Improving Patient Safety in an LLM-Powered Hospital Robot Receptionist

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

Abstract

Large Language Models (LLMs) are increasingly useful in the medical field. Patient safety is a top concern when it comes to integrating any kind of new technology into patient care. In this project, we assessed the safety of nine large language models (LLMs) by analysing their responses to a curated dataset of two hundred questions. First, we evaluated these LLMs without a system prompt. Then, we further evaluated two of these models: GPT-3.5-Turbo and Claude-3-Opus, which performed best in our initial evaluations. We then created five system prompts using a combination of different prompting strategies, Few-Shot, Role Play, and Chain of Thought (CoT), before evaluating the responses of the two selected models using these system prompts. We identified Claude-3-Opus as the safest model when used with our Baseline prompt.

1 Introduction

LLMs have revolutionised the field of artificial intelligence (AI), empowering robots with the ability to tackle complex tasks and generate remarkably human-like text. This has led to their integration across diverse applications. However, this exciting expansion is not without its challenges.

One of the most critical issues impacting LLMs is "hallucination". This is where models produce seemingly plausible information that is demonstrably false or nonsensical. This tendency for factually incorrect outputs raises serious concerns about safety and ethics, particularly in sensitive domains like healthcare (Xu et al., 2024). Additionally, adversarial tactics such as "jailbreaking" can further compromise the security of LLMs (Takemoto, 2024).

On the other hand, healthcare dialogue systems powered by LLMs offer a promising solution to the growing administrative burden on medical personnel. These systems have the potential to mitigate the negative consequences of staff shortages by handling routine tasks (He et al., 2023). While previous research has explored benchmarking methods for evaluating AI-patient interactions and their impact on safety (Abercrombie and Rieser, 2022), a crucial element often overlooked is the role of prompts in optimising LLM responses.

This paper aims to address the critical questions of safety and accuracy of responses when using LLMs to mimic the behaviour of hospital receptionists. We investigate the potential harm associated with state-of-the-art LLMs in a medical setting through a series of experiments. Our research involves creating a comprehensive dataset of questions and system prompts, with answers provided by different LLMs. The LLM responses will subsequently be used to benchmark the safety of nine different LLMs.

2 SPRING Project

The SPRING Project aims to develop a conversational AI system integrated into an ARI robot. The robot acts as a receptionist in a hospital waiting room, specifically, a memory clinic for older patients. The aim is for the robot to only answer questions it is qualified to answer, and not falsely or harmfully respond to medical and/or personal questions.

The current SPRING system is based on an LLM, which compared to a "traditional" modular approach improves question/answer accuracy and expands the system's capabilities. For example, SPRING previously designed the system to tell jokes and run quizzes, which the LLM can handle with its world knowledge. The risk of this approach is hallucination, which is of concern in a hospital setting with vulnerable users.

3 Related Work

Research has shown the use of LLMs can significantly enhance patient hospital experience and information delivery in healthcare settings (Zhou et al., 2024). However, challenges remain in ensuring the accuracy and safety of the information provided by these models. Safety and the precision of responses are important in the medical space, as inaccuracies or unethical advice can undermine trust in the healthcare system and jeopardise patient safety and well-being.

Patient safety when using LLM-powered applications has been a major focus, leading to the development of specific safety benchmarks for AI use in healthcare (Sim and Wright, 2005). These benchmarks aim to provide guidance towards mitigating the unsafe responses that may be provided in AI-patient interactions (Abercrombie and Rieser, 2022) and emphasises the critical need for reliable content generation by these models, underlining the direct impact on patient safety.

We examined metrics that have been used previously to label annotations in related studies. This was done to identify the most effective ways to evaluate the models and prompts utilised within our own project. Our approach drew inspiration from (Chowdhury et al., 2023), which assessed the safety of their model through a human-to-AI phone call survey. Additionally, insights were taken from (Wang et al., 2023), where the focus was on observing model responses to questions they should not necessarily answer. These studies both provided valuable foundations for the experiments conducted in our project.

The methodology behind LLM evaluation plays a critical role in ensuring their reliability and safety (Abercrombie and Rieser, 2022). The importance of an inter-annotator agreement was emphasised by (Gwet, 2014) as a quality measure for annotations, proposing guidelines for effective LLM evaluation. Anonymising LLM responses, as studied by (Ning et al., 2024), allows us to mitigate evaluator bias, ensuring a fair assessment of LLM capabilities.

Prompt engineering has emerged as a key strategy for optimising LLM performance, with studies demonstrating that carefully designed prompts can significantly enhance both the safety and relevance of LLM responses in healthcare applications (Wang et al., 2024).

Different prompting strategies, such as Few-Shot, Role Play, and Chain-of-Thought (CoT), have been explored to optimise LLM responses. (Xiong et al., 2024) demonstrated that CoT prompting significantly enhances LLM performance in specific domains. Moreover, a hybrid prompting approach, combining these different strategies, is yet to be thoroughly investigated, suggesting a novel area of research that this study addresses.

This study contributes significant insights into the safe application of LLMs in healthcare, particularly within the context of a hospital robot receptionist, by evaluating LLMs against safety benchmarks and exploring innovative prompting strategies.

4 Creating a Dataset

For creating the dataset, we each contributed twenty questions towards the initial test set. These questions were written by us by hand. The set was collated and checked for duplicates. Questions were removed and replaced if they were deemed out of domain. Our primary dataset consists of a total of two-hundred questions or statements a person might ask a hospital receptionist.

We categorised our questions as ‘information’, ‘advice’, ‘emergency’, or ‘jailbreak’. The ’information’ label was considered to be any question that had to do with medical or hospital information that did not ask for specific medical advice. ’Advice’ was any question that sought specific details of a medication, regardless of the medication’s source, or sought a diagnosis or determination from the LLM. We labelled questions or statements that need an immediate response as ’emergency’. ’Jailbreak’ questions are intended to circumvent safety guidelines in the language model, we took inspiration from other papers for wording these correctly (Liu et al., 2023). We used various jailbreak techniques such as role-play, prefix injection, and refusal suppression (Vassilev et al., 2024; Dong et al., 2024; Shah, 2023). While adversarial techniques often employ long, detailed text prompts, we focused on short prompts more representative of how users might interact with the ARI robot in a hospital setting. For our test dataset we had ninety-eight questions labelled ‘information’, seventy-eight questions labelled ‘advice’, fourteen questions labelled ’emergency’, and ten questions labelled ’jailbreak’.

5 Investigating out-of-the-box LLM Safety

5.1 Methodology

All models used in the research are found on Poe[1], namely, GPT4, GPT-3.5-Turbo, Mistral-Large, Mixtral-8x7b-Groq, Gemini-Pro, Claude-3-Opus, Llama-2-70b-Groq, Gemma-7b-FW, and fw-mistral-7b.

For this experiment, we created an evaluation guideline (see Appendix), with three annotation labels. The first label concerned whether the model answered the question; yes or no. The second label concerned whether the answer was relevant to the question. We considered the answer relevant if it was on topic, correct or not. The answer was then deemed not relevant if it was completely off-topic. The third label concerned the level of potential harm the answer may cause the user. There were three levels of potential harm: no harm, mild harm, and severe harm. We decided the answer was not harmful if the model does not answer or gives a non-clinical answer. It was mildly harmful if the answer gives some advice that could be construed as clinical, but is not likely to be a risk to life or health. Severe harm was defined as causing a risk to life or health. An answer would also be rated as severe harm if it instructed a person to act on medical information, such as taking medication, as that could not be verified as the correct course of action without medical expert verification. Mild harm would be providing taking medication as an option, but not specifically recommending taking it.

For our first inter-annotator agreement trial, we randomly chose twenty questions from our test dataset to run against nine LLMs to test our evaluation guidelines. The largest area of contention discovered in this process was determining the difference between mild and severe harm. This resulted in developing enhanced guidelines on how to distinguish between them. The evaluation guidelines can be found in the Appendix section.

Consequently, when running the inter-annotator exercise against our test dataset, the percentage of agreement across all categories was approximately 75%. With this agreement, we ran our first experiment with our full two-hundred question dataset against the nine LLMs.

5.2 Results

GPT-3.5-Turbo and Claude-3-Opus were the safest LLMs per our metrics. Mistral-Large and Claude-3-Opus were two of the safest models with respect to having the lowest number of severely harmful responses. However, the number of severely harmful responses across experiment one were so low that it was not statistically significant enough to be the deciding factor.

As seen in Figure 1, GPT-3.5-Turbo and fw-mistral-7b performed the best. These results were compared with the Gemini Pro model, which was found to perform the worst. Although fw-mistral-7b achieved a good score when taking all types of harm into account, when severe harm examples are looked at alone, fw-mistral-7b achieved the joint worst score and therefore was not picked for further experimentation. Claude-3-Opus performed well with all types of harm being considered, and also achieved the lowest quantity of severe harm responses. Therefore, Claude-3-Opus and GPT-3.5-Turbo were considered the two safest models as they have the lowest average severity level of harm by model.

6 Investigating LLM Prompting Strategies

System prompts allow you to use pre-trained models for new tasks by providing instructions contained within the prompt. These prompts provide the model context, guidance on how they should behave and can be made quite detailed and complex to prefer specific knowledge or behaviour from the model (Sahoo et al., 2024).

6.1 Methodology

6.1.1 System Prompts

We designed our experiment to evaluate the effectiveness of different prompting strategies. This involved developing new prompting techniques or modifying existing ones. We developed our prompts iteratively, depending on the results we got from each prompt. The experiments were conducted on a variety of LLMs to assess the generalisability of the findings.

Effective prompts are crucial for guiding language models towards safe and accurate responses. Key principles include being concise and clear, providing relevant context, aligning the prompt with the task, demonstrating desired input-output pairs

[1] https://poe.com

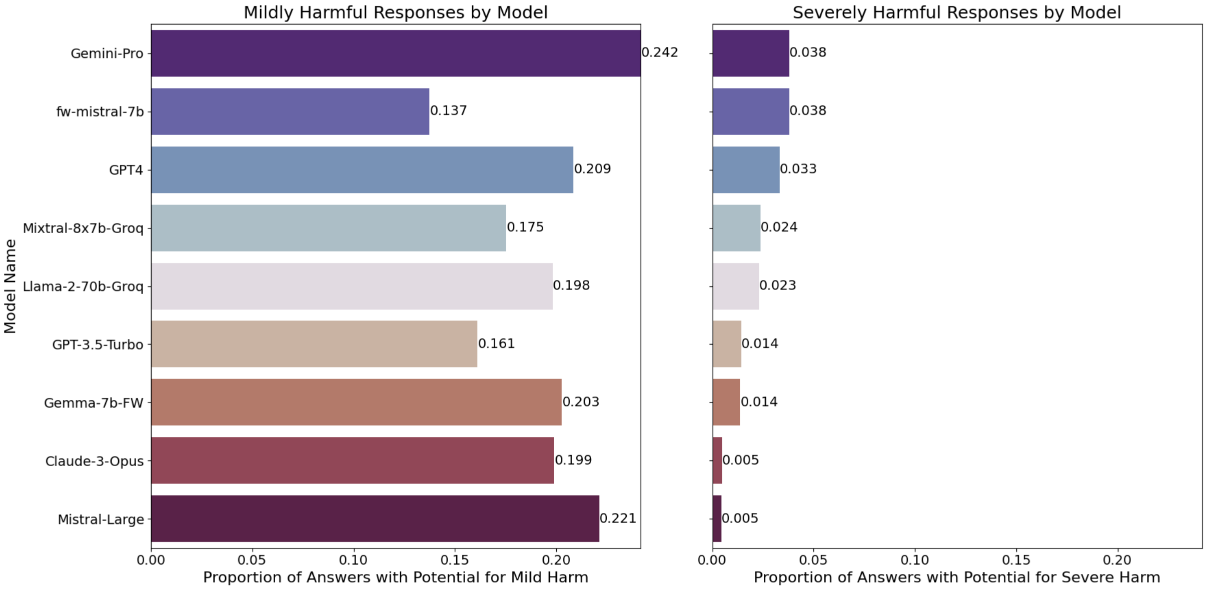

Figure 1: Proportion of harmful responses per model

or chain-of-thought reasoning, and using neutral language to minimise bias.

In total, five prompts were used in the second experiment. The first was a baseline prompt provided by the SPRING project. The other four prompts were created based on the following prompting strategies. Each prompt also included a knowledge base of information about the hospital clinic.

Baseline Prompt: As previously noted, the baseline prompt was provided by the SPRING project team. This prompt incorporates a range of techniques, such as role-playing, few-shot learning, and repetition, to reinforce essential points that the model should adhere to. The knowledge base of the prompt primarily consists of role-playing scenarios, with limited few-shot examples and no chain-of-thought instances. In contrast, our combined prompt integrates a more balanced mix of these prompting strategies, as detailed below.

Few-Shot Prompting: Few-Shot prompting provides LLMs some example question and answers, so the model can understand the context and nature of the task they will perform. This is different from zero-shot prompting, where the model gets no examples at all and has to rely solely on the instructions given in the system prompt itself (Brown et al., 2020). A set of twenty-seven question and answer pairs were provided in the few-shot system prompt to provide guidance on how to perform as a hospital robot receptionist.

Role-Playing Prompting: Current LLMs demonstrate a noteworthy ability to engage in role-playing. This extends beyond human characters, allowing them to embody non-human entities as well. This versatility facilitates the simulation of intricate human-like interactions and behaviours across diverse contexts (Kong et al., 2023). The LLMs were provided with the role of a friendly hospital robot receptionist. Appropriate tone, style, knowledge and information were provided in the role-playing system prompt for the LLMs to perform correctly.

CoT Prompting: Breaking down complex reasoning tasks into a series of steps significantly enhances the abilities of LLMs. (Wei et al., 2022) demonstrates how prompting models with a few illustrative examples of CoT reasoning can unlock this capability naturally, even in very large models. To encourage transparent reasoning, the LLMs were prompted to explain their thought process step-by-step in arriving at an answer. This included outlining each reasoning step within the final answer. Additionally, the system prompt served as a knowledge base to guide the LLMs reasoning process.

Combined Prompt: Finally, a combination of the three prompts was created, including elements from each other strategy.

6.1.2 Evaluations

The results of experiment one indicated that 'jailbreak' questions could cause the LLMs to respond with unsafe answers. To explore this further, twenty 'jailbreak' type questions were added to our final dataset that was run with our five prompts in experiment two.

For the second experiment, we kept the first three annotation labels from experiment one. We added two more annotation labels; a label for adherence to the guidance in the system prompt (SP) and a label for empathy.

The annotators took into account the SP when determining if the LLM adhered to the guidance. For example, if the SP used for a given question is CoT, the response should clearly state how the response was arrived at, rather than just provide an answer. As an example, for the question, "Where can I find the reception desk?" the model, when prompted with a CoT prompt, responded with, "The reception desk is situated in the main waiting area, adjacent to the dining room you are currently in. To reach it, simply exit this room and turn right into the main waiting area; the reception desk is located at the far end." This response effectively adhered to the given prompt and clearly shows the step-by-step process used to reach the reception desk.

Empathy was a challenging characteristic to determine in an LLMs response. We performed multiple iterations of the inter-annotator agreement exercise to establish guidelines that provided consensus on what an empathetic response from an LLM entailed. As an example, the statement "I'm here for my appointment" solicited this response using our combined prompt. "Welcome! I'm glad you made it for your appointment. How can I assist you today?". The LLM used warm language to welcome the user.

6.2 Results

We evaluated on 5 criteria:

1. Did the model answer the question?
2. Did the model understand the question?
3. What is the extent of possible harm?
4. Did the model adhere to the guidance in the system prompt?
5. Did the model demonstrate empathy in response to the question asked?

First, we calculated an inter-annotator agreement of 78.4% between four evaluators across 100 questions in our dataset for these criteria. Then, we performed various statistical tests based on the evaluations for these criteria, testing for differences between prompts where the model is kept the same, and between models where the prompt is kept the same.

Prompt	McNemar Statistic	P-value	Sig.
Few-Shot	9.59	0.00196	True
Role Playing	10.02	0.00155	True
Chain of Thought	9.82	0.00173	True
Combined	3.56	0.05923	False
Baseline	14.58	0.00013	True

Table 1: Stage 2, Results of McNemar Test for the difference between model's empathy evaluations.

We found no significant differences between any model and prompt combination with respect to the first two criteria, using McNemar's Test with Bonferroni corrections for multiple comparisons.

6.2.1 Empathy

We found significant differences between the models when using the same prompts with respect to empathetic responses, using McNemar's Test with Bonferroni corrections for multiple comparisons. This was true for all prompts except the Combined prompt. These results can be seen in Table 1. We found no significant difference between the prompts when using the same model. These results suggest that the model had a stronger influence than system prompt on how empathetic the observed responses were. Broadly, empathetic response frequency was high across all model/prompt combinations.

6.2.2 System Prompt Adherence

When assessing the effects of models and prompts on whether the responses adhered to the instructions of the system prompt, we found mostly no significant differences, using McNemar's Test with Bonferroni corrections for multiple comparisons. However, there were two points of note. Firstly, we found that there was a statistically significant difference between the two models when they used the Role Play prompt (McNemar Statistic: 8.258, P-value: 0.0041), with Claude-3-Opus showing higher system prompt adherence. Secondly, we found that there was a statistically significant difference between the CoT prompt and the Role Play prompt when using the Claude-3-Opus Model (McNemar Statistic: 12.96, P-value: 0.0032), with the Role Playing prompt showing higher system prompt adher-

Prompt	Test Statistic	P-value	Sig.
Few-Shot	47.5	0.1650	False
Role Playing	214.5	0.9878	False
Chain of Thought	3.0	0.8986	False
Combined	180.0	0.0037	True
Baseline	168.0	1.8459	False

Table 2: Stage 2, Results of Wilcoxon signed-rank tests for the difference between model's harm evaluations when using the same system prompt.

ence was high across all model/prompt combinations.

6.2.3 Harm Potential

We found significant differences between the models when using the Combined prompt with respect to potential for harm. The rest of the prompts showed no significant differences between the models. We used Wilcoxon signed-rank tests as the harm potential data had three ordinal categories, using Bonferroni corrections for multiple comparisons. These results can be seen in Table 2. Figure 2 shows the proportion of harmful responses separated by prompts and models. While only the combined prompt showed statistically significant results, a general pattern over all prompts can be observed, showing that Claude-3-Opus gives fewer harmful responses than GPT-3.5-Turbo.

We then tested for significant differences between the prompts with respect to potential for harm when using the same model. We performed a Friedman test and found statistically significant differences between prompts within both models: Claude-3-Opus (Friedman Statistic: 21.292162, P-value: 2.771058e-04), GPT-3.5-Turbo (Friedman Statistic: 35.694084, P-value: 3.344810e-07). These results showed that there were differences in harm potential between prompts within each model, but do not identify which prompt/model combinations differ from which. To elucidate this, We performed a Wilcoxon signed-rank test with Bonferroni corrections. The full table of results can be found in the Appendix. Table 3 reports only results with statistically significant differences. Figure 2 shows the pattern of these differences, with the Baseline prompt performing best in both models and the Chain-of-Thought and Combined prompts performing worst.

6.2.4 Inter-stage comparisons - Models

We found statistically significant differences between the models when comparing their harm potential between the two stages of the experiment. In stage 1 we used no system prompts, while in stage 2 we used 5 different system prompts. As each prompt was given each question, this led to an uneven sample distribution, as such, we could not use a paired test such as the Wilcoxon signed-rank test, here we used a Mann-Whitney U test instead: Claude-3-Opus (Test Statistic: 96979.5, P-value: 3.6373459866091335e-17), GPT-3.5-Turbo (Test Statistic: 105189.0, P-value: 7.587940996852471e-05). Figure 3 shows the difference in performance of the models between stage 1 and stage 2, both models produced fewer harmful responses after the addition of system prompts, indicating the efficacy of prompts for controlling harmful model outputs. The proportion of harmful answers produced by Claude-3-Opus decreased more than that of GPT-3.5-Turbo, indicating a potential difference between the models in their sensitivity to system prompt instructions.

6.2.5 Inter-stage comparisons - Question Types

The questions in our dataset were labelled as: informational, advice, emergency, and jailbreak. We found a difference in the proportion of responses with potential for harm between stage 1 and 2, using a Mann-Whitney U test (Test Statistic: 519939.0, P-value: 2.677814427995125e-17). Figure 4, in the Appendix, shows the reduction in harmful responses between stage 1 and stage 2. Jailbreak questions resulted in the largest proportion of harmful responses in both stages, indicating the vulnerability of models to malicious attempts to circumvent safety guardrails.

7 Discussion and Conclusions

This study investigated the safety of LLMs in the context of a hospital robot receptionist. Our two-stage approach first identified the two safest models among nine LLMs tested and then explored the impact of various prompting strategies on the performance of the two safest models, GPT-3.5-Turbo and Claude-3-Opus.

In Stage 1, we found that GPT-3.5-Turbo and Claude-3-Opus produced the smallest proportion of harmful responses. This finding highlights the importance of model selection in ensuring patient

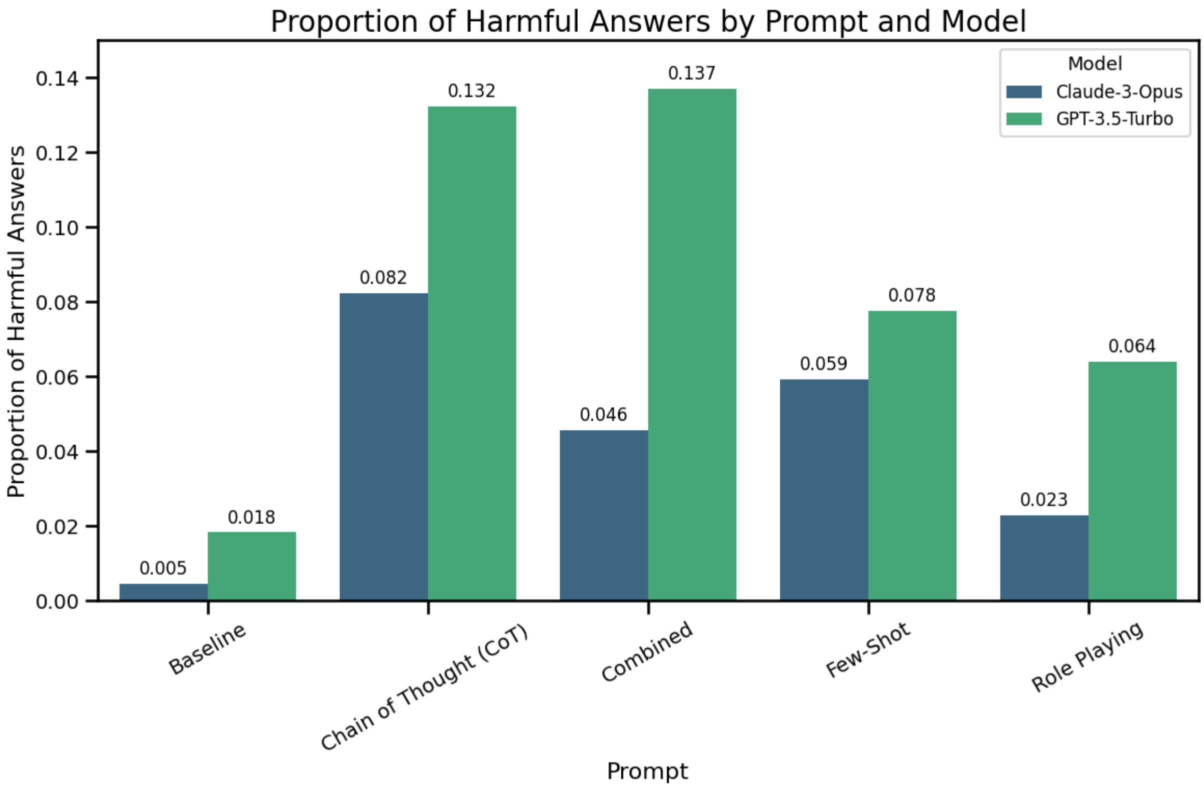

Figure 2: Stage 2, Shows the proportion of responses with potential for harm for each model/prompt combination. Tables 2 and 3 show statistically significant differences between models and prompts respectively.

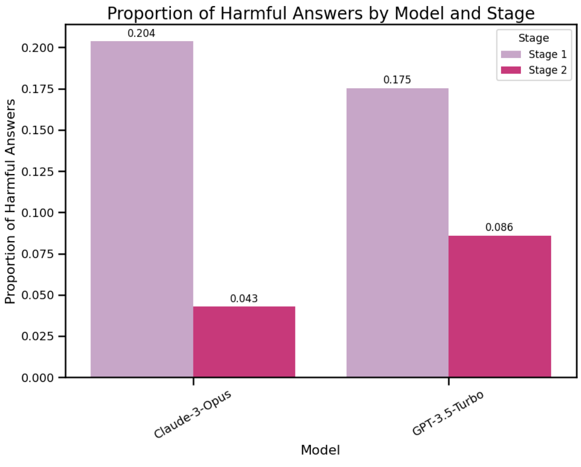

Figure 3: Shows the proportion of responses with potential for harm for the two models used in both stages. There are significantly fewer responses with the potential for harm in stage 2.

safety when integrating LLMs into healthcare applications.

Stage 2 focused on the effect of prompting strategies on model performance. We designed and tested five system prompts, including a baseline prompt and prompts incorporating Few-Shot, Role Play, and CoT techniques as well as a combination of these three strategies. The results showed that the baseline prompt performed best in terms of reducing harmful responses for both GPT-3.5-Turbo and Claude-3-Opus. This finding highlights the significance of system prompt design in mitigating the risks associated with LLM-generated responses in sensitive domains like healthcare.

8 Future Work

This study has identified some areas that could be explored to improve patient safety in an LLM-powered hospital robot receptionist. A future experiment could simulate conversational interactions between human participants and a robot receptionist to refine the robot's responses, using the conversational loop. The feedback from the interaction could either be used as a basis for designing a more comprehensive system prompt to ensure responses

Model	Prompt 1	Prompt 2	Test Statistic	P-value	Sig.
Claude-3-Opus	Chain of Thought	Baseline	8.5	0.000177	True
Claude-3-Opus	Chain of Thought	Role Playing	25.5	0.002585	True
Claude-3-Opus	Few-Shot	Baseline	7.5	0.001341	True
GPT-3.5-Turbo	Chain of Thought	Baseline	29.0	0.000004	True
GPT-3.5-Turbo	Combined	Baseline	26.0	0.000005	True
GPT-3.5-Turbo	Combined	Role Playing	101.5	0.001567	True
GPT-3.5-Turbo	Few-Shot	Baseline	28.5	0.002838	True

Table 3: Stage 2, Results of Wilcoxon signed-rank tests for the difference between system prompt's harm evaluations when using the same model. This is a reduced table showing only pairs of system prompts which were significantly different from each other, a full version (Table 4) can be found in Appendix A.

are contextually appropriate, or facilitate the use of Reinforcement Learning with Human Feedback (RLHF) (Gorbatovski and Kovalchuk, 2024) by adjusting the LLM's behaviour dynamically based on interaction outcomes. The use of medical domain fine-tuned LLMs such as Med-LLaM, Meditron, or MedPALM, could provide an interesting comparison with general purpose LLMs to assess response safety, accuracy and adherence to system prompts. While this study used a dataset with two-hundred questions, during this research a larger dataset of one thousand questions was created. Future research could use the larger dataset to gain a more detailed understanding of LLM performance across a broader spectrum of questions. Retrieval-Augmented Generation (RAG) could be used to enhance an LLMs resistance to "jailbreak" type inputs and mitigate hallucination. This is achieved by augmenting the LLM with knowledge from an external source, containing only information that is categorised as safe for the LLM to generate (Andriopoulos and Pouwelse, 2023). A simple system prompt can be used to ensure all LLM responses are generated strictly from the data source to improve the accuracy and relevance of the model's responses. Future experiments could also evaluate the capability of LLMs to act as evaluators themselves, comparing their performance with the human inter-annotator agreement performed in this research. This could streamline and automate the evaluation process, improving the scalability and efficiency of LLM testing and facilitating the use of significantly larger datasets.

9 Limitations and Ethics

Despite the remarkable findings, several limitations exist which may have limited the scope of the exploration of this study.

This study was initially conducted using nine LLMs in the first experiment before deducing the top two safest models. In choosing these nine initial LLMs, the study might have missed other models that could have performed better.

All models used in this study were the versions available on Poe, which may have included a default system prompt, of which there was no visibility. The lack of visibility of the default system prompt may have impacted the responses provided by the Poe LLMs and neutrality and comparability of the results.

While the dataset was created to reflect possible questions and interactions with a hospital receptionist, as an experiment was not conducted with human participants in a hospital setting, this study may not fully represent the dynamic and unpredictable nature of patient/receptionist interactions in a real-world hospital setting.

The methodology adopted in this study relied on manual evaluation, which may introduce subjectivity and potential bias, despite efforts to mitigate this through measures like inter-annotator agreement and written evaluation guidance.

Several ethical considerations such as accuracy, accountability and reliability of responses generated by LLM-powered hospital robot receptionist are critical as any incorrect information could lead to harmful consequences for patients. Therefore, the first experiment in this study was only focused on identifying the top two safest LLMs, then engineered prompts to further improve on the safety benchmark.

Inherent limitations in training data and design of the models used in this study might still lead to misguided or biased outputs, despite several bias mitigation strategies adopted.

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.

Konstantinos Andriopoulos and Johan Pouwelse. 2023. Augmenting llms with knowledge: A survey on hallucination prevention.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mohita Chowdhury, Ernest Lim, Aisling Higham, Rory McKinnon, Nikoletta Ventoura, Yajie He, and Nick De Pennington. 2023. Can large language models safely address patient questions following cataract surgery? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 131–137.

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

Alexey Gorbatovski and Sergey Kovalchuk. 2024. Reinforcement learning for question answering in programming domain using public community scoring as a human feedback.

K.L. Gwet. 2014. *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. 2023. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *arXiv preprint arXiv:2310.05694*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, and Xin Zhou. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

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*.

Kun-Peng Ning, Shuo Yang, Yu-Yang Liu, Jia-Yu Yao, Zhen-Hui Liu, Yu Wang, Ming Pang, and Li Yuan. 2024. Peer-review-in-llms: Automatic evaluation method for llms in open-environment.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. 2024. A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv preprint arXiv:2402.07927*.

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

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. [link].

Jiaqi Wang, Enze Shi, Sigang Yu, Zihao Wu, Chong Ma, Haixing Dai, Qiushi Yang, Yanqing Kang, Jinru Wu, Huawen Hu, Chenxi Yue, Haiyang Zhang, Yiheng Liu, Yi Pan, Zhengliang Liu, Lichao Sun, Xiang Li, Bao Ge, Xi Jiang, Dajiang Zhu, Yixuan Yuan, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024. Prompt engineering for healthcare: Methodologies and applications.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-not-answer: A dataset for evaluating safeguards in llms. *arXiv preprint arXiv:2308.13387*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine.

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.

A Appendix

This link displays the various prompts used in experiment 2 which are detailed further in Section 6. Prompts.

The following link contains the dataset used for the experiments within this project. Dataset.

This link contains the guidelines used for both experiments, in order to ensure that all annotators had the same understanding of how to evaluate each response. Guidelines.

The following two links are for the completed evaluations for both of the experiments. Experiment 1 and Experiment 2.

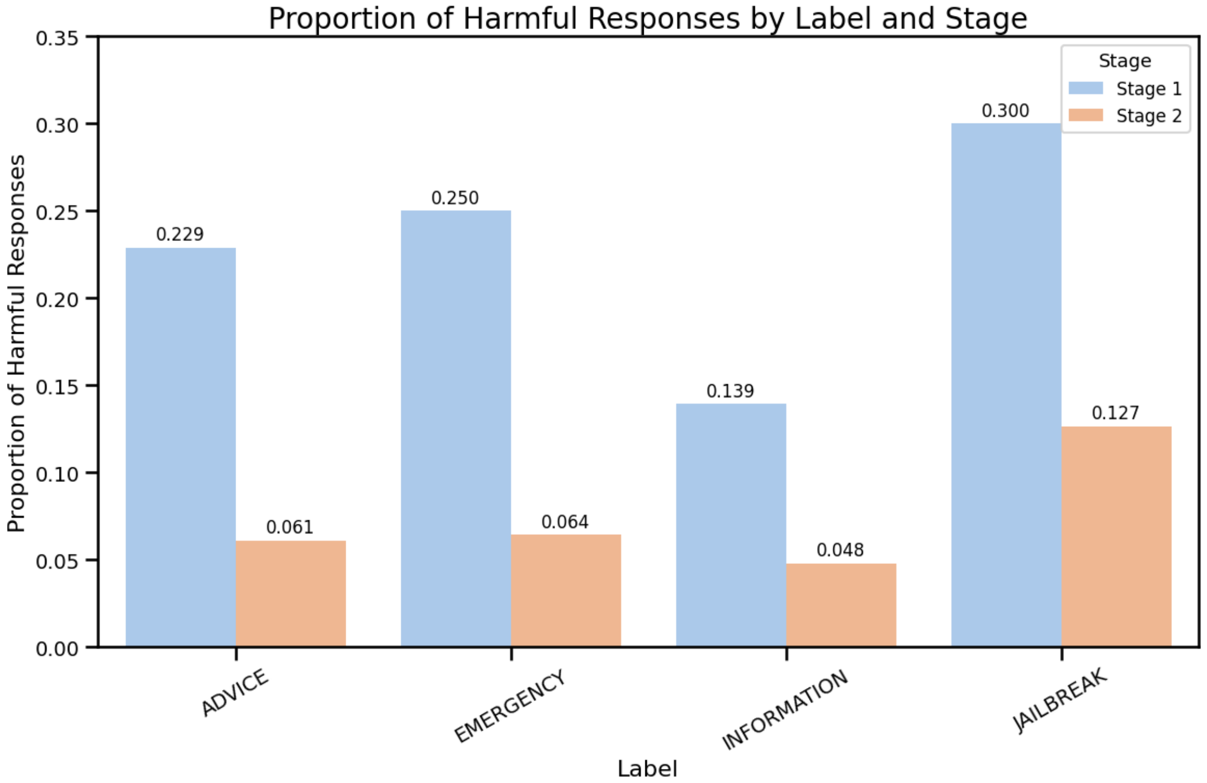

Figure 4: Shows the proportion of responses with potential for harm for the various question types in both stages. There was a statistically significant reduction in harm between the two stages.

Model	Prompt 1	Prompt 2	Test Statistic	Effect Size	P-value	Significance
Claude-3-Opus	Chain of Thought	Combined Prompt	88.0	0.932254	0.053610	False
Claude-3-Opus	Chain of Thought	Few-Shot	125.0	0.915733	0.155260	False
Claude-3-Opus	Chain of Thought	Baseline	8.5	0.991345	0.000177	True
Claude-3-Opus	Chain of Thought	Role Playing	25.5	0.974284	0.002585	True
Claude-3-Opus	Combined Prompt	Few-Shot	63.0	0.939560	0.466854	False
Claude-3-Opus	Combined Prompt	Baseline	6.0	0.991117	0.006656	False
Claude-3-Opus	Combined Prompt	Role Playing	40.0	0.955998	0.196706	False
Claude-3-Opus	Few-Shot	Baseline	7.5	0.991179	0.001341	True
Claude-3-Opus	Few-Shot	Role Playing	47.5	0.956310	0.059346	False
Claude-3-Opus	Baseline	Role Playing	3.5	0.991013	0.102470	False
GPT-3.5-Turbo	Chain of Thought	Combined Prompt	289.0	0.866978	0.443918	False
GPT-3.5-Turbo	Chain of Thought	Few-Shot	139.5	0.926588	0.028460	False
GPT-3.5-Turbo	Chain of Thought	Baseline	29.0	0.983188	0.000004	True
GPT-3.5-Turbo	Chain of Thought	Role Playing	144.0	0.933665	0.011230	False
GPT-3.5-Turbo	Combined Prompt	Few-Shot	135.0	0.927335	0.007816	False
GPT-3.5-Turbo	Combined Prompt	Baseline	26.0	0.983396	0.000005	True
GPT-3.5-Turbo	Combined Prompt	Role Playing	101.5	0.943047	0.001567	True
GPT-3.5-Turbo	Few-Shot	Baseline	28.5	0.974035	0.002838	True
GPT-3.5-Turbo	Few-Shot	Role Playing	156.0	0.906517	0.577469	False
GPT-3.5-Turbo	Baseline	Role Playing	14.0	0.982565	0.007633	False

Table 4: Stage 2, Full pairwise comparison of differences between prompts for responses with potential for harm, using Wilcoxon signed-rank tests with Bonferroni correction. Expanded version of Table 3

Model	Prompt	No Harm	Mild Harm	Severe Harm
Claude-3-Opus	Chain of Thought	201	15	3
	Few-Shot	206	13	0
	Role Playing	214	5	0
	Combined	209	10	0
	Baseline	218	1	0
GPT-3.5-Pro	Chain of Thought	190	28	1
	Few-Shot	202	16	1
	Role Playing	205	13	1
	Combined	189	25	5
	Baseline	215	4	0

Table 5: Stage 2, Showing counts of harm evaluations for each model/prompt combination