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

Kirsten Mayland
Accelerated Computational Linguistics, Dartmouth College
Advised by Prof. Sarah Masud Preum

Why This Research Matters

Efficient communication is critical in clinical care. Physicians send a significant portion of their messages outside regular hours, adding to workload strain. (Huang et al., 2022)

- 7.0% of daily messages sent by physicians occur between 5–7 PM.
- 5.4% of messages from nurse practitioners or physician assistants are sent in this window.

Most patient-provider exchanges are short, but **inefficiencies add up.**

- 95.1% of message threads end within four rounds (Huang et al., 2022).
- Extra back-and-forth wastes time and increases provider workload.

Solution: LLM-driven follow-up questions help patients provide essential details upfront, reducing unnecessary messaging and improving efficiency.

Existing Research on LLMs in Healthcare

Recent studies explore LLMs in patient-provider communication and clinical decision-making.

Open Medical-LLM Leaderboard (Ura et al., 2024)

- evaluates models providing direct medical advice.
- contrasting with my focus on prompting additional patient information.

Large Language Model Prompting Techniques (Shah et al., 2024)

- emphasizes prompt engineering to reduce errors
- aligning with my goal of eliciting relevant patient details (Shah et al., 2024).

MediQ Framework & *Medical Question-Generation for Pre-Consultation* (Li et al., 2024; Winston, 2024).

- use multi-turn systems for adaptive questioning
- whereas my approach is a simpler, single-response model.

Key Study and Project Contributions

Using LLMs to Guide Patients in Clinical Messaging tested LLM-generated follow-up questions using a patient portal dataset, evaluated by physicians (Liu et al., 2024). Found that LLMs show “a great potential for improving patient-provider communication.”

Our work builds on this by:

- Investigating **lightweight LLMs capability** for helpful question generation.
- Using **r/AskDocs data**, which is more varied than structured patient portal messages.

These studies provided crucial insights that shaped and informed my research.

Introduction to AI in Healthcare

As AI grows in prevalence, its role in healthcare is under scrutiny due to potential risks from errors or incorrect diagnoses.

This project focuses on a **low-risk application**: using AI for follow-up questions instead of diagnoses.

The goal is to **streamline the pre-consultation phase** by prompting patients for relevant information, reducing back-and-forth communication, and supporting clinical decision-making.

My approach:

- Fine-tuned LLMs on comprehensive database built from the **r/AskDocs subreddit** to generate relevant follow-up questions. Focusing on follow-up questions from verified medical professionals

Methodology and Results

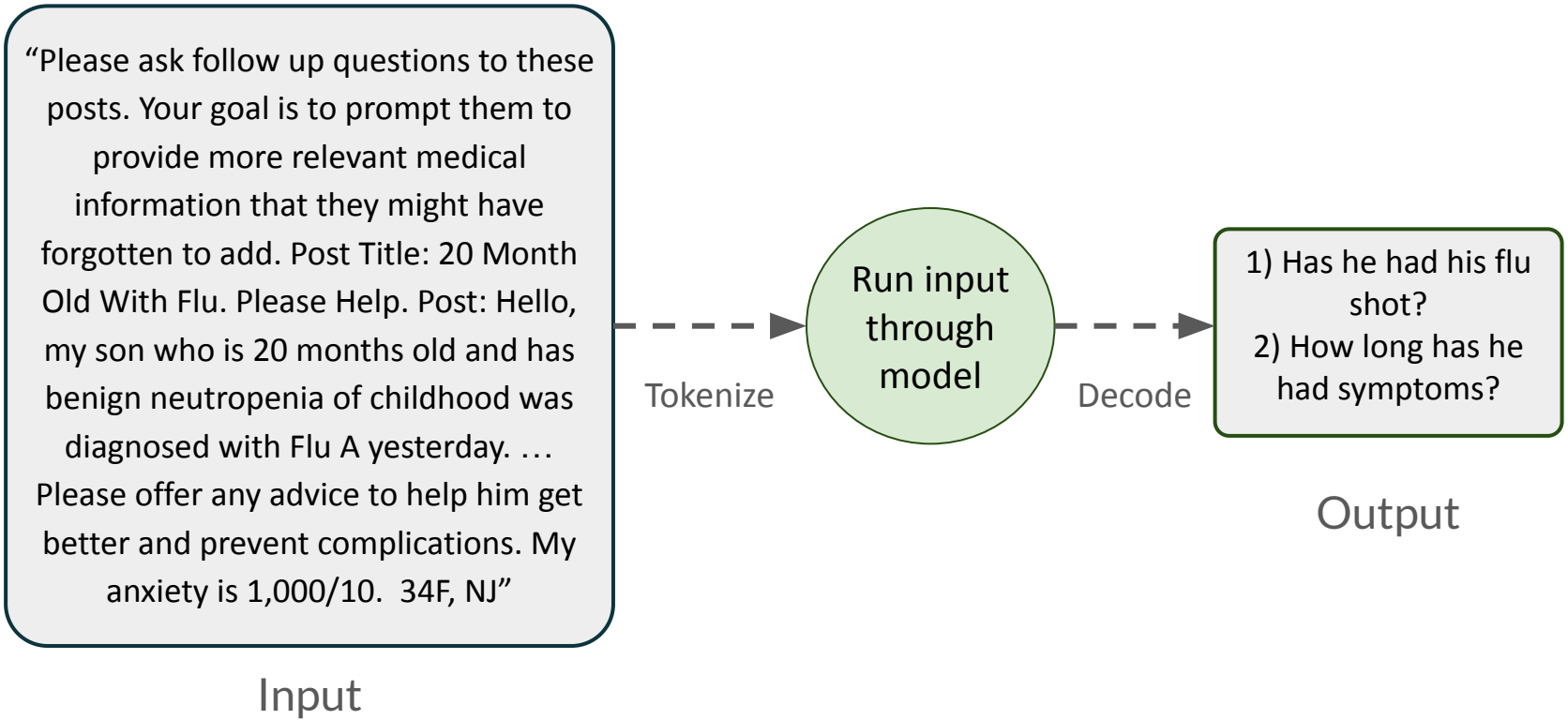
Evaluated the generated follow-ups and baseline (GPT-4o-mini) responses based on **four axes: Utility, Necessity, Completeness, and Clarity** using a five-point Likert scale.

- Used GPT-4o-mini-2024-07-18 as an evaluator, scoring responses on a five-point Likert scale.

Results

- **Model size > Fine-tuning:** Larger models consistently outperformed smaller ones, even when fine-tuned.
- **Fine-tuning helped but was not transformative**, with only minor improvements over untrained models.
- Baseline GPT-4o-mini remained the strongest performer, excelling across all evaluation criteria.
- Findings suggest that while **fine-tuned small models can be cost-effective, large models remain superior for high-quality medical follow-ups.**

System Demonstration



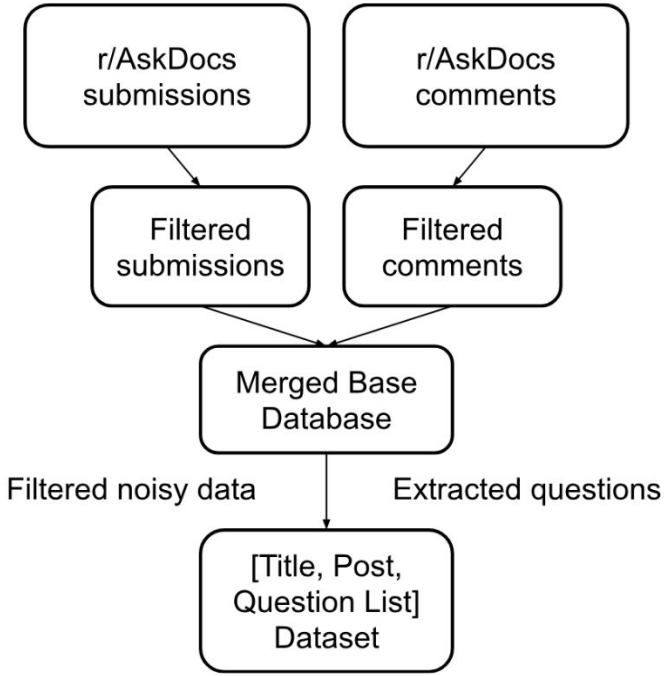
Methods: Dataset Preparation

Primary Dataset: r/AskDocs (2005–2024)

- Extracted key post and comment columns
- Removed posts without comments, comments lacking question marks, and non-medical professionals.
- Merged posts and comments based on submission identifier, converted HTML to plain text.
- Relevant questions were extracted, formatted into a question list for each post.

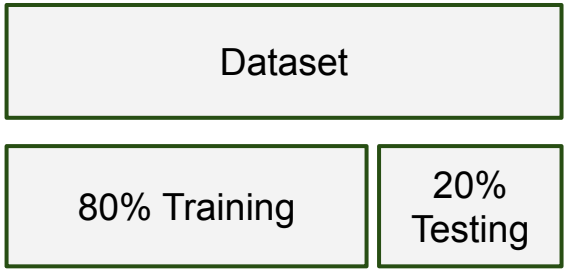
Yielded a dataset of 55,887 entries in [Title, Post, Follow-Up Questions] form.

Evaluation dataset: 25 entries of [Title, Post] form, hand-picked from the most recent posts, not in the training set.



Methods: Model Training

Models: google/flan-t5-small, google/flan-t5-base



Training Process:

- 80% training, 20% test split.
- Model trained for one epoch on the tokenized dataset.

Inputs: Brief prompt, post title, and body as seen in the system demonstration

Targets: Follow-up questions from the dataset.

Model parameters: Standard (e.g., temperature = 0.7).

Training environment: Google Colab, Python 3, Google Compute Engine backend (GPU), A100 GPU.

Methods: Model Prompting

Simple Prompts

- Goal: Test the efficacy of training rather than complex prompts.

Input prompt format:

- “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 training:

- Models were provided with follow-up questions from the dataset, preformatted.

Methods: Evaluation Strategy

Evaluation Metrics:

- **Utility:** Measures how helpful the question is for gathering relevant information.
- **Necessity:** Assesses if the question is essential for the patient to provide critical details.
- **Completeness:** Evaluates whether the question prompts enough information to form a comprehensive follow-up.
- **Clarity:** Judges the questions' clarity and ease of understanding for the patient.

Scoring Process:

- Scored by GPT-4o-mini using Chain-of-Thought reasoning and parsed output.
- Prompted with metric definitions, scoring rubrics, and example output to ensure consistent and accurate evaluations.

Experimental Setup Overview

Primary dataset from **r/AskDocs subreddit**, sourced from two public Reddit datasets created by u/pushshift.

- Combined and parsed the datasets into **55,887** data points, each containing:
- [Title, Body, Relevant Questions]

Data example:

- Title: “20 Month Old With Flu. Please Help.”
- Body: "Hello, my son who is 20 months old..."
- Relevant Question: "Has he had his flu shot?"

Input titles and bodies were capped at 512 tokens for computational constraints.

A smaller dataset for evaluations of 25 entries was created from the most recent posts, not included in the training set.

Experimental Setup: Baselines and Models

Baseline Model:

- GPT-4o-mini-2024-07-18

Aimed to compare the performance of smaller, fine-tuned models with larger generative models for follow-up question generation.

Models Tested:

- google/flan-t5-small (standard)
- google/flan-t5-small (reddit data trained)
- google/flan-t5-base (standard)
- google/flan-t5-base (reddit data trained)

Experimental Setup: Evaluation Metrics

Four Evaluation Metrics:

- Utility, Necessity, Completeness, Clarity

Scoring:

- 5-point Likert scale: 1 = strongly disagree, 5 = strongly agree
- Mean and standard deviation reported for each metric.
- Overall score: Averaged across all four metrics, with Utility used as the tie-breaker for ranking.

Evaluation Process:

- Scored by GPT-4o-mini using Chain-of-Thought reasoning across metrics
- Results based on qualitative, subjective judgments by LLM, aggregated for final analysis.

Results

Key Findings

- Model size significantly impacts quality—larger models like GPT-4o-mini consistently outperformed smaller models across all evaluation metrics.
- Fine-tuning improves performance, but the gains were marginal compared to the advantage of using a larger model.

Quantitative Results

- Fine-tuned models showed a small but consistent improvement over their untrained counterparts.
- GPT-4o-mini scored significantly higher on all four evaluation axes, demonstrating its superior ability to generate meaningful follow-ups.

	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

Overall score across all four metrics, all four metrics were relatively similar, (t) indicates it was fine-tuned

Error Analysis Overview

Primary Limitations:

- Time, Finances, and RAM Constraints
- Led to the use of lightweight models, which introduced inherent errors due to model size and parameters.

Training on Extracted Questions:

- Trained on questions pulled out of context rather than full responses.
- This may limit performance or result in models offering advice instead of follow-up questions. Further research is needed.

Additional Sources of Error

Subjectivity of Question Generation:

- Unlike classification, question generation is subjective, making precise error metrics difficult.
- Evaluations based on general performance rather than specific metrics.

Specialized Medical Knowledge:

- Misapplication of context-specific medical knowledge by general models could lead to irrelevant queries.
- Visually noticed as a common error

Absence of Visual Context:

- Many posts relied on photos not present in the training data, potentially causing confusion and errors in generating relevant questions.

In Conclusion

Key Takeaways

- **Model size is the primary determinant of performance**—larger models consistently generate higher-quality follow-ups.
- Fine-tuning improves responses but does not bridge the gap between small and large models.
- **Lightweight models can be optimized** for medical question generation but **remain less effective** than state-of-the-art LLMs.
- Using LLMs for pre-consultation follow-ups **can streamline patient-provider communication**, reducing unnecessary back-and-forth.

Future Directions

- Optimizing small models further to close the performance gap while maintaining efficiency.
- Incorporating human feedback for more clinically relevant and patient-friendly follow-ups.
- Exploring multi-turn interactions rather than single-response generation for more dynamic questioning.