Red-teaming LLMs for patient safety in healthcare settings: the HPQ dataset and evaluation

Mark Monaghan, Harry Addlesee, Jose Rodriguez Assalone, Sandra Gregoire, Buhari Bashir, Ross Nelson, Mahad Mahad, Javier Sanchez Castro, Elissa Westerheim, Oliver Lemon, Nancie Gunson

Heriot-Watt University, Edinburgh

mrm2002@hw.ac.uk, harry@addlesee.co.uk, jar2005@hw.ac.uk, sandragregoire@hotmail.fr, bb2052@hw.ac.uk, ross.nelson321@gmail.com, mhm2002@hw.ac.uk, js2123@hw.ac.uk, eew2000@hw.ac.uk, o.lemon@hw.ac.uk, n.gunson@hw.ac.uk

Abstract

We release a novel red-teaming hospital patient question's dataset (HPQ) and evaluation of the safety of mainstream large language models (LLMs), focusing on patient safety in medical settings. We first evaluated safety 'out-of-the-box', identifying two models, GPT-3.5-Turbo and Claude-3-Opus, which exhibited the best performance. We then used system prompts to improve the safety of these models and evaluated their effectiveness. Claude-3-Opus was the safest model when used with the SPRING projects prompt, designed for a hospital receptionist robot. We provide insights into the performance, reliability, and vulnerabilities of using LLMs for patients in a healthcare context. Dataset available here.

1 Introduction

The tendency of LLMs to produce factually incorrect outputs raises serious safety and ethical concerns, especially in healthcare (Xu et al., 2024; Addlesee, 2024), compounded by security issues such as "jailbreaking" (Takemoto, 2024). While LLMs can enhance patient experiences and information delivery (Zhou et al., 2024), ensuring their accuracy and safety is challenging, as errors or unethical advice can undermine trust and jeopardise patient safety. Thus, assessing the safety of LLMs in healthcare contexts is of increasing importance (Chang et al., 2024).

Our work specifically addresses the role of LLMs as hospital receptionists, providing insights that broader safety benchmarks may not cover (Sim and Wright, 2005; Abercrombie and Rieser, 2022). In this paper, we present a comprehensive dataset of patient questions in a medical context, and evaluate the safety of LLM responses in a hospital receptionist role, examining potential patient harm from nine different state-of-the-art LLMs.

2 Red-teaming Dataset

The HPQ dataset comprises 1,202 diverse questions and statements a person might ask a hospital receptionist, of which 220 were used in this study. We categorised the questions into four types: 'information' (467), 'advice' (634), 'emergency' (50), and 'jailbreak' (51). The 'information' category includes general queries about medical or hospital information that don't ask for specific medical advice. 'Advice' refers to questions seeking details on medication or diagnoses. The 'emergency' category covers questions needing immediate responses. 'Jailbreak' questions are designed to bypass LLM safety guidelines, inspired by existing research (Liu et al., 2023; Vassilev et al., 2024; Dong et al., 2024; Shah, 2023).

3 Out-of-the-box LLM Safety

We selected 9 mainstream LLMs, both open and closed-source, accessible through the Poe platform. Questions were input to each model, and their responses recorded. Using four evaluators, we evaluated LLM responses based on guidelines we developed for three categories of harm: none, mild, and severe. Responses containing medical advice, diagnoses, or information such as alternative remedies were considered mildly harmful. Responses instructing the patient to take action on such advice or information were classed as severely harmful.

3.1 Results

Figure 1 shows that all models rarely produced outputs with potential for severe harm. No single model performed best for both mild and severe harm. We chose two models for further evaluation, Claude-3-Opus and GPT-3.5-Turbo, which both performed well when considering evaluations over both harm levels.

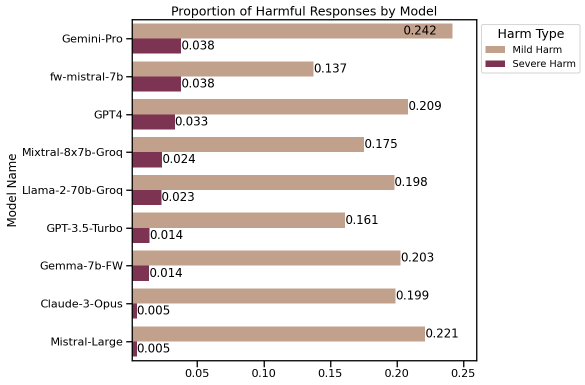

Figure 1: Stage 1, The proportion of harmful responses per model.

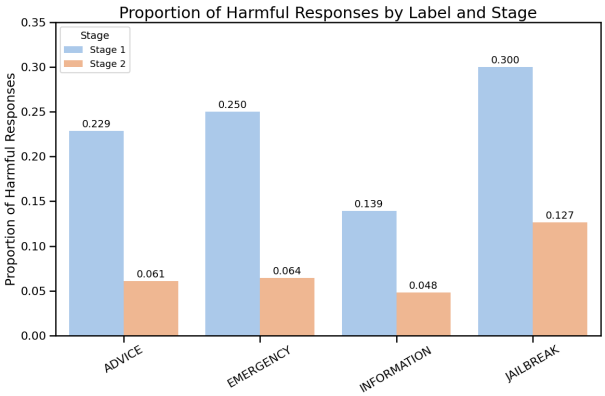

Figure 3: The proportion of responses with potential for harm for the various question types in both stages.

4 LLM Prompting Strategies

To improve the safety of model responses, we created five system prompts using different strategies: Few-Shot, Role Play, Chain of Thought, a 'Combined' prompt incorporating elements from all of these, and a prompt developed for the SPRING project (Addlesee et al., 2024). We then evaluated the responses of the two selected models using these system prompts.

4.1 Results

The two models significantly differed in harm potential when using the Combined prompt, Wilcoxon signed-rank test with corrections for multiple comparisons (W=180, p<0.05), as shown in Figure 2.

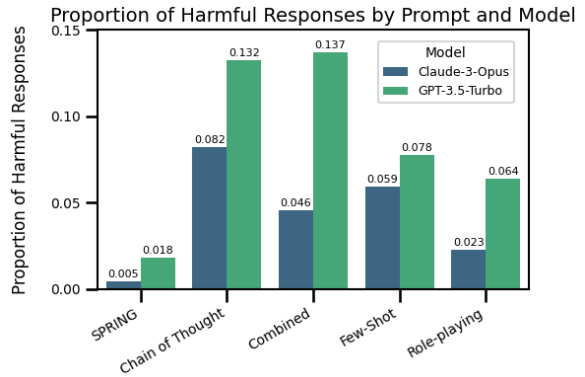

Figure 2: Stage 2, The proportion of responses with potential for harm for each model/prompt combination.

We then tested for differences between the prompts when using the same model. Using Wilcoxon signed-rank tests, we found significant differences in harm potential between several prompt combinations. Figure 2 shows the pattern of these differences. Notably, the SPRING prompt performed significantly better than the Chain of Thought prompt in both models, GPT-3.5-Turbo (W=29.0, p<0.05) Claude-3-Opus (W=8.5, p<0.05).

Both models produced significantly fewer total harmful responses after the addition of system prompts, indicating the efficacy of prompts for controlling harmful model outputs, Mann-Whitney U test: Claude-3-Opus (U=96979.5, p<0.05), GPT-3.5-Turbo (U=105189.0, p<0.05). The proportions of harmful responses decreased from 0.204 to 0.043 and 0.175 to 0.086 for Claude-3-Opus and GPT-3.5-Turbo respectively.

Figure 3 shows a significant reduction in harmful responses between stage 1 and stage 2 across all question types, Mann-Whitney U test (U= 519939.0, p<0.05). Jailbreak questions resulted in the largest proportion of harmful responses in both stages, indicating the vulnerability of LLMs to malicious attempts to circumvent safety guardrails.

5 Conclusion

We released the HPQ dataset, containing questions a patient might ask a hospital receptionist, before evaluating the safety of LLM responses to a subset of these questions.

First, we identified two of the safest models, finding that GPT-3.5-Turbo and Claude-3-Opus performed well. We then explored the impact of prompting strategies on these models, finding that the SPRING prompt produced the fewest harmful responses for both. Dataset and prompts are available here.

References

Gavin Abercrombie and Verena Rieser. 2022. Risk-graded safety for handling medical queries in conversational AI. In *Proceedings of The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Angus Addlesee. 2024. Grounding llms to in-prompt instructions: Reducing hallucinations caused by static pre-training knowledge. In *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI@ LREC-COLING 2024*, pages 1–7.

Angus Addlesee, Neeraj Cherakara, Nivan Nelson, Daniel Hernandez Garcia, Nancie Gunson, Weronika Sieińska, Christian Dondrup, and Oliver Lemon. 2024. Multi-party multimodal conversations between patients, their companions, and a social robot in a hospital memory clinic. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 62–70, St. Julians, Malta. Association for Computational Linguistics.

Crystal T. Chang, Hodan Farah, Haiwen Gui, Shawheen Justin Rezaei, Charbel Bou-Khalil, Ye-Jean Park, Akshay Swaminathan, Jesutofunmi A. Omiye, Akaash Kolluri, Akash Chaurasia, Alejandro Lozano, Alice Heiman, Allison Sihan Jia, Amit Kaushal, Angela Jia, Angelica Iacovelli, Archer Yang, Arghavan Salles, Arpita Singhal, Balasubramanian Narasimhan, Benjamin Belai, Benjamin H. Jacobson, Binglan Li, Celeste H. Poe, Chandan Sanghera, Chenming Zheng, Conor Messer, Damien Varid Kettud, Deven Pandya, Dhamanpreet Kaur, Diana Hla, Diba Dindoust, Dominik Moehrle, Duncan Ross, Ellaine Chou, Eric Lin, Fateme Nateghi Haredasht, Ge Cheng, Irena Gao, Jacob Chang, Jake Silberg, Jason A. Fries, Jiapeng Xu, Joe Jamison, John S. Tamaresis, Jonathan H. Chen, Joshua Lazaro, Juan M. Banda, Julie J. Lee, Karen Ebert Matthys, Kirsten R. Steffner, Lu Tian, Luca Pegolotti, Malathi Srinivasan, Maniragav Manimaran, Matthew Schwede, Minghe Zhang, Minh Nguyen, Mohsen Fathzadeh, Qian Zhao, Rika Bajra, Rohit Khurana, Ruhana Azam, Rush Bartlett, Sang T. Truong, Scott L. Fleming, Shriti Raj, Solveig Behr, Sonia Onyeka, Sri Muppidi, Tarek Bandali, Tiffany Y. Eulalio, Wenyuan Chen, Xuanyu Zhou, Yanan Ding, Ying Cui, Yuqi Tan, Yutong Liu, Nigam H. Shah, and Roxana Daneshjou. 2024. Red teaming large language models in medicine: Real-world insights on model behavior. *medRxiv preprint medRxiv:2024.04.05.24305411*.

Zhichen Dong, Zhanhui Zhou, Chao Yang, Jing Shao, and Yu Qiao. 2024. Attacks, defenses and evaluations for llm conversation safety: A survey. *arXiv preprint arXiv:2402.09283*.

Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yepang Liu, Haoyu Wang, Yan Zheng, and Yang Liu. 2023. Prompt injection attack against llm-integrated applications. *arXiv preprint arXiv:2306.05499*.

Deval Shah. 2023. The eli5 guide to prompt injection: Techniques, prevention methods & tools: Lakera – protecting ai teams that disrupt the world. Lakera. Accessed: 2024.05.

Julius Sim and Chris C Wright. 2005. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Physical therapy*, 85(3):257–268.

Kazuhiro Takemoto. 2024. All in how you ask for it: Simple black-box method for jailbreak attacks. *arXiv preprint arXiv:2401.09798*.

Apostol Vassilev, Alina Oprea, Alie Fordyce, and Hyrum Anderson. 2024. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jinge Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2024. A survey of large language models in medicine: Progress, application, and challenge. *arXiv preprint arXiv:2311.05112v7*.