MedQDx

Interim Presentation

Mai Werthaim & Maya Kimhi

Project Description

Clinical diagnosis is an interactive process that relies on asking the right questions in response to incomplete patient information.

While LLMs show promise in diagnostic tasks, existing benchmarks expose them to fully revealed cases, ignoring the role of strategic inquiry.

MedQDx is a novel benchmark that simulates realistic, partial clinical scenarios and evaluates LLMs' ability to reach a diagnosis through adaptive, question-driven reasoning.

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

Case Preparation

Input: Diseases and their symptoms
Output: Patient cases and their diseases
Task: Patient Case Creation

Benchmark Creation

Input: Partial patient case
Output: “Doctor” questions and diagnosis
Task: Doctor & patient role playing

Model's Evaluation

Input: Doctor's questions and diagnosis, Model's questions and diagnosis
Output: Zero-Shot Diagnostic Accuracy (ZDA), Mean Questions to Correct Diagnosis (MQD), and Interrogation Sequence Efficiency (ISE)
Task: Comparing models questions and diagnosis to MedQDx benchmark

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

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

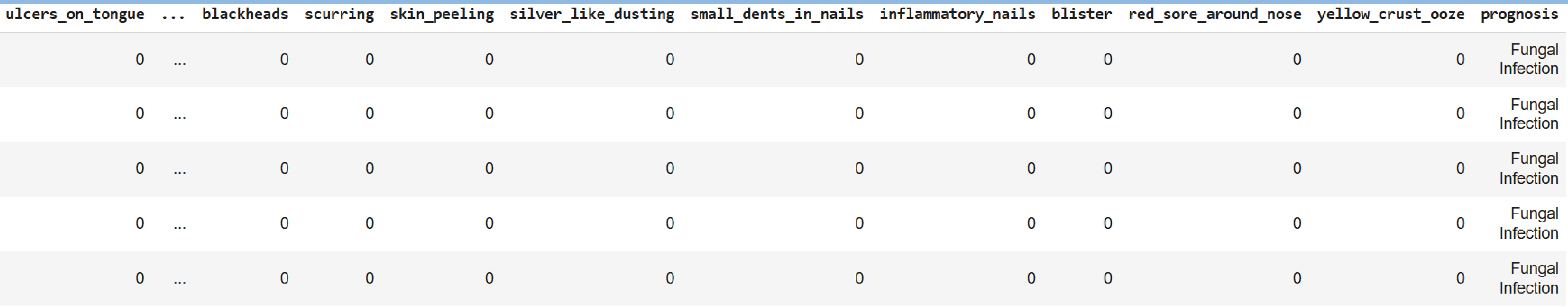
Model: GPT-4.1

Metric: ZDA and Mean similarity

Data exploration

Raw data - Diseases and their Symptoms

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



ulcers_on_tongue	...	blackheads	scurring	skin_peeling	silver_like_dusting	small_dents_in_nails	inflammatory_nails	blister	red_sore_around_nose	yellow_crust_ooze	prognosis
0	...	0	0	0	0	0	0	0	0	0	Fungal Infection
0	...	0	0	0	0	0	0	0	0	0	Fungal Infection
0	...	0	0	0	0	0	0	0	0	0	Fungal Infection
0	...	0	0	0	0	0	0	0	0	0	Fungal Infection
0	...	0	0	0	0	0	0	0	0	0	Fungal Infection

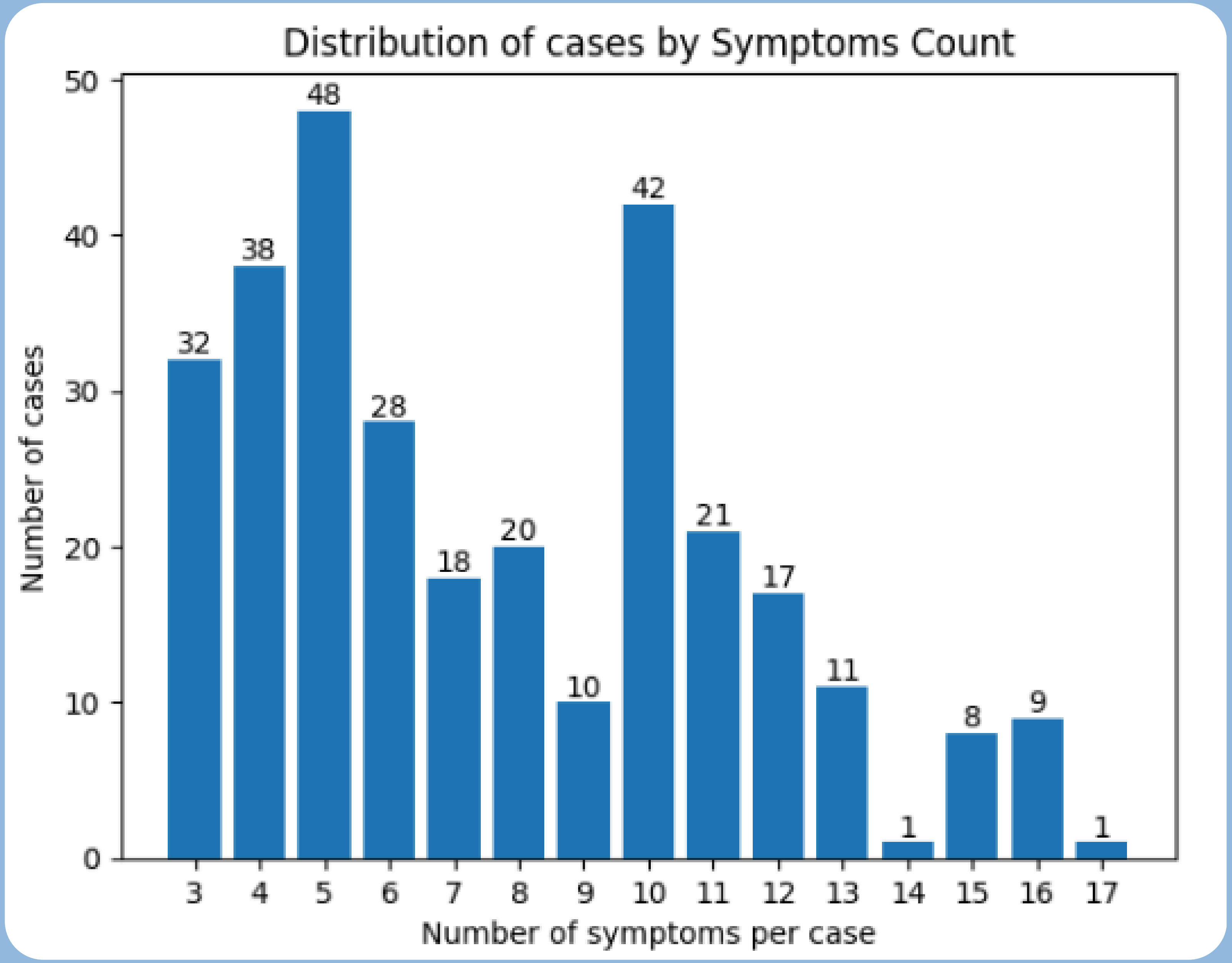
Data exploration

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

Data Treating:

- Duplicate deletion
- Removal of symptoms not associated with any disease
- Selection of cases with >=5 symptoms



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).

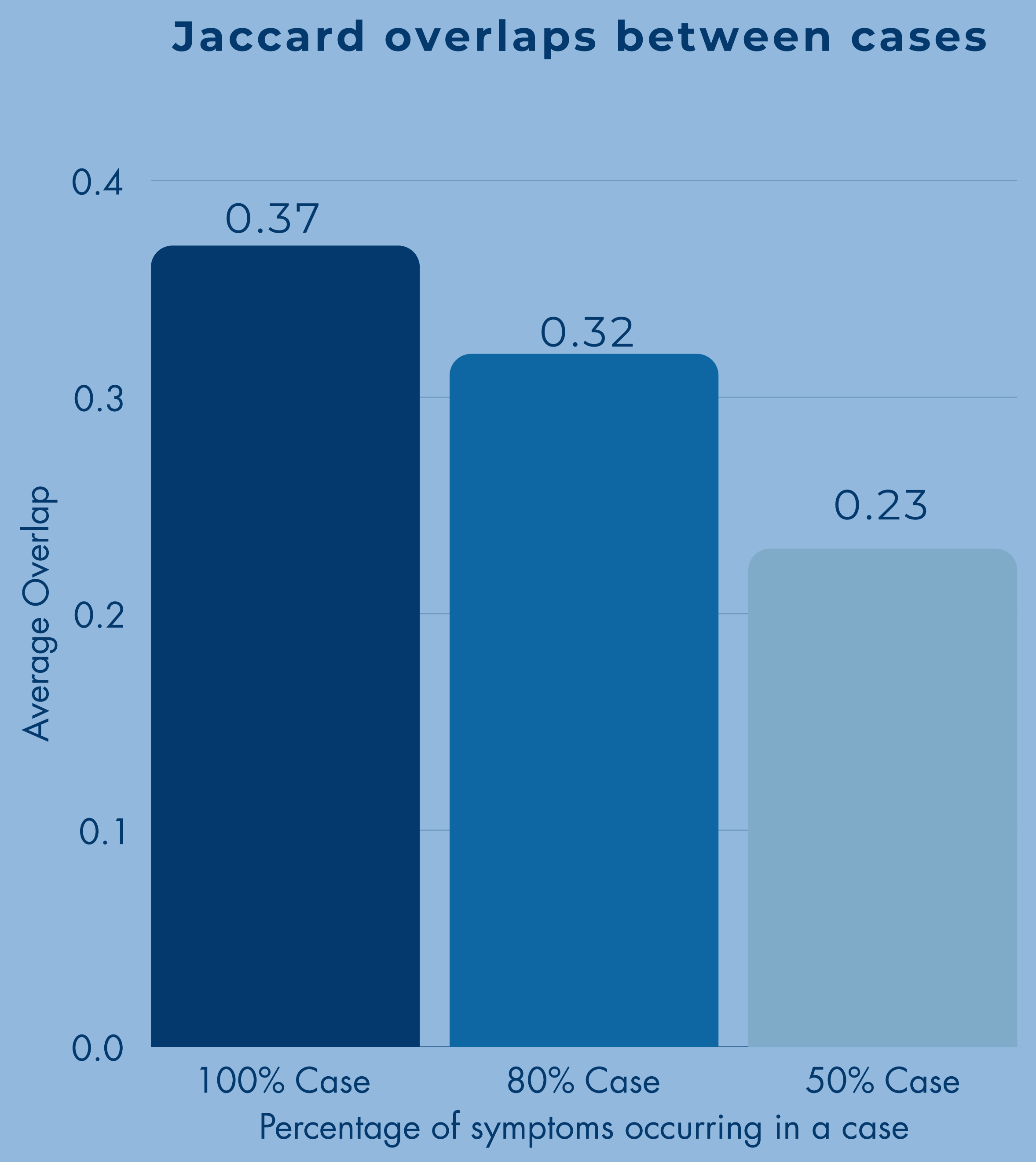
Each case includes:

- 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%

Baseline

	Example case (GERD)
100%	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%	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%	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.



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

Insights

The data source is rich enough to provide good patient cases for diagnosis

There is a relationship between the amount of exposure and accuracy.

Recommendations

Assessment whether dataset size can be reduced.

Zero-shot diagnosis for further evaluate the robustness of the generated cases.