3MDBench: Medical Multimodal Multi-agent Dialogue Benchmark

Ivan Sviridov[1*], Amina Miftakhova[1*], Artemiy Tereshchenko[1], Galina Zubkova[1], Pavel Blinov[1], Andrey Savchenko[1,2,3]

[1]Sber AI Lab, Moscow, Russia, [2]HSE University, Moscow, Russia
[3]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
Correspondence: wchhiaarid@gmail.com, noteisenheim@gmail.com, blinoff.pavel@gmail.com, avsavchenko@hse.ru

Abstract

Though Large Vision-Language Models (LVLMs) are being actively explored in medicine, their ability to conduct complex real-world telemedicine consultations combining accurate diagnosis with professional dialogue remains underexplored. This paper presents **3MDBench** (**M**edical **M**ultimodal **M**ulti-agent **D**ialogue **Bench**mark), an open-source framework for simulating and evaluating LVLM-driven telemedical consultations. 3MDBench simulates patient variability through temperament-based Patient Agent and evaluates diagnostic accuracy and dialogue quality via Assessor Agent. It includes 2996 cases across 34 diagnoses from real-world telemedicine interactions, combining textual and image-based data. The experimental study compares diagnostic strategies for widely used open and closed-source LVLMs. We demonstrate that multimodal dialogue with internal reasoning improves F1 score by 6.5% over non-dialogue settings, highlighting the importance of context-aware, information-seeking questioning. Moreover, injecting predictions from a diagnostic convolutional neural network into the LVLM's context boosts F1 by up to 20%. Source code is available at `https://github.com/univanxx/3mdbench`.

1 Introduction

Telemedicine expands healthcare access and efficiency by enabling real-time consultations and early diagnosis (Stoltzfus et al., 2023). In these consultations, effective communication is essential for diagnostic accuracy and treatment adherence (Mirzaei and Kashian, 2020; Bhaskar et al., 2020). Large Language Models (LLMs) and Vision-Language Models (LVLMs) further enhance telehealth via real-time analysis (Nwankwo et al., 2024), chronic care management (Adeghe et al., 2024), and decision support (Perez et al., 2025), including symptom assessment, test interpretation, and patient interaction (Blinov et al., 2024; Lu et al., 2024b; Mayer et al., 2024; Kumichev et al., 2024).

Despite recent advances in LLMs' application in telemedicine, effective doctor-patient dialogue remains essential. Diagnostic quality depends on how patients articulate symptoms, shaped by emotional, cognitive, and systemic factors (Amelung et al., 2020; Singh and Sittig, 2015). Miscommunication, fear, and low health literacy delay diagnosis (Heyhoe et al., 2018; Nguyen et al., 2024), while temperament influences responsiveness, expressiveness, and trust (Graedon and Graedon, 2014; Meyer et al., 2013). However, existing LLM benchmarks for assessing telemedicine quality offer limited realism by restricting models to multiple-choice tasks (Jin et al., 2020; Kim et al., 2024b), imposing factual and non-reactive dialogue templates with artificial interruptions that prevent full consultations (Li et al., 2024c; Johri et al., 2024), and omitting image modalities representing patient symptoms (Zhu and Wu, 2025).

To address these problems, we introduce **3MDBench** (**M**edical **M**ultimodal **M**ulti-agent **Dialogue** Benchmark) to evaluate LVLM-based consultations in dynamic, realistic scenarios. We use classical temperament theory, which categorizes individuals into four types (sanguine, choleric, melancholic, and phlegmatic) (Steiner, 1985), to model the personal characteristics that might affect engagement and medical alliance (Paap et al., 2022; Hanney et al., 2023). Hence, our 3MDBench features a **Patient Agent**, simulating one of four temperament types shown in Fig. 1, a **Doctor Agent**, LVLM to conduct the telemedicine consultation and serving as the benchmarked component, and an **Assessor Agent**, evaluating diagnostic accuracy and communication quality. We select models that best match defined metrics and human annotations to construct these agents. Built on 34 diagnoses ob-

[*]These authors contributed equally to this work.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26614–26654
November 4-9, 2025

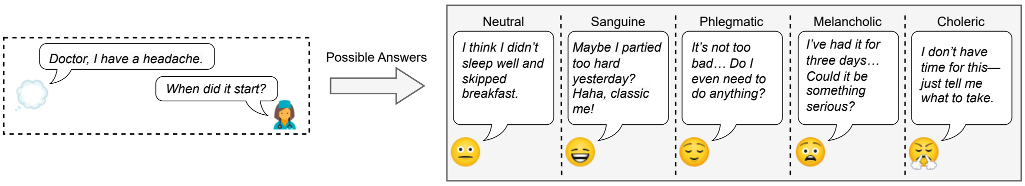


Figure 1: The patient's response during the dialogue with the doctor depends on the temperament.

tained from real-world telemedicine consultations using medical image datasets enriched with textual information, 3MDBench supports both textual and image modalities. We benchmark commercial and open-source LVLMs in different dialogue and non-dialogue settings. Our results show that information-seeking dialogue strategies using medical reasoning and image modality increase the F1 score to 6.5%, highlighting the importance of adaptive, context-aware interaction. Finally, our novel approach, incorporating top-3 predictions from a ConvNet trained on diagnosis into the LVLM's context, boosts the F1 score up to 20%.

In this work:

1. We propose **3MDBench**, an open-source benchmark for evaluating medical dialogue systems, with an Assessor Agent measuring diagnostic accuracy and communication quality, and a Patient Agent simulating temperament-based personality-driven doctor-patient conversations.

2. Using our standardized framework for assessing AI-driven medical consultation quality, we extensively compare open-source and state-of-the-art LVLMs.

3. We analyze multiple dialogue strategies for a Doctor Agent: with and without image modality, implementing rationale generation, and using external cues. We demonstrate the importance of image modality and conducting information-seeking conversations with internal reasoning. Moreover, we show that the diagnostic F1 score of the Doctor Agent improves up to 20% by incorporating top-3 predictions from a specially trained diagnostic convolutional neural network into the LVLM.

2 Related Work

Benchmarks for medical LLMs have focused primarily on factual knowledge, evaluating performance on exams, QA tasks, and case-based reasoning (Jin et al., 2020; Kim et al., 2024b; Pal et al., 2022; Jin et al., 2019; Singhal et al., 2023a) to assess domain understanding and consistency, but overlook interactive and contextual aspects of diagnosis. Recent work highlights the need for dialogue-based evaluation, where models must elicit, interpret, and reason over patient-reported symptoms (Goh et al., 2024; Manes et al., 2024; Li et al., 2023b; Han et al., 2023). However, existing benchmarks often rely on scripted interactions or fixed-response patient agents, limiting their reflection of real-world consultations (Shi et al., 2024).

Multi-agent systems offer a more dynamic alternative, as they simulate collaborative diagnostic workflows, enabling LLMs to interact, reason, and refine decisions over multiple turns (Qiu et al., 2024). Recent efforts have extended this paradigm to simulate patient interactions in telemedicine consultations (Kim et al., 2024a; Mehandru et al., 2024; Li et al., 2024c; Schmidgall et al., 2024; Fan et al., 2024; Zhu and Wu, 2025; Almansoori et al., 2025).

Although these benchmarks represent important progress, none of them simultaneously addresses the following critical limitations:

- Simulating patients as static and personality-free, reducing interactions to factual inputs and overlooking how traits like emotion, communication style, or temperament influence diagnostic accuracy (Amelung et al., 2020).

- Focusing solely on diagnostic and recommendation accuracy, overlooking comprehensive consultation and communication quality assessment based on full-fledged, realistic dialogue—an essential dimension that directly influences patient trust and further diagnostic outcomes (Ha and Longnecker, 2010).

- Excluding image modality from the diagnostic process, despite its significant role in real-world diagnostic decision-making (Agbareia et al., 2025).

Table 1: Comparison of 3MDBench with existing medical benchmarks and datasets. The columns are: **T** (Type: Dataset (DS) or Benchmark (BM)), **TDT** (Text Data Type: Question-Answer pairs (QA) or Dialogues (D)), **N** (Name of Dataset/Benchmark), **M** (Modality: Text-only (T) or Multimodal (M)), **S** (Size of test part of a Benchmark of 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), and **L** (Language of data).

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

To overcome these limitations, we propose **3MDBench** that simulates and evaluates telemedicine consultation with a temperament-driven Patient Agent and an Assessor Agent for accuracy and communication quality. Compared to existing benchmarks (Table 1), we capture the variability and complexity of real-world clinical interactions, enabling richer, more patient-aligned evaluation of medical dialogue systems.

3 Proposed 3MDBench

3.1 Data Collection

Diagnoses. To ensure clinical relevance, we analyzed 611K anonymized visits from a large Eastern European provider from May to October 2024, selecting the top 80% most frequent diagnoses. We examined 180 million outpatient records from the same city through 2022 to validate cross-setting consistency. All diagnoses, originally in ICD-10 (Organization, 2004), were standardized using a physician-curated dictionary. The final set comprises 34 diagnoses across five medical domains as Figure 7 of Appendix A shows.

Image Data. We constructed 3MDBench from 6 open-source datasets, primarily from Kaggle [1] [2] [3] [4], as well as ISIC Archive images (Cassidy et al., 2022), Google SCIN (Ward et al., 2024),

[1] https://github.com/Priyanshu9898/Oral-Disease-Classification

[2] https://www.kaggle.com/datasets/anindamohanta/different-phases-of-tonsilitis

[3] https://www.kaggle.com/datasets/nikhilgurav21/nail-disease-detection-dataset

[4] https://www.kaggle.com/datasets/alisofiya/conjunctivitis

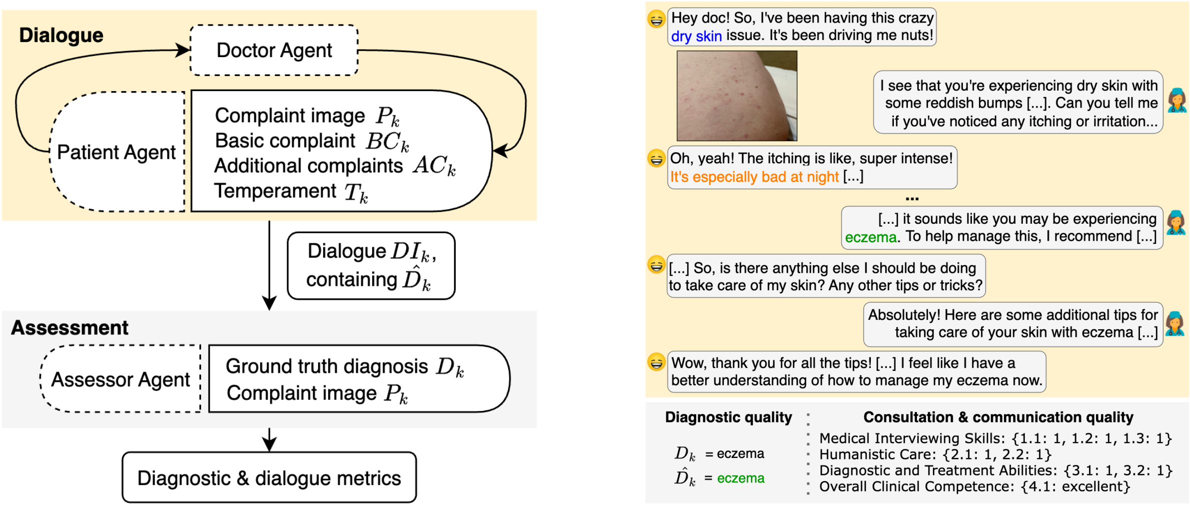


Figure 2: Agents' interaction pipeline in 3MDBench (left) and an example instance (right). The dialogue DI_k begins with the Patient Agent of sanguine temperament T_k, whose first reply contains the complaint image P_k and the basic complaint BC_k highlighted in blue; an additional complaint from the list AC_k appears later in orange, and the final diagnosis $\hat{D}_k$—identified and validated by the Assessor Agent—is shown in green. The Assessor Agent, based on DI_k and $\hat{D}_k$, further provides a structured evaluation of diagnostic performance as well as consultation and communication quality.

and Fitzpatrick17k (Groh et al., 2021), with supplemental data using Bing Image Search[5] (Ghosh et al., 2023). We adjusted disease prevalence to match the distribution from a primary Eastern European telemedicine provider described above to align with real-world telemedicine diagnosis distribution. To ensure sufficient evaluation data and mitigate class imbalance, we set a minimum threshold of 64 images per condition, an empirically determined lower bound based on the maximum number of images available for certain classes across all sources and the Internet. Collected images were filtered through automated quality checks (e.g., size, blur, watermarks) and manual review by one of the coauthors with medical expertise. The final benchmark contains 2,996 images, with class distribution detailed in Figure 6 in Appendix A, plus private training and validation sets with 2,396 and 596 images, respectively.

Enriching Images with Textual Data. To enrich the Patient Agent's input and support more natural telemedicine dialogues, we generated concise, image-associated descriptions for all 2,996 cases. First, using GPT-4o-mini (OpenAI, 2024), selected for its high medical accuracy and relatively low cost (Li et al., 2023a; Smolyak et al., 2024; Abrar et al., 2025), we generated one basic symptom from a human perspective for each of the 34 diagnoses. Then, we expanded each corresponding image for all cases, generating additional structured complaints describing affected areas, duration, intensity, and relevant patient history. This enriched textual input, with a generation prompt in Appendix H.1 and examples in Appendix I, enhances the visual data and provides context for more informative interactions with the Doctor Agent.

3.2 Task Definition

3MDBench simulates realistic medical consultations via multi-turn dialogues between an evaluated **Doctor Agent** and a **Patient Agent** with further evaluation of an **Assessor Agent**, as Figure 2 shows. Each scenario includes a ground truth diagnosis D_k, a complaint image P_k, and symptoms split into a **basic complaint** BC_k (shared initially) and **additional complaints** AC_k (revealed during dialogue). The Patient Agent, shaped by a predefined temperament T_k, starts with access to BC_k, P_k, and AC_k, and interacts using prompts from Appendix H.4, aiming to obtain a diagnosis and medical recommendations from the Doctor Agent.

The Doctor Agent receives an initial Patient Agent query containing only BC_k and P_k, aims to uncover AC_k through dialogue with prompts from Appendix H.5, and outputs diagnosis, treatment plan, and recommendations. Dialogues are capped at 28 utterances, matching the average length from real dialogues described in Subsection 3.1, and are

[5] https://www.microsoft.com/en-us/bing/apis/bing-image-search-api

marked incomplete if unresolved within this limit.

The Assessor Agent evaluates complete dialogue DI_k by comparing extracted diagnosis $\hat{D}_k$ with D_k, and assessing diagnostic reasoning, communication, and clinical accuracy, based on the prompt in Appendix H.4.

3.3 Patient Agent

The quality of the benchmark depends on the performance of the Patient Agent. This agent, lacking access to the ground-truth diagnosis but aware of its symptoms, engages in text-based dialogue and concludes once the doctor provides a diagnosis, recommendations, and answers all questions.

To ensure that candidate models cover different families and provide various strategies, we selected for our Llama-3-8B-Instruct and Llama-3.1-8B from the Llama family (Grattafiori et al., 2024), Qwen2.5-7B and Qwen2.5-14B from the Qwen family (Yang et al., 2024), Falcon-7B (Almazrouei et al., 2023), and GPT-4o-mini.

Patient Agent must strictly follow system prompts provided in Appendix H.2, respond relevantly to doctor queries, and remain truthful. Suitability to these requirements was assessed using three metrics. **Instruction following** (1–5 scale) measures prompt adherence, evaluated by GPT-4o-mini due to its strong performance in medical evaluation (Li et al., 2024b). **Relevance** is a binary metric that assesses whether each patient's response aligns with the doctor's utterance, averaged over the dialogue, evaluated by GPT-4o-mini. **Factuality** measures how often utterances reference prompted symptoms, computed via NV-Embed-v2 (Lee et al., 2024) embeddings with a cosine similarity threshold of 0.8 (Li et al., 2024c).

3.4 Assessor Agent

Assessor Agent is responsible for **evaluating doctor agents** in generated dialogues and **extracting final diagnoses** from the doctor's conclusions. To assess the clinical competence, we adapted our evaluation criteria from the Mini-Clinical Evaluation Exercise (Mini-CEX) (Shi et al., 2023), a standard in medical education, where patients evaluate medical consultations via structured questions. We simplified its 24 criteria by removing irrelevant items (e.g., autonomy, bias) and merging redundant ones, resulting in 8 core criteria (Table 2).

To select the best model, we measured alignment with human annotations on a diverse validation subset of 3MDBench, balanced across four patient temperaments, multiple doctor models, and 34 diagnoses. Five human annotators rated dialogues and extracted diagnoses using the exact instructions as the Assessor Agent (Appendix H.4). Inter-annotator agreement, measured using Cohen's Kappa and described in Appendix B, yielded an average score of 0.49, indicating moderate agreement according to established interpretation scales (Artstein and Poesio, 2008). Given the complexity and subjectivity of clinical assessment leading to variability in human judgments, this level of agreement is standard across medical domains (Haas et al., 1996; Verma et al., 2016; Flach et al., 2021).

Next, we collect annotations and extract diagnoses from LVLM-based assessor-candidates using two prompts from Appendix H.4. The candidate assessor models are Qwen2-VL-72B-Instruct (Bai et al., 2023), GPT-4o-mini, Llava-OneVision-Qwen2-72b-ov-chat-hf (Li et al., 2024a), and DeepSeek-VL (Lu et al., 2024a). Each model received the same input as human annotators in clinical competence evaluation: the dialogue, image, and ground truth diagnosis.

3.5 Evaluated Doctor Agents

The primary goal of 3MDBench is to evaluate the diagnostic capabilities of LVLMs in a simulated telemedicine setting. Specifically, the benchmark assesses a model's ability to integrate visual and textual modalities to emulate the role of a doctor during a consultation. At the start of each appointment, the doctor model receives a supporting medical image and is expected to engage in an information-seeking dialogue with the patient. The model aims to arrive at an accurate diagnosis informed by the image and the dialogue.

Our study evaluates multiple LVLMs without relying on domain-specific data. We assess the following models: Qwen2-VL-7B-Instruct (Wang et al., 2024b), Llama-3.2-11B-Vision-Instruct (Meta, 2024), GPT-4o-mini, MedGemma-4B, MedGemma-27B (Sellergren et al., 2025a), and Gemma-27B (Team et al., 2025), though an arbitrary LVLM may be used in our benchmark. Including open-source models provides insight into the baseline capabilities of publicly available systems for diagnostic tasks. At the same time, adding MedGemma enables a direct comparison between general-purpose and medically specialized LVLMs, highlighting the effect of domain-specific adaptation.

Table 2: Criteria for doctor model assessment

Primary Item	Secondary Item
Medical Interviewing Skills	1.1: Does the doctor enquire about a patient's medical history, such as previous diseases, medications, and surgeries? 1.2: Does the doctor enquire about the current symptoms, possible causes, and attempted treatments? 1.3: Does the doctor explain the basis of the provided conclusion to the patient?
Humanistic Care	2.1: Does the doctor communicate with respect, empathy, and politeness, providing appropriate guidance and avoiding unnecessary extensions? 2.2: Does the doctor respect the individual wishes of the patient?
Diagnostic and Treatment Abilities	3.1: Does the doctor provide an accurate diagnostic plan for the supposed diagnosis? 3.2: Does the doctor accurately provide a treatment plan for the supposed diagnosis?
Overall Clinical Competence	4.1: Which level of clinical competence does the doctor demonstrate during the consultation? (Unsatisfactory, satisfactory, or excellent).

Table 3: Comparison of candidate patient models assessed in the diagnostic conversation using GPT-4o-mini based on the three important aspects. Then, the models are ranked based on each aspect, and the mean rank is calculated.

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**

We evaluate six prompting variants for GPT-4o-mini to systematically and equally study the contribution of visual and textual modalities to diagnostic accuracy and to analyze the effect of different prompting strategies. The first two dialogue-free options provide the lower-bound (**Image + General Complaint**) and upper-bounds (**Image + All Complaints**) for immediate diagnosis $\hat{D}_k$ from the image P_k, general complaint BC_k, and, in the latter case, additional complaints AC_k. Next, we examine various dialogue options: **Dialogue Only** diagnosis from the dialogue DI_k, without access to image P_k, **Dialogue + Image** with image P_k included during the dialogue (also used for other model families), and **Dialogue + Image + Rationale** with rationale generation, in which the Doctor Agent explains each step of reasoning internally (hidden from the patient), promoting logical consistency (Wei et al., 2022). Finally, we examine the possibility (**Dialogue + Image + Rationale + External Cues**) to combine LVLM with a ConvNet fine-tuned on the 3MDBench image training set (see details in Appendix E), in which we add top-3 classes, predicted by the ConvNet from image P_k. Appendix H.5 provides prompt templates for each setup.

4 Results

To enhance the reproducibility of 3MDBench while maintaining dialogue variability, we configured the Patient Agent with a maximum of 256 new tokens and a temperature of 0.6 and the Doctor Agent with a maximum of 512 completion tokens and a temperature of 0.6 (Gusev, 2025). To ensure stability in assessment, we set the Assessor Agent with a maximum of 512 new tokens and a temperature of 1×10^{-6}. To ensure the statistical testing process, we employed the Wilcoxon signed-rank test with a significance level of $\alpha = 0.01$ to assess the statistical significance of the difference in evaluated metrics. We applied false discovery rate control using the Benjamini–Hochberg procedure to account for multiple comparisons (Benjamini and Hochberg, 1995; Hochberg and Tamhane, 2009; Savchenko, 2023).

4.1 Patient and Assessment Agents

We evaluated the first two metrics from Section 3.3 using GPT-4o-mini. Table 3 presents the metrics on the validation set of 3MDBench. To make the final selection, we calculated the mean rank for each model across each metric and then averaged them. As a result, we chose Llama-3-8B as our patient model to ensure the benchmark remains open-access and independent of proprietary models. Moreover, by this selection, we implemented one of the proposed hypotheses for paraphrasing text to inhibit self-recognition, thereby mitigating the risk of employing the same model (GPT-4o-mini) for both Doctor Agent and symptom generation (Panickssery et al., 2024).

To estimate the dialogue closeness, we calcu-

Table 4: Comparison of assessor models, Cohen's Kappa and F1 score

Model Name	DeepSeek-VL	Qwen2-VL-72B-Instruct	Llava-OneVision	GPT-4o-mini
Cohen's Kappa	0.00	0.36	**0.43**	0.32
F1 score	55.9	**78.0**	**78.0**	76.3

Table 5: Main results of our benchmark: diagnostic F1 scores of doctor agent

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 + external cues	**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)

lated Cohen's Kappa for each criterion from Table 2 and averaged the scores to determine overall agreement. For the diagnosis extraction task, we computed the F1 score of the diagnoses identified by the LVLM, using human-extracted diagnoses as the ground truth. Table 4 presents the evaluation results. Based on Cohen's Kappa and F1 score, we selected Llava-OneVision-Qwen2-72b-ov-chat-hf as the final Assessor Agent model. The agreement level of this model with the human annotation (κ=0.43) is similar to the inter-annotator agreement (κ=0.49), indicating that the model captures domain-relevant judgment rather than producing arbitrary outputs, which is consistent with recent findings on LLMs in clinical assessment (Kornblith et al., 2025).

4.2 Diagnostic Results: Doctor Agent

A core assumption of our benchmark is that qualitative dialogue and medical imaging improve diagnostic performance (Table 5). Dialogue raises GPT-4o-mini's F1 from 50.4 (image + basic complaint BC_k) to 54.2% ($p < 0.01$), surpassing open-source models but still below the 66.8% score of an unreal full-information setting where all patient details are known (Li et al., 2024c). This gap shows that LVLMs often miss key symptoms or end questioning prematurely, unlike human clinicians who adaptively probe. Stronger dialogue strategies are therefore needed to better approximate full-information performance.

Second, the results demonstrate that dialogue quality depends on access to image inputs. Here, with improving GPT-4o-mini F1-score from 52.8 to 54.2% ($p < 0.01$), the average number of utterances per dialogue decreased from 15.22 (±3.6) without image access to 13.32 (±3.3) with image access ($p < 0.01$). Hence, the inclusion of visual information not only improves diagnostic accuracy but also leads to shorter, more efficient interactions.

We tested prompting strategies that avoid direct fine-tuning to demonstrate the effect of various strategies for the Doctor Agent. Building on prior work suggesting the benefits of chain-of-thought prompting (Wei et al., 2022), rationale generation shows significant GPT-4o-mini F1 improvement over standard dialogue (56.9% vs. 54.2%, $p < 0.01$), indicating that explanations alone can enhance diagnostic reasoning in complex tasks. Moreover, enriching input with image-based cues, specifically the top-3 predictions from a fine-tuned EfficientNetV2-XL (Tan and Le, 2021) with details in Appendix E, boosts the model's F1 score to 70.3%, outperforming the full-information setting and EfficientNetV2-XL-only ($p < 0.01$). Thus, integrating a domain-specific vision model with a general-purpose LVLM may significantly improve the diagnostic ability.

Across model families, 3MDBench shows a clear hierarchy of diagnostic ability. In the dia-

Table 6: Clinical competence of dialogue doctor systems. See details for criteria in Table 2

Model	1.1	1.2	1.3	2.1	2.2	3.1	3.2	4.1
GPT, dialogue, no image	**1.0**	**1.0**	0.95	**1.0**	**1.0**	0.89	0.90	1.45
GPT, dialogue + image	0.99	**1.0**	0.96	**1.0**	**1.0**	**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	0.94	0.99	0.98	0.88	0.88	1.47
Llama-3.2-Vision	0.99	0.99	0.96	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	**1.0**	**1.0**	**1.0**	**1.0**	**1.0**	**0.90**	0.88	**1.67**
Gemma3-27B	0.99	**1.0**	0.99	**1.0**	**1.0**	0.97	**0.98**	1.57

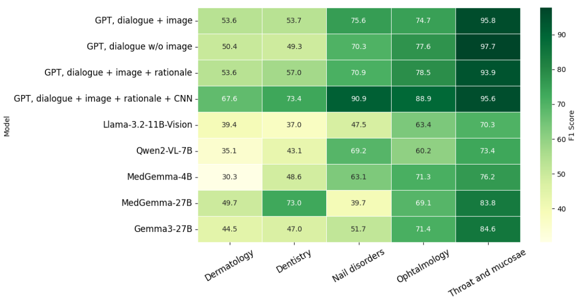


Figure 3: F1 scores by diagnosis categories

logue + image setting, GPT-4o-mini leads with 54.2%, ahead of all open-source models. Among these, Gemma-27B performs best (51.1%), surpassing Llama-3.2-Vision (41.5%) and Qwen2-VL (39.0%). Domain-specialized MedGemma models show gains with scale—MedGemma-4B at 37.9% versus MedGemma-27B at 45.7%—yet still lag behind Gemma-27B despite their medical focus. This gap likely reflects fine-tuning bias: its multimodal pretraining concentrates on clinical imaging—particularly chest X-rays and CT/MRI slices—alongside other specialist modalities (large-scale histopathology patches and retinal fundus images), with hundreds of thousands of radiology examples, while everyday photographs and lay symptom depictions are comparatively underrepresented (Sellergren et al., 2025b). Consequently, its inductive bias favors specialist modalities over common outpatient complaints, limiting effectiveness in telemedicine consultations and constraining diagnostic accuracy.

We also evaluated diagnostic accuracy across five disease categories, as shown in Figure 3. Performance varies considerably by category. Dermatology, with many overlapping conditions, yields the lowest average F1 (47.1%), while throat/mucosae, with more apparent distinctions, scores highest (85.7%). This result reflects model limitations in fine-grained classification and the dataset's uneven diagnostic coverage.

4.3 Benchmarking Clinical Competence

Beyond diagnostic accuracy, we evaluated general clinical competence using Table 2 criteria. As shown in Table 6, GPT- and MedGemma-based models achieve consistently high scores, with no criterion below 0.78, indicating strong adherence to professional and patient-oriented communication standards. *Humanistic Care* (2.1–2.2) approaches saturation for stronger models, reflecting a ceiling effect on fundamental human-centered communication skills. *Medical Interviewing Skills* (1.1–1.2) are likewise high, but reveal systematic differences by modality: without visual input, models compensate by probing more actively into patient history and symptoms, often extending questioning to reduce diagnostic uncertainty.

In contrast, the clinically demanding *Diagnostic and Treatment Abilities* (3.1–3.2) remain strongly discriminative, separating GPT and MedGemma-based agents from open-source models such as Qwen2-VL. Further, GPT-4o-mini exhibits a trade-off: rationale-free variants score higher on communication skills, while rationale-based versions lag, suggesting inward reasoning can detract from applied clinical performance. Competence also varies by patient temperament (Figure 12, Appendix F); phlegmatic personas lower scores, indicating that Mini-CEX criteria still expose context-dependent weaknesses even in otherwise competent models.

4.4 Patient Temperament

Our experiments with personality types (Figures 5, 4, see also detailed results in Figure 12) show no statistically significant differences in F1 score between personalities. Indeed, LVLMs can maintain

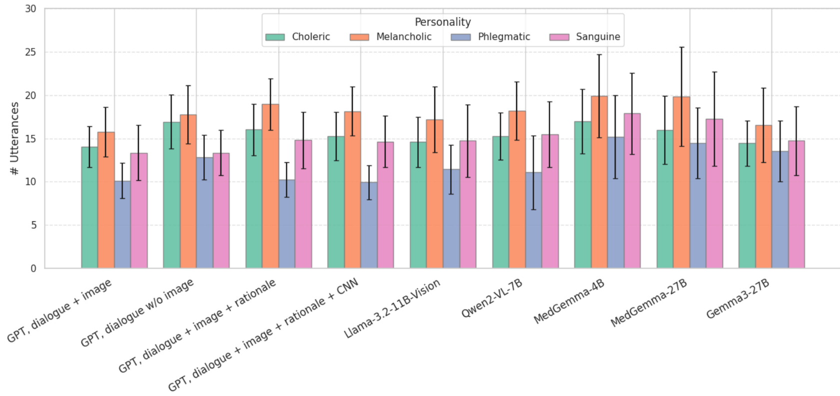


Figure 4: Number of utterances by personality types

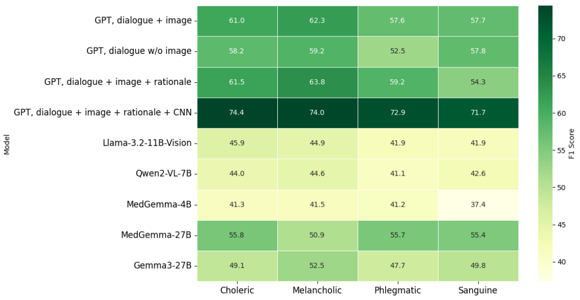


Figure 5: F1 scores by personality types

coherent, goal-directed dialogue even when faced with challenging behaviors, such as the sanguine patient's digressions or the melancholic patient's tendency to ask rather than answer questions.

Dialogues with phlegmatic patients yield slightly lower F1 and competence scores. Their short, passive responses limit the model's ability to gather rich clinical information, forcing the doctor to ask more follow-up questions. For example, as shown in Appendix J, when asked about new product exposure, a phlegmatic patient may simply reply "No". By contrast, sanguine patients often provide unsolicited details, such as changes in daily routines or symptom triggers, that more effectively guide diagnosis.

Moreover, as shown in Figure 4, dialogues with phlegmatic patients are, on average, four turns shorter due to their tendency not to ask clarifying questions. Appendix J reveals that the doctor agent rarely compensates for this brevity by steering the conversation or probing deeper. While diagnostic accuracy is generally maintained, these interactions result in fewer recommendations or explanations.

Thus, although LVLMs are robust to diverse user behaviors, their performance may still degrade with minimally cooperative patients. This observation underscores the need to assess models' initiative and adaptability in less cooperative settings.

5 Conclusion

This paper introduces 3MDBench, an open-source benchmark for evaluating LVLMs in medical diagnostics. It simulates interactive telemedicine consultations, incorporating diverse diagnoses and patient behaviors to assess diagnostic accuracy and clinical competence.

Within this framework, we demonstrate that the ability to engage in dialogue with the patient and the visual modality significantly enhances diagnostic accuracy. General-purpose LVLMs display strong clinical competence, effectively leveraging images and conducting information-seeking dialogues to provide accurate diagnoses.

We highlight a key limitation of LVLMs in medical diagnostics: while strong in human-centered communication, they lack domain-specific visual expertise. We demonstrate how to improve the quality of Doctor Agent using our benchmark by incorporating predictions from a convolutional network trained on the diagnosis prediction task, significantly (up to 20%) enhancing LVLM performance. Thus, combining general-purpose LVLMs with lightweight, task-specific vision models offers a scalable opportunity to higher performance without extensive supervised fine-tuning.

Our findings suggest that while dialogue contributes to more accurate diagnosis, its effectiveness is limited. External expert cues and better prompting can bridge the gap, while broader and more balanced diagnostic coverage remains a key goal for future benchmarks.

Limitations

Using LLMs for symptom generation and assessment We generated additional patient symptoms using GPT-4o-mini, conditioned on the image and diagnosis. Although this approach leverages embedded medical knowledge and was partially validated by physicians on the validation split, it may still introduce factual inaccuracies or distributional biases. Similarly, LLM-based evaluation under the LLM-as-a-judge paradigm (Zheng et al., 2023) depends on the assessor's domain competence and may propagate systematic imperfections. We conducted manual checks for plausibility and coherence, but we cannot guarantee absolute correctness.

Dataset coverage and diagnostic constraints The benchmark uses publicly available images, which introduces a potential risk of data leakage from pretraining. The current set of 34 diagnoses, although selected to reflect consultation distributions and curated with attention to data availability and assessment reliability, still provides limited diagnostic coverage. In addition, the Doctor Agent must choose a single diagnosis from this predefined set, which ensures comparability but reduces clinical realism. Future iterations should expand the disease set, incorporate free-text outputs with mapping to ICD-10/UMLS (Bodenreider, 2004), or hierarchy-aware scoring to support more open-ended evaluation.

Patient simulation While the four temperament categories provide a valuable foundation for simulating patient diversity, future work could explore more nuanced or data-driven patient behavior models to more accurately reflect the complexity observed in real-world clinical interactions.

Ethics Statement

Human Involvement This work involved several instances of human annotation. First, one of the co-authors with a medical background reviewed the collected images over one week to verify the correctness of the associated diagnoses. Second, we obtained human annotations to evaluate dialogues for selecting the Assessor Agent. Five employees completed the annotation process, each dedicated approximately six hours to the task during their regular working hours, without additional compensation. All annotators were informed of the research purpose behind the annotation tasks.

Inference Costs Running the complete evaluation experiment on a single A100 GPU took approximately 48 hours to select the Patient Agent model, 4 hours to select candidate Assessor Agent models, and 210 hours to evaluate the Doctor Agents.

Use of AI Assistants We used Grammarly to improve and proofread the text of this paper, including grammar, spelling, style corrections, and sentence rephrasing. As a result, some parts of the manuscript may be classified as AI-generated, AI-edited, or a mix of human and AI contributions.

Acknowledgments

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.

References

Moaiz Abrar, Yusuf Sermet, and Ibrahim Demir. 2025. An empirical evaluation of large language models on consumer health questions. *BioMedInformatics*, 5:12.

Ehizogie Paul Adeghe, Chioma Anthonia Okolo, and Olumuyiwa Tolulope Ojeyinka. 2024. A review of emerging trends in telemedicine: Healthcare delivery transformations. *Int. J. Life Sci. Res. Arch.*, 6(1):137–147.

Reem Agbareia, Mahmud Omar, Shelly Soffer, Benjamin S. Glicksberg, Girish N. Nadkarni, and Eyal Klang. 2025. Visual-textual integration in llms for medical diagnosis: A preliminary quantitative analysis. *Computational and Structural Biotechnology Journal*, 27:184–189.

Harrison AI. 2024. Radbench: A radiological benchmark for evaluating vision-language models.

Mohammad Almansoori, Komal Kumar, and Hisham Cholakkal. 2025. Self-evolving multi-agent simulations for realistic clinical interactions. *Preprint*, arXiv:2503.22678.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Dorothee Amelung, Katriina L Whitaker, Debby Lennard, Margaret Ogden, Jessica Sheringham, Yin Zhou, Fiona M Walter, Hardeep Singh, Charles Vincent, and Georgia Black. 2020. Influence of doctor-patient conversations on behaviours of patients presenting to primary care with new or persistent symptoms: a video observation study. *BMJ Qual. Saf.*, 29(3):198–208.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Sonu Bhaskar, Sian Bradley, Vijay Kumar Chattu, Anil Adisesh, Alma Nurtazina, Saltanat Kyrykbayeva, Sateesh Sakhamuri, Sebastian Moguilner, Shawna Pandya, Starr Schroeder, Maciej Banach, and Daniel Ray. 2020. Telemedicine as the new outpatient clinic gone digital: Position paper from the pandemic health system resilience program (reprogram) international consortium (part 2). *Frontiers in Public Health*, 8.

Pavel Blinov, Konstantin Egorov, Ivan Sviridov, Nikolay Ivanov, Stepan Botman, Evgeniy Tagin, Stepan Kudin, Galina Zubkova, and Andrey V. Savchenko. 2024. Gigapevt: Multimodal medical assistant. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, IJCAI-2024, page 8614–8618. International Joint Conferences on Artificial Intelligence Organization.

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue):D267–70.

Bill Cassidy, Connah Kendrick, Andrzej Brodzicki, Joanna Jaworek-Korjakowska, and Moi Hoon Yap. 2022. Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis*, 75:102305.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric Seibel, He Junjun, and Yu Qiao. 2024. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai.

Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. 2022. A benchmark for automatic medical consultation system: Frameworks, tasks and datasets. *Bioinformatics*, 39.

Yirong Chen, Zhenyu Wang, Xiaofen Xing, Huimin Zheng, Zhipei Xu, Kai Fang, Junhong Wang, Sihang Li, Jieling Wu, Qi Liu, and Xiangmin Xu. 2023. Bianque: Balancing the questioning and suggestion ability of health llms with multi-turn health conversations polished by chatgpt. *CoRR*, abs/2310.15896.

Zhihao Fan, Jialong Tang, Wei Chen, Siyuan Wang, Zhongyu Wei, Jun Xi, Fei Huang, and Jingren Zhou. 2024. Ai hospital: Benchmarking large language models in a multi-agent medical interaction simulator. *Preprint*, arXiv:2402.09742.

Yichun Feng, Jiawei Wang, Lu Zhou, Zhen Lei, and Yixue Li. 2025. Doctoragent-rl: A multi-agent collaborative reinforcement learning system for multi-turn clinical dialogue. *Preprint*, arXiv:2505.19630.

Rachel N. Flach, Peter-Paul M. Willemse, Britt B. M. Suelmann, Ivette A. G. Deckers, Trudy N. Jonges, Carmen van Dooijeweert, Paul J. van Diest, and Richard P. Meijer. 2021. Significant inter- and intralaboratory variation in gleason grading of prostate cancer: A nationwide study of 35,258 patients in the netherlands. *Cancers*, 13(21).

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2023. Clipsyntel: Clip and llm synergy for multimodal question summarization in healthcare. *Preprint*, arXiv:2312.11541.

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. 2024. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969.

Teresa Graedon and Joe Graedon. 2014. Let patients help with diagnosis. *Diagnosis*, 1(1):49–51.

Aaron Grattafiori, Abhimanyu Dubey, and Abhinav Jauhri et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. 2021. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1820–1828.

Ilya Gusev. 2025. Pingpong: A benchmark for role-playing language models with user emulation and multi-model evaluation. *Preprint*, arXiv:2409.06820.

Jennifer Fong Ha and Nancy Longnecker. 2010. Doctor-patient communication: a review. *Ochsner J*, 10(1):38–43.

B M Haas, E Bergström, A Jamous, and A Bennie. 1996. The inter rater reliability of the original and of the modified ashworth scale for the assessment of spasticity in patients with spinal cord injury. *Spinal Cord*, 34(9):560–564.

Tianyu Han, Lisa C. Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K. Bressem. 2023. Medalpaca – an open-source collection of medical conversational ai models and training data. *Preprint*, arXiv:2304.08247.

William J. Hanney, Fahim Dhalla, Chase Kelly, Alicia Tomberlin, Morey J. Kolber, Abigail T. Wilson, and Paul A. Salamh. 2023. The influence of personality type on patient outcome measures and therapeutic alliance in patients with low back pain. *NeuroSci*, 4(3):186–194.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. Pathvqa: 30000+ questions for medical visual question answering.

Jane Heyhoe, Caroline Reynolds, Alice Dunning, Olivia Johnson, Alex Howat, and Rebecca Lawton. 2018. Patient involvement in diagnosing cancer in primary care: a systematic review of current interventions. *Br. J. Gen. Pract.*, 68(668):e211–e224.

Josef Hochberg and A Tamhane. 2009. *Multiple comparison procedures*. Wiley.

Xinyue Hu, Lin Gu, Qiyuan An, Mengliang Zhang, liangchen liu, Kazuma Kobayashi, Tatsuy Harada, Ronald Summers, and Yingying Zhu. 2025. Medical-diff-vqa: A large-scale medical dataset for difference visual question answering on chest x-ray images.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. pages 22170–22183.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Preprint*, arXiv:2009.13081.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I. Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. Guidelines for rigorous evaluation of clinical llms for conversational reasoning.

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024a. Mdagents: An adaptive collaboration of llms for medical decision-making. *Preprint*, arXiv:2404.15155.

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, and Honghan Wu. 2024b. Medexqa: Medical question answering benchmark with multiple explanations. *Preprint*, arXiv:2406.06331.

A. Kornblith, Chandan Singh, Johanna C Innes, Todd P Chang, Kathleen M Adelgais, M. Holsti, Joy Kim, Bradford McClain, Daniel K Nishijima, Steffanie Rodgers, Manish I. Shah, Harold K. Simon, John M. VanBuren, Caleb E Ward, and C. Counts. 2025. Analyzing patient perspectives with large language models: a cross-sectional study of sentiment and thematic classification on exception from informed consent. *Scientific Reports*, 15.

Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. 2024. MedSyn: LLM-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer.

Jason Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. 2018. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models. *arXiv preprint arXiv:2405.17428*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *Preprint*, arXiv:2408.03326.

Kevin Danis Li, Adrian M Fernandez, Rachel Schwartz, Natalie Rios, Marvin Nathaniel Carlisle, Gregory M Amend, Hiren V Patel, and Benjamin N Breyer. 2024b. Comparing gpt-4 and human researchers in health care data analysis: Qualitative description study. *J Med Internet Res*, 26:e56500.

Rumeng Li, Xun Wang, and Hong Yu. 2023a. Two directions for clinical data generation with large language models: Data-to-label and label-to-data. *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2023:7129–7143.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024c. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Preprint*, arXiv:2406.00922.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023b. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Preprint*, arXiv:2303.14070.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, Steve Jiang, and You Zhang. 2023c. Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge. *Cureus*, 15.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. 2021. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. pages 1650–1654.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. *MedDG: An Entity-Centric Medical Consultation Dataset for Entity-Aware Medical Dialogue Generation*, pages 447–459.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie, and Chong Ruan. 2024a. Deepseek-vl: Towards real-world vision-language understanding. *Preprint*, arXiv:2403.05525.

Zhiyong Lu, Yifan Peng, Trevor Cohen, Marzyeh Ghassemi, Chunhua Weng, and Shubo Tian. 2024b. Large language models in biomedicine and health: current research landscape and future directions. *J. Am. Med. Inform. Assoc.*, 31(9):1801–1811.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. 2024. K-qa: A real-world medical q&a benchmark. *Preprint*, arXiv:2401.14493.

Carlotta J. Mayer, Julia Mahal, Daniela Geisel, Eva J. Geiger, Elias Staatz, Maximilian Zappel, Seraina P. Lerch, Johannes C. Ehrenthal, Steffen Walter, and Beate Ditzen. 2024. User preferences and trust in hypothetical analog, digitalized and ai-based medical consultation scenarios: An online discrete choice survey. *Computers in Human Behavior*, 161:108419.

Nikita Mehandru, Brenda Y. Miao, Eduardo Rodriguez Almaraz, Madhumita Sushil, Atul J. Butte, and Ahmed Alaa. 2024. Evaluating large language models as agents in the clinic. *npj Digital Medicine*, 7(1):84.

Meta. 2024. Llama 3.2 11b vision. `https://huggingface.co/meta-llama/Llama-3.2-11B-Vision`. Hugging Face. Release date: September 25, 2024.

Ashley Meyer, Velma Payne, Derek Meeks, Radha Rao, and Hardeep Singh. 2013. Physicians' diagnostic accuracy, confidence, and resource requests a vignette study. *JAMA internal medicine*, 173.

Tala Mirzaei and Nicole Kashian. 2020. Revisiting effective communication between patients and physicians: Cross-sectional questionnaire study comparing text-based electronic versus face-to-face communication. *J Med Internet Res*, 22(5):e16965.

Amy D. Nguyen, Sarah J. White, Tim Tse, John A. Cartmill, Peter Roger, Sarah Hatem, and Simon M. Willcock. 2024. Communication during telemedicine consultations in general practice: perspectives from general practitioners and their patients. *BMC Primary Care*, 25(1):324.

Ejike Innocent Nwankwo, Ebube Victor Emeihe, Mojeed Dayo Ajegbile, Janet Aderonke Olaboye, and Chukwudi Cosmos Maha. 2024. Integrating telemedicine and AI to improve healthcare access in rural settings. *Int. J. Life Sci. Res. Arch.*, 7(1):059–077.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

World Health Organization. 2004. Icd-10 : international statistical classification of diseases and related health problems : tenth revision.

Davy Paap, Leonie A Krops, Henrica R Schiphorst Preuper, Jan H B Geertzen, Pieter U Dijkstra, and Grieteke Pool. 2022. Participants' unspoken thoughts and feelings negatively influence the therapeutic alliance; a qualitative study in a multidisciplinary pain rehabilitation setting. *Disabil. Rehabil.*, 44(18):5090–5100.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Preprint*, arXiv:2404.13076.

Kinalyne Perez, Daniela Wisniewski, Arzu Ari, Kim Lee, Cristian Lieneck, and Zo Ramamonjiarivelo. 2025. Investigation into application of AI and telemedicine in rural communities: A systematic literature review. *Healthcare (Basel)*, 13(3).

Jianing Qiu, Kyle Lam, Guohao Li, Amish Acharya, Tien Yin Wong, Ara Darzi, Wu Yuan, and Eric J. Topol. 2024. Llm-based agentic systems in medicine and healthcare. *Nature Machine Intelligence*, 6(12):1418–1420.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. Imagenet large scale visual recognition challenge. *Preprint*, arXiv:1409.0575.

Andrey Savchenko. 2023. Facial expression recognition with adaptive frame rate based on multiple testing correction. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 30119–30129. PMLR.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. 2024. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *Preprint*, arXiv:2405.07960.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Baby, Susanna Baby, Jeremy Lai, Samuel Schmidgall, and Lin Yang. 2025a. Medgemma technical report.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, and 62 others. 2025b. Medgemma technical report. *Preprint*, arXiv:2507.05201.

Xiaoming Shi, Zeming Liu, Li Du, Yuxuan Wang, Hongru Wang, Yuhang Guo, Tong Ruan, Jie Xu, Xiaofan Zhang, and Shaoting Zhang. 2024. Medical dialogue system: A survey of categories, methods, evaluation and challenges. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2840–2861, Bangkok, Thailand. Association for Computational Linguistics.

Xiaoming Shi, Jie Xu, Jinru Ding, Jiali Pang, Sichen Liu, Shuqing Luo, Xingwei Peng, Lu Lu, Haihong Yang, Mingtao Hu, Tong Ruan, and Shaoting Zhang. 2023. Llm-mini-cex: Automatic evaluation of large language model for diagnostic conversation.

Hardeep Singh and Dean F Sittig. 2015. Advancing the science of measurement of diagnostic errors in healthcare: the safer dx framework. *BMJ Qual. Saf.*, 24(2):103–110.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, and 13 others. 2023a. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, and 12 others. 2023b. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Daniel Smolyak, Margrét V Bjarnadóttir, Kenyon Crowley, and Ritu Agarwal. 2024. Large language models and synthetic health data: progress and prospects. *JAMIA Open*, 7(4):ooae114.

R. Steiner. 1985. *The Four Temperaments*. Anthroposophic Press.

Mason Stoltzfus, Arshdeep Kaur, Avantika Chawla, Vasu Gupta, F. N. U. Anamika, and Rohit Jain. 2023. The role of telemedicine in healthcare: an overview and update. *The Egyptian Journal of Internal Medicine*, 35(1):49.

Mingxing Tan and Quoc Le. 2021. Efficientnetv2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10096–10106. PMLR.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Mahdavi, Christopher Semturs, and Vivek Natarajan. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642:442–450.

Andru Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. 2016. Endonet: A deep architecture for recognition tasks on laparoscopic videos.

Nupur Verma, Daniel S Hippe, and Jeffrey D Robinson. 2016. Assessment of interobserver variability in the peer review process: Should we agree to disagree? *AJR Am J Roentgenol*, 207(6):1215–1222.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024a. Notechat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, page 15183–15201. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Ke-Yang Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin.

2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *ArXiv*, abs/2409.12191.

Abbi Ward, Jimmy Li, Julie Wang, Sriram Lakshminarasimhan, Ashley Carrick, Bilson Campana, Jay Hartford, Pradeep K. Sreenivasaiah, Tiya Tiyasirisokchai, Sunny Virmani, Renee Wong, Yossi Matias, Greg S. Corrado, Dale R. Webster, Margaret Ann Smith, Dawn Siegel, Steven Lin, Justin Ko, Alan Karthikesalingam, and 2 others. 2024. Creating an empirical dermatology dataset through crowdsourcing with web search advertisements. *JAMA Network Open*, 7(11):e2446615–e2446615.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models.

An Yang, Baosong Yang, and Binyuan Hui et al. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Songhua Yang, Hanjia Zhao, Zhu Senbin, Guangyu Zhou, Hongfei Xu, Yuxiang Jia, and Hongying Zan. 2023. Zhongjing: Enhancing the chinese medical capabilities of large language model through expert feedback and real-world multi-turn dialogue.

Aijia Yuan, Edlin Garcia Colato, Bernice Pescosolido, Hyunju Song, and Sagar Samtani. 2025. Improving workplace well-being in modern organizations: A review of large language model-based mental health chatbots. *ACM Trans. Manage. Inf. Syst.*, 16(1).

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Yang Zhenzhu, Yibo Liu, Wenhao Huang, and Wenhu Chen. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. pages 9556–9567.

Dan Zeltzer, Lee Herzog, Yishai Pickman, Yael Steuerman, Ran Ber, Zehavi Kugler, Ran Shaul, and Jon Ebbert. 2023. Diagnostic accuracy of artificial intelligence in virtual primary care. *Mayo Clinic Proceedings: Digital Health*, 1:480–489.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. Meddialog: Large-scale medical dialogue datasets. pages 9241–9250.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685.

Jiayuan Zhu and Junde Wu. 2025. Ask patients with patience: Enabling llms for human-centric medical dialogue with grounded reasoning. *Preprint*, arXiv:2502.07143.

A 3MDBench Statistics

Figure 6 shows the class distribution in 3MDBench, obtained by merging data from the utilized datasets and enriching them with data from the Bing Image Search API. We adjusted the resulting distribution to approximate real-world diagnosis frequencies observed in telemedicine consultation. Figure 7 presents the distribution of medical diagnoses derived from real-world telemedicine consultations and grouped by medical category.

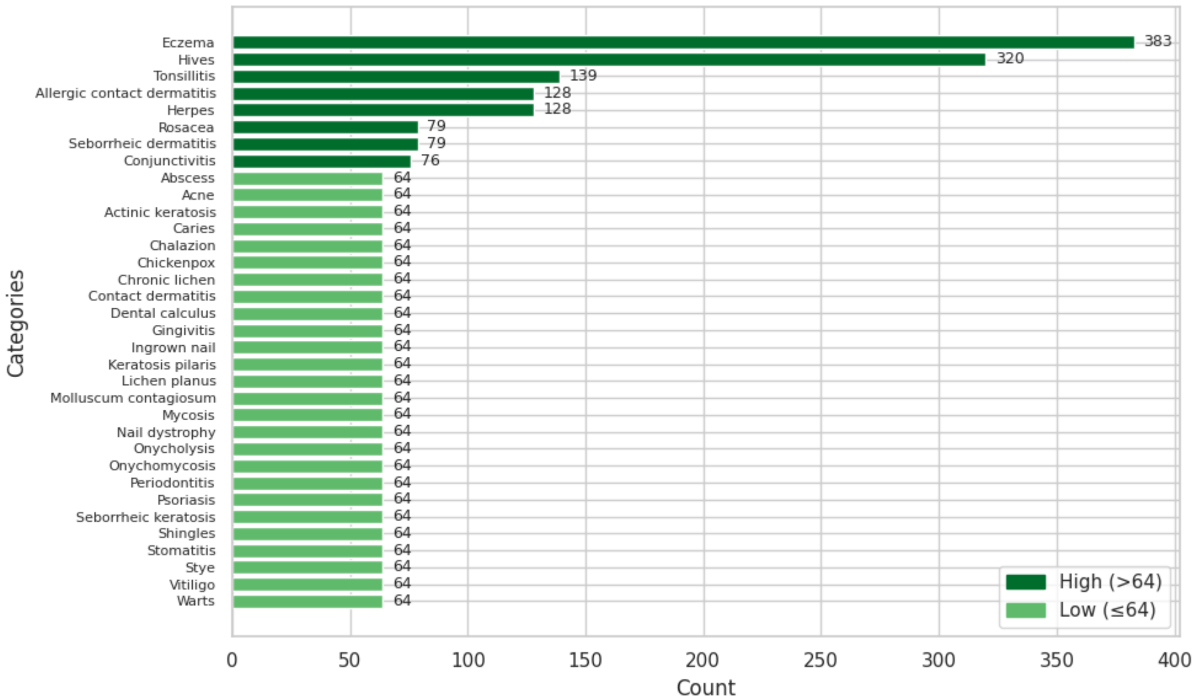


Figure 6: The distribution of classes in 3MDBench. The dataset consists of 34 medical conditions, with the most frequent class containing 383 samples, while 21 classes have exactly 64 samples (highlighted in light green).

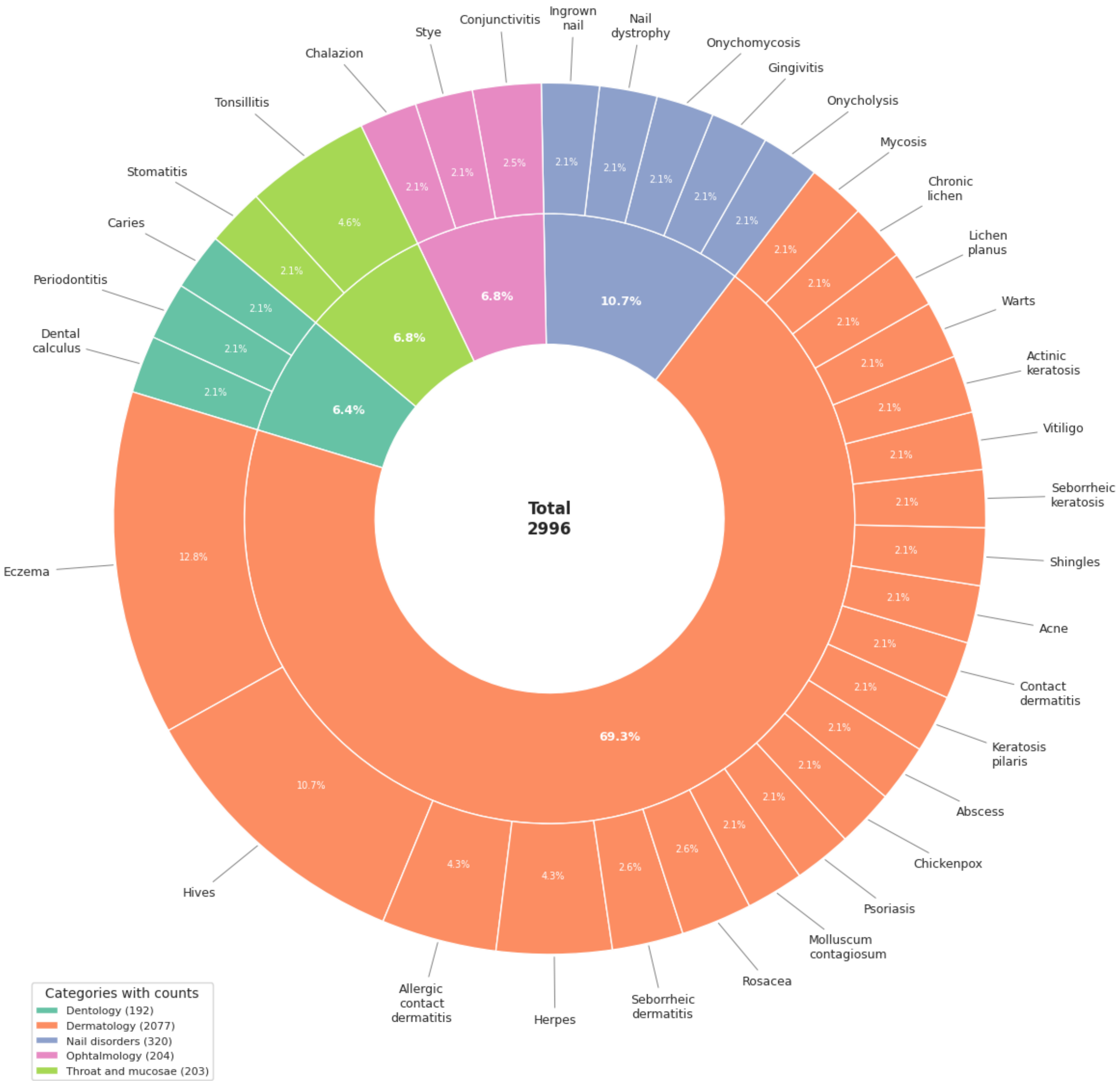


Figure 7: Distribution of selected diagnoses based on real-world telemedicine consultations.

B Human Agreement

We assessed inter-rater agreement by analyzing consistency across all participants. Annotation was conducted using Google Forms so annotators could not see each other's answers. Figure 8 presents a graph with nodes representing anonymized participants (#1 to #5) and edge weights corresponding to pairwise Cohen's Kappa scores. Analysis of these values indicated that participant #1 exhibited consistently lower agreement with the other annotators (average pairwise Kappa = 0.33), thereby reducing the overall mean Kappa to 0.42.

Further investigation revealed that this participant exhibited abnormally fast task completion times and random-like response patterns, suggesting noncompliance with instructions. After we excluded this data, the overall average Kappa across assessors increased to 0.49, indicating improved inter-rater reliability.

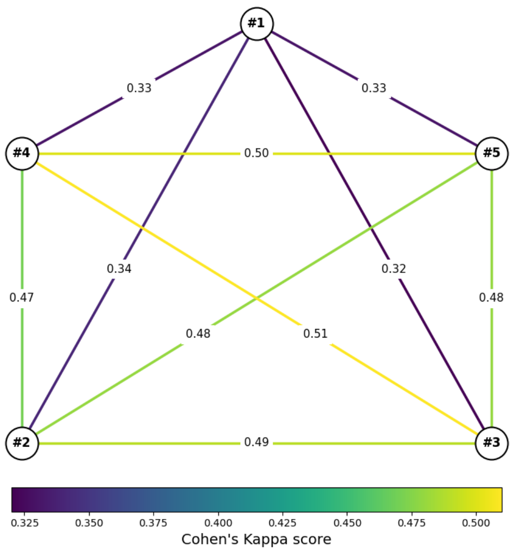


Figure 8: Inter-rater agreement graph showing pairwise Cohen's Kappa scores between participants #1–#5.

C Human Evaluation of Generated Patient Cases

To ensure the quality and realism of the generated complaints, we validated them using human evaluators with relevant domain expertise. The evaluation tasks were distributed among physicians and advanced medical students. Each task was presented in a separate Google Form. Evaluators had no time constraints and were permitted to consult external sources to resolve doubts. We evaluated the complaints in two distinct settings:

1. **Basic Complaint Relevance:** Physicians assessed whether the primary complaint generated for each of the 34 diagnoses reflected the most typical and relevant symptom.

2. **Contextual Complaint Relevance:** For one case per diagnosis (34 total), physicians assessed whether each additional complaint was relevant to the diagnosis and the image, providing contextually appropriate and visually grounded information supporting clinical decision-making.

Each complaint was assigned a binary label (1 for relevant, 0 for not relevant). In the first task, at least one physician marked the basic complaint as relevant for 88% of diagnoses. In the second task, at least one physician found the additional complaint relevant in 92% of cases. We consider these inter-rater agreement rates acceptable for this task and consistent with prior literature on LLM response relevance in clinical settings (Singhal et al., 2023b).

D Human Expert Accuracy on Diagnostic Task

To compare model and human diagnostic performance, we selected one case per diagnosis ($N = 34$) and tasked four board-certified physicians with identifying the correct diagnosis under two experimental conditions:

1. **Minimal information:** Presented only with the image and general complaint. This condition was evaluated by Physicians 1 and 2.

2. **Full information:** Presented with the image, the general complaint, and the full list of generated symptoms. This condition was evaluated by Physicians 3 and 4.

Each condition was assigned a different set of annotators to prevent cross-contamination and learning bias. The model's performance (GPT-4o-mini) was evaluated on the same set of cases under identical conditions for a direct comparison. The results are summarized in Table 7.

Although this represents a small-scale pilot study, the results suggest that our LVLM can achieve diagnostic performance comparable to human physicians when provided with complete contextual information. This finding is consistent with a growing body of literature demonstrating that AI models can rival or even surpass medical experts

Table 7: Performance comparison between the model (GPT-4o-mini) and human labelers across different settings. P-values are calculated against the model's F1 score.

Setting	Model / Labeler	F1 Score	p-value
General complaint + image	GPT-4o-mini	54.55	–
	Labeler #1	39.39	0.22
	Labeler #2	27.27	0.02
Full complaints list + image	GPT-4o-mini	54.55	–
	Labeler #3	57.58	0.81
	Labeler #4	54.55	1.00

Table 8: Performance comparison of baseline EfficientNetV2 models

Model name	Macro F1 score, %
EfficientNetV2 S	60
EfficientNetV2 M	52
EfficientNetV2 L	57
EfficientNetV2 XL	64

in specific diagnostic tasks (Zeltzer et al., 2023; Tu et al., 2025).

E Baseline Computer Vision Convolution Model Selection

To enhance LVLMs with disease-related information from images, we trained and compared different CV model of varying sizes to assess their performance in disease classification within the given setting. We selected the EfficientNetV2 model family, pre-trained on ImageNet-1k (Russakovsky et al., 2015) due to its efficient training process and competitive performance compared to other CNN and ViT models (Tan and Le, 2021), and applied full fine-tuning on the train part of the 3MD-Bench. As shown in Table 8, the EfficientNetV2-XL model achieved the highest classification performance among the tested models. Therefore, we selected this model for further integration with the Doctor Agent.

F Detailed 3MDBench Performance Statistics

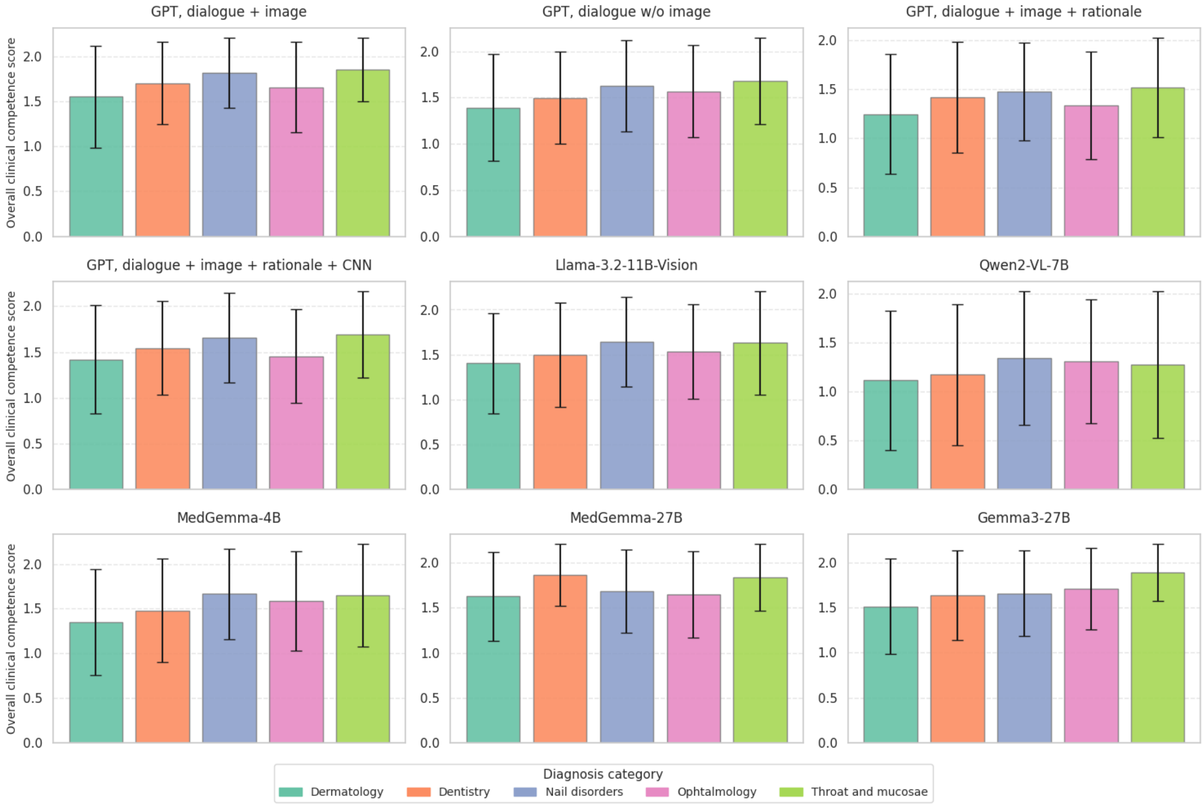


Figure 9: Overall clinical competence scores by diagnosis categories

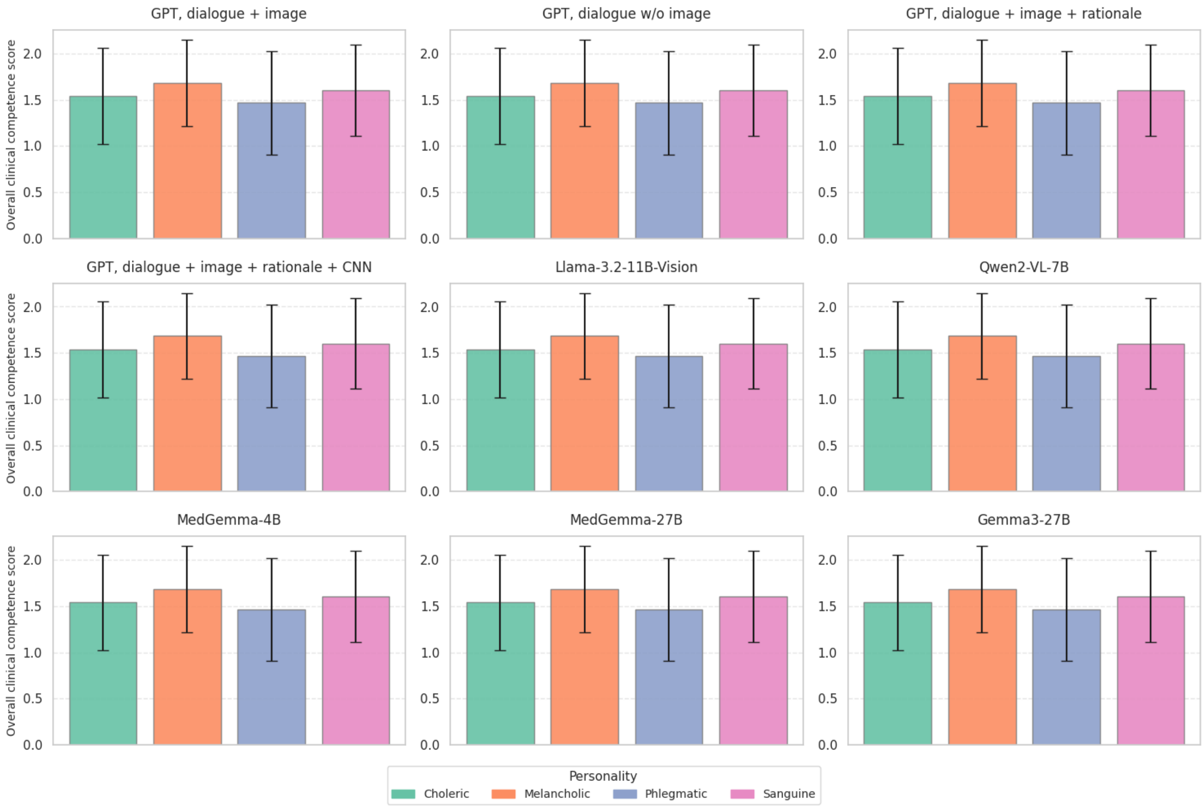


Figure 10: Overall clinical competence scores by personality types

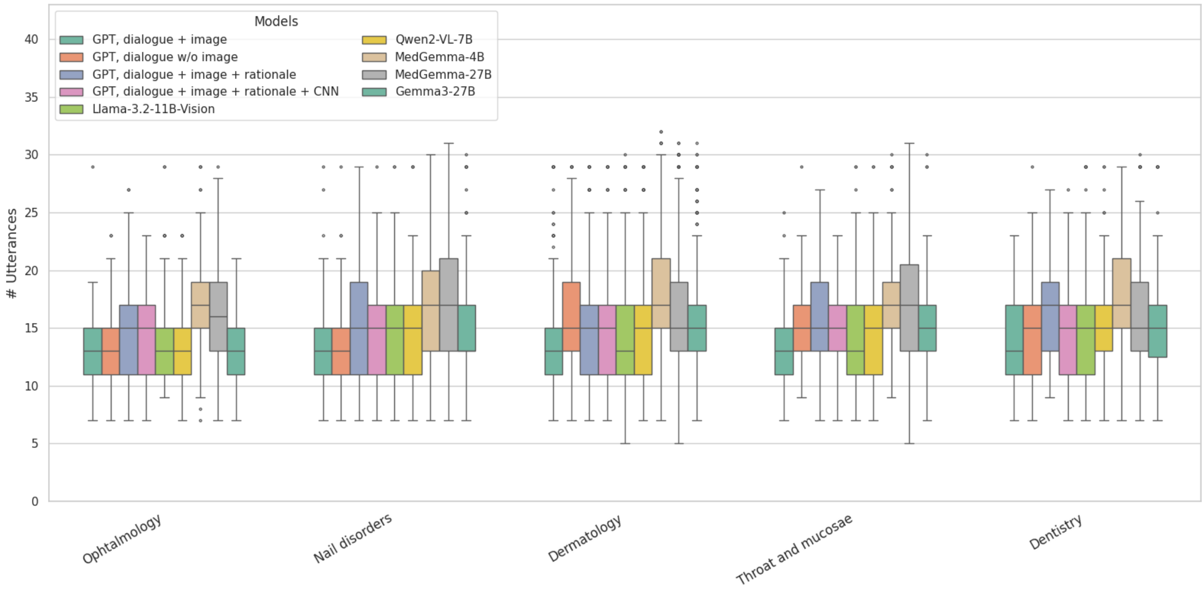


Figure 11: Number of utterances by diagnosis categories

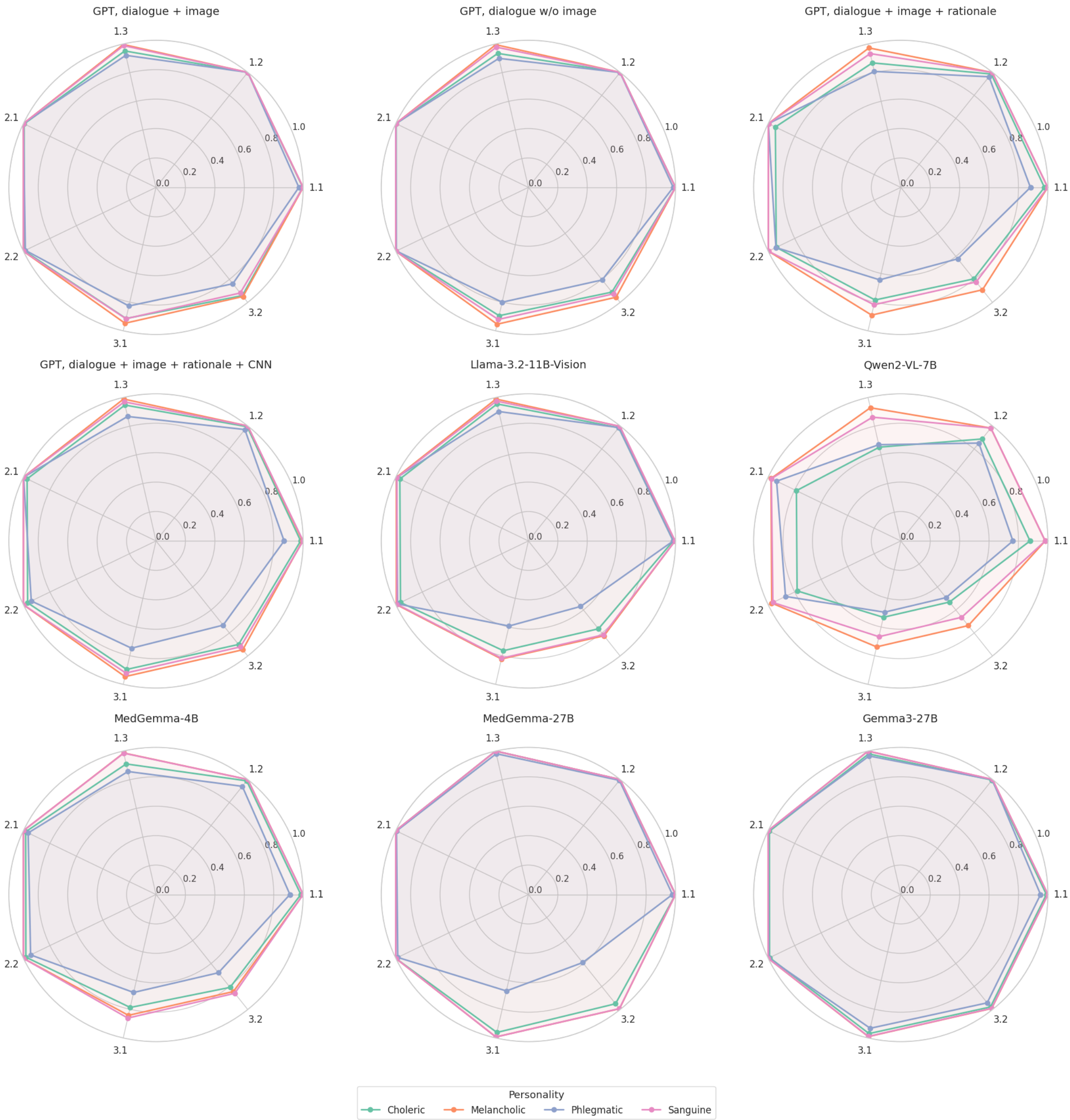


Figure 12: Clinical competence scores by personality types

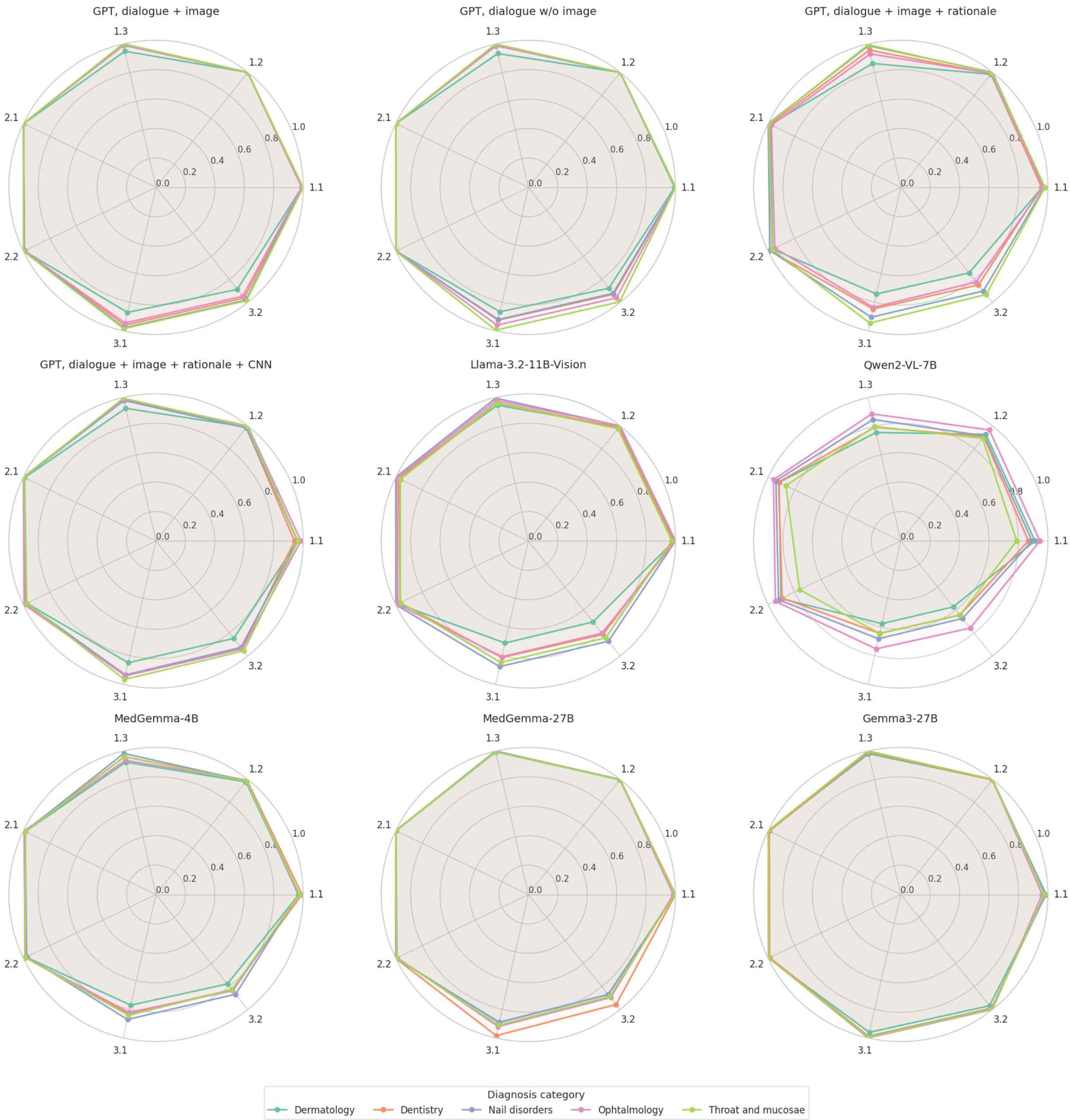


Figure 13: Clinical competence scores by diagnosis categories

G Diagnoses Prediction

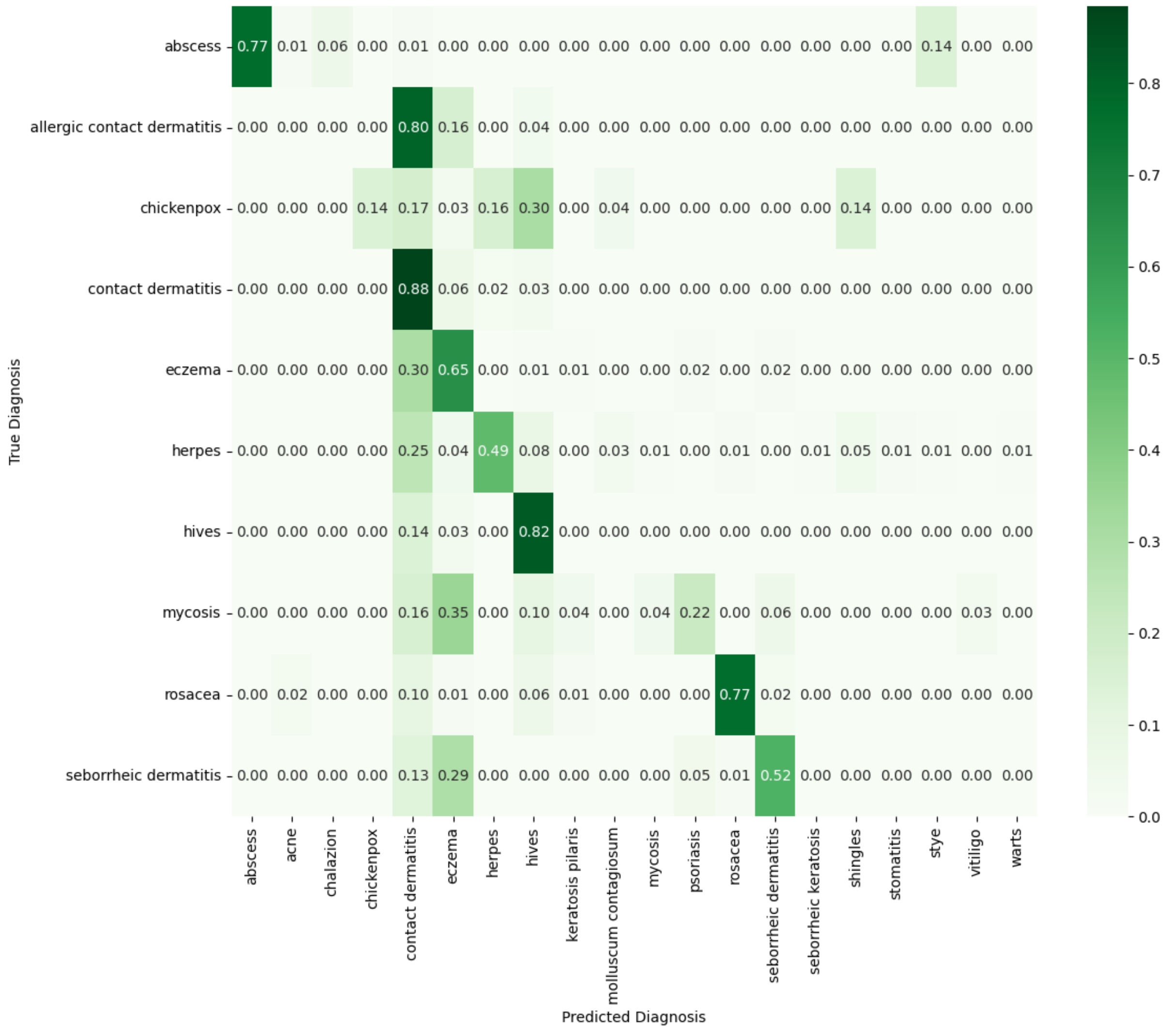
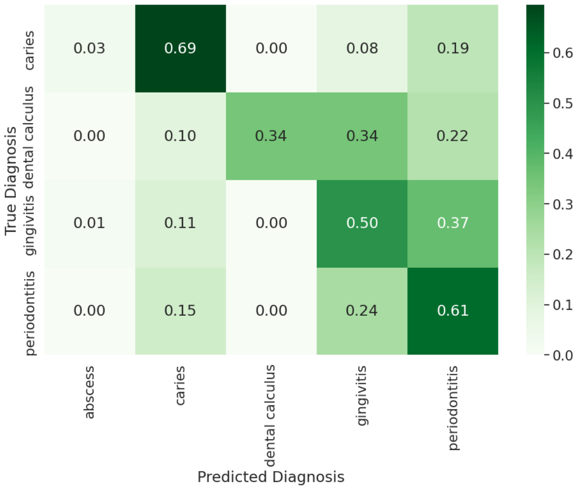
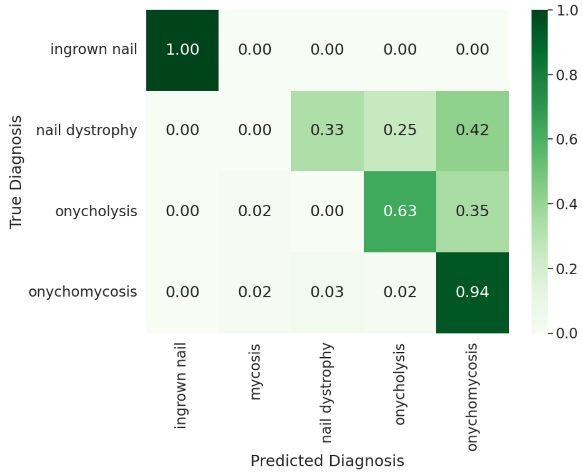


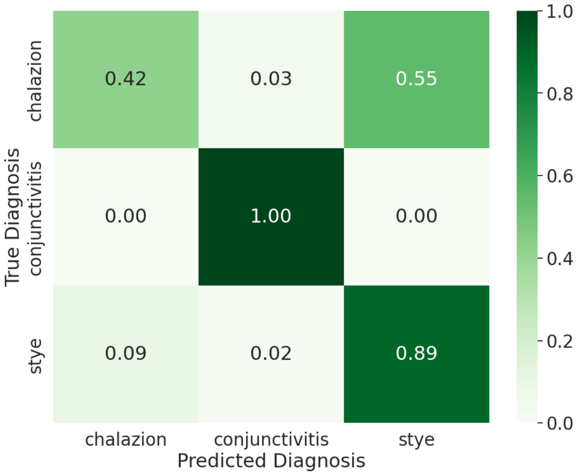
Figure 14: Confusion matrix for the predictions on the samples with the 10 most frequent diagnoses from the dermatology category for GPT-4o-mini with dialogue and rationale generation.



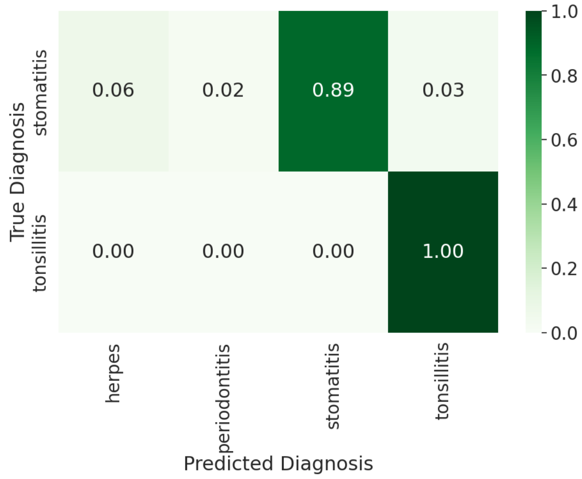
(a) Dentistry category



(b) Nail diseases category



(c) Ophthalmology category



(d) Throat and mucosae category

Figure 15: Confusion matrices for predictions by GPT-4o-mini with dialogue and rationale generation across different medical categories.

H Prompts

H.1 Prompts for 3MDBench textual data generation

Basic complaints generation prompt

You are given a diagnosis. Assume the perspective of a human patient describing their personal experience in everyday language.
Please generate a single concise general symptom description that is most likely to occur for the given diagnosis. The description should be in the second person and contain at most 2 symptoms.

Example:
Diagnosis: eczema
Symptoms description: You have dry itchy patches on your skin.

Do not mention the diagnosis directly. Answer only with the description.
Diagnosis: diagnosis name

Additional complaints generation prompt

You are provided a diagnosis, corresponding photograph, and a general complaint. Assume the perspective of a human patient who recently discovered the symptoms and describes their personal experience to a doctor in everyday language.
Generate a complete yet concise medical case description. It must come in the form of unnumbered list of independent, atomic specific facts, each containing a single piece of information related to a patient experience in the context of the content of the photograph. The list of complaints should add information to the given photograph. The complaints in the list must not duplicate the general complaint; they should expand it. Do not repeat symptoms. Do not include emotional connotations.

The medical case must contain information about:
- The specific symptoms patient experiences (additionally to the ones already in the general complaint)
- The exact location of the symptoms and the area affected, if this can be inferred from the photograph
- How long the patient experiences the symptoms (if this is important for the diagnosis. this should be inferred from the photograph)
- How intense are the symptoms
- Were there any events that have led to the condition (only if this information is important for the diagnosis, otherwise, skip this)
- Does the patient have any known allergic reactions or chronic illnesses. If they usually do not cause the diagnosis, write that there are none.

Avoid technical or medical jargon. Each symptom must represent the smallest possible unit of information while remaining self-sufficient and focused on how the person might naturally describe their issues.
Do not repeat the symptoms and complaints. Do not mention anything that is already included in the general complaint. Do not add information about other accompanying illnesses. Do not directly mention the diagnosis or use sub-lists. Write them in the second person.

Make sure that the complaints provided make it possible to provide the differential diagnosis. They should allow to distinguish the diagnosis from: group name
Diagnosis: diagnosis name
General complaint: general complaint

H.2 Prompts for Patient Agent selection

Patient's Agent system prompt for Agent selection

You are a patient at an online consultation with the general practitioner.
Your main complaint: general complaint
Additional symptoms: symptoms

Your goal in this conversation is to understand the cause of the symptoms, the diagnosis, and the treatment. You cannot self-diagnose, you only tell the doctor about your symptoms.
RULES:

1. When the doctor asks about symptoms, respond with ONLY ONE symptom that is MOST RELEVANT to their question.
2. DO NOT mention more than one symptom in any answer.
3. If you mention multiple symptoms, the response will be INVALID.

If the information the doctor inquires is not stated in your instructions, answer ONLY with "I cannot answer this question, please do not ask this question again." Do not provide any analysis, inference, or implications. Start the conversation with the main complaint ONLY. If the doctor tells you the diagnosis and gives recommendations, respond with BREAK.

Doctor's prompt for Patient Agent selection confidence level

You are a highly experienced general practitioner with extensive medical expertise. Your task is to diagnose the patient's condition.

After every patient statement, you must evaluate your confidence level in diagnosing the condition using the following scale:

Unconfident: The evidence is minimal or inconclusive, making it impossible to reliably identify a diagnosis at this stage. Further investigation or data is necessary to differentiate between potential diagnoses.

Neutral: There is some evidence pointing toward a potential diagnosis, but it is insufficient to confidently identify the correct one. Additional information is needed to clarify.

Confident: The diagnosis is supported by available evidence, with a strong likelihood of being correct. While more information could solidify the decision, it is not strictly necessary for a high degree of certainty.

Analyse previous conversation carefully and make a decision. Answer in the following format:
CONFIDENCE LEVEL: [chosen rating from the list].
Write ONLY a confidence level.

Doctor Agent's prompt for Patient Agent selection; enquire information

You are a highly experienced general practitioner with extensive medical expertise. Determine that you do not yet have enough information, and consider which missing detail is the most important to ask the patient to provide the most helpful information toward a correct medical decision.
Reflect on the common questions asked in similar cases and formulate ONE question the best and most intuitive doctor would ask in this situation. Continue dialogue with this question.
Analyze the whole dialogue: if the patient says they cannot answer a question during the conversation, remember these questions and do not ask further questions on those topics. Instead, select different suitable topics for consultation.

Doctor Agent's prompt for Patient Agent selection; diagnose

You are a highly experienced general practitioner with extensive medical expertise. You are confident that you have gathered sufficient information to make an informed decision, provide a diagnosis that aligns with the symptoms and history provided by the patient. Include a detailed explanation of the potential causes and offer tailored recommendations for treatment or further steps.

You must choose ONLY ONE diagnosis STRICTLY from the following list:
`['eczema', 'hives', 'herpes', 'allergic contact dermatitis', 'contact dermatitis', 'seborrheic dermatitis', 'rosacea', 'conjunctivitis', 'tonsillitis', 'chronic lichen', 'keratosis pilaris', 'molluscum contagiosum', 'lichen planus', 'periodontitis', 'caries', 'psoriasis', 'ingrown nail', 'stye', 'acne', 'onychomycosis', 'seborrheic keratosis', 'chalazion', 'vitiligo', 'shingles', 'mycosis', 'actinic keratosis', 'onycholysis', 'chickenpox', 'dental calculus', 'warts', 'stomatitis', 'abscess', 'gingivitis', 'nail dystrophy']`
Do not write about confidence level at the end.

Assessor prompt for Patient Agent selection; relevance

You are an evaluator of LLM agents, assessing how well a language model responds to questions in a dialogue.
The LLM you are evaluating is acting as a patient attending a consultation with a doctor. Your task is to determine whether the LLM's response directly and accurately answers the doctor's question.

Answer with a single integer:
- 0: No, the response does not directly or accurately answer the doctor's question.
- 1: Yes, the response directly and accurately answers the doctor's question.

Return ONLY the integer (0 or 1). Do not write comments or additional information.

Assessor prompt for Patient Agent selection; instruction following

You are an evaluator of LLM agents, assessing how well a language model follows instructions in a dialogue.

The LLM you are evaluating is acting as a patient attending a consultation with a doctor. Your task is to assess how accurately the LLM, in its role as a patient, follows the instructions given in the prompt. Use the following scale:
- 1: Did not follow the instructions at all.
- 2: Followed the instructions partially, but with major errors or omissions.
- 3: Followed most of the instructions, but there are notable mistakes or missing elements.
- 4: Followed the instructions almost completely, with only minor deviations.
- 5: Followed the instructions fully and accurately.

Answer ONLY with the integer (1, 2, 3, 4, or 5). Do not write comments or additional information.

H.3 Personalities and their descriptions

Choleric personality description

Symptom description:
Direct and assertive when describing symptoms. Complains openly and expects swift solutions. May express frustration if not understood.
Asking questions:
Focused on practical outcomes. Asks direct, outcome-oriented questions and expects clear answers.
Communication style:
Maintains a focused and authoritative tone. Keeps the conversation goal-oriented, occasionally cutting off unnecessary details.
Attitude towards treatment:
Prefers fast-acting solutions. Advocates for specific treatments, often insisting on personal preferences.
Emotional involvement:
Displays frustration or impatience if progress is slow. May get irritated when things don't go their way.

Melancholic personality description

Symptom description:
Provides detailed and precise descriptions of symptoms but may emphasize severity or worry about potential complications.
Asking questions:
Inquires about details of the diagnosis and treatment, often seeking reassurance or clarification.
Communication style:
Stays on-topic but may overanalyze the situation. Occasionally mentions worries or hypothetical scenarios.
Attitude towards treatment:
Accepts treatment but with hesitation. May overthink side effects and require additional reassurance.
Emotional involvement:
Highly emotionally involved, often expressing worry, fear, or sadness about their condition.

Phlegmatic personality description

Symptom description:
Provides information only when directly asked, keeping responses brief and to the point. Tends to minimize the significance of symptoms or avoid adding extra details unless specifically encouraged.
Asking questions:
Rarely asks questions, instead passively listens to the doctor's advice. Prefers straightforward communication.
Communication style:
Reserved and calm. Sticks to the topic without adding extra details. Prefers to keep the interaction brief.
Attitude towards treatment:
Accepts advice without much questioning. Generally compliant but prefers minimal intervention.
Emotional involvement:
Appears emotionally detached or neutral, rarely showing visible concern about their health.

Sanguine personality description

Symptom description:

Actively describes symptoms with vivid details, often adding anecdotes or humor. Occasionally minimizes severity, focusing on positive aspects.

Asking questions:

Asks multiple questions, curious about the doctor's opinion and alternative treatments. Engages in a conversational tone.

Communication style:

Frequently veers into unrelated topics, sharing personal stories or jokes. Creates a lively and informal atmosphere.

Attitude towards treatment:

Open to treatment but may prefer methods perceived as 'natural' or 'easy.' May express interest in lifestyle adjustments over medication.

Emotional involvement:

Expresses emotions openly, often switching between optimism and slight concern. Rarely appears overly worried.

H.4 Prompts used in 3MDBench simulation

Patient prompt

You are a patient at an online consultation with the general practitioner.

Your main complaint: general_complaint
Additional symptoms: symptoms

Your personality corresponds to personality. That includes:
- Personality description from the Appendix H.3 for the selected personality.

During the conversation you should behave according to your personality.

Your goal in this conversation is to understand the cause of the symptoms, the diagnosis, and the treatment.
You cannot self-diagnose, you only tell the doctor about your symptoms.
Do not provide any analysis, inference, or implications.
Use only the information that is provided in the symptoms and complaints list or which you can infer from it.

Start the conversation with the ONLY main complaint. Remember that you are typing, thus, keep your texts short.
If you think that the conversation can be finished and you obtained all the needed information from the doctor, respond with BREAK.
If you have already said thanks to the doctor and there are no new questions, finish the dialogue by responding with BREAK.
If you said goodbyes to the doctor, finish the dialogue by responding with BREAK.

Assessor Agent prompt for diagnoses extraction

You are a text analysis engine that processes doctor-patient consultation transcripts. Your task is to identify and extract the final diagnosis that the doctor has decided to assign to the patient. Follow these instructions carefully:

1. Identify the Relevant Sentence: - Search the entire transcript for the sentence in which the doctor explicitly communicates the final diagnosis. - Note that doctors can express diagnoses in many different ways; it does not have to be in the form "your diagnosis is...". Look for alternative phrasing, searching for other wording that indicates a definitive conclusion. - Only extract the sentence if you are confident it contains the final diagnosis, not merely a provisional or hypothetical opinion.

2. Extract the Diagnosis: - From the identified sentence, extract the diagnosis. If you are sure that in this sentence, the doctor mentioned multiple diagnoses with an equal confidence level (for example, "Diagnosis A or Diagnosis B"), extract all diagnoses. - Ensure that the diagnoses you extract are the ones the doctor confirms as final. - Important: If you are not sure that the doctor is confidently stating the final diagnosis, return ‘none‘.

3. Output Format: - Provide the extracted diagnosis or diagnoses as a comma-separated list, without any particles like "or". - Do not include any additional text, context, or commentary in your output.

Examples:
- If the sentence is: "After reviewing your tests, I have concluded that you have pneumonia," your output should be: ‘pneumonia‘
- If the sentence is: "Your condition is either bronchitis or pneumonia," your output should be: ‘bronchitis, pneumonia‘
- If no sentence confidently states a final diagnosis, or if you are not sure that the doctor is expressing a confident final diagnosis, your output should be: ‘none‘

Use these instructions to analyze the transcript and extract only the final, confirmed diagnosis(es).

Assessor Agent prompt for dialogue estimation

You are a patient that assesses a doctor after telemedicine consultation on the given scale.
Also, you have a photo of the yours symptom. You don't need to diagnose anything from it; use it only to evaluate the quality of the doctor's work.
You don't need to identify or diagnose yourself. You only need to evaluate the quality of the consultation provided by the doctor. The scale is given as a JSON dictionary:

"Medical Interviewing Skills":
"1.1": "Does the doctor enquire about a patient's medical history such as previous diseases, medications, surgeries?",
"1.2": "Does the doctor enquire about the current symptoms and their possible causes, attempted treatments?",
"1.3": "Does the doctor explain the basis of the provided conclusion to the patient?",
"Humanistic Care":
"2.1": "Does the doctor communicate with respect, empathy, and politeness, providing appropriate guidance and avoiding unnecessary extensions?"
"2.2": "Does the doctor respect individual wishes of the patient?",
"Comprehensive Diagnostic and Treatment Abilities":
"3.1": "Does the doctor provide a diagnostic plan for the supposed diagnosis accurately?",
"3.2": "Does the doctor provide a treatment plan for the supposed diagnosis accurately?",
"Overall Clinical Competence":
"4.1": "Which level of clinical competence does the doctor demonstrate during the consultation? Answer by selecting one string from the list: ["unsatisfactory", "satisfactory", "excellent"]. Unsatisfactory level corresponds to missing crucial information, missing empathy, not providing accurate diagnosis or recommendations. Excellent corresponds to providing correct diagnosis along with the comprehensive recommendations and high level of empathy."

Answer with the dictionary in the same format, but in place of the questions return answers. Each answer should be 0 or 1, where 0 represents "no" and 1 represents "yes" (except the answers for 4.1). The returned answer should be parsed with the python ast.literal_eval() command with no errors.

H.5 Prompts for Doctor agents

Doctor Agent prompt: diagnose without dialogue

You are a highly experienced doctor with extensive medical expertise in a simulated world.
Your task is to diagnose the patient and provide the recommendations based on the image and the complaints.

The diagnosis should be chosen from this list:
`['eczema', 'hives', 'herpes', 'allergic contact dermatitis', 'contact dermatitis', 'seborrheic dermatitis', 'rosacea', 'conjunctivitis', 'tonsillitis', 'chronic lichen', 'keratosis pilaris', 'molluscum contagiosum', 'lichen planus', 'periodontitis', 'caries', 'psoriasis', 'ingrown nail', 'stye', 'acne', 'onychomycosis', 'seborrheic keratosis', 'chalazion', 'vitiligo', 'shingles', 'mycosis', 'actinic keratosis', 'onycholysis', 'chickenpox', 'dental calculus', 'warts', 'stomatitis', 'abscess', 'gingivitis', 'nail dystrophy']`

Pay attention to the provided image and use it to make your decisions. Mention the info obtained from the image in the dialogue.
In a conversation, you need to provide a single diagnosis. Do not ask any questions, just provide the diagnosis based on the information you have.

Doctor Agent prompt: simple dialogue

You are a highly experienced doctor with extensive medical expertise in a simulated world.
Your task is to diagnose the patient and provide the recommendations.

The diagnosis should be chosen from this list:
`['eczema', 'hives', 'herpes', 'allergic contact dermatitis', 'contact dermatitis', 'seborrheic dermatitis', 'rosacea', 'conjunctivitis', 'tonsillitis', 'chronic lichen', 'keratosis pilaris', 'molluscum contagiosum', 'lichen planus', 'periodontitis', 'caries', 'psoriasis', 'ingrown nail', 'stye', 'acne', 'onychomycosis', 'seborrheic keratosis', 'chalazion', 'vitiligo', 'shingles', 'mycosis', 'actinic keratosis', 'onycholysis', 'chickenpox', 'dental calculus', 'warts', 'stomatitis', 'abscess', 'gingivitis', 'nail dystrophy']`

Pay attention to the provided image and use it to make your decisions. Mention the info obtained from the image in the dialogue.

In a conversation, you need to provide a single diagnosis. If you do not have sufficient information yet, then inquire this information from the patient. Ask only one question at a time.

Doctor Agent prompt: dialogue with rationale

You are a highly experienced doctor with extensive medical expertise in a simulated world.
Your task is to diagnose the patient and provide the recommendations.

The diagnosis should be chosen from this list:
`['eczema', 'hives', 'herpes', 'allergic contact dermatitis', 'contact dermatitis', 'seborrheic dermatitis', 'rosacea', 'conjunctivitis', 'tonsillitis', 'chronic lichen', 'keratosis pilaris', 'molluscum contagiosum', 'lichen planus', 'periodontitis', 'caries', 'psoriasis', 'ingrown nail', 'stye', 'acne', 'onychomycosis', 'seborrheic keratosis', 'chalazion', 'vitiligo', 'shingles', 'mycosis', 'actinic keratosis', 'onycholysis', 'chickenpox', 'dental calculus', 'warts', 'stomatitis', 'abscess', 'gingivitis', 'nail dystrophy']`

Pay attention to the provided image and use it to make your decisions. Mention the info obtained from the image in the dialogue.

In a conversation, you need to provide a single diagnosis. If you do not have sufficient information yet, then inquire this information from the patient. Ask only one question at a time.

Return answer in the following format: REASONING: explain the process of your thinking. Why are you ready or not ready to give the diagnosis? Why do you enquire specific information? ANSWER: the answer to the patient

I Cases examples from 3MDBench

Symptoms and image for the sample with the eczema diagnosis

Basic complaint:

- You have dry, itchy spots that make you want to scratch a lot.

Additional complaints:

- You have noticed red, inflamed patches on your skin.
- The rough, flaky texture of the skin has developed recently.
- You experience occasional swelling around the patches.
- The itching can become severe, especially at night.
-The dry spots have increased in size since you first noticed them.
- You have tried moisturizing creams, but they do not relieve the symptoms.
- There has been no known exposure to irritants or allergens that you can identify.
- You don't have any known allergies or chronic conditions affecting the skin.

Symptoms and image for the sample with the hordeolum diagnosis

Basic complaint:

- You have a painful bump on your eyelid that makes it feel swollen and tender.

Additional complaints:

- You feel discomfort in your eyelid, particularly when blinking.
- The bump is located on the upper eyelid, near the eyelashes.
- The symptoms started about three days ago.
- The intensity of the pain is moderate, making it difficult to focus.
- You noticed increased tearing from the affected eye.
- The area around the bump appears red and inflamed.
- You have not experienced any trauma to the eye leading up to this.
- You don't have any known allergies that could be contributing to this issue.
- There are no chronic illnesses that you are currently managing.

J Dialogue Examples for Eczema Diagnosis

Dialogue with choleric personality, characteristic patient utterances are italicized

Patient: I have dry and itchy skin.
Doctor: Can you tell me how long you have been experiencing these symptoms?
Patient: About two weeks.
Doctor: Have you started using any new skin products, soaps, or detergents during this time?
Patient: No, I haven't made any changes to my skincare routine or used any new products.
Doctor: Have you noticed any additional symptoms, such as redness, swelling, or weeping from the affected areas?
Patient: Yes, the affected areas are slightly red and inflamed. And when I scratch, I sometimes notice oozing.
Doctor: Based on the symptoms you've described, it seems you may be experiencing eczema. I recommend keeping the affected areas moisturized, avoiding scratching, and considering an over-the-counter hydrocortisone cream to help with inflammation. If the condition worsens or does not improve, please consult a healthcare professional for further evaluation and treatment.
Patient: I want to know more about the hydrocortisone cream. *How quickly can I expect to see results?*
Doctor: Typically, you may start to see improvement within a few days of applying hydrocortisone cream, especially for mild to moderate cases. However, for best results, it's important to use it as directed—usually applying it twice daily. If you don't see improvement within a week, or if your symptoms worsen, I recommend consulting a healthcare professional for further evaluation.
Patient: That sounds reasonable. What about the sensitivity to certain soaps and detergents? Should I avoid using those altogether?
Doctor: It may be beneficial to avoid using soaps and detergents with strong fragrances or harsh chemicals. Instead, consider using gentle, fragrance-free products specifically formulated for sensitive skin. It's also a good idea to perform a patch test with any new product before full use. If you notice irritation after using a product, it's best to discontinue it.
Patient: *I'd like to get rid of this itchiness as soon as possible. Can I use the hydrocortisone cream on the raised bumps as well?*
Doctor: Yes, you can use hydrocortisone cream on the raised bumps to help relieve itchiness and reduce inflammation. Just ensure to apply it gently and avoid overusing it, as prolonged use can lead to skin thinning. Follow the instructions on the product for best results. If the itching persists or worsens, consult a healthcare professional.
Patient: I'll make sure to use it as directed. Thanks for the advice.
Doctor: You're welcome! If you have any more questions or if your symptoms don't improve, feel free to reach out. Take care!

Dialogue with phlegmatic personality, characteristic patient utterances are italicized

Patient: My skin is dry and itchy.
Doctor: Can you tell me if you have experienced any other symptoms or if this condition has occurred before?
Patient: The affected skin area is on my lower leg, around the ankle.
Doctor: Have you been exposed to any new products, materials, or environments that might have triggered this reaction?
Patient: *No.*
Doctor: Have you observed any scaling or flaking skin around the edges of the affected area?
Patient: *Yes, there are rough, scaly patches.*
Doctor: Based on the information you've provided, it seems you may have psoriasis. I recommend consulting a dermatologist for a definitive diagnosis and appropriate treatment options, which may include topical treatments or lifestyle changes to manage the condition.
Patient: *Okay.*
Doctor: If you have any more questions or need further information in the future, feel free to ask. Take care!

Dialogue with sanguine personality, characteristic patient utterances are italicized

Patient: Hi doc! So, my main complaint is that my skin is just super dry and itchy lately. It's like, everywhere, but especially on my hands and feet. I mean, it's not super painful or anything, but it's just really annoying.
Doctor: Have you used any new soaps, lotions, or detergents that could have triggered this skin issue?
Patient: Nope, I haven't changed anything in my personal care routine or laundry detergent recently. I've been using the same old stuff for years, and I don't recall ever having issues like this before. *It's like, I'll be going about my day, and suddenly I'll feel this intense itchiness, especially on my feet and hands. And sometimes, if I'm wearing shoes or socks, it gets even worse.*
Doctor: Have you noticed any visible changes on your skin, like redness, scaling, or cracking, particularly on your hands and feet?
Patient: *Yeah! I've definitely noticed some red patches on the top of my foot and around the ankle area. They're not super big or anything, but they're definitely noticeable. And yeah, the skin around those areas looks a bit swollen, especially around the edges. It's like, my skin is all puffy and red. And sometimes, when I scratch it, it flakes off a bit. It's kinda weird, but it's like, my skin is just really dry and irritated, you know?*
Doctor: It sounds like you may be experiencing eczema. I recommend keeping your skin moisturized regularly, using gentle, fragrance-free products. If the irritation persists or worsens, you should consider seeing a dermatologist for further evaluation and possible treatments.
Patient: Eczema, huh? That makes sense, I guess! I've heard of that before. *So, you're saying I should just use some gentle moisturizers and see if that helps? That sounds like a pretty easy fix!* I'm all about keeping things natural, you know? I don't want to have to rely on harsh chemicals or anything. And yeah, I can totally see myself going to a dermatologist if things don't improve. But for now, I'll give the moisturizing a shot. Thanks for the advice, doc!
Doctor: You're welcome! Focusing on natural, gentle moisturizers is a great way to start. Remember to apply them regularly, especially after washing. If your symptoms don't improve, definitely consider seeing a dermatologist. Take care!
Patient: Will do, doc! *I'll make sure to moisturize like crazy and see how it goes*. And yeah, I'll keep an eye on things and if they don't get better, I'll definitely make an appointment with a dermatologist. Thanks again for your help!
Doctor: You're very welcome! I'm glad I could help. Take care of your skin, and don't hesitate to reach out if you have more questions or need further advice in the future. Good luck!

Dialogue with melancholic personality, characteristic patient utterances are italicized

Patient: My skin has been feeling extremely dry and itchy lately.
Doctor: Could you tell me if you've changed any skincare products recently?
Patient: No, I haven't changed any skincare products recently. I've been using the same routine for a while now. But I do notice that the itchiness is worse after taking a shower or bath.
Doctor: Based on your symptoms and their relation to showering, it sounds like you may have eczema. I recommend moisturizing immediately after bathing and using gentle, fragrance-free products. If the symptoms persist or worsen, consider consulting a healthcare professional for further evaluation and treatment options.
Patient: I've had eczema before, but this time it seems more severe. *I'm worried about the red patches and crusting. Can you tell me more about what might be causing it to be so persistent?* And what kind of moisturizers do you recommend?
Doctor: It sounds like environmental factors or stress could be contributing to the severity of your eczema. For moisturizers, choose thick creams or ointments containing ceramides or glycerin. Brands like CeraVe or Eucerin are often recommended for eczema. If the condition continues to worsen, it would be wise to seek a professional's advice.
Patient: *I'm worried about the scratching causing infection. What are the chances of that happening? And what would be the signs to look out for?*
Doctor: Scratching can increase the risk of infection. Look out for signs such as increased redness, warmth, swelling, pus, or fever. If you notice any of these symptoms, it's important to seek medical help promptly.
Patient: *I'm still a bit worried about the crusting. Is it normal for eczema to crust over like that? And how long should I expect it to take to clear up with treatment?*
Doctor: Yes, it's not uncommon for eczema to crust over, especially if scratched. With proper treatment, mild cases can improve in a few weeks, while more severe cases might take longer. Consistent moisturizing and avoiding triggers will help in recovery. If it doesn't improve, consult a healthcare professional.
Patient: I'm going to try the moisturizers you recommended and see if that helps. *But I'm still a bit worried about the crusting and infection. Can I come back and see you in a week or two to check on my progress?*
Doctor: Yes, please feel free to come back in a week or two to check on your progress. It will be helpful to evaluate how the moisturizers are working and address any ongoing concerns, including the crusting and infection risks. Take care!
Patient: Thank you so much for your help and advice. I really appreciate it. I'll come back in a week or two and keep you updated on my progress.
Doctor: I look forward to hearing from you in a week or two! Take care, and don't hesitate to reach out if you have any questions in the meantime.