3MDBench: Medical Multimodal Multi-agent Dialogue Benchmark

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, Andrey Savchenko

SBER AI Lab

fall into ML 2025
4th conference on machine learning & AI

1. MOTIVATION

Telemedicine is reshaping access to healthcare with the usage of Large Vision-Language Models (LVLMs). However, **most existing telemedicine benchmarks for LVLMs are limited**:

- ✖ They focus on static QA or multiple-choice tasks
- ✖ Lack of multi-turn, interactive dialogue
- ✖ Ignore patient personality and behavioral variation
- ✖ Rarely include visual clinical inputs

We introduce **3MDBench – Medical Multimodal Multi-agent Dialogue Benchmark** that:

- Simulates dialogue consultations with **Doctor Agent** using image modality;
- Introduces **Patient Agent** with different temperament-dictated behaviours;
- Evaluates diagnostic and communication quality via **Assessor Agent**;
- Benchmarks **different LVLMs** as Doctor Agents across **multiple strategies**.

2. DATA COLLECTION

1. Forming a diagnosis list

- **> 611K** real telemedicine consultations
- **180M outpatient records** for the distribution validation
- **34 diseases** across five domains

2. Obtaining images

- **2996 clinical images** (public datasets, Kaggle, Bing, etc.)
- **≥64 images/class** for balance
- Filtered via automation + manual review

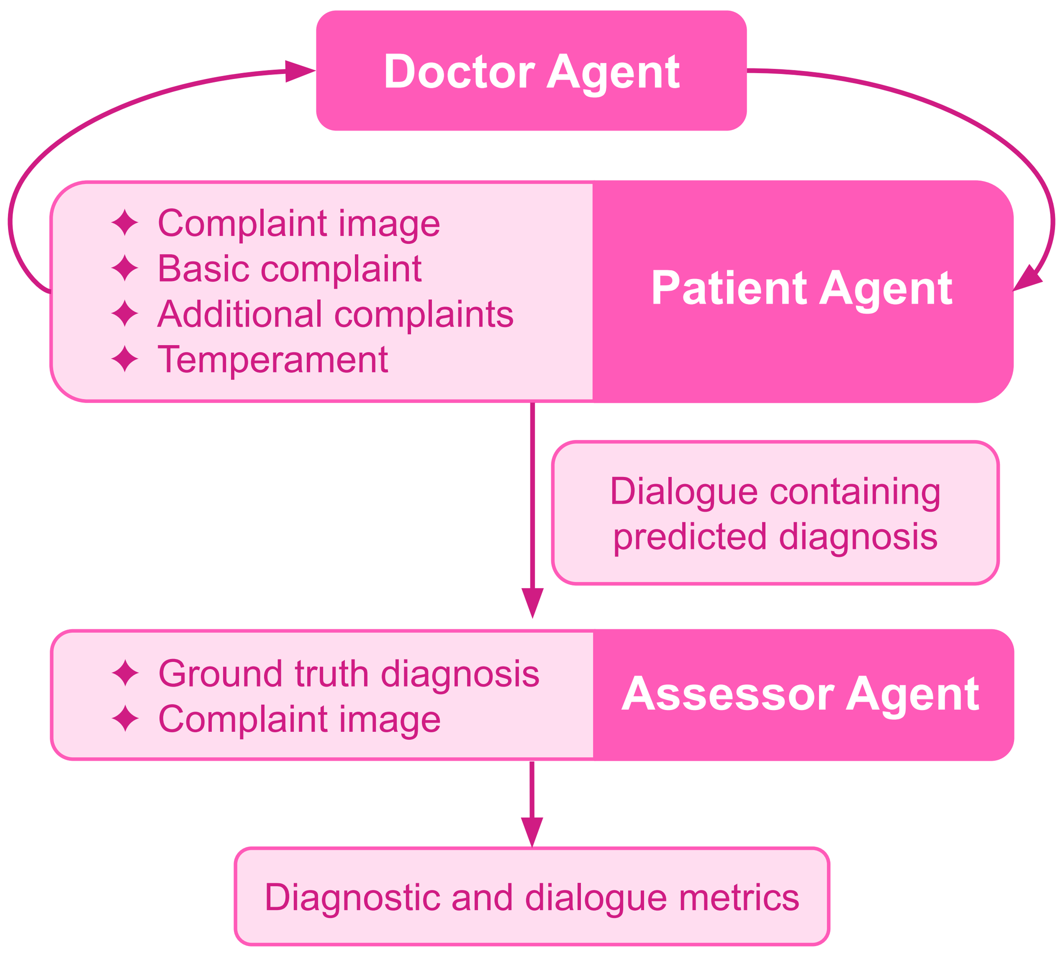
3. Generating complaints

- Generated via **GPT-4o-mini**
- **One basic complaint** per diagnosis
- **List of additional complaints** per image: duration, intensity, history

4. Ensuring multimodality

- ✔ Each case = image + basic + additional symptoms
- ✔ We ensured medical validation of the generated symptoms
- ✔ We obtained private train/val parts

3. SIMULATION FLOW



4. AGENTS DESIGN

Patient Agent

Llama-3-8B

- Complains, reports symptoms, asks questions
- Expects to discover its diagnosis and recommendations on what to do
- Selected based on:
  - **Instruction following** (0-5 LLM judge score)
  - **Answer relevance** (0-5 LLM judge score for each answer)
  - **Factuality** (embedding closeness to the actual symptoms) for each answer

Assessor Agent

Qwen2-VL-72B-Instruct

- Evaluates clinical competence using the adapted Mini-CEX scale[1]
  - Medical interviewing skills
  - Humanistic care
  - Treatment abilities
- Extracts the final diagnosis to assess the diagnostic accuracy
- Selected based on:
  - **Alignment with human assessments** for clinical competence via Cohen's k
  - **F1-score** for the diagnostics

[1] Shi, Xiaoming et al. "LLM-Mini-CEX: Automatic Evaluation of Large Language Model for Diagnostic Conversation." ArXiv abs/2308.07635 (2023)

Doctor Agent

Open-source and proprietary models with multiple strategies

- Has a goal of **determining the diagnosis and providing recommendations** on treatment and further diagnostics
- Receives the basic complaint and the image as the first message
- Should conduct the diagnostic dialogue: ask clarifying questions regarding symptoms
- After diagnostics, it should answer the patient's questions

5. RESULTS

CLINICAL COMPETENCE

Model	1.1	1.2	1.3	2.1	2.2	3.1	3.2	4.1
GPT, dialogue, no image	**1.00**	**1.00**	0.95	**1.00**	**1.00**	0.89	0.90	1.45
GPT, dialogue + image	0.99	**1.00**	0.96	**1.00**	**1.00**	0.90	0.91	1.61
GPT, dialogue + image + rationale	0.96	0.99	0.89	0.99	0.97	0.78	0.78	1.31
GPT, dialogue + image + rationale + external cues	0.96	0.99	0.96	0.99	0.98	0.88	0.88	1.47
Llama-3.2-Vision	0.99	0.99	0.94	0.99	0.99	0.75	0.74	1.45
Qwen2-VL	0.90	0.93	0.78	0.92	0.90	0.61	0.61	1.16
MedGemma-4B	0.97	0.98	0.94	0.99	0.98	0.79	0.80	1.42
MedGemma-27B	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	0.90	0.88	**1.67**
Gemma3-27B	0.99	**1.00**	0.99	**1.00**	**1.00**	**0.97**	**0.98**	1.57

- Diagnostic and treatment abilities (3.1 and 3.2) demonstrate how **domain-specific models are better aligned for telemedicine than the general-domain ones**.

CLINICAL COMPETENCE CRITERIA

Primary item	Secondary item
Medical Interviewing Skills	1.1. Enquiry about medical history
	1.2. Enquiry about current symptoms
	1.3. Explaining the basis of conclusions
Humanistic Care	2.1. Communicating with respect and empathy
	2.2. Respecting the individual wishes
Diagnostic and Treatment Abilities	3.1. Providing accurate diagnostic plan
	3.2. Providing accurate treatment plan
Overall Clinical Competence	4.1. Level of clinical competence: unsatisfactory, satisfactory, excellent

DIAGNOSTIC RESULTS

Model name	Configuration	F1 Score	Number of utterances
EfficientNetV2-XL	Fine-tuned on the train part	61.0	-
GPT 4o-mini	No dialogue, image + general complaint	50.4	-
	No dialogue, image + all complaints	66.8	-
	Dialogue, no image	52.8	15.22 (±3.63)
	Dialogue + image	54.2	13.32 (±3.33)
	Dialogue + image + rationale	56.9	14.99 (±4.23)
	Dialogue + image + rationale + cues from pretrained CNN	**70.3**	14.48 (±3.97)
Llama-3.2-Vision	Dialogue + image	41.5	14.49 (±4.02)
Qwen2-VL	Dialogue + image	39.0	15.11 (±4.39)
MedGemma-4B	Dialogue + image	37.9	17.48 (±4.84)
MedGemma-27B	Dialogue + image	45.7	16.88 (±5.25)
Gemma3-27B	Dialogue + image	51.1	14.81 (±3.81)

- **Dialogue improves diagnostic accuracy**
  - However, F1-score remains below full-information levels;
  - Using **cues from a pretrained CNN** improves F1-score to 20%.
- **General-purpose models outperform domain-specialized ones**, likely due to training biases toward specific imaging tasks or structured QA formats.
- **The visual channel shortens and refines the dialogue**.

6. DIALOGUE EXAMPLE

True diagnosis	eczema
Predicted diagnosis	eczema
Diagnostic F1	1.0
1.1	1
1.2	1
1.3	1
2.1	1
2.2	1
3.1	1
3.2	1
4.1	excellent

7. CONCLUSIONS

- **3MDBench** – a multi-agent, multimodal benchmark simulating doctor–patient dialogue with varying temperaments and consultation assessment.
- **Multiple models and strategies** assessment.
- We demonstrate that:
  - **Dialogue and expert visual cues** enhance F1-score;
  - **Domain tuning** does not always improve multi-turn diagnostic accuracy;
  - **There should be a balance** between clinical competence and diagnostic accuracy.

TEST 3MDBENCH: