MedQDx

DIAGNOSE LIKE A DOCTOR

Mai Werthaim & Maya Kimhi

INTRODUCTION

Since patients rarely present a complete clinical picture at first, physicians must actively ask targeted questions and investigate to obtain all the information needed for an accurate diagnosis and appropriate treatment decisions. Therefore, the ability to conduct a dynamic dialogue and ask the right questions is essential for high-quality diagnosis.

In recent years, large language models (LLMs) have increasingly been integrated into clinical workflows and diagnostic support systems, leveraging their advanced language capabilities to assist in patient assessment, medical reasoning, and clinical decision-making.

Despite the promise of LLMs in diagnosis, current benchmarks do not assess their ability to conduct strategic, adaptive inquiry, as they rely on fully revealed patient cases.

MedQDx is a benchmark that its' main goal is to simulates realistic diagnostic uncertainty, for evaluating large language models (LLMs) in the context of interactive diagnosis through patient interrogation.

FORMAL TASKS

To create a novel benchmark, we proceed through three primary missions:

Case Preparation
Task: Generating Patient Cases by transforming numericl readings into plain text
Input: Diseases and their symptoms
Output: Patient cases and their diseases

Benchmark Creation
Task: Simulating Doctor & patient Question - Answer - Diagnosis dialogue
Input: Partial patient case
Output: “Doctor” questions and diagnosis

Model’s Evaluation
Task: Comparing similarity between doctor’s diagnosis and real prognosis
Input: Doctor’s questions and diagnosis
Output: Zero-Shot Diagnostic Accuracy (ZDA), Mean Question-based Diagnostic Similarity (MQD) and Mean of Max Similarity Across Row (MMS)

PRIOR ART

Name	Med-PaLM 2	AMIE	ClinicalGPT-R1
Source	Singhal, K., et al. (2025). Toward expert-level medical question answering with large language models. Nature..	Tu, T., et al. (2025). Towards conversational diagnostic artificial intelligence. Nature.	Lan, W., et al. (2025). ClinicalGPT-R1: Pushing reasoning capability of generalist disease diagnosis with large language model. arXiv.
Goal	Enhance reasoning and grounding in long-form medical question answering through ensemble refinement and chain-of-retrieval strategies	Conduct AI-driven diagnostic dialogue by simulating clinician–patient interactions	Improve generalist disease diagnosis
Approach	Transformer+ fine-tuning on medical data; uses prompt tuning & ensemble refinement for reliable answers	Vignette generator Dialogue simulator Self-play loops	Synthetic Data Generation Two-Stage Fine-Tuning
Data	USMLE-style questions (MedQA), medical research (PubMedQA), MedMCQA, and clinical topics in MMLU	Real-world transcripts (~99 K conversations from MIMIC-III) and a self-play multi-agent to synthesize new case	Real EHR records with long-chain CoT prompts
Metrics	Accuracy	Clinicians scored AMIE's history-taking and diagnostic reasoning using PACES-style criteria	Accuracy
Results	86.5 % accuracy on MedQA (+19 % over Med-PaLM)	Generated ~12K dialogues AMIE matched or exceeded benchmarks on key axes	Outperforms GPT-4o in Chinese diagnosis tasks and matches GPT-4o in English on MedBench-Hard

DATA DESCRIPTION

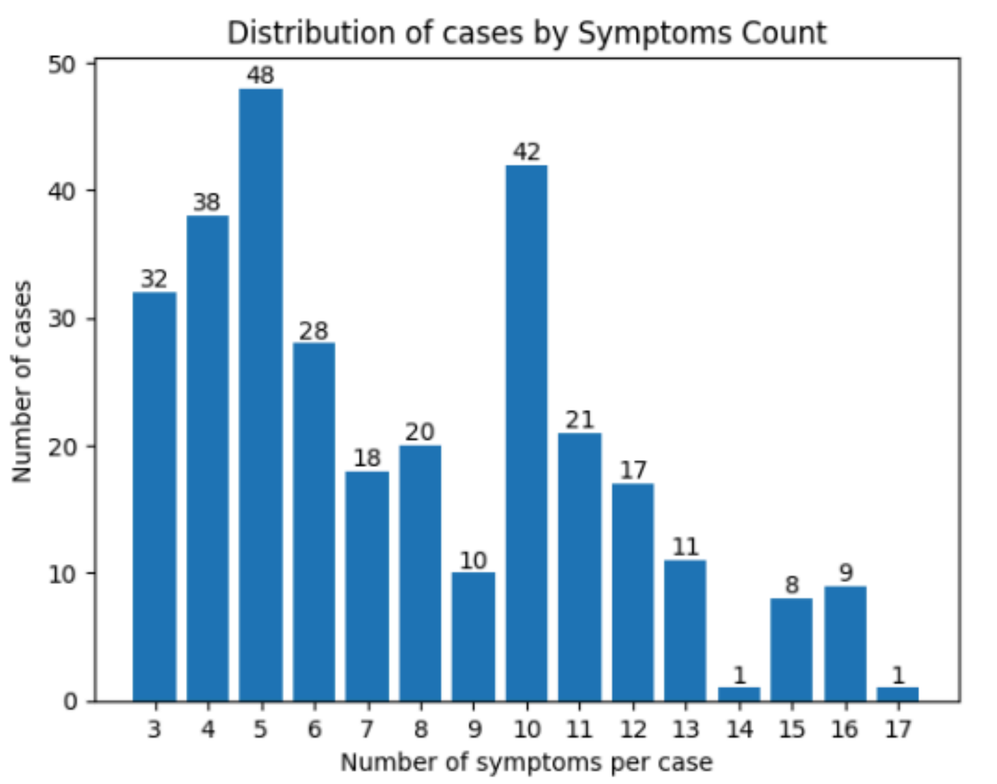
Dataset: Symptom-Disease Prediction Dataset (SDPD)
Labels: Prognosis

Raw data - Diseases and their Symptoms

- 4961 rows
- 132 symptoms
- 41 unique diseases
- 4657 duplicate rows

Raw data statistics:

- Average rows per disease: 7.41
- Disease with most rows: Chickenpox (10 rows)
- Disease with fewest rows: AIDS (5 rows)
- Each disease have 3-17 symptomes



NLP PIPELINE

Case Preparation

Input: Raw Data (100 random sample of Symptoms and prognosis)

Output: Table of 5 columns (Diagnosis, symptoms, full patient case, 80% case, 50% case)

Task: Patient Case Creation

Model: GPT-4o-mini

Metric: Jaccard overlaps

Benchmark Generation

Input: Table of 5 columns (Diagnosis, symptoms, full patient case, 80% case, 50% case)

Output: For each case, 3 quartets of columns (doctor's question, patient answer, doctor's diagnosis)

Task: Doctor & patient role playing

Model: GPT-4.1 as doctor & GPT-4o-mini as patient, text-embedding-3-small

Metric: Disease - Diagnosis Similarity

Evaluation

Input: Doctor’s diagnosis

Output: Similarity between doctor’s diagnosis and real prognosis

Task: Comparing between doctor’s diagnosis and real prognosis

Model: GPT-4.1

Metric: ZDA , MQD and MMS

CODE ORGANIZATION

MedQDx GitHub - https://github.com/MaiWert/MedQDx

- EDA & Baseline
- Doctor and Patient Dialogue
- Benchmark
- Evaluation - results files

BASELINE

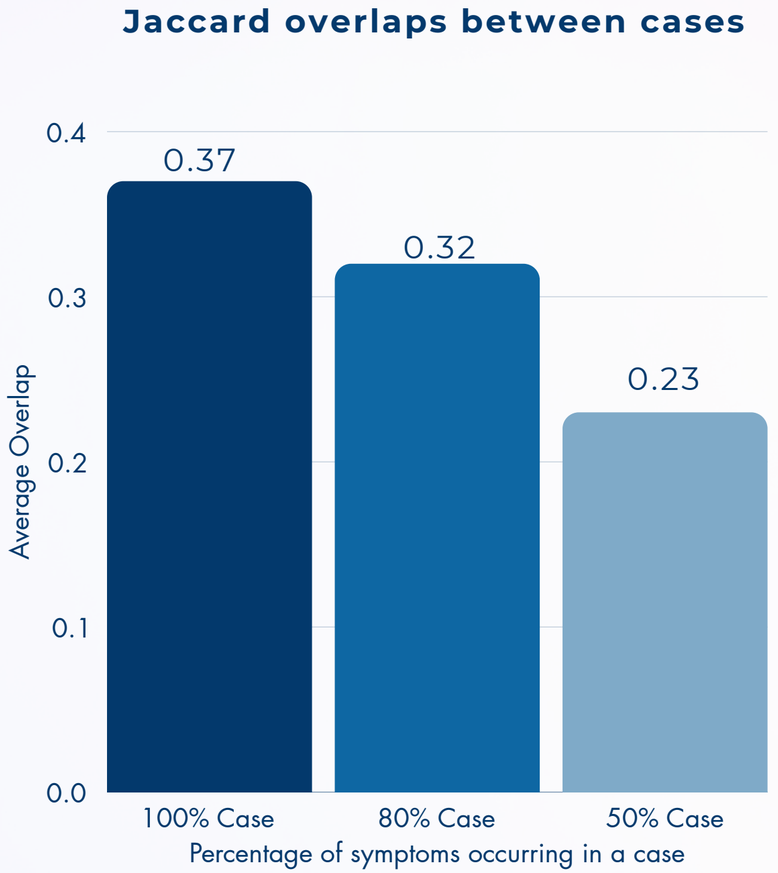
Random Sampling:
A random sample of 100 rows is selected from the original dataset.
Each row represents a real disease profile with associated symptoms.

Patient Case Generation:
For each selected disease instance, a synthetic patient case is generated using a language model (gpt-4o-mini).

Each case includes:

- Full Case: All symptoms associated with the disease.
- 80% Case: Approximately 80% of the symptoms.
- 50% Case: Approximately 50% of the symptoms.
-

Text generation evaluation:
Computes Jaccard overlaps between the true and extracted symptom sets, and checks that overlap decreases from 100% → 80% → 50%



As expected, the overlap decreases from 100% → 80% → 50% as the percentage of available data in the case descends.

EXAMPLE

Prognosis: GERD

Symptoms:
stomach_pain, acidity, ulcers_on_tongue, vomiting, cough

100% Case:
A 35-year-old male presents with a history of chronic stomach pain and frequent episodes of acidity. He reports having developed painful ulcers on his tongue over the past week, which has made it difficult for him to eat and drink. Alongside these symptoms, he has experienced bouts of vomiting, particularly after meals, and a persistent cough that seems to worsen at night. He denies any recent travel or changes in diet but admits to high-stress levels at work.

80% Case:
A 35-year-old male presents with chronic stomach pain and frequent episodes of acidity. He mentions painful ulcers on his tongue that have persisted for about a week, making it uncomfortable to eat. Additionally, he has experienced occasional vomiting, especially after meals, but does not report a significant cough. He attributes some of his discomfort to high-stress levels at work.

50% Case:
A 35-year-old male reports experiencing chronic stomach pain and frequent episodes of acidity. He has been dealing with these symptoms for the past couple of months. Although he does not mention ulcers on his tongue or vomiting specifically, he notes that the stomach discomfort has made it challenging for him to manage his diet. He also mentions that he has been under considerable stress lately.

EXAMPLE

Question_1:
Have you noticed if your stomach pain or acidity symptoms are related to eating specific foods or occurring at particular times of the day?

Answer_1:
I have not noticed that.

Diagnosis_1:
Dyspepsia

Similarity_1:
0.478660937811607

Question_2:
Have you experienced any unintentional weight loss, changes in appetite, or difficulty swallowing along with your stomach pain?

Answer_2:
No, I have not noticed any unintentional weight loss, changes in appetite, or difficulty swallowing.

Diagnosis_2:
Gastritis

Similarity_2:
0.512374242089334

Question_3:
Have you had any episodes of black or tarry stools, blood in your stool, or noticed any changes in your bowel habits recently?

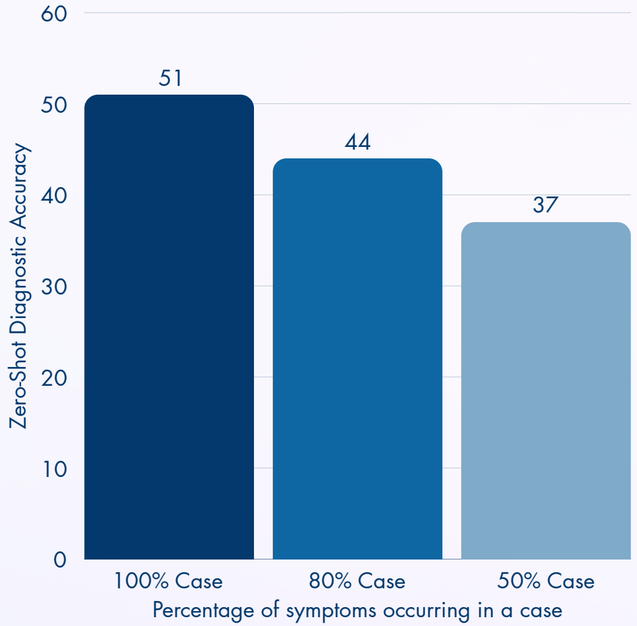
Answer_3:
No, I have not noticed any episodes of black or tarry stools, blood in my stool, or changes in my bowel habits recently.

Diagnosis_3:
Gastroesophageal reflux disease (GERD)

Similarity_3:
1.00000

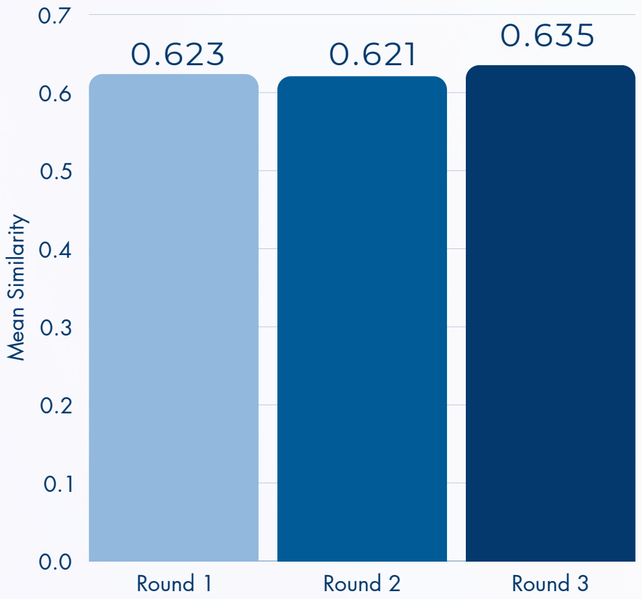
RESULTS

Zero shot accuracy between cases



As expected, the accuracy of the diagnosis increases as the percentage of available data in the case rises.

Mean similarity between rounds

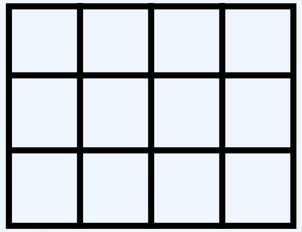


Mean of Max Similarity Across Row
0.657

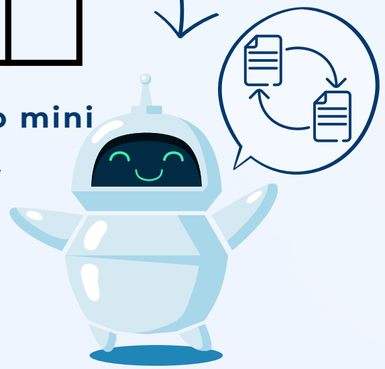
The mean similarity steadily increases across rounds, suggesting that successive doctor–patient interactions improve alignment with the true prognosis.

Patient Case Creation

Symptom-Disease Prediction Dataset



GPT 4o mini



Patient cases

Jaccard Symptom Overlap

Benchmark Creation

3 Rounds of Doctor - Patient dialogue
Question-Answer-Diagnosis-Similarity

Benchmark Evaluation

Zero-Shot Diagnostic Accuracy (ZDA)

Mean Question-based Diagnostic Similarity (MQD)

Mean of Max Similarity Across Row