

SNOMED Prediction Benchmark Framework

Author: Youngwook Yang (youngwook.yang@gmail.com)

Introduction

Accurately identifying patient medical conditions is one of the most critical steps in healthcare delivery. In many real-world settings—especially in low-resource environments—clinicians must rely on iterative question-and-answer exchanges with patients to arrive at a diagnosis. While large language models (LLMs) have demonstrated impressive performance in static benchmarks, few benchmarks evaluate their ability to interactively gather information by asking relevant questions before making a clinical inference, such as predicting SNOMED codes.

To address this gap, I propose a novel interactive benchmarking framework that evaluates LLMs as virtual clinicians. In this setup, a target LLM attempts to identify a patient's conditions by conversationally querying a simulated patient, powered by a fine-tuned BioMistral-7B model acting as an oracle. The benchmark tests whether LLMs can engage in informative, efficient, and accurate diagnostic conversations, and ultimately produce a list of SNOMED codes that reflect the patient's conditions.

This is an important challenge because:

- Medical Q&A tasks often require context-specific reasoning, not just factual retrieval.
- Multiturn dialogue better simulates how clinicians work with patients compared to one-shot prompts.
- SNOMED coding is widely used in clinical decision support and documentation, making it a practical prediction target.

This project contributes an interactive benchmark dataset, a lightweight oracle model that mimics patients in lay language, and a head-to-head comparison of multiple LLMs (e.g., GPT-3.5-turbo, grok-3-mini) on their ability to gather relevant information and make accurate diagnoses. This benchmark framework lays the foundation for future work in using LLMs as interactive clinical assistants.

Related Work

Recent progress in medical large language models (LLMs) has led to promising results across a range of clinical tasks, including diagnosis, triage, and medical question answering. However,

most benchmarks assess models in static, single-turn settings, lacking the interactive dynamics typical of real clinical encounters. This project builds on and extends this growing body of work by focusing on multiturn, conversational inference of structured codes such as SNOMED.

SDBench: Sequential Diagnosis Benchmark (Microsoft, 2025)

<https://microsoft.ai/new/the-path-to-medical-superintelligence/>

SDBench, introduced by Microsoft in The Path to Medical Superintelligence, is a pioneering benchmark designed to evaluate a model's ability to ask questions interactively in order to arrive at a diagnosis. SDBench uses a simulated "Oracle" to answer questions about patients. My benchmark shares this goal but focuses specifically on predicting SNOMED codes from conversations and uses a fine-tuned BioMistral-7B model as a lightweight oracle.

MultiMedQA (Singhal et al., 2022)

<https://arxiv.org/abs/2212.13138>

MultiMedQA is a comprehensive benchmark that evaluates LLMs across six existing medical QA datasets as well as a new set of consumer-searched health questions. It measures model performance across axes such as factuality, reasoning, potential harm, and bias. However, the format is limited to single-turn responses and does not capture the back-and-forth dialog that occurs in clinical interviews. In contrast, my benchmark emphasizes multiturn interaction, testing the model's ability to gather information through follow-up questions before making structured predictions.

SNOMED AutoCoding Models

<https://arxiv.org/abs/1912.11975>

Much of the prior work on SNOMED or ICD code prediction involves supervised learning on discharge summaries or structured EHR data. Models like ClinicalBERT (Huang et al., 2019) process large amounts of text to directly output codes. My benchmark departs from this approach by testing whether a model can conversationally elicit relevant patient details to drive its predictions—an approach that more closely mirrors clinical reasoning.

Methodology

To evaluate the ability of large language models to interactively identify patient conditions, I built a benchmark framework where an LLM assumes the role of a virtual clinician and engages in a dialog with a simulated patient (the "oracle") to predict a list of SNOMED codes.

1. Oracle Construction (Patient Simulator)

1.1 Data Preparation

To simulate realistic medical conversations, I fine-tuned the [BioMistral-7B](#) model using synthetic question-answer data generated with GPT-3.5-turbo. Each QA sample is grounded in a synthetic patient record derived from [Synthea](#), an open-source patient simulator that generates structured EHR-like data.

However, the raw Synthea output (see below) is not directly usable for conversational fine-tuning. It contains structured observations, diagnoses (e.g., SNOMED codes), and vital signs but lacks question-answer pairs in natural language.

```
{
  "patient_id": "339144f8-50e1-633e-a013-f361391c4cff",
  "oracle_context": "Age: 27\nGender: M\nLocation: Boston, Massachusetts\nConditions: Sprain of ankle; Hypertension; Acute viral pharyngitis (disorder); Viral sinusitis (disorder); Refugee (person)\nObservations:\n- Body Height: 166.8\n- Pain severity - 0-10 verbal numeric rating [Score] - Reported: 3.0\n- Body Weight: 70.9\n- Body Mass Index: 25.5\n- Body mass index (BMI) [Percentile] Per age and gender: 94.0\n- Diastolic Blood Pressure: 82.0\n- Systolic Blood Pressure: 129.0\n- Heart rate: 87.0\n- Respiratory rate: 15.0\n- Tobacco smoking status NHIS: Never smoker",
  "snomed_gold": [
    "44465007 - Sprain of ankle",
    "59621000 - Hypertension",
    "195662009 - Acute viral pharyngitis (disorder)",
    "444814009 - Viral sinusitis (disorder)",
    "446654005 - Refugee (person)"
  ]
},
```

To bridge this gap, I used GPT-3.5-turbo to generate conversational data by prompting it to simulate a doctor asking questions and a patient replying in layperson language. Each synthetic sample includes:

- 5 positive QA pairs related to the patient's known conditions
- 2 negative QA pairs asking about unrelated conditions or symptoms

The negative examples are crucial. They train the oracle model to respond with uncertainty or say "I'm not sure" when a symptom is not part of the patient's profile—thereby mimicking realistic patient behavior. This process was repeated for 500 patients, yielding over 3,500 QA pairs. The final dataset was formatted in a JSONL structure compatible with Hugging Face's fine-tuning tools, where each entry contains:

- The patient context
- The doctor's question
- The patient's response
- An instruction prompt guiding the model to act as a patient

```
{"instruction": "Answer the user's question based on the patient context. Use lay language only.", "snomed_gold": ["19169002 - Miscarriage in first trimester", "444814009 - Viral sinusitis (disorder)", "68496003 - Polyp of colon"], "input": "Patient Info:\nAge: 59\nGender: F\nLocation: Wenham, Massachusetts\nConditions: Miscarriage in first trimester; Viral sinusitis (disorder); Polyp of colon\nObservations:\n- Body Height: 176.6\n- Pain severity - 0-10 verbal numeric rating [Score] - Reported: 3.0\n- Body Weight: 86.5\n- Body Mass Index: 27.7\n- Diastolic Blood Pressure: 81.0\n- Systolic Blood Pressure: 128.0\n- Heart rate: 62.0\n- Respiratory rate: 15.0\n- Total Cholesterol: 194.7\n- Triglycerides: 148.6\nQuestion: How are you feeling after the miscarriage?", "output": "Physically and emotionally drained."}
```

1.2 Model Fine Tuning

To train the patient simulator (oracle model), I fine-tuned the BioMistral-7B model using LoRA (Low-Rank Adaptation), a parameter-efficient technique that enables fine-tuning large language models with limited computational resources. This approach was chosen primarily due to the hardware constraints of my environment—specifically, a single NVIDIA RTX 4070 SUPER GPU (12GB VRAM).

Rather than updating all of BioMistral's 7 billion parameters, LoRA injects small trainable matrices into selected layers (e.g., attention projections), drastically reducing memory usage while still achieving strong adaptation. This fine-tuning setup was deliberately kept lightweight so that other researchers or developers with a single modern GPU can reproduce it easily, without relying on expensive cloud compute or distributed training.

The training process ran for nearly five days, with a batch size of 1 and sequence length capped at 512 tokens. Despite the limited setup, the resulting model demonstrated significantly more realistic and context-aware behavior than the base BioMistral model. It was particularly effective at:

- Responding naturally in lay language
- Handling both positive and negative symptom queries
- Avoiding hallucinated details not present in the patient context

This made the oracle a reliable interactive proxy for simulating patients in downstream benchmarking experiments. Here is a comparison.

Before Fine-Tuning: You can see that it is not really acting like a patient.

```
Input Prompt:
### Instruction:
You are a patient. Answer the doctor's question based on the patient context given.
Use lay language only. Please be concise and answer in one or two short sentences.

### Input:
Patient Info:
Age: 7
Gender: F
Location: Somerville, Massachusetts
Observations:
- Body Height: 49.4
- Pain severity - 0-10 verbal numeric rating [Score] - Reported: 2.0
- Body Weight: 3.1
- Weight-for-length Per age and sex: 32.4
- Head Occipital-frontal circumference: 33.2
- Diastolic Blood Pressure: 82.0
- Systolic Blood Pressure: 117.0
- Heart rate: 67.0
- Respiratory rate: 14.0
- Leukocytes [#volume] in Blood by Automated count: 8.9
Question: Are you feeling tired or energetic lately?

### Response:
...
77 year old male patient with chronic congestive heart failure, hypertension, acute bacterial sinusitis and viral sinusitis. Patient is a non-smoker and has no history of alcohol or drug abuse.

Expected (Fine-Tuned) Answer:
I've been feeling pretty tired and a bit short of breath.
```

After Fine-Tuning: You can see that it is now more like a real patient.

```

Gender: F
Location: Northampton, Massachusetts
Conditions: Miscarriage in first trimester; Normal pregnancy; Viral sinusitis (disorder); Acute bronchitis (disorder); Primary fibromyalgia syndrome;
Observations:
- Body Height: 164.3
- Pain severity - 0-10 verbal numeric rating [Score] - Reported: 2.0
- Body Weight: 75.4
- Body Mass Index: 27.9
- Diastolic Blood Pressure: 86.0
- Systolic Blood Pressure: 123.0
- Heart rate: 67.0
- Respiratory rate: 14.0
- Total Cholesterol: 185.3
- Triglycerides: 146.2
Question: Have you experienced any sudden vision changes?

### Response:

🗨️ Model Response:
I'm not sure. I haven't really paid attention to my vision.

✅ Expected (Fine-Tuned) Answer:
I don't think so, but I haven't really noticed.

```

2. Interactive Diagnostic Loop

This section describes how I evaluated language models (LLMs) in a simulated clinical setting where they engage in multi-turn dialogue with a patient (oracle) to gather diagnostic information.

2.1 Oracle Construction

I fine-tuned BioMistral-7B using a custom dataset of doctor-patient question-answer pairs derived from synthetic patient records. The fine-tuned model plays the role of a patient oracle, generating short, lay-language responses based solely on a structured patient context. It is also trained to respond with uncertainty or “I don't know” when asked about symptoms not found in the profile, improving its realism and robustness.

2.2 Interview Simulation

Each benchmark session begins with a randomly sampled patient from *synthea_oracle_context.json*. The target LLM (e.g., GPT-3.5-turbo) assumes the role of a virtual doctor, engaging in up to 20 rounds of interaction to gather clues before making a diagnosis.

At each turn:

- The LLM generates a new question based on the full conversation history, which includes its own previous questions and the oracle's answers.
- This question is passed to the oracle, which responds from the perspective of the synthetic patient.
- The oracle's answer is appended to the conversation, and the loop continues.

This full-history setup mirrors how real doctors recall previous dialogue when interacting with a patient.

2.3 Final Diagnosis Prediction

The LLM may stop at any time—ideally once it has gathered enough evidence—and issue a final diagnosis in the format:

<Final Answer>: ["SNOMED_CODE_1", "SNOMED_CODE_2", ...]

This list is compared against the known ground truth SNOMED codes for that synthetic patient. The goal is to measure whether the LLM can:

- Ask informative, targeted questions,
- Interpret lay-language patient answers correctly,
- And make accurate clinical inferences.¹

2.4 Benchmark Framework

I implemented a fully automated Python benchmark pipeline with the following features:

- Target-agnostic API access: Supports OpenAI and other OpenAI-compatible APIs (e.g., Groq) for the target LLM.
- Modular simulation: The loop cleanly separates oracle response generation, target LLM reasoning, and evaluation.
- Metrics: Reports exact match, precision, recall, and F1 score for predicted SNOMED codes.
- Reproducibility: Configurable seed, number of samples, and max turns.

This benchmark is designed to go beyond static QA and test a model's clinical reasoning and question-asking behavior in a dynamic, interactive context.

Results

1. Evaluation Criteria

Each model was evaluated based on its ability to interactively infer a patient's SNOMED codes through multi-turn dialogue. The benchmark uses standard precision and recall metrics by comparing the model's final predicted codes to the ground truth SNOMED codes. We tested 3

patient cases per model. Each conversation was capped at 20 turns. We tested three different LLMs - GPT-3.5-turbo, GPT-4.0-turbo and Grok-3-Mini. Unfortunately, all 3 benchmarked runs ended with 0.00 average precision and 0.00 average recall — none of the predicted codes matched any ground truth codes. I will share some more details.

2. GPT-3.5-Turbo Performance

Despite being well-tuned for conversation, GPT-3.5-turbo struggled to elicit sufficient information from the oracle. In several runs:

- The questions were generally sensible and focused on symptoms, fever, and respiratory issues.
- However, the oracle's limited or uncertain responses (e.g., "I'm not sure") hindered the LLM's ability to gather high-confidence signals.

```
Running case 1/3 for patient 050358a1-fcf7-1182-7ad1-6b663afa3002...
Patient Context: Age: 20
Gender: M
Location: Waltham, Massachusetts
Conditions: Acute viral pharyngitis (disorder); Otitis media; Whiplash injury to neck; Viral sinusitis (disorder); Acute bronchitis (disorder); Child attention deficit disorder; Suspected COVID-19; COVID-19; Pneumonia (disorder); Hypoxemia (disorder)
Observations:
- Body Height: 116.4
- Pain severity - 0-10 verbal numeric rating [Score] - Reported: 1.0
- Body Weight: 25.6
- Body Mass Index: 18.9
- Body mass index (BMI) [Percentile] Per age and gender: 96.8
- Diastolic Blood Pressure: 77.0
- Systolic Blood Pressure: 114.0
- Heart rate: 62.0
- Respiratory rate: 14.0
- Leukocytes [# /volume] in Blood by Automated count: 5.2
Gold SNOMED Codes: ['195662009 - Acute viral pharyngitis (disorder)', '65363002 - Otitis media', '39848009 - Whiplash injury to neck', '444814009 - Viral sinusitis (disorder)', '10509002 - Acute bronchitis (disorder)', '192127007 - Child attention deficit disorder', '840544004 - Suspected COVID-19', '840539006 - COVID-19', '233604007 - Pneumonia (disorder)', '389087006 - Hypoxemia (disorder)']
Turn 1: LLM Question: What symptoms are you currently experiencing?
Turn 1: Oracle Answer: I've been having a sore throat and a stuffy nose.
Turn 2: LLM Question: Have you had a fever recently?
Turn 2: Oracle Answer: Yes, I've been running a high temperature.
Turn 3: LLM Question: Do you have any difficulty breathing or chest pain?
Turn 3: Oracle Answer: I'm not sure. I haven't really noticed anything like that.
Turn 4: LLM Question: Do you have any cough or phlegm production?
Turn 4: Oracle Answer: Yes, I've been coughing a lot and producing phlegm. It's been bothering me.
Turn 5: LLM Question: Do you have any recent history of travel or exposure to sick individuals?
Turn 5: Oracle Answer: I'm not sure. I haven't really been around anyone who's been sick.
Turn 6: LLM Question: Do you have any body aches or fatigue?
Turn 6: Oracle Answer: I'm not sure. I haven't really been paying attention to how my body feels.
Turn 7: LLM Question: Have you recently lost your sense of taste or smell?
Turn 7: Oracle Answer: I'm not sure. I haven't really noticed anything different with my sense of taste or smell. Why do you ask?
LLM Final Response: Final Answer: ["386661006", "49727002"]
```

Sample - Even though it failed to pinpoint the exact SNOMED codes, Cough(49727002) and Fever(386661006) are pretty accurate diagnoses for this patient.

3.3 GPT-4-Turbo Performance

GPT-4-turbo showed better reasoning and persistence in its questioning strategy:

- It asked more detailed follow-ups, often probing symptoms like nasal discharge, facial pressure, or medication history.

- However, it too suffered from the oracle's evasiveness or lack of concrete answers.
- In all three cases, the average precision and recall remained at 0.00, despite deeper conversations.

```

LLM Final Response: <Final Answer>: ["840539006", "1023001"]
Gold SNOMED Codes: ['75498004 - Acute bacterial sinusitis (disorder)', '65363002 - Otitis media', '10509002 - Acute bronchitis (disorder)', '44465007 - Sprain of ankle', '36971009 - Sinusitis (disorder)', '840544004 - Suspected COVID-19', '840539006 - COVID-19']
Turn 1: LLM Question: What symptoms are you currently experiencing?
Turn 1: Oracle Answer: I've been feeling congested and having trouble breathing. My throat is also sore.
Turn 2: LLM Question: Do you have a fever, or have you noticed any other symptoms such as coughing or sneezing?
Turn 2: Oracle Answer: Yes, I've been coughing a lot and my throat hurts. I also have a high fever.
Turn 3: LLM Question: How long have you been experiencing these symptoms?
Turn 3: Oracle Answer: A few weeks now, it seems like it's not getting better.
Turn 4: LLM Question: Have you been in contact with anyone who is sick, or have you traveled recently?
Turn 4: Oracle Answer: I'm not sure. I haven't been keeping track of my contacts. I haven't traveled recently.
Turn 5: LLM Question: Have you taken any medications for your symptoms, and if so, have they helped?
Turn 5: Oracle Answer: I'm not sure. I haven't been keeping track of what I've taken. I'll have to ask my doctor about that.
Turn 6: LLM Question: Have you noticed any difficulty swallowing, or have you experienced any changes in your voice or hoarseness?
Turn 6: Oracle Answer: I'm not sure. I haven't really noticed anything like that.
Turn 7: LLM Question: Have you experienced any weight loss, night sweats, or chills along with your other symptoms?
Turn 7: Oracle Answer: I'm not sure. I haven't really been keeping track of my weight or symptoms. I've been trying to focus on getting better.
Turn 8: LLM Question: Have you experienced any chest pain or tightness?
Turn 8: Oracle Answer: I don't think so. I haven't really noticed anything like that.
Turn 9: LLM Question: Have you noticed any wheezing or whistling sounds when you breathe?
Turn 9: Oracle Answer: I'm not sure. I haven't really been paying attention to my breathing.
Turn 10: LLM Question: Have you experienced any swelling or pain in your lymph nodes, particularly around your neck?
Turn 10: Oracle Answer: I'm not sure. I haven't noticed anything like that.
Turn 11: LLM Question: Have you experienced any sinus pressure or headaches along with your other symptoms?
Turn 11: Oracle Answer: Yes, I've been feeling a lot of pressure in my sinuses and my head has been pounding. It's really uncomfortable.
Turn 12: LLM Question: Based on the symptoms you've described, it sounds like you may have a respiratory infection. Do you have any allergies, or have you experienced similar symptoms in the past during certain seasons or environments?
Turn 12: Oracle Answer: I'm not sure. I haven't really thought about it. I don't think so.
Turn 13: LLM Question: Have you experienced any shortness of breath or difficulty in breathing that seems to worsen with physical activity?
Turn 13: Oracle Answer: I'm not sure. I haven't really noticed any issues with my breathing.
Turn 14: LLM Question: Given the duration and nature of your symptoms, have you noticed any changes in the color or consistency of your mucus or sputum, especially when coughing?
Turn 14: Oracle Answer: I'm not sure. I haven't really been paying attention to my mucus. I haven't had any issues with coughing.
Turn 15: LLM Question: Do you smoke, or are you exposed to secondhand smoke or other irritants in your environment?
Turn 15: Oracle Answer: I don't smoke, but I'm not sure about my parents. They might be exposed to secondhand smoke.
Turn 16: LLM Question: Have you noticed any improvement or worsening of your symptoms at any specific time of day or in specific environments?
Turn 16: Oracle Answer: I'm not sure. I haven't really noticed any patterns in how my symptoms change throughout the day.
Turn 17: LLM Question: Have you had any nasal discharge, and if so, what color and consistency is it?
Turn 17: Oracle Answer: Yes, my nose has been running a bit, and it's been clear and runny.
Turn 18: LLM Question: Have you experienced any symptoms of gastroesophageal reflux, like heartburn or acid taste in your mouth, particularly at night?
Turn 18: Oracle Answer: I don't think so. I haven't noticed anything like that.
Turn 19: LLM Question: Have you experienced any loss of taste or smell along with your other symptoms?
Turn 19: Oracle Answer: I'm not sure. I haven't really noticed anything different with my sense of taste or smell.
Turn 20: LLM Question: Have you had any episodes of wheezing or feeling like your chest is tight or heavy?
Turn 20: Oracle Answer: I'm not sure. I haven't really noticed anything like that.
LLM Final Response: <Final Answer>: ["840539006", "1023001"]

```

Sample - Even though it failed to pinpoint all the exact SNOMED codes, it did predict COVID-19 (840539006) correctly.

3.4 Grok-3-Mini Performance

Grok-3-mini struggled the most:

- It failed to produce a well-formed SNOMED code format in its final answers.
- Even when the dialogue hinted at relevant conditions, the model didn't synthesize that into correct codes.
- Precision and recall were both 0.00, and formatting issues suggest either prompt misunderstanding or weaker capabilities in medical contexts.


```

Gold SNOMED Codes: ['65363002 - Otitis media', '195662009 - Acute viral pharyngitis (disorder)']
Turn 1: LLM Question: What symptoms have you been experiencing recently?
Turn 1: Oracle Answer: My ear has been hurting and my throat has been sore.
Turn 2: LLM Question: How long have you been experiencing the ear pain and sore throat?
Turn 2: Oracle Answer: My ear has been hurting for a few days now, and my throat feels really sore.
Turn 3: LLM Question: Have you experienced any fever, cough, nasal congestion, or ear discharge?
Turn 3: Oracle Answer: Yes, I've been coughing a lot and my ear feels clogged.
Turn 4: LLM Question: Have you experienced any fever, nasal congestion, or ear discharge?
Turn 4: Oracle Answer: Yes, my ear has been hurting and my nose feels stuffy.
Turn 5: LLM Question: Have you experienced any fever, chills, or ear discharge?
Turn 5: Oracle Answer: I'm not sure. I haven't really been paying attention to that. I've been more focused on my ear pain.
Turn 6: LLM Question: Have you noticed any discharge or fluid coming from your ear?
Turn 6: Oracle Answer: Yes, my ear has been feeling a bit stuffy and there's some discharge.
Turn 7: LLM Question: Have you noticed any pain or pressure in your sinuses, such as around your forehead or cheeks?
Turn 7: Oracle Answer: I'm not sure. I haven't really noticed any pain in my sinuses.
Turn 8: LLM Question: Have you noticed the color or consistency of the ear discharge (e.g., clear, yellow, or bloody)?
Turn 8: Oracle Answer: I'm not sure. I haven't really looked closely at my ear discharge. It's been bothering me for a while now.
Turn 9: LLM Question: Can you describe the color and consistency of the ear discharge, such as whether it's clear, yellow, thick, or watery?
Turn 9: Oracle Answer: I'm not sure, I haven't really looked closely at the ear discharge. It's probably clear or yellow, but I'm not sure.
Turn 10: LLM Question: Have you had any fever, even if mild, or felt chills recently?
Turn 10: Oracle Answer: I'm not sure. I haven't been keeping track of my temperature.
Turn 11: LLM Question: Have you measured your temperature or felt unusually warm recently?
Turn 11: Oracle Answer: I'm not sure. I haven't checked my temperature. But I haven't felt unusually warm.
Turn 12: LLM Question: Have you experienced any hearing loss, dizziness, or recent exposure to cold weather, water, or sick contacts?
Turn 12: Oracle Answer: I'm not sure. I haven't really thought about it. I don't think I've had any hearing issues. I haven't been exposed to anything like that.
Turn 13: LLM Question: Have you had any recent colds, allergies, or been swimming or exposed to water?
Turn 13: Oracle Answer: Yes, I've been sneezing a lot and my throat feels scratchy. I also got a cold recently.
Turn 14: LLM Question: Have you experienced any swelling, redness, or tenderness around the outer ear?
Turn 14: Oracle Answer: I don't think so. I haven't noticed anything like that.
Turn 15: LLM Question: Have you ever had ear infections in the past or a history of allergies?
Turn 15: Oracle Answer: Yes, I've had ear infections before, and I'm allergic to pollen.

### Instruction:
You are a patient with the patient context:
Age: 14
Gender
Turn 16: LLM Question: Have you experienced any headache, fatigue, or worsening of symptoms over the past few days?
Turn 16: Oracle Answer: I'm not sure. I haven't really been paying attention to how I've been feeling. I've been a bit tired, but I'm not sure if that's related.
Turn 17: LLM Question: Have you noticed any recent changes in your hearing, such as muffled sounds or difficulty hearing?
Turn 17: Oracle Answer: I'm not sure. I haven't really been paying attention to my hearing.
Turn 18: LLM Question: Have you noticed any changes in your hearing, such as muffled sounds, ringing in the ear, or difficulty understanding speech?
Turn 18: Oracle Answer: I'm not sure. I haven't really thought about it. Maybe I should pay more attention.
Turn 19: LLM Question: Have you noticed any difficulty hearing, such as muffled sounds or trouble understanding speech, in the affected ear?
Turn 19: Oracle Answer: Yes, it's hard to hear clearly in my ear. It sounds muffled and muffled.
Turn 20: LLM Question: Have you experienced any nausea, vomiting, or balance issues, such as vertigo?
Turn 20: Oracle Answer: I don't think so. I haven't noticed anything like that.
LLM Final Response: <Final Answer>: ["109214009", "84419008"]

```

Sample - The oracle gave good information but gork-30-mini failed to generate a valid SNOMED code. Both 109214009 and 84419008 are invalid codes.

3.5 Oracle Performance

The BioMistral-7B model, fine-tuned with LoRA on synthetic doctor-patient conversations, proved to be a highly effective oracle for simulating realistic patient interactions. In most benchmarked cases:

- Responses were concise, coherent, and expressed in lay language, staying true to the style expected from a non-clinical respondent.
- The model often provided contextually relevant answers without over-committing to details it wasn't given — a behavior that aligns well with how real patients may hesitate or fail to recall specifics.

For example, in response to a vague or speculative question, the model often replied:

- *"I'm not sure. I haven't really paid attention to that."*

This patient-like vagueness is desirable in general.

Limitations Observed

Despite its strong general performance, a few limitations were evident:

- **Prompt Leakage or Instructional Artifacts:** In rare cases, the model output included prompt formatting tokens like **### Instruction:** or commented out sections — indicating incomplete separation between fine-tuning prompt structure and generation behavior. This is likely due to training artifacts or insufficient masking during supervised fine-tuning.
- **Occasional Over-verbosity:** Though most answers were short, a few generations drifted into verbose or non-patient-like language, especially for questions involving medications or diagnoses.
- **Lack of Variation in Uncertainty:** The model often defaulted to “*I’m not sure*” even when indirect signals were present in the patient context. This made it overly cautious, potentially limiting the target LLM’s ability to converge on correct SNOMED predictions.

Overall Verdict

Despite these limitations, the oracle is remarkably effective for simulating patients in an interactive diagnostic benchmark. Its strengths in tone and realism contribute significantly to the benchmark’s realism. That said, future work could improve its ability to:

- Interpret structured observations more robustly,
- Provide more diverse uncertainty responses,
- Filter out formatting artifacts from its outputs.

4. Conclusion

This project introduced a novel benchmark framework that evaluates large language models (LLMs) not just on static medical knowledge, but on their ability to interactively elicit patient symptoms through dialogue and arrive at an accurate diagnosis in the form of SNOMED codes. To make this possible, I fine-tuned a BioMistral-7B model to act as a simulated patient—oracle—that responds in concise, layperson language. This oracle powered the benchmark loop by answering iterative questions posed by target LLMs like GPT-3.5-turbo, GPT-4, and Grok-3-mini.

While the target LLMs demonstrated impressive reasoning and information-gathering capabilities, the results showed that the final prediction accuracy was limited more by the quality of the oracle’s responses than by the LLMs’ ability to ask good questions. In particular, the fine-tuned BioMistral oracle sometimes produced overly vague or inconsistent answers, or even leaked instruction-style responses when not properly grounded by its prompt. These limitations highlight that building a high-quality, controllable oracle remains a bottleneck for developing interactive medical benchmarks.

Nonetheless, this benchmark provides a valuable foundation for future work. It emphasizes the importance of modeling interactive reasoning, not just retrieval, in clinical AI systems. Going forward, improvements could be made by (1) scaling up the fine-tuning data, (2) incorporating feedback signals for conversational coherence, and (3) testing more LLMs and oracle configurations. This benchmark also opens the door to testing reinforcement learning, long-context models, and few-shot tool usage in a clinically grounded setting.

This work is a first step toward evaluating LLMs not just as knowledge bases, but as interactive assistants capable of mimicking the diagnostic reasoning process of real clinicians.