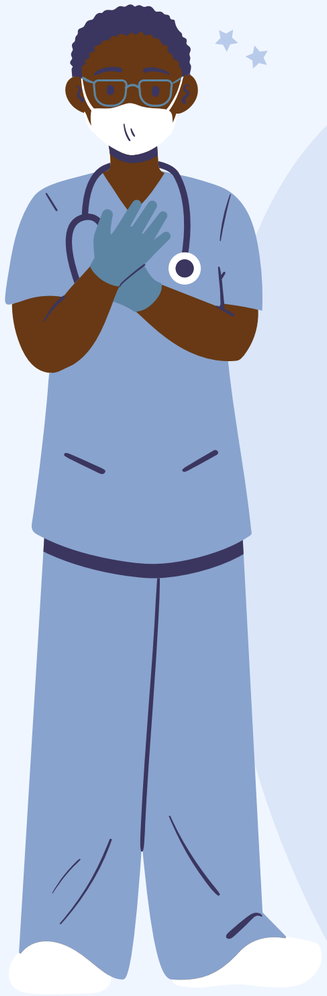
Specialist Recommendation

Problem Understanding

This project aims to build a dynamic medical questionnaire that adapts to the patient's answers and recommends the most appropriate specialist based on the collected information.

Research Questions



How can we adapt questions dynamically based on previous answers? , what is the best method?

How can we accurately recommend the most appropriate medical specialist based solely on the patient's answers?

Related Work

- **Qsnail**: A Questionnaire Dataset for Sequential Question Generation (2024)
- **Information Seeking Question Generation** using Dynamic Meta-Information Retrieval and Knowledge Graphs (2021)
- **From General to Specific**: Tailoring Large Language Models for Personalized Healthcare (2024)
- **Automatic question generation**: a review of methodologies, datasets, evaluation metrics, and applications (2023)

Medical Question-Generation for Pre-Consultation with LLM In-Context Learning

Medical Question-Generation for Pre-Consultation with LLM In-Context Learning

Caleb Winston
Stanford University, USA
calebwin@cs.stanford.edu

Cleah Winston
University of Washington, USA
cleahw@uw.edu

Claris Winston
University of Washington, USA
clarisw@uw.edu

Chloe Winston
University of Pennsylvania, USA
chloe.winston@pennmedicine.upenn.edu

Abstract

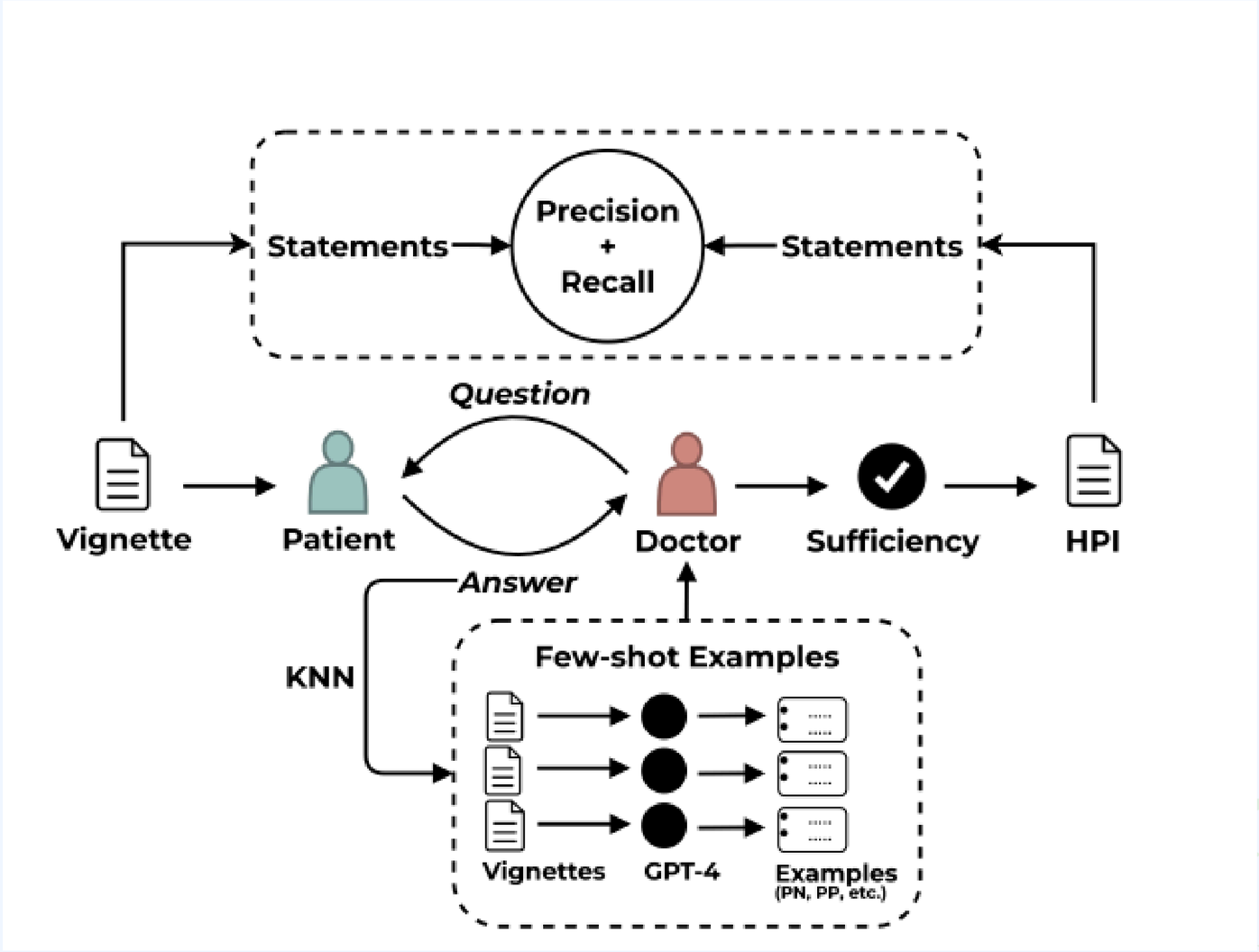
Pre-consultation gives healthcare providers a history of present illness (HPI) prior to a patient's visit, streamlining the visit and promoting shared decision making. Compared to a digital questionnaire, large language model (LLM)-powered AI agents have proven successful in providing a more natural interface for pre-consultation. But general LLM-based approaches struggle to ask productive follow-up questions and require complex prompts to guide the consultation. While effective automated prompting strategies exist for medical question-answering LLMs, the task of question generation for pre-consultation lacks effective strategies. In this study, we develop a methodology for evaluating existing approaches to medical pre-consultation, using prior datasets of HPIs and patient-doctor dialogues. We propose a novel approach of converting clinical note data into question generation examples and then retrieving relevant examples for in-context learning. We find this approach to question generation for pre-consultation achieves a higher recall of facts in a ground truth consultation than baseline approaches across a range of simulated patient personalities.

1 Introduction

The utility of large language models (LLMs) in healthcare is rapidly expanding. Before implementing LLMs in clinical practice, the safety-critical nature of various medical tasks must be carefully considered. Some have proposed the use of LLMs as aids to the clinician, for example in-diagnostic reasoning, rather than independent providers [22]. How to safely incorporate LLMs in more patient-facing tasks such as question answering and complete medical consultations is still unclear, but medical pre-consultation potentially represents a safe and beneficial applications of LLMs in medicine [15]. LLM-collected histories can be quickly verified by clinicians, who can then focus the visit on exploring potential diagnoses and therapeutic options, similar to how clinicians rely on trainees' reports of patient histories. The present work focuses on improving the capability of general purpose LLMs at efficiently collecting history from a patient and robustly evaluating the quality of LLM-led conversations.

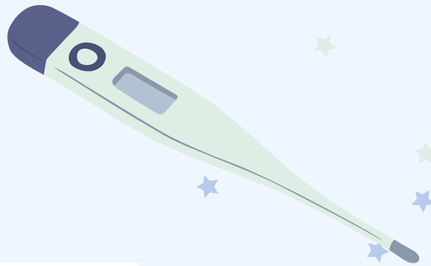
The clinical history critically leads the evaluation and treatment of a patient's presenting symptom,

Methodology



Dataset

Histories of Present Illness (HPIs)



"A 54-year-old male presents with chest pain for two days. The pain worsens with exertion and improves with rest. He denies shortness of breath, dizziness, nausea, or vomiting. His medical history includes hypertension and a long-term smoking habit (10 cigarettes/day for 20 years). His father had a history of a heart attack at age 60. Physical examination shows elevated blood pressure (140/90 mmHg) with normal heart and lung sounds. The differential diagnosis includes angina, gastroesophageal reflux disease (GERD), and musculoskeletal chest pain."

Limitations

Limitations

Dependence on In-Context Learning (ICL)

No Explicit Symptom-Disease Relationship Modeling

No Specialist Assignment

Meet Your Doctor Estena3y

A smarter way to connect patients with the right care. Doctor Estena3y uses advanced technology to understand your symptoms and guide you to the appropriate specialist.

Try Doctor Estena3y Now

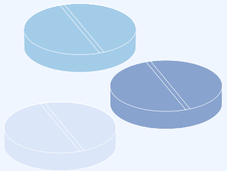
Our aim is to label medical dialogue data and use the labeled information to recommend the most appropriate medical specialist for each patient.

Our Contribution

1.Data Labeling

2. Fine-Tuning Approach

3. Knowledge Graph Approach



Dataset

ID	section_he	section_text	dialogue
0	GENHX	The patient is a 76-year-	Doctor: What brings you back into the clinic
1	GENHX	The patient is a 25-year-	Doctor: How're you feeling today?
2	GENHX	This is a 22-year-old fem	Doctor: Hello, miss. What is the reason for
3	MEDICATIC	Prescribed medications	Doctor: Are you taking any over the counter
4	CC	Burn, right arm.	Doctor: Hi, how are you?
5	PASTMEDI	Asthma.	Doctor: How's your asthma since you
6	PASTMEDI	The patient denies high b	Doctor: Do you smoke?
7	ALLERGY	No known drug allergies	Doctor: Any know drug allergies?
8	FAM/SOCH	His mother died of comp	Doctor: Hi there, sir! How are you today?

ID: Unique identifier for each case.

Section Header: Categorizes different types of medical history

Section Text: Contains specific details of the patient's medical history

Dialogue: Includes the conversation between the doctor and the patient, simulating real-life consultations.

Dataset- Section Header categorization

The full list of normalized section headers:

1. fam/sochx [FAMILY HISTORY/SOCIAL HISTORY]
2. genhx [HISTORY of PRESENT ILLNESS]
3. pastmedicalhx [PAST MEDICAL HISTORY]
4. cc [CHIEF COMPLAINT]
5. pastsurgical [PAST SURGICAL HISTORY]
6. allergy
7. ros [REVIEW OF SYSTEMS]
8. medications
9. assessment
10. exam
11. diagnosis
12. disposition
13. plan
14. edcourse [EMERGENCY DEPARTMENT COURSE]
15. immunizations
16. imaging
17. gynhx [GYNECOLOGIC HISTORY]
18. procedures
19. other_history
20. labs

Dataset Labeling

Data Labeling evaluation

We evaluated the labeled data using **a prompt-based label evaluation method**, where a medical domain LLM acted as an expert to judge if the assigned specialist matched the patient's information. The model reviewed each case and classified it as correct or incorrect, allowing us to calculate an overall labeling accuracy score.

System Evaluation

We use the LLaMA 3.2 model to evaluate medical dialogues by scoring the quality of questions and checking if the assigned specialist is correct. We then calculate the overall specialist accuracy and average question quality score to ensure the appropriateness of the questions and specialist assignments.

Data Labeling

1 Key Phrase Extraction + LLM

- Extracted key phrases using KeyBERT with the BioBERT medical encoder.
- Predicted specialists using LLaMA 3.2-3B-Instruct (small LLM) based on extracted keywords.
- Result: ~30% accuracy.

2 Diagnosis-Only Semantic Matching

- Used only the Diagnosis section text.
- Semantic similarity comparison with specialist keywords (using BioBERT embeddings).
- Result: ~58% accuracy (data was small: ~20 cases contained diagnosis).

3 Full Section Text Semantic Matching

- Combined all section texts per patient into one profile.
- Semantic comparison with specialist profiles.
- Result: ~60% accuracy.

4 Two-Stage Labeling with Keywords and LLaMA 3.2

- Matched section text against specialist keyword sets.
- Used LLaMA 3.2-3B-Instruct to predict the specialist from the section text.
- If keyword matching and LLaMA agreed, the label was accepted.
- If uncertain, applied the same process to the dialogue field.
- Final specialist assigned based on the first confident match.
- Result: ~63% accuracy.

Weak Labeling

Keyword Mapping

Augment Keywords via TF-IDF

Snorkel label function (Keyword Matching)

Snorkel label function (Zero shot)

Train Label Model

Finetuning

Finetuning

1. **Labeling the Dialogs**
   - **Weak supervision** (Snorkel) combined simple keyword rules and a zero-shot classifier to assign each patient dialogue a “pseudo” specialty label.
   - We then mapped those numeric labels back to human-readable specialties (e.g. Cardiology, Neurology).

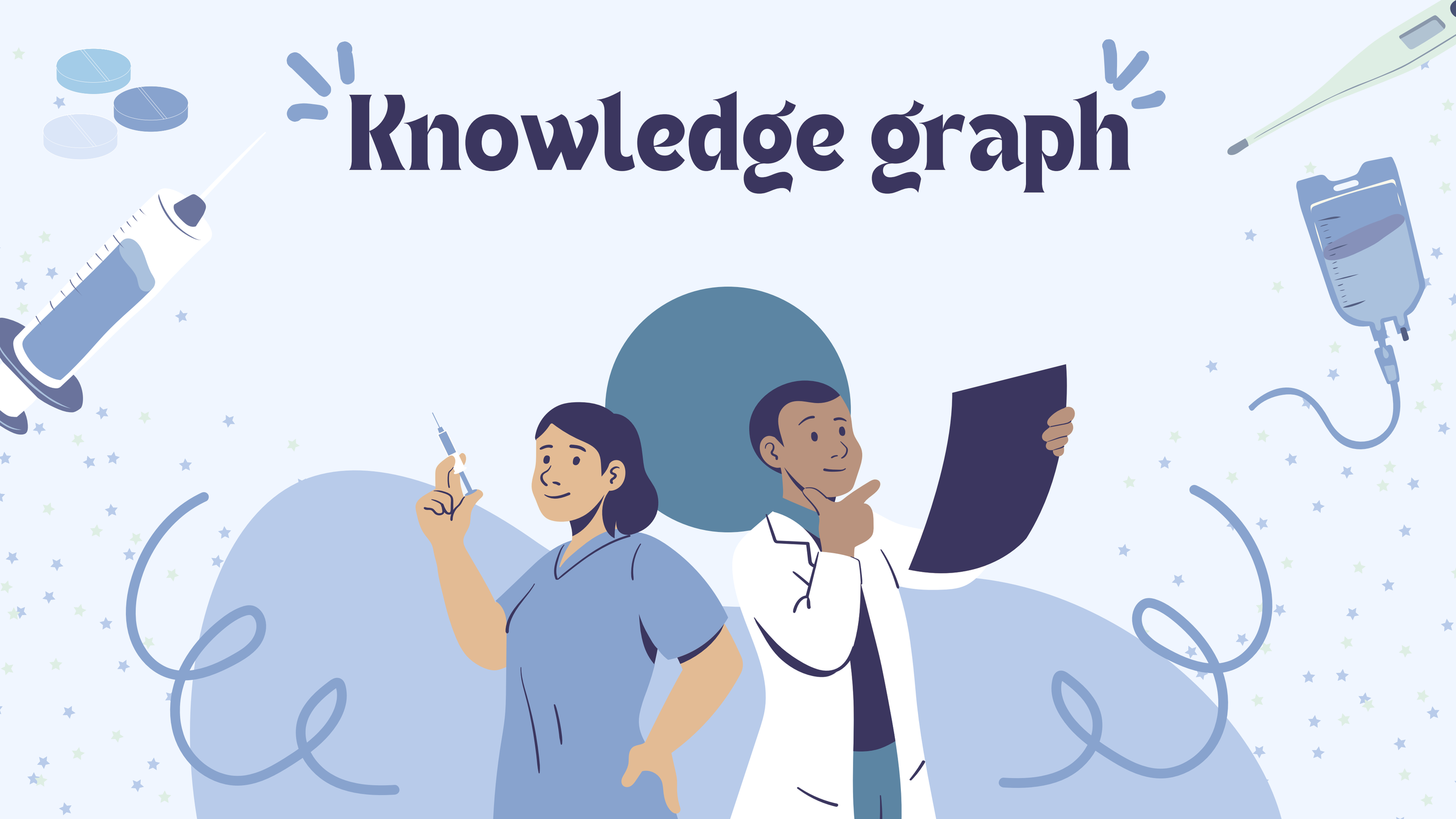
2. **Formatting for Fine-Tuning**
   - Split the labeled data into train/validation sets.
   - For each example, created a “prompt” containing the patient’s section_text and dialogue, and a “completion” of the specialty name.
   - Saved these as JSONL files in Hugging-Face format.

3. **LoRA-Based Adapter Tuning**
   - Loaded a pre-trained 3 B-parameter instruct model (Llama-3.2-3B-Instruct) in half-precision (FP16) on GPU.
   - Added a small Low-Rank Adaptation (LoRA) “adapter” module—only these adapter weights are updated, keeping the base model frozen.
   - This drastically reduces memory and compute cost compared to full fine-tuning.

4. **Training Setup**
   - Tokenized prompts and completions with padding/truncation.
   - Used a language-model collator so inputs and labels align perfectly.
   - Ran 3 epochs with a small batch size and gradient accumulation, logging and saving every few hundred steps.

5. **Deployment & Testing**
   - Combined the frozen base model + trained LoRA adapter at inference.
   - Wrapped it in a text-generation pipeline that, given a new patient dialog, outputs the recommended specialist.
   - We then built a simple back-and-forth loop (or even an LLM-to-LLM simulation) to verify adaptive Q&A and final recommendations.

Knowledge graph



Why a Knowledge Graph Instead of a Generative LLM?

1. Hallucination: May generate unsafe or incorrect answers.
2. No Explicit Symptom-Disease Relationship Modeling
3. Black box: Difficult to trace how decisions are made.

A Graph-Based Medical Dialogue System (Built with Neo4j)

1. Nodes:
   - Case, Symptom, Question, Response, Specialist
2. Relationships:
   - HAS_SYMPTOM, CONTAINS_DIALOGUE, FOLLOWED_BY, RESPONSE_TO, REFERRED_TO

How It Works

- Patient describes symptoms ("fever and cough").
- System retrieves the most similar past Case.
- Replays the exact question flow from that case.
    - “Do you have a sore throat?”
- If responses match → continue.
- If not → switch to a better-matching case.
- Final step → Recommend the same specialist.(“ENT Specialist”)

Comparison

Fine-Tuned

Overall specialist accuracy: 65

Average question quality score: 7

Knowledge graph

Overall specialist accuracy: 70

Average question quality score: 6

Comparison

Metrics for Fine-Tuned LLM:

- Average Question Quality: 88.67 (out of 100)
- Specialist Accuracy: 62.00% correct specialist assignments

Metrics for Knowledge Graph:

- Average Question Quality: 78.67 (out of 100)
- Specialist Accuracy: 66.33% correct specialist assignments

Thank you for your attention