EMNLP 2025
Suzhou, China | 中国苏州

3MDBench: Medical Multimodal Multi-agent Dialogue Benchmark

Andrey V. Savchenko
[1]Scientific director at Sber AI Lab
[2]Full. Prof., Leading Researcher at HSE University

Email: avsavchenko@hse.ru
URL: www.hse.ru/en/staff/avsavchenko

27.10.2025

Motivation

GigaDoc: Sber’s Multimodal medical assistant

Overview

Telemedicine is reshaping access to healthcare, enabling remote diagnostics through dialogue and images. Recent advances in large vision-language models (LVLMs) make them promising candidates for virtual medical assistants. However, most existing benchmarks for medical LVLMs are limited:

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

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

- Simulates **doctor-patient dialogue consultations** with 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;
- Promotes scalable, multimodal medical AI grounded in real-world interaction.

Comparison with existing benchmarks

T	TDT	N	M	S	D	A	P	CQ	F	L
DS	D	MedDialog-EN (Zeng et al., 2020)	T	300K	+	-	-	-	-	EN
DS	D	MedDialog-CN (Zeng et al., 2020)	T	1100K	+	-	-	-	-	CN
DS	D	MedDG (Liu et al., 2022)	T	18K	+	-	-	-	-	CN
DS	D	CMtMedQA (Yang et al., 2023)	T	70K	+	-	-	-	-	CN
DS	D	Icliniq-10K (Li et al., 2023b)	T	10K	+	-	-	-	-	EN
DS	D / QA	BianQueCorpus (Chen et al., 2023)	T	2437K	+	-	-	-	-	CH
DS	D / QA	HealthCareMagic-100k (Li et al., 2023c)	T	100K	+	-	-	-	-	EN
DS	D / QA	Psych8k (Yuan et al., 2025)	T	8K	+	-	-	-	-	EN
DS	D	IMCS-21 (Chen et al., 2022)	T	811	+	+	-	-	-	CN
DS	D	NoteChat (Wang et al., 2024a)	T	30K	+	+	-	-	-	EN
DS	D	MTMedDialog (Feng et al., 2025)	T	10.1K	+	+	-	-	-	EN
BM	QA	Cholec80-VQA (Twinanda et al., 2016)	M	9K	-	-	-	-	-	EN
BM	QA	VQA-RAD (Lau et al., 2018)	M	3.5K	-	-	-	-	-	EN
BM	QA	PathVQA (He et al., 2020)	M	6K	-	-	-	-	-	EN
BM	QA	SLAKE (Liu et al., 2021)	M	2K	-	-	-	-	-	EN
BM	QA	RadBench (AI, 2024)	M	137K	-	-	-	-	-	EN
BM	QA	MMMU (H & M) (Yue et al., 2024)	M	11.5K	-	-	-	-	-	EN
BM	QA	OmniMedVQA (Hu et al., 2024)	M	128K	-	-	-	-	-	EN
BM	QA	GMAI-MMBench (Chen et al., 2024)	M	26K	-	-	-	-	-	EN
BM	QA	Medical-Diff-VQA (Hu et al., 2025)	M	70K	-	-	-	-	-	EN
BM	D	MediQ (Li et al., 2024c)	T	1.2K	+	+	-	-	-	EN
BM	D	AgentClinic (Schmidgall et al., 2024)	M	457	+	+	-	-	-	EN
BM	D	MedAgentSim (Almansoori et al., 2025)	M	637	+	+	-	-	-	EN
BM	D	AI Hospital (Fan et al., 2024)	M	506	+	+	+	+	-	CN
BM	D	Dr.APP (Zhu and Wu, 2025)	T	1.5K	+	+	+	+	-	EN
BM	D	3MDBench (Ours)	M	3K	+	+	+	+	+	EN

- **T (Type)**: Dataset (DS) / Benchmark (BM)
- **TDT (Text Data Type)**: Question-Answer pairs (QA) / Dialogues (D)
- **N (Name)** of Dataset/Benchmark
- **M (Modality)**: Text-only (T) / Multimodal (M)
- **S (Size)** of test part of a Benchmark or full size of a Dataset
- **D (Dialogues present)**
- **A (multi-Agent** approach used)
- **P (Personality** modeling used)
- **CQ (Consultation and communication qualities** tested)
- **F (Full-fledged** consultation simulated until both agents naturally conclude the dialogue)
- **L (Language)** of data

Data collection

1. Forming diagnosis list

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

2. Obtaining images

- ☐ **2996 clinical images** (ISIC, Kaggle, etc.) for the test part.
- ☐ **≥64 images/class** for balance.
- ☐ Filtered via automation + manual review.

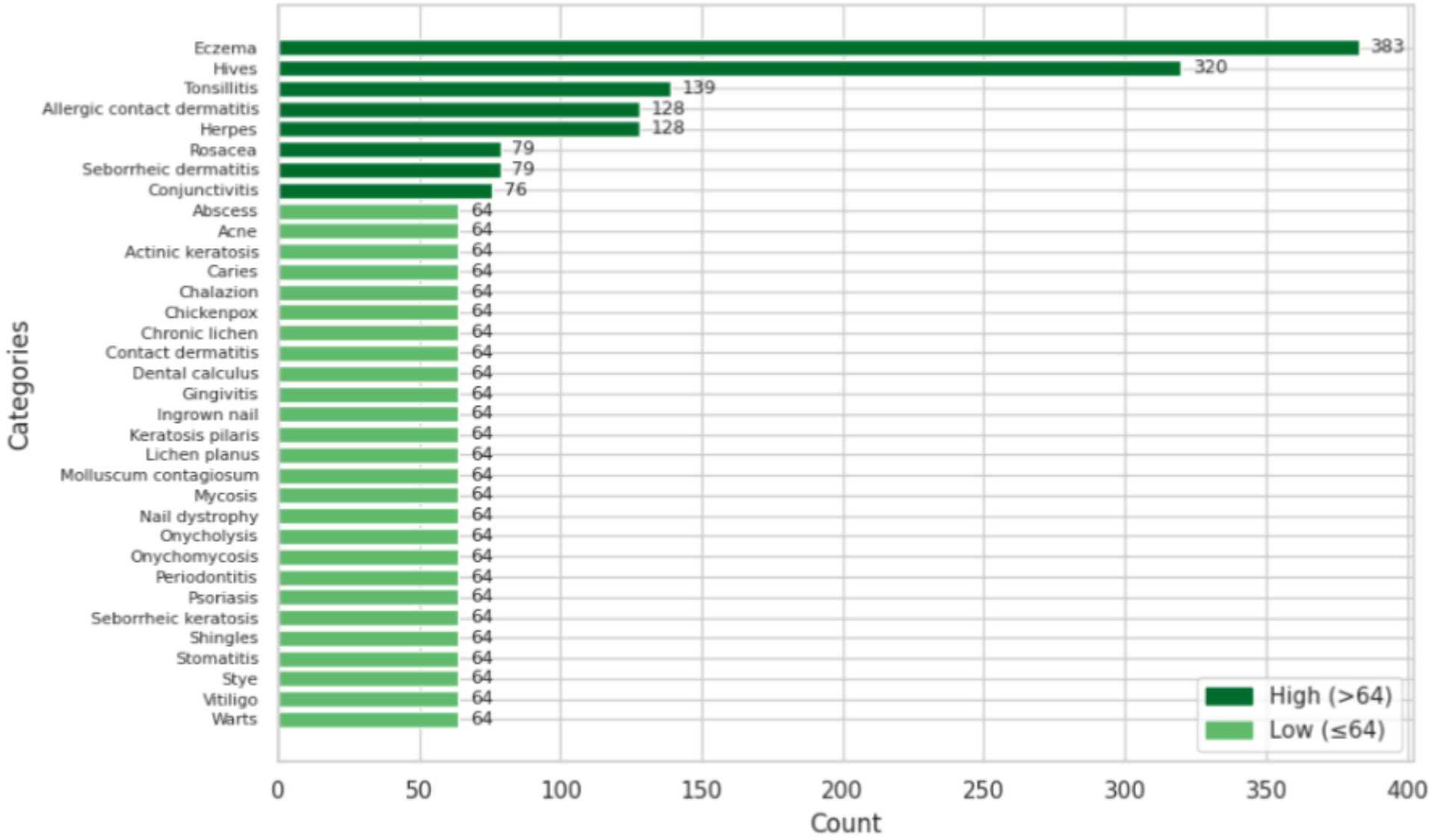
3. Generating complaints

- ☐ Generated via **GPT-4o-mini.**
- ☐ **One general complaint** per disease.
- ☐ **List of structured symptoms** per image: duration, intensity, history.

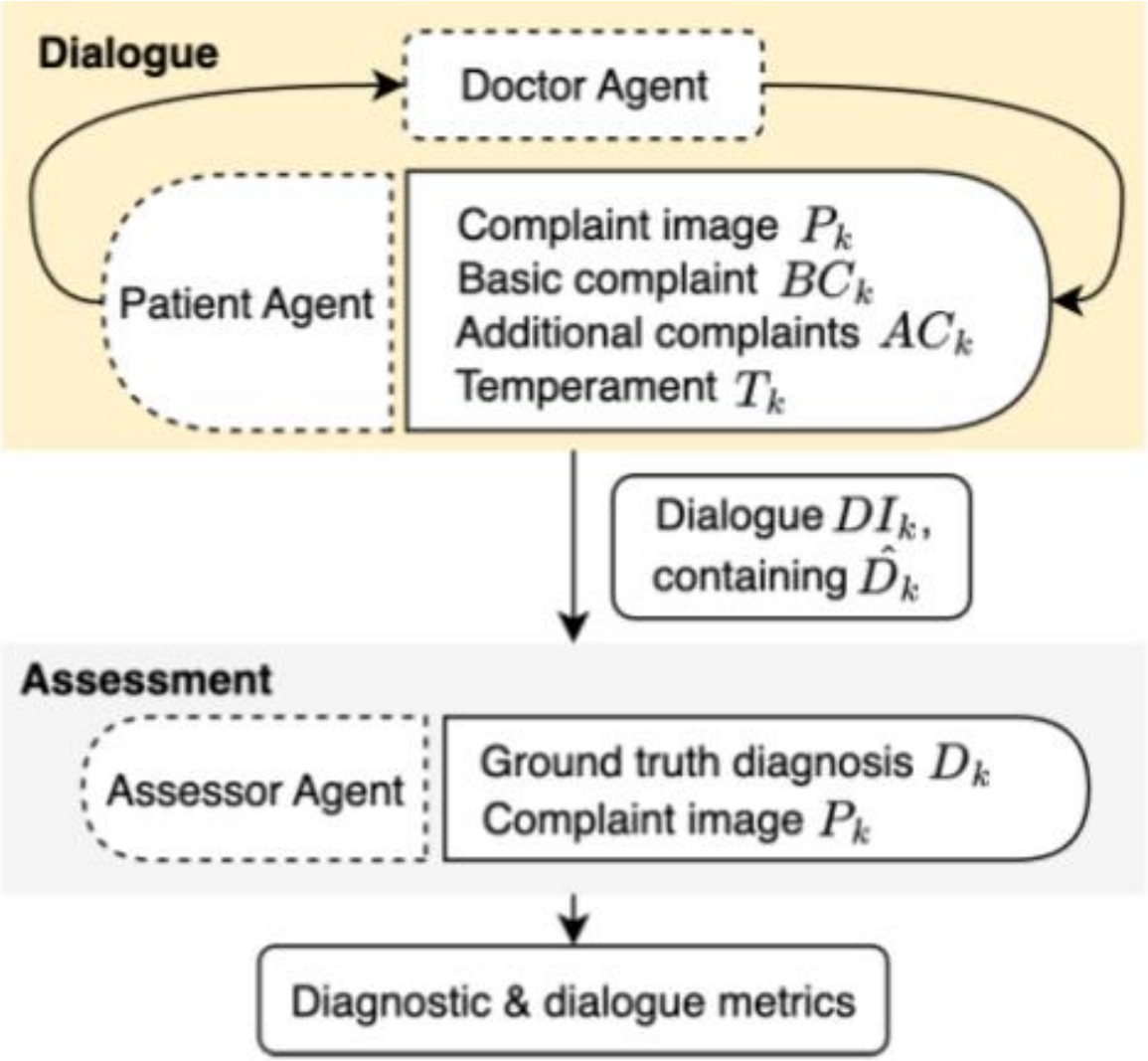
4. Ensuring multimodality

- ✔ Each case = image + general + structured symptoms.
- ✔ We also obtained private train and validation parts.

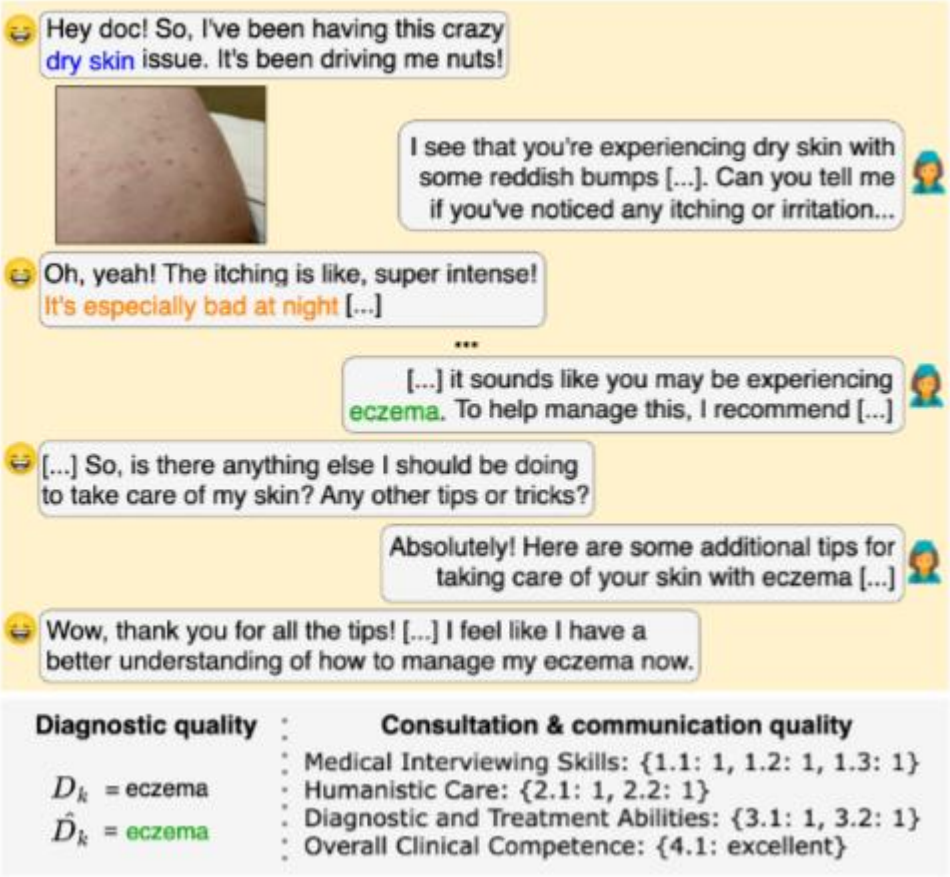
Distribution of classes



Agents' interaction pipeline



Dialogue example



- The dialogue DI_k begins with the Patient Agent of sanguine temperament T_k, whose first reply contains the complaint image P_k and the basic complaint BC_k highlighted in blue

- An additional complaint from the list AC_k appears later in orange, and the final diagnosis D_k - identified and validated by the Assessor Agent - is shown in green.

- The Assessor Agent, based on DI_k and D_k, further provides a structured evaluation of diagnostic performance as well as consultation and communication quality.

True diagnosis	Eczema
Predicted diagnosis	eczema
Diagnostic F1	1.0

Agent design

Assessor Agent	Patient Agent	Doctor 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	**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	● **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

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

Patient Agent

Temperament

Choleric

Never had such a headache: feels like I have a very tight headband!

Sanguine

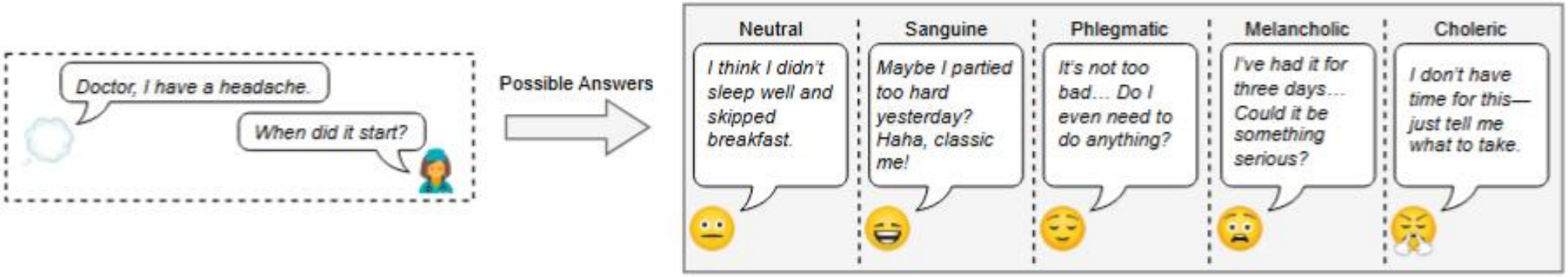
What can I take to cure a headache?

Melancholic

I have a terrible headache… Am I dying?

Phlegmatic

I have a headache.



Model Name	Llama-3-8b	Llama-3.1-8b	Qwen2.5-7B	Qwen2.5-14B	Falcon-7B	GPT-4o-mini
Instruction following	4.72	**4.74**	4.71	4.59	4.37	4.38
Relevance	0.65	0.59	0.84	0.76	**0.90**	0.82
Factuality	0.79	0.77	0.67	0.78	0.59	**0.98**
Mean Rank	**3.00**	3.67	3.33	3.67	4.33	**3.00**

Diagnostic evaluation

- **Dialogue improves diagnostic accuracy**
  - However, F1-score remains below full-information levels;
  - Using **cues from a pretrained CNN improves F1-score to 20%.**

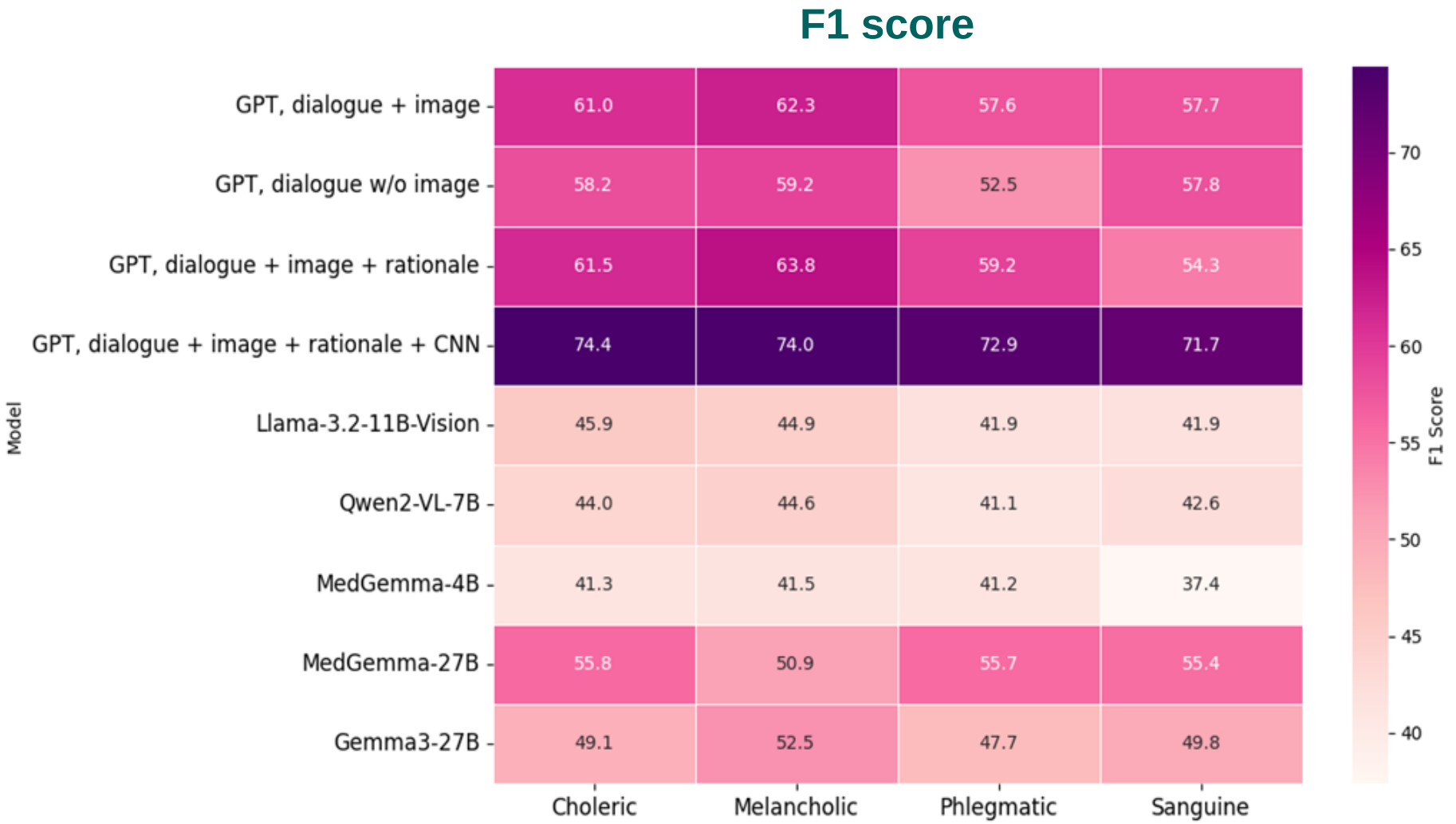
- **General-purpose models outperform domain-specialized ones, likely due to** training biases toward specific imaging tasks or structured QA formats.

- **The visual channel shortens and refines the dialogue.**

Model	Configuration	F1 Score	No. 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
	Dialogue + image	54.2	13.32
	Dialogue + image + rationale	56.9	14.99
	Dialogue + image + rationale + external cues	**70.3**	14.48
Llama-3.2-Vision	Dialogue + image	41.5	14.49
Qwen2-VL	Dialogue + image	39.0	15.11
MedGemma-4B	Dialogue + image	37.9	17.48
MedGemma-27B	Dialogue + image	45.7	16.88
Gemma3-27B	Dialogue + image	51.1	14.81

Personality types

- LVLMs handle diverse patient temperaments with stable diagnostic accuracy.
- dialogues with phlegmatic patients are shorter and less informative, slightly lowering competence scores.
- These cases reveal that while models stay robust, they struggle to compensate for passive or minimally cooperative behavior, highlighting the need to evaluate adaptability in challenging interactions.



Clinical competence evaluation

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

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

Models show strong overall clinical competence, with GPT- and MedGemma- based agents excelling in communication and professionalism.
Diagnostic and treatment abilities (3.1 and 3.2) demonstrate how **domain-specific models are better aligned for telemedicine than the general-domain ones.**

Conclusion

● 3MDBench – a multi-agent, multimodal benchmark simulating doctor-patient dialogue with varying temperaments and consultation assessment.

● It supports 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.

Source code

arxiv

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, Andrey Savchenko, 3MDBench: Medical Multimodal Multi-agent Dialogue Benchmark, EMNLP25 (main track)

Thank you!

The work of A. Savchenko was supported by a grant, provided by the Ministry of Economic Development of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000C313925P4G0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated June 20, 2025 No. 139-15-2025-011.