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: Diseases and their Symptoms
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	Benchmark Generation	Model Examination	Evaluation
Input: Raw Data: 100 random sample of Symptoms and diagnoses.	**Input:** A table of 4 columns - Diagnosis, full patient case, 80% case, 50% case.	**Input:** Full patient case, 80% patient case, 50% patient case.	**Input:** Doctor's question & diadnosis, model's question & diagnosis
Output: Table of 4 columns - Diagnosis, full patient case, 80% case, 50% case.	**Output:** K pairs of columns (doctor's question, doctor's diagnosis)	**Output**: K pairs of columns (model question and diagnosis)	**Output:** Similarity between doctor questions & student questions
Task: Patient Case Creation	**Task:** Doctor & patient role playing	**Task:** model & patient role playing.	**Task:** Comparing models to doctor
Model: MedLlama2	**Model:** Me-LLaMA 13B as doctor & DeepSeek-R1 as patient	**Model:** Models & DeepSeek-R1 as patient	**Model:** None (NLP Metrics)
Metric: Model-based evaluation (PubMedBERT)	**Metric:** Zero-Shot Diagnostic, Accuracy, AUC	**Metric:** Zero-Shot Diagnostic, Accuracy, AUC	**Metric:** ZDA, MDQ, and ISE

Data exploration

Raw data - Diseases and their Symptoms

- 2564 rows
- 400 symptoms
- 133 unique diseases
- 13 duplicate rows

	Unnamed: 0	pain chest	shortness of breath	dizziness	asthenia	fall	syncope	vertigo	sweat sweating increased	palpitation	...	prodrome	hypoproteinemia	alcohol binge episode	abdomen acute	air fluid level	catching breath	large-for-dates fetus	immobile	homicidal thoughts	prognosis
0	0	0	1	0	0	0	1	0	1	1	...	0	0	0	0	0	0	0	0	0	hypertensive disease
1	0	0	1	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	diabetes
2	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	depression mental , depressive disorder
3	0	1	1	0	0	0	0	0	1	0	...	0	0	0	0	0	0	0	0	0	coronary arteriosclerosis ,coronary heart disease
4	0	0	1	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	pneumonia
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2559	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	affect labile
2560	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	affect labile
2561	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	affect labile
2562	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	affect labile
2563	0	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	decubitus ulcer

Data exploration

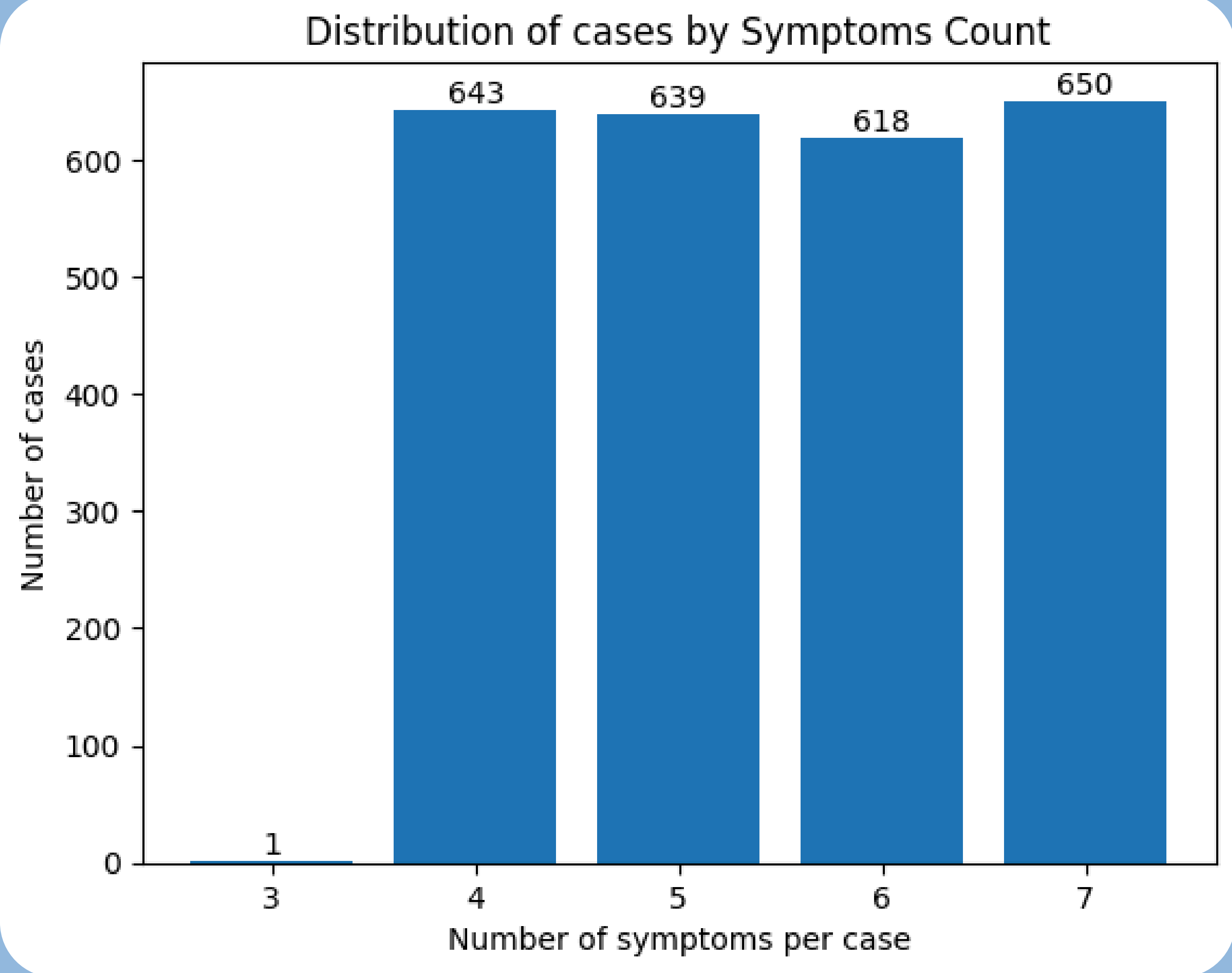
Raw data statistics:

- Average rows per disease: 19.18
- Disease with most rows: bipolar disorder (43 rows)
- Disease with fewest rows: decubitus ulcer (3 rows)

- Each disease have 3-7 symptomes
- The most common symptom is pain (323 cases)
- The least common symptom is dizzy spells (1 case)

Data Treating:

- Duplicate deletion
- Removal of symptoms not associated with any disease
- Selection of cases with >=4 symptomes



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 (MedLlama2).
Each case includes:

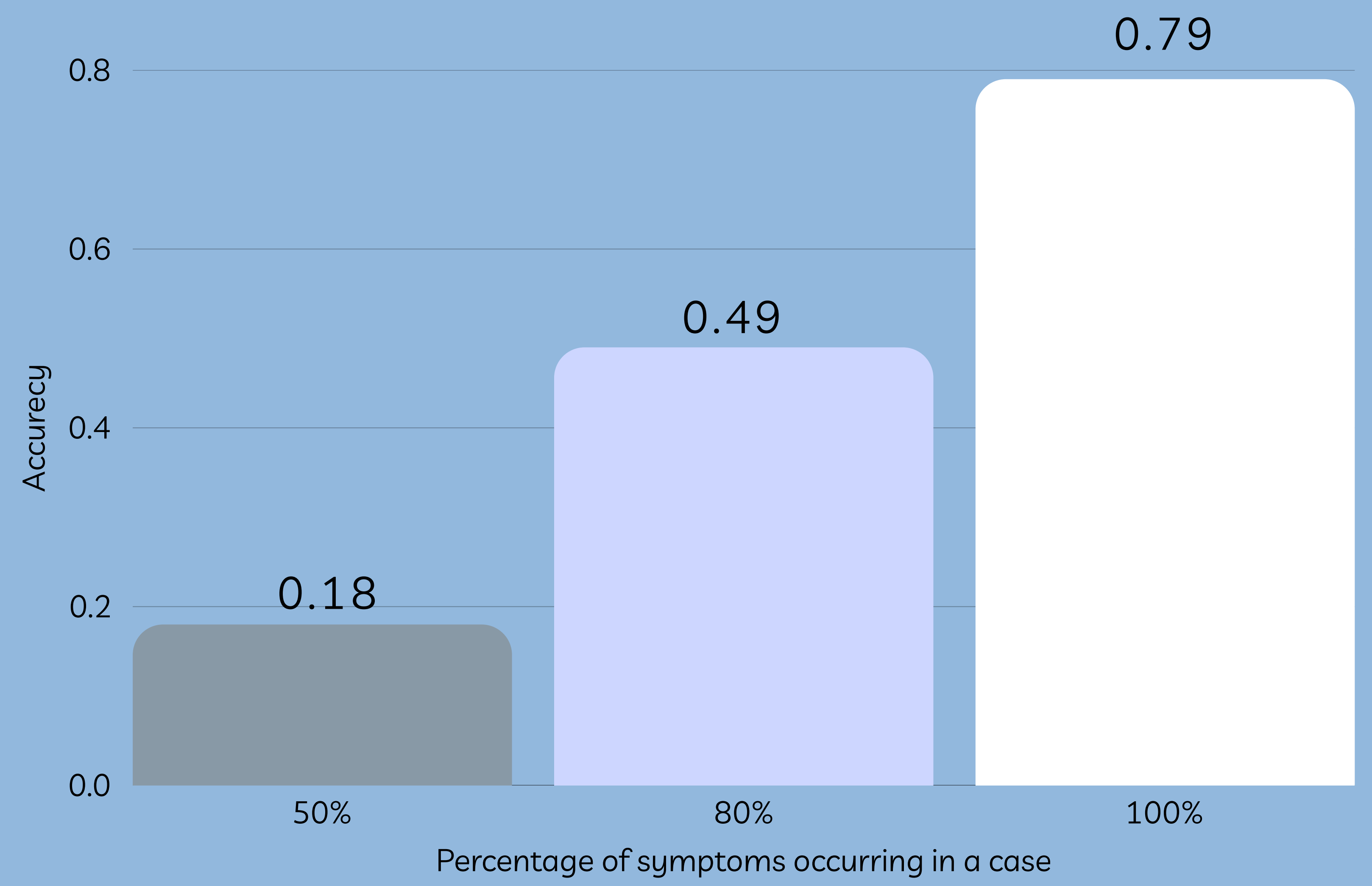
- Full Case: All symptoms associated with the disease.
- 80% Case: Approximately 80% of the symptoms.
- 50% Case: Approximately 50% of the symptoms.

Text-Based Diagnosis Modeling as doctor:
PubMedBERT is fine-tuned on the generated case descriptions to simulate a doctor's diagnosis.

Accuracy Comparison Across Case Levels:
Accuracy is measured for each level to assess how case completeness affects diagnosis quality.

Baseline

	Example case (Asthma)
100%	A 20-year-old female presents with shortness of breath, productive cough, distress respiratory, symptom aggravating factors.
80%	A 20-year-old female presents with symptom aggravating factors, distress respiratory, productive cough.
50%	A 20-year-old female presents with distress respiratory, shortness of breath.



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

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.