An Evaluation of Lightweight LLM Approaches to Enhancing Pre-Consultation Patient Questioning

Kirsten Mayland
Dartmouth College
kirsten.r.mayland.25@dartmouth.edu

1 Introduction

As artificial intelligence has become prevalent in modern society, its presence in the medical field has been closely looked at as errors and incorrect diagnosis have the potential to result in great harm to someone. In this project, I looked at one aspect of the medical field where artificial intelligence has the possibility in our current state of development to help without the potential for large negative consequences—that of follow-up questions instead of diagnoses. The project looks at language learning model's capability to prompt a person seeking medical answers for more relevant information to diminish the potential back and forth between a patient and a medical provider—effectively streamlining the pre-consultation phase and supporting better clinical decision-making.

Looking at related studies, the MediQ framework employs a dual-system that alternates between patient queries and expert evaluation to guide follow-up questions, illustrating the value of structured information gathering in clinical settings (Li et al., 2024). Additionally, research on patient messaging emphasizes how efficient communication can alleviate provider burnout, reinforcing the need for streamlined, targeted inquiries (Huang et al., 2022). While these studies often focus on diagnostic decision-making or extensive multi-turn dialogues, my approach diverges by utilizing a lightweight, single-response model solely aimed at prompting patients for critical information, thereby complementing and extending current methodologies in medical dialogue systems.

I approached the problem by first building a comprehensive database from the r/AskDocs subreddit, compiling a list of titles and posts with corresponding follow-up questions asked by verified medical professionals. I then fine-tuned LLMs on this data and asked them to generate questions based on posts. Then, I took the generated follow-ups as well as my baseline follow-ups of GPT 4o-mini's questions, and had GPT 4o-mini-2024-07-18 evaluate the strength of the generated responses on four different axes (Utility, Necessity, Completeness, and Clarity) using a five-point Likert scale, where one is minimally embodying that feature and five is greatly embodying that feature.

While fine-tuning improved performance, model size was the primary factor in generating high-quality follow-up questions. Smaller models, even when trained, lagged behind GPT-4o-mini, which consistently outperformed them across all metrics. Fine-tuning provided slight gains, but large models remained significantly superior for this task.

Key Contributions

- Lightweight LLM exploration for cost-effective medical follow-up question generation.
- Fine-tuning on r/AskDocs data, leveraging real patient-provider interactions.

- Comprehensive evaluation framework using GPT-4o-mini for multi-metric assessment.
- Trade-off insights between model size and fine-tuning, highlighting large models' superiority.

2 Related Work

Recent research on large language models (LLMs) in healthcare has explored various methods to enhance patient-provider communication and clinical decision-making. The "Open Medical-LLM Leaderboard" compares LLMs' performance across medical fields, with models providing direct advice, yielding an interesting insight into the state of medical-focused LLMs and their Q&A based approach (Ura et al., 2024). Unlike these models, my approach focuses on prompting additional patient information to support human decision-making rather than offering diagnoses or recommendations.

Additionally, "Large Language Model Prompting Techniques for Advancement in Clinical Medicine" highlights the importance of prompt engineering to minimize errors and optimize model performance (Shah et al., 2024). My work also prioritizes prompt accuracy but focuses on eliciting relevant patient information to aid human decision-making.

The MediQ framework introduces an adaptive question-asking system where an expert system evaluates available information and requests further questions if needed (Li et al., 2024). The study "Medical Question-Generation for Pre-Consultation with LLM In-Context Learning" addresses challenges in using general-purpose LLMs for medical dialogues, particularly their difficulty in asking targeted follow-up questions (Winston, 2024). Both studies focus on a multi-turn model which contrasts with our simpler, single-layer query system focused on prompting relevant patient information for human review.

The primary study that inspired my research was "Using large language model to guide patients to create efficient and comprehensive clinical care message" by Liu et al (2024). Their approach to the feasibility of a pre-consultation, follow-up question LLM bot was to test three models' (an LLM trained on patient portal, GPT4-simple prompt, and GPT4-complex prompt) aptitude for question generation. The results were judged in a blinded study by five physicians. My deviation and contribution to this field of inquiry is two fold: firstly, I investigated the efficacy of lightweight, small parameter LLMs at this task and secondly, I based my training data off of r/AskDocs posts, which are more varied in content and style than the more formal environment of a patient portal.

All of these studies provided valuable insights into LLMs in healthcare and helped to shape and inform my research.

3 Methods

3.1 Input/Output

A Text2Text LLM takes in a title and post related to a medical question and outputs at minimum, one follow-up question designed to prompt the asker to provide more relevant medical information that they might have forgotten to add.

3.2 Dataset

The primary training dataset was generated by processing Reddit posts and comments from the r/AskDocs subreddit, from 2005 to 2024. First, relevant files containing posts and comments were extracted and reduced down to their

essential data. For the posts that included the "name," "title," "selftext," and "num_comments," and for comments, that included the "author," "author_flair_text," "body," and "link_id." Posts without comments were removed. Comments lacking question marks, replying to other comments, or being from non-medical professionals were also removed.

Posts and comments were then merged based on their submission identifier, and HTML was converted to plain text. Relevant questions were extracted from comments, ensuring they were genuine inquiries (e.g., excluding suggestions, references, or nondescript questions) and formatted into a question list for each post. Finally, a CSV file was created containing post titles, bodies, and relevant questions, yielding a dataset with 55,887 entries.

For evaluating the success of the model I created a second dataset of 25 entries, hand-picked as the most recent post to r/AskDocs and therefore not in our training set. It is only composed of [Title, Body].

3.3 Model Training

The model was trained on a 80% train, 20% test split of the dataset. It runs for one epoch, on the tokenized_train set and is evaluated using the tokenized_test set. The inputs were composed of a brief prompt, the title, and body of a post as detailed below in the prompting section. The targets were composed of follow-up questions to said post. The title, body, and questions all came from the training dataset. The model parameters were kept standard (eg. temperature = 0.7).

All of this training was done in Google Colab, Python 3 Google Compute Engine backend (GPU), utilizing a runtime type of A100 GPU.

3.4 Model Prompting

My prompts were very simple as I wanted to test the efficacy of the training as opposed to the prompting. The models were simple Text2Text LLMs, so I constructed it as follows. The inputs equaled, “Please ask follow up questions to these posts. Your goal is to prompt them to provide more relevant medical information that they might have forgotten to add. Post Title: {insert title}. Post: {insert post}". For generation, that was all they received, but for training, the models were also given follow-up questions straight from the data set, and they had already been preformatted.

3.5 Evaluation Strategy

The success of the models were evaluated on four different metrics (Utility, Necessity, Completeness, and Clarity) as described below. The scoring was done by GPT 4o-mini using Chain-of-thought reasoning, after having been prompted with the metric definition, scoring rubric, and example output.

3.6 Figure Overview

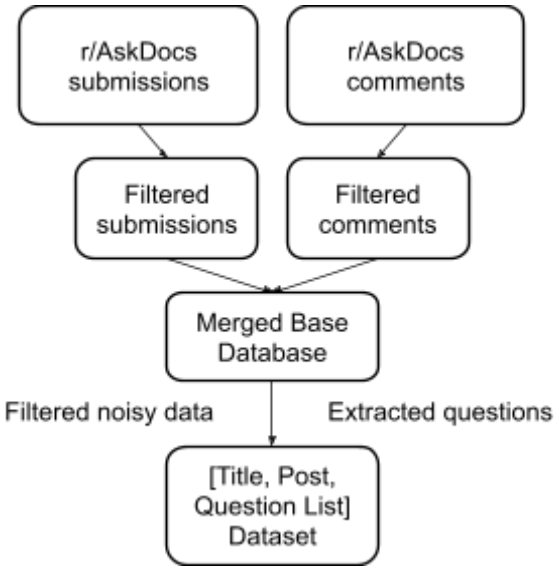

Fig 1. Database Construction

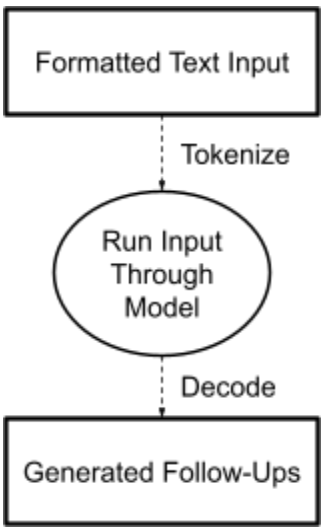

Fig 2. System Format

4 Experimental Setup

4.1 Dataset

I created my primary training dataset from two datasets constructed by u/pushshift on Reddit of public data on the subreddit r/AskDocs, the comments and the post. Their datasets were massive and as I detailed above in the Methods section, I condensed the two together and parsed the data cleanly into the dataset used for this model.

It has 55,887 data points with each point containing three values, [Title, Body, Relevant Questions]. For instance, [“20 Month Old With Flu. Please Help.”, “Hello, my son who is 20 months old and has benign neutropenia of childhood was diagnosed with Flu A yesterday. I am completely freaking out and terrified. Everything I read online talks about how dangerous this is for children under 2. He is taking tamiflu. Please offer any advice to help him get better and prevent complications. My anxiety is 1,000/10. 34F, NJ”, “1) Has he had his flu shot?”]. The inputed titles and bodies were capped at 512 tokens in length due to computational power.

As noted in the Methods section, I also created a smaller dataset of 25 entries of [Title, Body] that were pulled from the most recent posts on r/AskDocs and therefore were not in the training dataset.

4.2 Baselines

GPT 4o-mini-2024-07-18 was the baseline model in evaluations, as the aim of this study was to see if smaller, easier-to-train Text2Text models that were fine-tuned on comprehensive datasets were capable of emulating the success of larger generative AI models, for his specific task.

4.3 Models

I ran this study on four models, google/flan-t5-small (standard), google/flan-t5-small (reddit data trained), and google/flan-t5-base (standard), and google/flan-t5-base (reddit data trained).

4.4 Metrics

The evaluation metrics were copied from “Using large language model to guide patients to create efficient and comprehensive clinical care message.” (Liu et al., 2024) as I felt them to be representative and complete metrics. The success of the model in producing follow-up questions was evaluated across four dimensions (utility, necessity, completeness, and clarity).

Utility is defined as “The follow-up questions would be useful to a healthcare provider in responding to the patient message”. Necessity is defined as “All the follow-up questions are necessary for a healthcare provider in addressing the patient’s concern”. Completeness is defined

as “The follow-up questions are not missing important information necessary for a healthcare provider in addressing the patient’s concern”. Finally, clarity is defined as “The follow-up questions are easy to understand and answer by patients.”

As I am judging generated text and is thus more qualitative and subjective, these four metrics will be evaluated over a five-point Likert scale, where one equates to ‘strongly disagree’ and five equates to ‘strongly agree’.

As in Liu et al., 2024, for each metric, the mean and standard deviation were reported. The overall score was determined by averaging the scores from all four metrics. If the overall scores were the same, then the score of “utility” metric would be used to determine the final ranking.

5 Results

Note, for the rest of the paper (t) will indicate the trained version of a model.

5.1 Utility

	Mean Score	SD
GPT 4o mini	4.93	2.49
Flan-5t-s	1.67	0.38
Flan-5t-s (t)	1.21	0.41
Flan-5t-b	1.58	0.97
Flan-5t-b (t)	1.83	1.31

Fig 3. On a scale from 1-5, how useful were the questions to a healthcare provider in responding to the patient message

5.2 Necessity

	Mean Score	SD
GPT 4o mini	4.85	2.48
Flan-5t-s	1.26	0.88
Flan-5t-s (t)	1.46	0.66
Flan-5t-b	1.71	1.00
Flan-5t-b (t)	2.17	1.35

Fig 4. On a scale from 1-5, how necessary were the questions for a healthcare provider in addressing the patient’s concern

5.3 Completeness

	Mean Score	SD
GPT 4o mini	5.0	2.21
Flan-5t-s	1.08	0.28
Flan-5t-s (t)	1.13	0.41
Flan-5t-b	1.67	0.48
Flan-5t-b (t)	1.38	0.49

Fig 5. On a scale from 1-5, were the questions not missing important information necessary for a healthcare provider in addressing the patient’s concern

5.4 Clarity

	Mean Score	SD
GPT 4o mini	4.85	1.88
Flan-5t-s	1.42	0.50
Flan-5t-s (t)	1.65	0.97
Flan-5t-b	1.96	0.55
Flan-5t-b (t)	2.50	1.44

Fig 6. On a scale from 1-5, how easy were the questions to understand and answer by patients

5.5 Overall score

	Mean Score	SD
GPT 4o mini	3.32	2.32
Flan-5t-s	1.22	0.57
Flan-5t-s (t)	1.33	0.68
Flan-5t-b	1.60	0.83
Flan-5t-b (t)	1.95	1.26

The results are relatively unsurprising, even when trained, smaller parameter LLMs struggle to compete against a large generative AI model such as GPT 4o mini.

There is an interesting and definitive trend across all metrics, where larger models surpass smaller models, fine-tuned models surpass non-fine tuned models, and larger non-fined tuned models surpass smaller fine tuned models.

In other words, when looking at Text2Text generating models, specifically those designed to ask questions, parameter size is more important than fine-tuning, but both make marked differences in improving the model.

Additionally, as the model grows, so does the standard deviation of its results. This is most likely because the quality of the results wants to fall in a gaussian pattern but due to the limited output ([1, 2, 3, 4, 5]), models that consistently fall at one extreme have a standard deviation that is forced unnaturally low; however, this is just a theory and it does invite room for further study.

6 Error Analysis

The primary limitations of this study were time, finances, and RAM. I wanted to train/fine-tune models on my computer, but due to RAM and time constraints, I was forced to work with more lightweight models. Therefore, there is a decent degree of inherent error due to the model size and their parameters.

A potential source of error is the fact that the LLMs were trained on questions that were pulled out of context. If they had been trained on the full response instead, there is a chance they could have performed better. There is also a chance that they could have simply started providing advice instead of asking questions. This warrants further research.

As question generation is a more subjective field of study than rote classification, I do not have specific error metrics, just general performance results.

Some more potential sources of error that could have affected the models are for one, that the medical field is very fast and therefore contains a lot of specific knowledge which is only applicable in certain situations. A model which is trained more generally could misapply this knowledge easily and query for information that is not relevant. Finally, a lot of the posts relied on photos that were not present in the training as this is a text based model. This could result in some confusion in the training and thus error when applied.

7 Conclusion

This study explored the feasibility of using lightweight, fine-tuned LLMs to generate follow-up medical questions, aiming to streamline patient-provider communication. By training models on a dataset of r/AskDocs posts and evaluating their performance across utility,

necessity, completeness, and clarity, we assessed their effectiveness compared to GPT-4o mini.

Results showed that while fine-tuning improved question generation, model size remained the most significant factor in performance. Larger models consistently outperformed smaller ones, highlighting the limitations of parameter-constrained LLMs for this task. These findings suggest that while fine-tuning enhances specialized capabilities, robust question generation in medical contexts still demands larger-scale architectures. Future work could explore refining training methodologies and leveraging multimodal data to improve performance in real-world applications.

Acknowledgments

Thank you to Dartmouth College for the resources and opportunity to undertake this research. Thank you to Prod. Sarah Masud Preum for her consultation and advice on this project.

References

Huang, M., Fan, J., Prigge, J., Shah, N. D., Costello, B. A., & Yao, L. (2022). *Characterizing patient-clinician communication in Secure Medical Messages: Retrospective Study*. Journal of Medical Internet Research.

Li, S. S., Tsvetkov, Y., Koh, P. W., Pierson, E., Ilgen, J., Feng, S., & Balachandran, V. (2022, June). *Mediq: Question-asking LLMS for adaptive and reliable clinical reasoning.*

Liu, S., Wright, A. P., Mccoy, A. B., Huang, S. S., Genkins, J. Z., Peterson, J. F., Kumah-Crystal, Y. A., Martinez, W., Carew, B., Mize, D., Steitz, B., & Wright, A. (2024, August 1). *Using large language model to guide patients to create efficient and comprehensive clinical care message*. Journal of the American Medical Informatics Association : JAMIA.

Shah, K., Xu, A. Y., Sharma, Y., Daher, M., McDonald, C., Diebo, B. G., & Daniels, A. H. (2024, August 28). *Large language model prompting techniques for advancement in Clinical Medicine*. Journal of clinical medicine.

Ura, A., Minervini, P., & Fourrier, C. (2024). *The open medical-LLM leaderboard: Benchmarking large language models in healthcare*. Hugging Face – The AI community building the future.

Winston, C. (2024, November 11). *Medical question-generation for pre-consultation with LLM...* OpenReview.