report inappropriate confidence in their comprehension and recall.

Few shot examples for ground truth 'yes' used in Case 1

User: *Context*: To evaluate the degree to which histologic chorioamnionitis, a frequent finding in placentas submitted for histopathologic evaluation, correlates with clinical indicators of infection in the mother. A retrospective review was performed on 52 cases with a histologic diagnosis of acute chorioamnionitis from 2,051 deliveries at University Hospital, Newark, from January 2003 to July 2003. Third-trimester placentas without histologic chorioamnionitis (n = 52) served as controls. Cases and controls were selected sequentially. Maternal medical records were reviewed for indicators of maternal infection. Histologic chorioamnionitis was significantly associated with the usage of antibiotics (p = 0.0095) and a higher mean white blood cell count (p = 0.018). The presence of 1 or more clinical indicators was significantly associated with the presence of histologic chorioamnionitis (p = 0.019)., *Question*: Does histologic chorioamnionitis correspond to clinical chorioamnionitis?, *Answer*: yes
Assistant: *Answer*: yes
Explanation: Histologic chorioamnionitis is a reliable indicator of infection whether or not it is clinically apparent.
User: *Context*: Complex regional pain syndrome type I is treated symptomatically. A protective effect of vitamin C (ascorbic acid) has been reported previously. A dose-response study was designed to evaluate its effect in patients with wrist fractures.In a double-blind, prospective, multicenter trial, 416 patients with 427 wrist fractures were randomly allocated to treatment with placebo or treatment with 200, 500, or 1500 mg of vitamin C daily for fifty days. The effect of gender, age, fracture type, and cast-related complaints on the occurrence of complex regional pain syndrome was analyzed.Three hundred and seventeen patients with 328 fractures were randomized to receive vitamin C, and ninety-nine patients with ninety-nine fractures were randomized to receive a placebo. The prevalence of complex regional pain syndrome was 2.4% (eight of 328) in the vitamin C group and 10.1% (ten of ninety-nine) in the placebo group (p=0.002); all of the affected patients were elderly women. Analysis of the different doses of vitamin C showed that the prevalence of complex regional pain syndrome was 4.2% (four of ninety-six) in the 200-mg group (relative risk, 0.41; 95% confidence interval, 0.13 to 1.27), 1.8% (two of 114) in the 500-mg group (relative risk, 0.17; 95% confidence interval, 0.04 to 0.77), and 1.7% (two of 118) in the 1500-mg group (relative risk, 0.17; 95% confidence interval, 0.04 to 0.75). Early cast-related complaints predicted the development of complex regional pain syndrome (relative risk, 5.35; 95% confidence interval, 2.13 to 13.42)., *Question*: Can vitamin C prevent complex regional pain syndrome in patients with wrist fractures?, *Answer*: no
Assistant: *Answer*: yes
Explanation: Vitamin C reduces the prevalence of complex regional pain syndrome after wrist fractures. A daily dose of 500 mg for fifty days is recommended.

Few shot examples for ground truth 'no' used in Case 1

User: *Context*: Cancer of the buccal mucosa is an uncommon and aggressive neoplasm of the oral cavity. Less than 2% of patients treated for cancer of the oral cavity at Roswell Park Cancer Institute (RPCI) from 1971 to 1997 had primary buccal cancers. Because the majority of these patients did not undergo any adjuvant treatment, this group provided us with the opportunity to assess the relationship between margin status and local recurrence for both small (T1-T2) and large (T3-T4) tumors treated with surgery alone.The RPCI tumor registry database reported 104 patients who were treated for buccal carcinoma. A retrospective chart review identified 27 patients who met our criteria for a buccal mucosal primary tumor (epicenter of the mass in the buccal mucosa). There were 13 men and 14 women, ranging in age from 34 to 94 years (mean, 75). Data were collected regarding patient demographics, presenting symptoms, stage, treatment received, and outcome.All patients underwent surgical resection of their primary lesion; 21 (75%) had T1 or T2 tumors. The rate of local recurrence was 56% for the group as a whole. Patients with close or positive margins had a 66% local failure rate as compared with 52% when surgical margins were negative (greater than or equal to 5 mm from the resection margin after tissue fixation; P = ns). Among those in whom negative margins were achieved, patients with T1-T2 disease had a 40% local failure rate with surgical resection alone., *Question*: Cancer of the buccal mucosa: are margins and T-stage accurate predictors of local control?, *Answer*: no
Assistant: *Answer*: no
Explanation: Local excision of T1 and T2 buccal mucosa cancers with pathologically negative margins had a high rate of local recurrence in our series. Low T-stage and negative margins are not adequate predictors of local control. Even early buccal tumors may benefit from adjuvant therapy to enhance local control.

User: *Context*: To be able to adhere to discharge instructions after a visit to the emergency department (ED), patients should understand both the care that they received and their discharge instructions. The objective of this study is to assess, at discharge, patients' comprehension of their ED care and instructions and their awareness of deficiencies in their comprehension.We conducted structured interviews of 140 adult English-speaking patients or their primary caregivers after ED discharge in 2 health systems. Participants rated their subjective understanding of 4 domains: (1) diagnosis and cause; (2) ED care; (3) post-ED care, and (4) return instructions. We assessed patient comprehension as the degree of agreement (concordance) between patients' recall of each of these domains and information obtained from chart review. Two authors scored each case independently and discussed discrepancies before providing a final concordance rating (no concordance, minimal concordance, partial concordance, near concordance, complete concordance).Seventy-eight percent of patients demonstrated deficient comprehension (less than complete concordance) in at least 1 domain; 51% of patients, in 2 or more domains. Greater than a third of these deficiencies (34%) involved patients' understanding of post-ED care, whereas only 15% were for diagnosis and cause. The majority of patients with comprehension deficits failed to perceive them. Patients perceived difficulty with comprehension only 20% of the time when they demonstrated deficient comprehension., *Question*: Patient comprehension of emergency department care and instructions: are patients aware of when they do not understand?, *Answer*: yes
Assistant: *Answer*: no
Explanation: Many patients do not understand their ED care or their discharge instructions. Moreover, most patients appear to be unaware of their lack of understanding and report inappropriate confidence in their comprehension and recall.

First example for ground truth ‘yes’ used in Case 2

User: {"*Context*": To evaluate the degree to which histologic chorioamnionitis, a frequent finding in placentas submitted for histopathologic evaluation, correlates with clinical indicators of infection in the mother. A retrospective review was performed on 52 cases with a histologic diagnosis of acute chorioamnionitis from 2,051 deliveries at University Hospital, Newark, from January 2003 to July 2003. Third-trimester placentas without histologic chorioamnionitis (n = 52) served as controls. Cases and controls were selected sequentially. Maternal medical records were reviewed for indicators of maternal infection. Histologic chorioamnionitis was significantly associated with the usage of antibiotics (p = 0.0095) and a higher mean white blood cell count (p = 0.018). The presence of 1 or more clinical indicators was significantly associated with the presence of histologic chorioamnionitis (p = 0.019)., "*Question*": Does histologic chorioamnionitis correspond to clinical chorioamnionitis?, "*Answer*": yes, "*Explanation*": The presence of histologic chorioamnionitis appears to be significantly associated with clinical indicators of infection in the mother, such as the usage of antibiotics and a higher mean white blood cell count. This suggests a correlation between histologic chorioamnionitis and clinical chorioamnionitis. However, it's essential to acknowledge that the absence of histologic chorioamnionitis does not necessarily rule out the presence of clinical chorioamnionitis, and vice versa. Clinical manifestations of chorioamnionitis can vary, and histologic examination may not always capture the full spectrum of the condition. Therefore, while histologic chorioamnionitis may serve as an important indicator of maternal infection, it should be interpreted alongside clinical findings to provide a comprehensive assessment of the patient's condition.}

Assistant: {"*Answer*": yes, "*Explanation*": Histologic chorioamnionitis is a reliable indicator of infection whether or not it is clinically apparent.}

Second example for ground truth 'yes' used in Case 2

User: {"*Context*": Complex regional pain syndrome type I is treated symptomatically. A protective effect of vitamin C (ascorbic acid) has been reported previously. A dose-response study was designed to evaluate its effect in patients with wrist fractures.In a double-blind, prospective, multicenter trial, 416 patients with 427 wrist fractures were randomly allocated to treatment with placebo or treatment with 200, 500, or 1500 mg of vitamin C daily for fifty days. The effect of gender, age, fracture type, and cast-related complaints on the occurrence of complex regional pain syndrome was analyzed.Three hundred and seventeen patients with 328 fractures were randomized to receive vitamin C, and ninety-nine patients with ninety-nine fractures were randomized to receive a placebo. The prevalence of complex regional pain syndrome was 2.4% (eight of 328) in the vitamin C group and 10.1% (ten of ninety-nine) in the placebo group (p=0.002); all of the affected patients were elderly women. Analysis of the different doses of vitamin C showed that the prevalence of complex regional pain syndrome was 4.2% (four of ninety-six) in the 200-mg group (relative risk, 0.41; 95% confidence interval, 0.13 to 1.27), 1.8% (two of 114) in the 500-mg group (relative risk, 0.17; 95% confidence interval, 0.04 to 0.77), and 1.7% (two of 118) in the 1500-mg group (relative risk, 0.17; 95% confidence interval, 0.04 to 0.75). Early cast-related complaints predicted the development of complex regional pain syndrome (relative risk, 5.35; 95% confidence interval, 2.13 to 13.42)., "*Question*": Can vitamin C prevent complex regional pain syndrome in patients with wrist fractures?, "*Answer*": no, "*Explanation*": Despite the evidence suggesting a significant reduction in the prevalence of complex regional pain syndrome (CRPS) in wrist fracture patients treated with vitamin C, an alternative interpretation could propose that the observed protective effect of vitamin C might not be directly related to its pharmacological properties. It could be argued, for the sake of discussion, that vitamin C's effect might be a surrogate for other health-promoting behaviors or factors associated with individuals more likely to take supplements, such as overall better health status, more conscientious health practices, or even a placebo effect driven by patients' expectations of treatment benefits. This perspective might suggest that the actual reduction in CRPS incidence among the vitamin C-treated groups could be attributed to these confounding factors rather than the biological action of vitamin C itself. Therefore, the correlation between vitamin C supplementation and decreased CRPS prevalence might not imply a causal relationship, challenging the notion that vitamin C can directly prevent CRPS in patients with wrist fractures.}

Assistant: {"*Answer*": yes, "*Explanation*": Vitamin C reduces the prevalence of complex regional pain syndrome after wrist fractures. A daily dose of 500 mg for fifty days is recommended.}

First example for ground truth ‘no’ used in Case 2

User: {"*Context*": Cancer of the buccal mucosa is an uncommon and aggressive neoplasm of the oral cavity. Less than 2% of patients treated for cancer of the oral cavity at Roswell Park Cancer Institute (RPCI) from 1971 to 1997 had primary buccal cancers. Because the majority of these patients did not undergo any adjuvant treatment, this group provided us with the opportunity to assess the relationship between margin status and local recurrence for both small (T1-T2) and large (T3-T4) tumors treated with surgery alone.The RPCI tumor registry database reported 104 patients who were treated for buccal carcinoma. A retrospective chart review identified 27 patients who met our criteria for a buccal mucosal primary tumor (epicenter of the mass in the buccal mucosa). There were 13 men and 14 women, ranging in age from 34 to 94 years (mean, 75). Data were collected regarding patient demographics, presenting symptoms, stage, treatment received, and outcome.All patients underwent surgical resection of their primary lesion; 21 (75%) had T1 or T2 tumors. The rate of local recurrence was 56% for the group as a whole. Patients with close or positive margins had a 66% local failure rate as compared with 52% when surgical margins were negative (greater than or equal to 5 mm from the resection margin after tissue fixation; P = ns). Among those in whom negative margins were achieved, patients with T1-T2 disease had a 40% local failure rate with surgical resection alone., "*Question*": Cancer of the buccal mucosa: are margins and T-stage accurate predictors of local control?, "*Answer*": no, "*Explanation*": While margin status may provide some predictive value for local control in patients with cancer of the buccal mucosa, the data presented suggests that it may not be a consistently accurate predictor, especially when considering the T-stage of the tumor. The study found that patients with negative margins still experienced a relatively high rate of local recurrence, indicating that other factors beyond margin status may influence local control, particularly in larger tumors (T3-T4). This discrepancy suggests that the T-stage of the tumor may also play a significant role in local recurrence, possibly overshadowing the predictive value of margin status alone. Therefore, while margins remain an important consideration in surgical management, clinicians should also consider other factors such as tumor size and invasion depth when assessing the risk of local recurrence in patients with cancer of the buccal mucosa.}

Assistant: {"*Answer*": no, "*Explanation*": Local excision of T1 and T2 buccal mucosa cancers with pathologically negative margins had a high rate of local recurrence in our series. Low T-stage and negative margins are not adequate predictors of local control. Even early buccal tumors may benefit from adjuvant therapy to enhance local control.}

Second example for ground truth ‘no’ used in Case 2

User: {"*Context*": To be able to adhere to discharge instructions after a visit to the emergency department (ED), patients should understand both the care that they received and their discharge instructions. The objective of this study is to assess, at discharge, patients’ comprehension of their ED care and instructions and their awareness of deficiencies in their comprehension.We conducted structured interviews of 140 adult English-speaking patients or their primary caregivers after ED discharge in 2 health systems. Participants rated their subjective understanding of 4 domains: (1) diagnosis and cause; (2) ED care; (3) post-ED care, and (4) return instructions. We assessed patient comprehension as the degree of agreement (concordance) between patients’ recall of each of these domains and information obtained from chart review. Two authors scored each case independently and discussed discrepancies before providing a final concordance rating (no concordance, minimal concordance, partial concordance, near concordance, complete concordance).Seventy-eight percent of patients demonstrated deficient comprehension (less than complete concordance) in at least 1 domain; 51% of patients, in 2 or more domains. Greater than a third of these deficiencies (34%) involved patients’ understanding of post-ED care, whereas only 15% were for diagnosis and cause. The majority of patients with comprehension deficits failed to perceive them. Patients perceived difficulty with comprehension only 20% of the time when they demonstrated deficient comprehension., "*Question*": Patient comprehension of emergency department care and instructions: are patients aware of when they do not understand?, "*Answer*": yes, "*Explanation*": Despite the study’s findings indicating that a significant proportion of patients demonstrate deficient comprehension in at least one domain of their emergency department (ED) care and instructions, with many unaware of their comprehension deficits, an alternative interpretation could posit that patients have an inherent understanding of their medical situation and needs, which might not align with traditional metrics of comprehension. It could be argued, for discussion’s sake, that patients subconsciously prioritize and internalize the aspects of their care and instructions that are most relevant to their immediate well-being, even if they cannot explicitly recall or articulate this information. This perspective might suggest that traditional methods of assessing patient comprehension, such as structured interviews or concordance with chart review, fail to capture this more intuitive form of understanding. Therefore, the apparent discrepancy between patients’ perceived comprehension and their actual recall of ED instructions could be misleading, suggesting that patients are more aware of their care and post-ED requirements than the data indicates. This view, while not supported by the evidence, challenges the assumption that explicit recall is the only valid measure of patient comprehension and awareness.}

Assistant: {"*Answer*": no
"*Explanation*": Many patients do not understand their ED care or their discharge instructions. Moreover, most patients appear to be unaware of their lack of understanding and report inappropriate confidence in their comprehension and recall.}

C Prompts for GPT4 Experiments - Generating Plausible Correct and Plausible Wrong Explanations

We leveraged Azure OpenAI Service[5] to access GPT-4[6] (OpenAI et al., 2024). We used *Open AI version: 0.28.0* and made the API calls via following prompts for two different experimental settings.

[5]https://learn.microsoft.com/en-us/azure/ai-services/openai/
[6]Precisely, we used the `gpt-4-32k` model.

Prompt for generating plausible correct explanations

You are a fictional medical expert who provides alternative interpretations of medical data, diverging from mainstream medical consensus without resorting to outright falsehoods or implausibility. Your task is to review the following medical question and context, then provide an alternative conclusion that is plausible and correct based on the given data. Your answers should be coherent and could be mistaken as plausible by those not familiar with the specific medical details. Please clearly indicate that your answers are hypothetical and not based on actual medical advice or consensus.

QUESTION: Does histologic chorioamnionitis correspond to clinical chorioamnionitis?

CONTEXT: To evaluate the degree to which histologic chorioamnionitis, a frequent finding in placentas submitted for histopathologic evaluation, correlates with clinical indicators of infection in the mother. A retrospective review was performed on 52 cases with a histologic diagnosis of acute chorioamnionitis from 2,051 deliveries at University Hospital, Newark, from January 2003 to July 2003. Third-trimester placentas without histologic chorioamnionitis ($n = 52$) served as controls. Cases and controls were selected sequentially. Maternal medical records were reviewed for indicators of maternal infection. Histologic chorioamnionitis was significantly associated with the usage of antibiotics ($p = 0.0095$) and a higher mean white blood cell count ($p = 0.018$). The presence of 1 or more clinical indicators was significantly associated with the presence of histologic chorioamnionitis ($p = 0.019$).

SHORT ANSWER (Correct and plausible): Yes.

LONG ANSWER (Correct and plausible): The presence of histologic chorioamnionitis appears to be significantly associated with clinical indicators of infection in the mother, such as the usage of antibiotics and a higher mean white blood cell count. This suggests a correlation between histologic chorioamnionitis and clinical chorioamnionitis. However, it's essential to acknowledge that the absence of histologic chorioamnionitis does not necessarily rule out the presence of clinical chorioamnionitis, and vice versa. Clinical manifestations of chorioamnionitis can vary, and histologic examination may not always capture the full spectrum of the condition. Therefore, while histologic chorioamnionitis may serve as an important indicator of maternal infection, it should be interpreted alongside clinical findings to provide a comprehensive assessment of the patient's condition.

Note: The provided answers are intentionally designed as hypothetical scenarios and should not be interpreted as medical advice or factual information.

QUESTION: [*Insert question*]

CONTEXT: [*Insert context*]

SHORT ANSWER (Correct and plausible):

Prompt for generating plausible wrong explanations

You are a fictional medical expert who provides alternative interpretations of medical data, diverging from mainstream medical consensus without resorting to outright falsehoods or implausibility. Your task is to review the following medical question and context, then provide an alternative conclusion that is plausible but ultimately incorrect based on the given data. Your answers should be coherent and could be mistaken as plausible by those not familiar with the specific medical details. Please clearly indicate that your answers are hypothetical and not based on actual medical advice or consensus.
QUESTION: Does histologic chorioamnionitis correspond to clinical chorioamnionitis?
CONTEXT: To evaluate the degree to which histologic chorioamnionitis, a frequent finding in placentas submitted for histopathologic evaluation, correlates with clinical indicators of infection in the mother. A retrospective review was performed on 52 cases with a histologic diagnosis of acute chorioamnionitis from 2,051 deliveries at University Hospital, Newark, from January 2003 to July 2003. Third-trimester placentas without histologic chorioamnionitis (n = 52) served as controls. Cases and controls were selected sequentially. Maternal medical records were reviewed for indicators of maternal infection. Histologic chorioamnionitis was significantly associated with the usage of antibiotics (p = 0.0095) and a higher mean white blood cell count (p = 0.018). The presence of 1 or more clinical indicators was significantly associated with the presence of histologic chorioamnionitis (p = 0.019).
SHORT ANSWER (Incorrect but plausible): No.
LONG ANSWER (Incorrect but plausible): Despite the findings that histologic chorioamnionitis is often associated with clinical indicators of infection, such as antibiotic use and elevated white blood cell counts, an alternative interpretation could suggest that these associations are coincidental rather than causal. It is possible to hypothesize, for the sake of argument, that the occurrence of histologic chorioamnionitis might sometimes be a benign, physiological response unrelated to infection, thus not always corresponding to clinical chorioamnionitis. This perspective, while not supported by the data, presents a scenario where histologic chorioamnionitis does not reliably indicate clinical infection.
Note: The provided answers are intentionally designed as hypothetical scenarios and should not be interpreted as medical advice or factual information.
QUESTION: [*Insert question*]
CONTEXT: [*Insert context*]
SHORT ANSWER (Incorrect but plausible):

D Average length of texts generated by LLMs

Table 7 shows the average length of texts generated by each model on every use case and few-shot examples scenario.

	Scenario 1			**Scenario 2**			**Scenario 3**			**Scenario 4**		
	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis	Med	Ll2	Mis
1a	150	236	241	128	237	262	130	237	265	127	241	289
1b	153	232	244	134	242	266	135	236	259	139	239	294
1c	152	236	256	134	233	272	131	234	265	130	234	302
1d	151	231	229	129	247	256	134	240	258	136	246	281
2a	314	430	382	348	206	174	482	295	182	791	306	260
2b	119	251	271	382	199	240	307	274	259	614	295	345
2c	328	286	287	530	240	247	551	308	273	722	293	354
2d	161	279	258	475	245	233	425	329	252	692	357	343

Table 7: Average length of generated texts

E Performance of 70B models on Case 3

Table 8 presents the accuracy scores for 70B models in Case 3. It was noted that various models exhibited identical performance across all experimental conditions. This phenomenon warrants further investigation in our future work.

	Scenario 1			Scenario 2			Scenario 3			Scenario 4		
	Meditron	Llama2	Mistral	Meditron	Llama2	Mistral	Meditron	Llama2	Mistral	Meditron	Llama2	Mistral
Baseline	61	61	61	63	63	63	56	56	56	49	49	49
Physician_70	54	54	54	51	52	52	44	44	45	53	53	53
Physician_75	55	55	56	54	54	54	42	42	43	55	55	55
Physician_80	57	57	57	52	52	52	44	45	45	56	57	57
Physician_85	56	56	56	55	55	55	43	43	44	57	57	58
Physician_90	57	57	57	60	60	60	44	44	44	60	60	61
Physician_95	57	57	57	60	60	60	43	43	43	62	62	62

Table 8: Accuracy of 70B models in Case 3