EVERYDAYMEDQA: A CONSUMER-ORIENTED MULTIMODAL BENCHMARK FOR MEDICAL QUESTION ANSWERING IN EVERYDAY SETTINGS

Anonymous authors
Paper under double-blind review

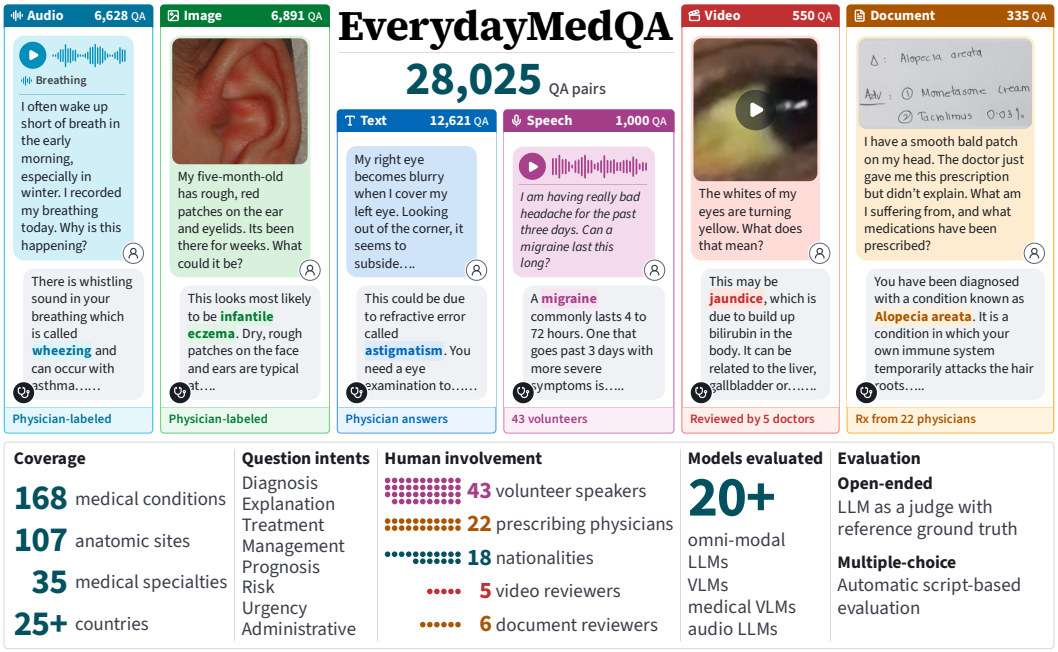

Figure 1: Overview of **EverydayMedQA**. The benchmark comprises 28,025 question-answer pairs across six consumer-oriented input types: text, image, audio, speech, video, and medical documents. Representative examples illustrate the inputs and corresponding questions and answers, while the bottom panels summarize clinical and geographic coverage, question intents, human involvement, evaluated model families, and the open-ended and multiple-choice evaluation settings.

ABSTRACT

Conversational AI assistants are increasingly used for healthcare guidance, both before and after consulting a clinician, to prepare for appointments, clarify medical advice, and seek additional information. Recent large multimodal models (LMMs), including *omni-modal* systems that process text, image, audio, and video, can now respond to medical queries expressed through many forms of input. Yet most medical AI benchmarks remain clinician-oriented: they use expert terminology, rely on clinical-grade inputs, cover narrow medical domains, and typically evaluate a single modality. As a result, how these models handle the informal, consumer-generated queries common in real-world telemedicine remains poorly understood. We introduce **EverydayMedQA**, a consumer-oriented, multimodal medical question-answering (QA) benchmark of **28,025** question–answer pairs across **six input types**: text, image, audio, speech, video, and document. Every question is phrased as a layperson query and paired with consumer-grade media rather than clinical instrumentation, spanning a broad range of medical conditions, specialties, anatomical regions, geographic contexts, and consumer question intents. Certified physicians contributed to benchmark construction through

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

data provision and review of cases requiring clinical judgment. Each item is available in both multiple-choice and open-ended formats, evaluated using a reference-based protocol.

Evaluating a comprehensive set of 23 leading models spanning omnimodal models, LLM, VLM, medical VLM, and audio LLMs, we observe that medical question answering performance does not transfer uniformly across input modalities. An omni-modal model that performs well on text may still struggle with patient questions presented in speech, images, audio, video, or documents. We also find that multiple-choice and open-ended questions probe different abilities: a model may select the right options but still give an incomplete free-form medical answer for the same sample. In many cases, models give fluent and relevant-sounding answers that remain clinically incomplete or incorrect. These results suggest that clinician-curated, single-modality benchmarks may overestimate model readiness for consumer-facing telemedicine. To our knowledge, EverydayMedQA is the first consumer-oriented medical QA benchmark to span these six input types under a unified evaluation protocol. Our benchmark and evaluation script will be made publicly available.

1 INTRODUCTION

Large language models (LLMs) and their multimodal successors (LMMs) are increasingly used as interfaces for seeking health information. These systems now pass medical licensing examinations, generate clinical documentation, and triage patient messages (Kasagga et al., 2025; Liu et al., 2025b; Paruchuri et al., 2025). People have started using AI assistants *before, during, and after* clinical consultations to interpret everyday symptoms and health artifacts, such as a rash, a cough, a swollen eye, or a prescription label (Cao et al., 2026; Alshehri et al., 2026). This pattern is especially important in remote and underserved settings, where limited access to physicians can make AI systems a practical first source of medical information rather than merely a supplement to clinic-based care. As these systems move toward consumer-facing deployment, their behavior on informal, consumer-generated health questions must be measured carefully because these answers may shape how people interpret symptoms, decide whether to seek care, or follow medical advice.

Despite this shift, most benchmarks used to evaluate medical AI remain designed around clinician-facing tasks. Resources such as PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2021), and SLAKE (Liu et al., 2021) draw heavily from board-style questions, biomedical literature, radiology reports, and curated pathology or clinical images. These datasets have driven important progress, but they evaluate a different input distribution from consumer health use: expert vocabulary rather than lay language, clinical-grade imagery rather than phone-captured media, and narrow clinical sources rather than everyday complaints. A model can therefore perform well on established medical benchmarks while remaining untested on the inputs and questions that consumer-facing health systems are likely to encounter.

We introduce **EverydayMedQA**, a multi-modal medical question-answering benchmark for everyday health information seeking. Each question is framed as a layperson query, similar to what a user might submit to a telemedicine chat, symptom-checking tool, or general-purpose AI assistant. The associated signals approximate consumer telemedicine inputs, including spoken questions, phone- or web-recorded respiratory audio, mobile images, self-recorded videos, and photographed or scanned medical documents, which are often difficult for non-expert users to interpret.

EverydayMedQA contains 28,025 question-answer pairs across six modalities: text (12,621), speech (1,000), image (6,891), audio (6,628), video (550), and document (335). Samples are provided in both open-ended and multiple-choice question formats to support complementary evaluation via LLM judging and exact-match scoring. Metadata span 35 medical specialties and 86 medical subspecialties, 107 anatomical locations, and 168 clinical categories. Combining 20 raw data sources with our newly collected samples, the benchmark contains Samples map to over 25 geographic countries , with the some drawn from global or web-based sources. Questions span eight consumer question intents: diagnosis, treatment, prognosis, urgency, explanation, management, risk, and administrative queries. To improve benchmark reliability, human review is used throughout curation. Certified physicians tackle cases that require medical judgment, such as checking whether a ques-

tion, answer, label, or media input is clinically correct. A non-expert reviewer handles technical issues such as transcription errors, OCR mistakes, and media-quality problems. The amount and type of review vary by source, modality, and likely error type. Figure 1 provides a benchmark overview.

We summarize our contributions as follows:

- **Consumer-facing multi-modal medical QA.** EverydayMedQA evaluates layperson-framed medical questions across consumer-grade text, speech, image, audio, video, and document inputs under a unified question-answering format.
- **Broad clinical and consumer-question coverage.** The benchmark spans 35 medical specialties, 86 medical subspecialties, 107 anatomic-location labels, 168 medical-condition labels, and eight consumer question intents, extending beyond the narrow domain focus of prior consumer multimodal benchmarks.
- **Human and medical expert involvement.** Human contributors and certified physicians participate throughout benchmark development, including speech collection, prescription contribution, source and disease selection, annotation, and QA review and refinement.
- **Dual-format evaluation.** Each item is evaluated in both open-ended and multiple-choice formats, scored using an LLM-as-judge protocol and a deterministic exact-match script respectively, enabling comparison under a shared evaluation protocol.
- **Systematic evaluation of leading models.** Evaluating 23 leading open and proprietary models, including omnimodal models, LLMs, VLMs, medical VLMs, and audio LLMs, we observe performance variation across input modalities and between open-ended and multiple-choice answer formats.

2 Related Work

Medical QA Benchmarks. Evaluation of large language models in medicine began with expert-level text tasks. Med-PaLM and Med-PaLM 2 (Singhal et al., 2025) demonstrated that instruction-tuned LLMs could match or exceed the USMLE passing threshold, and general-purpose models such as GPT-4 proved competitive on medical examinations without domain-specific fine-tuning (Kung et al., 2023; Nori et al., 2023). Standard text benchmarks including MedQA (Jin et al., 2021), PubMedQA (Jin et al., 2019), MedMCQA (Pal et al., 2022), and the medical subsets of MMLU (Hendrycks et al., 2021), consolidated in MultiMedQA (Singhal et al., 2023), remain common reference points for clinical reasoning. Multimodal evaluation followed with visual QA datasets such as VQA-RAD (Lau et al., 2018), PathVQA (He et al., 2021), SLAKE (Liu et al., 2021), and PMC-VQA (Zhang et al., 2023), alongside medically adapted LMMs such as LLaVA-Med (Li et al., 2023), Med-Flamingo (Moor et al., 2023), BiMediX2 (Mullappilly et al., 2025), MedGemma (Sellergren et al., 2025), and Lingshu (LASA Team et al., 2025). These resources have driven important progress, but they are overwhelmingly expert-oriented, sourced from clinical examinations, radiology reports, or curated diagnostic images, and they evaluate either text alone or static image-text pairs.

Consumer-Oriented Medical Benchmarks. A smaller body of work targets the consumer setting, where questions are informal and grounded in the limited information available to a patient. Text-only consumer QA resources include MedQuAD (Ben Abacha & Demner-Fushman, 2019), K-QA (Manes et al., 2024), and MedRedQA (Nguyen et al., 2023), the latter sourcing consumer question–physician answer pairs from r/AskDocs. More recent work extends geographic and linguistic coverage: AfriMed-QA (Nimo et al., 2025) spans 32 specialties across 16 African countries, and PerMedCQA (Jamali et al., 2025) covers 68,138 pairs from a Persian-language medical forum. Multimodal consumer benchmarks are narrower in clinical scope. DermaVQA (Yim et al., 2024a) and DermaVQA-DAS (Yim et al., 2025b) address skin image QA only; WoundcareVQA (Yim et al., 2025a) covers wound care; RA-QA (Bertolino et al., 2026) and MedMosaic (Rajgarhia et al., 2026) target respiratory and physiological audio rather than spoken patient questions, and 3MDBench (Sviridov et al., 2025) simulates text-and-image telemedicine dialogues. Medical video benchmarks such as MedVidQA (Gupta et al., 2023), CMIVQA (Li et al., 2025), and M3-Med (Liu et al., 2025a) use clinician-produced instructional content rather than self-recorded patient footage. For medical documents, RJUA-MedDQA (Jin et al., 2024) pairs 2,000 Chinese-language hospital report images

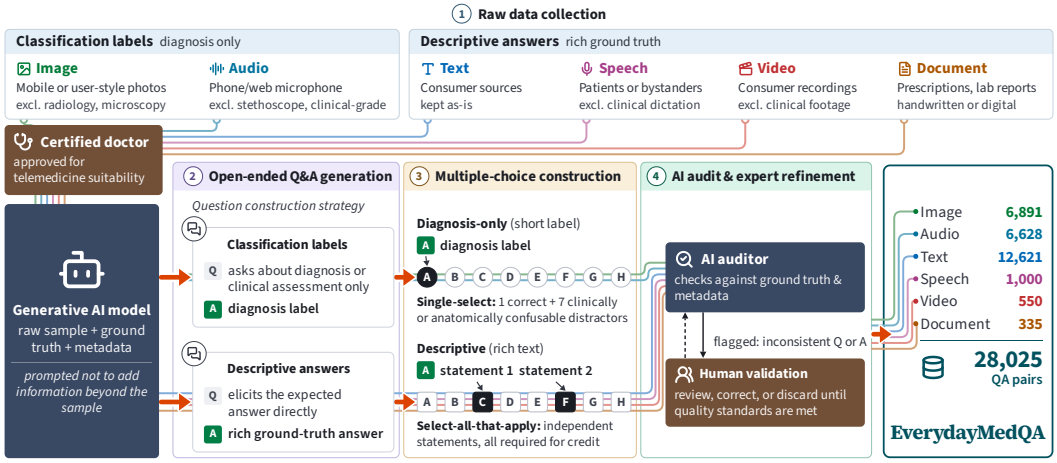

Figure 2: Overview of **EverydayMedQA** benchmark construction pipeline, from data collection to open-ended question-answer generation, multiple-choice construction, AI-assisted auditing and human validation.

with clinical QA, while MedOCR-Vision (Hussain, 2025) addresses text recognition in medical documents without a consumer QA framing. Consequently, no existing benchmark provides a unified consumer-facing QA evaluation across text, speech, image, audio, video, and document.

Omni-Modal Models and Healthcare Evaluation. The most recent generation of foundation models supports native omni-modal interaction: Gemini (Gemini Team et al., 2023), and the Qwen2.5-Omni and Qwen3-Omni families (Xu et al., 2025a;b) can process text, images, audio, including speech, and video within a single unified architecture. This capability is directly relevant to consumer healthcare, where users may submit a typed symptom description, a spoken question, a skin photograph, a cough recording, or a photographed lab report depending on their situation and available means. However, existing medical benchmarks evaluate these models on only one or two modalities, and almost always on clinician-curated inputs, making it difficult to assess how well they perform across the full range of consumer health queries they are likely to encounter in deployment. EverydayMedQA addresses this by providing a consumer-health QA benchmark that spans six input types under a consistent layperson-query framing, allowing systematic comparison of model capability across the inputs that arise in real-world telemedicine.

3 EVERYDAYMEDQA

EverydayMedQA is constructed through a four-stage pipeline: data collection, question-answer generation, multiple-choice construction, and AI-assisted auditing with human medical expert refinement when necessary. We do not apply a uniform re-annotation procedure across all sources; instead, the level of medical review depends on the provenance of the original annotation and the extent to which a sample is newly constructed or modified in EverydayMedQA. Table 1 lists the source datasets and clarifies the medical expert involvement, and Figure 2 shows the benchmark construction pipeline.

Stage 1: Consumer-oriented source selection and data acquisition. We select or collect inputs that could plausibly be submitted to a consumer-facing health system without specialist acquisition equipment. Text sources contain naturally written patient questions with medically verified answers. Image sources use standard cameras or mobile phones, excluding radiology and microscopy, while respiratory audio is recorded with phone or web microphones rather than clinical devices. Video sources use publicly shared, conventionally recorded clips showing visible medical findings, and documents comprise scanned or photographed medical records. A certified physician reviewed candidate sources and representative samples for clinical relevance and suitability for consumer telemedicine.

Table 1: Data sources used in EverydayMedQA, highlighting medical expert involvement in source annotation and benchmark construction.

Data source	Source data	Modality	Samples	Source annotation / validation	QA process
iCliniq (Mrini et al., 2021)	Q&A	Text, Speech	7,321 T / 815 S	Physician-written answers	■
MedRedQA (Nguyen et al., 2023)	Q&A	Text, Speech	5,099 T / 152 S	Physician-written answers	■
K-QA (Manes et al., 2024)	Q&A	Text, Speech	201 T / 33 S	Physician-written answers	■
COUGHVID (Orlandic et al., 2021)	Audio + Label	Audio	2,675	Physician-labeled	●
ICBHI 2017 (Rocha et al., 2017)	Audio + Label	Audio	233	Physician-labeled	●
Coswara (Bhattacharya et al., 2023)	Audio + Label	Audio	2,636	Physician-labeled + lab-confirmed	●
CODA TB (Cattamanchi et al., 2024)	Audio + Label	Audio	1,084	Physician-labeled + lab-confirmed	●
PAD-UFES-20 (Pacheco et al., 2020)	Image + Label	Image	2,298	Physician-labeled + lab-confirmed	●
Eye Diseases (Bitto et al., 2025)	Image + Label	Image	1,773	Physician-validated labels	●
MSLD v2.0 (Ali et al., 2024; 2022)	Image + Label	Image	755	Physician-validated labels	●
DermaVQA-IIYI (Yim et al., 2024a)	Image + Q&A	Image	998	Physician-written answers	▲
Mouth & Oral Diseases (Rashid et al., 2024)	Image + Label	Image	516	Physician-labeled	●
Oral Images (B R et al., 2021)	Image + Label	Image	323	Physician-labeled	●
External Eye Photographs (Yoo, 2025)	Image + Label	Image	128	Public labels	●
Tonsillitis Phases (Mohanta et al., 2021)	Image + Label	Image	100	Public labels	●
TikTok Videos	Patient + physician video	Video	324	Physician speech	◆
YouTube Videos	Public video	Video	226	Title-derived label	◆
MedOCR-Vision (Hussain, 2025)	Document + OCR	Document	178	OCR	◆
Prescription	Scanned document	Document	116	OCR	◆
Synthetic Australian Docs (Webb, 2026)	Document + OCR	Document	41	OCR/document content	◆

● Label-grounded QA, checked against source GT ▲ Consumer rewrite, checked against original QA
◆ Source-grounded QA, physician-reviewed/refined ■ Original QA retained + speech recorded

To address gaps in suitable public data, we conducted additional benchmark-specific data acquisition for three input types. For speech, we selected questions from the text sources and constructed a 1,000-item set comprising 500 human recordings and 500 samples synthesized with Kokoro-82M (Hexgrad, 2025). The synthetic samples span multiple male and female voices with varied accents, and gender-specific questions were assigned a corresponding voice when applicable. The human recordings were collected from 43 consenting adult volunteers of 18 nationalities using a dedicated collection tool. Participants voiced each question naturally, as if speaking to a clinician, and recordings were accepted only when they preserved the original clinical meaning. The participants included women and men from 12 first languages, aged between 18 and 44, introducing variation in phrasing, pronunciation, and recording conditions beyond TTS alone. Participants could also choose voice masking to reduce identifiability. For video, we carefully curated relevant public content from TikTok and YouTube. TikTok videos are built from physician *stitch* responses to patient-recorded clips. The patient clips are the consumer media and the physician response in the stitched videos are used as the source for answer generation. YouTube videos come from a physician-operated public channel documenting visible medical findings with patient consent. The doctor owning the channel uploads the video and provides a title that contains the label of the uploaded medical condition. For documents, we screened public sources, excluded medically invalid synthetic prescriptions, and added 116 prescriptions contributed by 22 physicians across medical specialties. These were either prescriptions from clinical practice that physicians de-identified before transfer or new prescriptions written by physicians using a template designed by a physician on our team.

Stage 2: Semi-automated open-ended question–answer generation. QA construction approach varies depending on the modality and the available annotation. A proprietary generative model (GPT or Gemini, depending on the modality) is used to convert the source label and available metadata into a layperson question and a label-grounded answer. For video and document sources, we use Gemini 3.1 Flash-Lite Preview, which supports native video input and OCR-capable document understanding. For image and audio sources, we use GPT-5.5. Both models were evaluated on pilot samples before selecting the choice of model per modality.

Existing consumer text QA pairs are retained to preserve their natural patient register and physician-authored answers. Speech uses a subset of text items; however, written questions are made speech-friendly by removing formatting artifacts and expanding abbreviations and units (e.g., ”kg” to ”kilograms”) before recording or synthesis.

Image and respiratory audio sources are mostly classification-based datasets, so the generative model converts the available metadata into diagnosis- or assessment-oriented questions, and then answers are constructed to be grounded by the source label. Existing image DermaVQA samples are rewritten only when needed to match the consumer framing. The QA construction prompt instructs the

model to preserve relevant context, such as age, sex, symptoms, and anatomical site, while prohibiting unsupported clinical details.

Video QA is grounded in source-specific evidence. For TikTok, the physician's *stitch* response transcription provides the evidential answer for the accompanying patient clip. From 108 patient–physician video pairs, the model constructs three QA pairs and corresponding evidence intervals per video, yielding 324 samples. For YouTube, video titles provide the medical condition labels. Clips are trimmed to exclude on-screen diagnostic text. For all videos, non-medical audio is muted to avoid unnecessary attached sounds like music or jokes, while clips with relevant audio, such as coughing or abnormal breathing, are kept unmuted.

For documents, the generative model uses OCR text together with the document to construct two question types: direct extraction of information such as medication names, dosages, or laboratory values, and extraction followed by medical reasoning.

Stage 3: Multiple-choice construction. Each question is also converted into a multiple-choice variant, with the answer options generated by a generative model using an instruction prompt designed to produce clinically plausible and question-aligned alternatives. Diagnosis-only image and respiratory audio items use a single-select format with seven clinically or anatomically plausible diagnosis distractors, requiring the model to distinguish among genuinely confusable conditions. In contrast, text, speech, video, and document items use a stricter select-all-that-apply format because their typically longer, descriptive ground truth answers make conventional single-select MC particularly vulnerable to answer-choice cues. In such settings, the correct option may be identifiable from its style, specificity, or completeness, while constructing equally detailed and clinically plausible distractors is difficult. A model may therefore select the most polished or question-aligned option without verifying all required clinical facts. To reduce this shortcut, descriptive reference answers are decomposed into shorter, independent statements distributed across the options, and credit is awarded only when the exact set of valid statements is selected.

Stage 4: AI-assisted auditing and human/expert refinement. Generated question–answer pairs are first passed to a second AI model acting as an independent auditor, which flags inconsistencies between the question, answer, source label, and associated metadata. Flagged samples are then manually reviewed and, when medical judgment is required, referred to certified medical experts for refinement or removal. The extent of expert review is adapted to the modality and the type of error identified.

For video samples, all generated question–answer pairs and their corresponding trimmed evidence clips are independently checked by five certified medical doctors, who refine or remove samples as needed. For medical documents, the auditor model is asked to review the provided document QA sample alongside the OCR text; non-clinical issues such as OCR errors are first corrected by a human editor; samples that require medical judgment are then referred to six certified physicians for verification, refinement, or removal. In total, 79 document samples were escalated for physician review. Human speech samples were first transcribed by the audio model (Whisper Large-v3 (Radford et al., 2023)) and then checked for semantic similarity with the ground truth text question. Samples that have a significant semantic difference are then passed to a human for manual checks. Out of 23 flagged samples, 10 speech samples were rejected due to missing critical key information, while the rest were rejected due to the Whisper model's failure to understand human pronunciation and accents. Image and audio QA samples are also checked by the AI auditor and human against the GT labels for consistency.

4 Evaluation Protocol

EverydayMedQA is evaluated in two complementary formats: open-ended QA, where models generate free-form answers without candidate options, and multiple-choice QA, which tests whether models recognize the correct clinical conclusion when options are supplied. Reporting both separates free-form clinical generation from constrained answer recognition. For each item, the model receives the question and its associated input signal: text, speech, image, respiratory audio, video, or document.

4.1 Open-Ended Evaluation

In the open-ended setting, models generate unrestricted natural-language answers, requiring them to identify relevant evidence, respond in patient-facing language, and avoid unsafe or unsupported advice. Because exact string matching is inappropriate for such responses, we use a reference-based LLM-as-judge protocol.

Each generated answer is judged against the reference answer by Gemma-4-31B-IT and Gemma-4-12B-IT using the same answer-type-specific rubric. Judges score whether the candidate preserves the reference clinical meaning, rather than rewarding word overlap, style, or plausible but unsupported extra information. We report Gemma-4-31B-IT scores in the main paper and provide Gemma-4-12B-IT results and the full rubric prompts in the appendix.

Diagnosis rubric. Diagnosis-exclusive image and respiratory audio items are scored on a single 0–10 criterion, *Diagnosis Correctness*. A score of 0 indicates a response that contradicts the reference clinical state; intermediate scores capture related but incorrect diagnoses, correct parent categories, or missing qualifiers; and 10 is reserved for the exact diagnosis or a fully equivalent synonym with all required qualifiers. Bedside manner, treatment advice, and disclaimers are ignored unless they contradict the diagnostic answer.

Clinical information rubric. Text, speech, video, and document items are scored on three independent 0–10 criteria: *Clinical Correctness*, which measures preservation of the reference conclusion, interpretation, recommendation, or uncertainty; *Safety and Triage*, which measures preservation of urgency, escalation advice, medication guidance, and warnings; and *Question Fulfillment*, which measures whether the response answers the user's question and important sub-questions. The judge may not use its own medical knowledge to fill gaps in the reference or reward responses for being detailed, polite, or plausible beyond what the reference supports.

Score aggregation. For diagnosis-only items, the Diagnosis Correctness score is divided by 10 and used as the sample-level OE score. For clinical-information items, the sample-level OE score is a weighted average of the three 0–10 rubric criteria: Clinical Correctness (60%), Safety and Triage (25%), and Question Fulfillment (15%), normalized to [0,1]. Modality-level OE scores are then reported as mean sample-level scores multiplied by 100.

4.2 Multiple-Choice Evaluation

Multiple-choice evaluation is fully automated with deterministic scoring and uses the same input signal and question as the open-ended setting, with answer choices added. As described in Section 3, diagnosis-only image and respiratory audio items use single-select questions with eight options, scored correct only when the selected option matches the ground truth. Text, speech, video, and document items use multi-select questions built from independent statements in the reference answer; full credit requires selecting exactly all valid options and no invalid ones.

The reported MC score is exact-match accuracy within each input type. For multi-select items, we also compute recall, the fraction of reference options recovered, and selection bias, the average number of predicted minus reference options, to distinguish missing information from over- or under-selection.

4.3 Results

Table 2 reports model performance across the evaluated input types. Gemini 3.1 Pro Preview obtains the highest six-modality Overall scores for both OE and MC (61.2 and 65.7). Among open-source models, Qwen3-Omni-30B-A3B-Thinking has the strongest Overall scores (50.3 OE and 32.1 MC). Performance varies substantially by modality. At the input-type level, respiratory audio has the lowest mean open-ended score (29.3), whereas documents have the highest mean open-ended scores (60.1). The gap between the best proprietary and best open-source model is small for text OE (2.4 points), but larger for video (14.7 points) and document inputs (16.9 points). A similar separation appears for document MC, where Gemini reaches 82.7 and the strongest open-source model reaches 29.2.

Table 2: EverydayMedQA results across six input modalities. **OE** denotes open-ended scores judged by Gemma-4-31B-IT, while **MC** denotes exact-match accuracy. **Overall** is the unweighted macro-average across modalities. The **Mean** row averages the displayed model scores, and the best score within each model group is in **bold**. Additional details are available at Table 4 and Table 5 in the Appendix. † GPT-4o and GPT-4o mini lack audio support and were not evaluated on speech or respiratory audio; their Overall scores average the four supported modalities and are therefore not directly comparable to six-modality Overall scores.

	Text		Speech		Image		Resp. Audio		Video		Document		Overall	
Model	OE	MC	OE	MC	OE	MC	OE	MC	OE	MC	OE	MC	OE	MC
Proprietary models														
Gemini 3.1 Pro Preview	**52.7**	**66.5**	**62.4**	**67.8**	**66.9**	**68.0**	**36.2**	**43.4**	**67.9**	**65.6**	**80.9**	**82.7**	**61.2**	**65.7**
GPT-4o†	51.0	24.3	–	–	52.2	52.1	–	–	63.0	53.5	70.3	48.7	59.1	44.7
GPT-4o mini†	49.0	15.3	–	–	33.5	44.4	–	–	53.3	35.3	63.8	24.2	49.9	29.8
Open-source models														
Qwen3-Omni-30B-A3B-Instruct	**50.3**	**24.3**	58.4	**19.1**	51.9	46.8	30.3	26.4	50.1	35.5	57.6	11.6	49.8	27.3
Qwen3-Omni-30B-A3B-Thinking	45.8	22.4	55.9	18.7	**53.4**	**49.5**	29.7	22.9	**53.2**	**49.8**	**64.0**	**29.2**	**50.3**	**32.1**
Gemma-4-E4B	42.7	15.7	52.5	11.8	39.9	36.5	24.2	30.4	51.9	31.6	63.7	20.3	45.8	24.4
MiniCPM-o 2.6	41.9	16.5	45.4	14.4	44.4	16.1	**39.1**	25.7	44.3	12.0	50.4	18.8	44.2	17.2
MiniCPM-o 4.5	38.6	12.5	46.7	7.1	39.2	37.9	28.6	27.9	50.1	40.9	63.3	11.6	44.4	23.0
HumanSense-Omni-Reasoning	43.3	14.2	50.3	14.7	42.8	34.8	32.0	29.4	45.0	27.4	54.4	23.6	44.6	24.0
Qwen2.5-Omni-7B	43.3	15.2	50.4	16.1	42.2	34.4	32.1	29.4	45.3	27.4	53.8	20.9	44.5	23.9
Nemotron-Nano-Omni-30B-A3B	23.8	16.8	**60.3**	13.3	43.2	29.6	26.9	29.1	47.5	8.6	58.3	20.6	43.3	19.7
Qwen2.5-Omni-3B	39.8	6.5	45.1	7.8	37.8	23.0	30.0	**46.9**	38.8	18.0	48.4	14.3	40.0	19.4
Gemma-4-E2B	40.9	12.7	38.8	8.4	30.7	37.2	13.5	38.4	46.6	10.0	52.1	5.4	37.1	18.7
Mean	43.3	20.2	51.5	18.1	44.5	39.0	29.3	31.8	50.6	32.0	60.1	25.5	47.2	28.5

Table 3: Evaluation results for modality-specific models. **OE** denotes open-ended scores judged by Gemma-4-31B-IT, while **MC** denotes exact-match accuracy for the multiple-choice/multi-select task and Grey shade marks unsupported modalities.

	Text		Speech		Image		Resp. Audio		Video		Document	
Model	OE	MC	OE	MC	OE	MC	OE	MC	OE	MC	OE	MC
LLMs												
K2-Horizon-MoVA-36B	42.7	12.2										
Granite 4.2 30B	55.1	18.4										
Nemotron 3.5 Lightning	53.7	11.8										
Audio LLMs												
Audio Flamingo 3			23.4	4.4			30.3	35.1				
Voxtral Small 24B			54.2	16.2			31.7	36.8				
Non-medical VLMs												
Qwen3.8-27B	54.6	34.8			52.3	64.7			62.5	52.4	72.0	50.4
Muse Glimmer 30B	34.3	24.0			49.7	49.5			54.9	32.7	77.5	49.3
Medical VLMs												
Lingshu-32B	51.1	29.6			47.5	44.0			59.3	32.5	61.7	40.3
Hulu-Med-30B-A3B	48.4	18.5			59.4	53.3			62.8	25.8	64.1	18.5
MedGemma 1.5	48.1	0.9			53.8	30.5					53.6	3.0

Table 3 shows that several modality-specific models are competitive with, and in some cases exceed, omni-modal models on their supported inputs. On text, Granite 4.2 30B obtains the highest OE score overall (55.1), above Gemini (52.7), while Qwen3.8-27B exceeds all omni models on text OE (54.6) and all open-source models on text MC (34.8). Modality-specific VLMs also outperform the best open-source omni models on image, video, and document inputs: Hulu-Med-30B-A3B leads image OE (59.4), Qwen3.8-27B leads image and video MC among open-source omni models (64.7 and 52.4), and Muse Glimmer 30B achieves the highest non-Gemini document OE score (77.5).

Table 2 reports OE scores from Gemma-4-31B-IT, while the full Gemma-4-31B-IT & Gemma-4-12B-IT comparison is provided in the appendix Table 6 & Table 7. The two OE judges produce broadly similar model rankings. Across the Overall columns, Gemma-4-31B-IT and Gemma-4-12B-IT have Spearman and Pearson correlations of 0.94 and 0.98, respectively. Agreement is strongest

for document and speech inputs ($\rho = 0.99$ each), followed by video ($\rho = 0.93$), image ($\rho = 0.86$), text ($\rho = 0.85$), and respiratory audio ($\rho = 0.58$).

OE and MC rankings are less consistently aligned than the two OE judges, especially when examined by modality. The highest OE–MC rank correlations occur for video ($\rho = 0.79$) and document inputs ($\rho = 0.71$), followed by speech ($\rho = 0.70$), text ($\rho = 0.68$), image ($\rho = 0.45$), and respiratory audio ($\rho = -0.14$). This divergence is visible for document inputs: Qwen3-Omni-30B-A3B-Thinking has the strongest open-source document OE score (64.0), but its document MC score is 29.2.

We also performed a physician check on the open-ended response rankings. One expert physician reviewed 370 anonymized model responses from a stratified set of 30 samples, with five samples from each evaluated modality, and ranked the eligible model responses within each sample. Gemma-4-31B-IT agreed with the physician on 74.0% of pairwise response comparisons, with Cohen's $\kappa = 0.481$.

5 Discussion

EverydayMedQA highlights that strong medical question answering remains uneven across input modalities. Omni-modal models that perform well on text or document inputs do not necessarily transfer that performance to speech, respiratory audio, or video. This suggests that everyday medical AI systems should not be evaluated only on textual clinical knowledge, since clinically relevant information can be communicated through spoken descriptions, audio, images, documents, and short videos.

Overall, the strongest proprietary model consistently outperforms open-source models, especially on video and document inputs. However, the gap is not uniform across modalities. Text performance is comparatively closer, while modalities requiring richer perceptual grounding show larger separation. This pattern suggests that current open-source omni-modal models have made progress in broad instruction following but still lag in integrating modality-specific evidence into generating clinically reliable answers.

The modality-specific results show that model specialization and omni-modal coverage solve different parts of the problem. Specialized models can be strong on the inputs they support, sometimes matching or exceeding omni-modal models, but this strength does not replace the need for broad modality coverage in everyday telemedicine. At the same time, medical specialization alone is not enough: medical VLMs trained largely on clinician-facing text and clinical images are competitive on several visual and document tasks, but they do not consistently surpass general-domain VLMs on consumer-style health questions and media.

The results also show that open-ended and multiple-choice evaluation are not interchangeable. MC can reward recognition of the correct option, while open-ended evaluation tests whether the model can generate the clinical answer without cues. The submetrics further show that fluent, responsive answers can still be clinically incomplete or incorrect, motivating separate reporting of clinical correctness, safety, and question fulfillment.

Overall, EverydayMedQA shows that omni-modal medical QA should be evaluated as a set of modality-specific capabilities rather than a single aggregate score. Aggregate results are useful for summarizing model behavior, but the modality, submetric, and selection-bias analysis reveal different failure modes.

6 Conclusion

EverydayMedQA provides a consumer-oriented benchmark for testing medical QA across the input types people use in everyday health seeking: text, images, respiratory audio, speech, video, and documents. Across more than 20 leading models, we find that performance does not transfer reliably across modalities, and that open-ended and multiple-choice formats reveal different model behaviors. Fluent, relevant-sounding answers can still miss clinically important information, underscoring the need for evaluations grounded in patient-generated inputs rather than only clinician-centered or text-only tasks. EverydayMedQA provides a benchmark for tracking progress toward more reliable multimodal medical assistants.

AI USE STATEMENT

Generative AI tools were used throughout benchmark development and evaluation to assist with source-grounded QA generation and rewriting, multiple-choice construction, automated consistency checking, dataset cleaning and reformatting, evaluation-protocol and judge-prompt development and refinement, and qualitative analysis of evaluation results. The models generated or transformed QA from source-provided information, including ground-truth labels, existing QA, OCR, physician speech, and video metadata, rather than independently determining clinical labels from their internal knowledge. AI tools were additionally used for manuscript drafting and editing. AI assistance was not used for the generation or modification of figure components, mathematical claims, proofs, or translation. All AI-assisted outputs incorporated into the benchmark, analysis, code, figures, or manuscript were reviewed by the authors, with medically substantive benchmark content further reviewed and refined by physicians where applicable. We take responsibility for the final content of this work, including all text, claims, code, data, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our dataset is under the CC-BY-NC-SA-4.0 license. All raw image and audio samples are released under open licenses by their original authors, several of whom explicitly document participant consent and, where applicable, IRB approval (see Table 10 in the Appendix). Text datasets are either de-identified by their creators (K-QA), licensed under Apache 2.0 (iCliniq), or released under CC BY-NC-SA 4.0 (MedRedQA). The question–answer pairs generated by our pipeline corresponding to those datasets stay under our license.

Raw TikTok and YouTube media will not be redistributed. Following the pointer-and-annotation model used by established web-video benchmarks (Wang et al., 2025; Carreira et al., 2019; Soomro et al., 2012; Caba Heilbron et al., 2015; Miech et al., 2019), we release only source pointers, temporal boundaries, and benchmark annotations; the source media remains under the control of its publisher and platform, and our license covers only the released annotations and metadata.

All human speech recordings were contributed by consenting adults who explicitly agreed to their recordings, coded participant IDs, and associated demographic categories being stored, used for research, and publicly released as part of the benchmark. Participants could independently choose natural-voice or voice-masked release; when masking was selected, only the filtered recording is stored and released. Participants were informed that both natural and, in some cases, masked voices may remain identifiable and that fictional prompts may contain sensitive health terminology. No names or contact information were collected.

REPRODUCIBILITY STATEMENT

To support reproducibility, we will release the EverydayMedQA benchmark, including the open-ended and multiple-choice question and answer files and media files. Additionally, the evaluation scripts used to compute all reported scores, as well as the prompts and configuration files for open-ended judging, multiple-choice scoring, and result aggregation, will be included. The release will also include the video-processing scripts used to construct the trimmed video inputs. The main paper describes the benchmark construction and evaluation protocol, while the appendix provides the full judge prompts and detailed result tables.

REFERENCES

Shams Nafisa Ali, Md. Tazuddin Ahmed, Joydip Paul, Tasnim Jahan, S. M. Sakeef Sani, Nawsabah Noor, and Taufiq Hasan. Monkeypox skin lesion detection using deep learning models: A feasibility study, 2022. URL https://arxiv.org/abs/2207.03342.

Shams Nafisa Ali, Md. Tazuddin Ahmed, Tasnim Jahan, Joydip Paul, S.M. Sakeef Sani, Nawsabah Noor, Anzirun Nahar Asma, and Taufiq Hasan. A web-based mpox skin lesion detection system using state-of-the-art deep learning models considering racial diversity. *Biomedical Signal Processing and Control*, 98:106742, 2024. ISSN 1746-8094. doi: 10.1016/j.bspc.2024.106742. URL https://www.sciencedirect.com/science/article/pii/S1746809424008000.

Nawaf Alshehri, Samar Aljohani, Sadeem Alruwaili, Lama Binsalamh, Taif Alshehri, Jenan Aljammaz, and Abdullah AlRafee. Public use of AI chatbots for symptom and laboratory interpretation and risky health behavior: a cross-sectional study in Saudi Arabia. *International Journal of Medicine in Developing Countries*, pp. 692–700, 2026. ISSN 1658-7820. doi: 10.24911/ijmdc.51-1768244580. URL http://dx.doi.org/10.24911/IJMDC.51-1768244580.

Nanditha B R, Geetha Kiran A, Chandrashekar H S, Dinesh M S, and Murali S. An ensemble deep neural network approach for oral cancer screening. *International Journal of Online and Biomedical Engineering (iJOE)*, 17(02):121–134, February 2021. ISSN 2626-8493. doi: 10.3991/ijoe.v17i02.19207. URL http://dx.doi.org/10.3991/ijoe.v17i02.19207.

Asma Ben Abacha and Dina Demner-Fushman. A question-entailment approach to question answering. *BMC Bioinformatics*, 20(1), October 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-3119-4. URL http://dx.doi.org/10.1186/s12859-019-3119-4.

Gaia A. Bertolino, Yuwei Zhang, Tong Xia, Domenico Talia, and Cecilia Mascolo. RA-QA: A benchmarking system for respiratory audio question answering under real-world heterogeneity, 2026. URL https://arxiv.org/abs/2602.18452.

Debarpan Bhattacharya, Neeraj Kumar Sharma, Debottam Dutta, Srikanth Raj Chetupalli, Pravin Mote, Sriram Ganapathy, C. Chandrakiran, Sahiti Nori, K. K. Suhail, Sadhana Gonuguntla, and Murali Alagesan. Coswara: A respiratory sounds and symptoms dataset for remote screening of SARS-CoV-2 infection. *Scientific Data*, 10(1), 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02266-0. URL http://dx.doi.org/10.1038/s41597-023-02266-0.

Abu Kowshir Bitto, Marzia Ahmed, Md Ruhul Rabbi, Arif Hossen, Abdul Kadar Muhammad Masum, and Dewan Md. Farid. Explainable AI based transfer learning with for multiclass eye disease classification from external eye images. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, pp. 1–7. IEEE, 2025. doi: 10.1109/qpain66474.2025.11171713. URL http://dx.doi.org/10.1109/QPAIN66474.2025.11171713.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 961–970, June 2015. doi: 10.1109/CVPR.2015.7298698.

Yancheng Cao, Yishu Ji, Yue Fu, Sahiti Dharmavaram, Meghan Turchioe, Natalie C Benda, Lena Mamykina, Yuling Sun, and Xuhai Xu. More than decision support: Exploring patients' longitudinal usage of large language models in real-world healthcare-seeking journeys. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722783. doi: 10.1145/3772318.3791946. URL https://doi.org/10.1145/3772318.3791946.

Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the Kinetics-700 human action dataset, 2019. URL https://arxiv.org/abs/1907.06987.

Adithya Cattamanchi, Simon Grandjean Lapierre, Grant Theron, William Worodria, Alfred Andama, Charles Yu, Nguyen Viet Nhung, DJ Christopher, Mihaja Raberahona, Rivonirina Rakotoarivelo, Issa Lyimo, Omar Lweno, Sage Bionetworks, and The CODA TB DREAM Challenge Consortium. CODA TB DREAM Challenge – data and benchmarking, 2024. URL https://repo-prod.prod.sagebase.org/repo/v1/doi/locate?id=syn31472953&type=ENTITY.

Gemini Team et al. Gemini: A family of highly capable multimodal models, 2023. URL https://arxiv.org/abs/2312.11805.

Deepak Gupta, Kush Attal, and Dina Demner-Fushman. A dataset for medical instructional video classification and question answering. *Scientific Data*, 10(1), March 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02036-y. URL http://dx.doi.org/10.1038/s41597-023-02036-y.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 708–718, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-short.90. URL https://aclanthology.org/2021.acl-short.90/.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.03300.

Hexgrad. Kokoro-82M text-to-speech model, 2025. URL https://huggingface.co/hexgrad/Kokoro-82M. Hugging Face repository.

Syed Naazim Hussain. MedOCR-Vision: Medical OCR dataset with balanced domain training, 2025. URL https://huggingface.co/datasets/naazimsnh02/medocr-vision-dataset.

Naghmeh Jamali, Milad Mohammadi, Danial Baledi, Zahra Rezvani, and Hesham Faili. PerMedCQA: Benchmarking large language models on medical consumer question answering in Persian language. *ArXiv*, abs/2505.18331, 2025. URL https://api.semanticscholar.org/CorpusID:278905891.

Congyun Jin, Ming Zhang, Weixiao Ma, Yujiao Li, Yingbo Wang, Yabo Jia, Yuliang Du, Tao Sun, Haowen Wang, Cong Fan, Jinjie Gu, Chenfei Chi, Xiangguo Lv, Fangzhou Li, Wei Xue, and Yiran Huang. RJUA-MedDQA: A multimodal benchmark for medical document question answering and clinical reasoning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, pp. 5218–5229, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671644. URL https://doi.org/10.1145/3637528.3671644.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021. ISSN 2076-3417. doi: 10.3390/app11146421. URL http://dx.doi.org/10.3390/app11146421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2567–2577, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1259. URL https://aclanthology.org/D19-1259/.

Alousious Kasagga, Aayam Sapkota, Gichin Changaramkumarath, Jane M Abucha, Mekdes M Wollel, Nethra Somannagari, Malik Y Husami, Kirubel T Hailu, and Ecem Kasagga. Performance of ChatGPT and large language models on medical licensing exams worldwide: A systematic review and network meta-analysis with meta-regression. *Cureus*, October 2025. ISSN 2168-8184. doi: 10.7759/cureus.94300. URL http://dx.doi.org/10.7759/cureus.94300.

Tiffany H. Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, February 2023. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000198. URL http://dx.doi.org/10.1371/journal.pdig.0000198.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning, 2025. URL https://arxiv.org/abs/2506.07044.

Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):180251, 2018. doi: 10.1038/sdata.2018.251.

Bin Li, Shenxi Liu, Yixuan Weng, Yue Du, Yuhang Tian, and Shoujun Zhou. Overview of the NLPCC 2025 shared task 4: Multi-modal, multilingual, and multi-hop medical instructional video question answering challenge. In Xian-Ling Mao, Zhaochun Ren, and Muyun Yang (eds.), *Natural Language Processing and Chinese Computing (NLPCC 2025)*, Lecture Notes in Computer Science, pp. 367–379. Springer Nature Singapore, 2025. doi: 10.1007/978-981-95-3352-7_31.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654. IEEE, 2021. doi: 10.1109/ISBI48211.2021.9434010.

Shenxi Liu, Kan Li, Mingyang Zhao, Yuhang Tian, Bin Li, Shoujun Zhou, Hongliang Li, and Fuxia Yang. M^3-Med: A benchmark for multi-lingual, multi-modal, and multi-hop reasoning in medical instructional video understanding, 2025a. URL https://arxiv.org/abs/2507.04289.

Siru Liu, Aileen P Wright, Allison B McCoy, Sean S Huang, Bryan Steitz, and Adam Wright. Detecting emergencies in patient portal messages using large language models and knowledge graph-based retrieval-augmented generation. *Journal of the American Medical Informatics Association*, 32(6):1032–1039, April 2025b. ISSN 1527-974X. doi: 10.1093/jamia/ocaf059. URL http://dx.doi.org/10.1093/jamia/ocaf059.

Itay Manes, Naama Ronn, David Cohen, Ran Ilan Ber, Zehavi Horowitz-Kugler, and Gabriel Stanovsky. K-QA: A real-world medical Q&A benchmark. In Dina Demner-Fushman, Sophia Ananiadou, Makoto Miwa, Kirk Roberts, and Junichi Tsujii (eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 277–294, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.22. URL https://aclanthology.org/2024.bionlp-1.22/.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Aninda Mohanta, Nobonita Saha, and Tamal Joyti Roy. Different phases of tonsilitis, 2021. URL https://www.researchgate.net/doi/10.13140/RG.2.2.15352.98562. https://www.kaggle.com/datasets/anindamohanta/different-phases-of-tonsilitis.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-Flamingo: a multimodal medical few-shot learner. In Stefan Hegselmann, Antonio Parziale, Divya Shanmugam, Shengpu Tang, Mercy Nyamewaa Asiedu, Serina Chang, Tom Hartvigsen, and Harvineet Singh (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 353–367. PMLR, 10 Dec 2023. URL https://proceedings.mlr.press/v225/moor23a.html.

Khalil Mrini, Franck Dernoncourt, Walter Chang, Emilia Farcas, and Ndapa Nakashole. Joint summarization-entailment optimization for consumer health question understanding. In Chaitanya Shivade, Rashmi Gangadharaiah, Spandana Gella, Sandeep Konam, Shaoqing Yuan, Yi Zhang, Parminder Bhatia, and Byron Wallace (eds.), *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pp. 58–65, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlpmc-1.8. URL https://aclanthology.org/2021.nlpmc-1.8/.

Sahal Shaji Mullappilly, Mohammed Irfan Kurpath, Sara Pieri, Saeed Yahya Alseiari, Shanavas Cholakkal, Khaled M Aldahmani, Fahad Shahbaz Khan, Rao Muhammad Anwer, Salman Khan, Timothy Baldwin, and Hisham Cholakkal. BiMediX2 : Bio-medical EXpert LMM for diverse medical modalities. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14051–14071, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.756. URL https://aclanthology.org/2025.findings-emnlp.756/.

Vincent Nguyen, Sarvnaz Karimi, Maciej Rybinski, and Zhenchang Xing. MedRedQA for medical consumer question answering: Dataset, tasks, and neural baselines. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 629–648, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.42. URL https://aclanthology.org/2023.ijcnlp-main.42/.

Charles Nimo, Tobi Olatunji, Abraham Toluwase Owodunni, Tassallah Abdullahi, Emmanuel Ayodele, Mardhiyah Sanni, Ezinwanne C. Aka, Folafunmi Omofoye, Foutse Yuehgoh, Timothy Faniran, Bonaventure F. P. Dossou, Moshood O. Yekini, Jonas Kemp, Katherine A Heller, Jude Chidubem Omeke, Chidi Asuzu Md, Naome A Etori, Aïmérou Ndiaye, Ifeoma Okoh, Evans Doe Ocansey, Wendy Kinara, Michael L. Best, Irfan Essa, Stephen Edward Moore, Chris Fourie, and Mercy Nyamewaa Asiedu. AfriMed-QA: A pan-African, multi-specialty, medical question-answering benchmark dataset. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1948–1973, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.96. URL https://aclanthology.org/2025.acl-long.96/.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *ArXiv*, abs/2303.13375, 2023. URL https://api.semanticscholar.org/CorpusID:257687695.

Lara Orlandic, Tomas Teijeiro, and David Atienza. The COUGHVID crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1), 2021. ISSN 2052-4463. doi: 10.1038/s41597-021-00937-4. URL http://dx.doi.org/10.1038/s41597-021-00937-4.

Andre G.C. Pacheco, Gustavo R. Lima, Amanda S. Salomão, Breno Krohling, Igor P. Biral, Gabriel G. de Angelo, Fábio C.R. Alves Jr, José G.M. Esgario, Alana C. Simora, Pedro B.C. Castro, Felipe B. Rodrigues, Patricia H.L. Frasson, Renato A. Krohling, Helder Knidel, Maria C.S. Santos, Rachel B. do Espírito Santo, Telma L.S.G. Macedo, Tania R.P. Canuto, and Luíz F.S. de Barros. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data in Brief*, 32:106221, October 2020. ISSN 2352-3409. doi: 10.1016/j.dib.2020.106221. URL http://dx.doi.org/10.1016/j.dib.2020.106221.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Akshay Paruchuri, Maryam Aziz, Rohit Vartak, Ayman Ali, Best Uchehara, Xin Liu, Ishan Chatterjee, and Monica Agrawal. "what's up, doc?": Analyzing how users seek health information in large-scale conversational AI datasets. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2312–2336, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.125. URL https://aclanthology.org/2025.findings-emnlp.125/.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Harshit Rajgarhia, Shuubham Ojha, Asif Shaik, Akhil Pothanapalli, Rachuri Lokesh, Abhishek Mukherji, and Prasanna Desikan. MedMosaic: A challenging large scale benchmark of diverse medical audio, 2026. URL https://arxiv.org/abs/2605.00969.

Javed Rashid, Bilal Shabbir Qaisar, Muhammad Faheem, Arslan Akram, Riaz ul Amin, and Muhammad Hamid. Mouth and oral disease classification using InceptionResNetV2 method. *Multimedia Tools and Applications*, 83(11):33903–33921, March 2024. ISSN 1573-7721. doi: 10.1007/s11042-023-16776-x. URL http://dx.doi.org/10.1007/s11042-023-16776-x.

B. M. Rocha, D. Filos, L. Mendes, I. Vogiatzis, E. Perantoni, E. Kaimakamis, P. Natsiavas, A. Oliveira, C. Jácome, A. Marques, R. P. Paiva, I. Chouvarda, P. Carvalho, and N. Maglaveras. A respiratory sound database for the development of automated classification. In Nicos Maglaveras, Ioanna Chouvarda, and Paulo de Carvalho (eds.), *Precision Medicine Powered by pHealth and Connected Health*, IFMBE Proceedings, pp. 33–37. Springer Singapore, November 2017. ISBN 9789811074196. doi: 10.1007/978-981-10-7419-6_6. URL http://dx.doi.org/10.1007/978-981-10-7419-6_6.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. MedGemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

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, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-06291-2. URL http://dx.doi.org/10.1038/s41586-023-06291-2.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R. Pfohl, Heather Cole-Lewis, Darlene Neal, Qazi Mamunur Rashid, Mike Schaekermann, Amy Wang, Dev Dash, Jonathan H. Chen, Nigam H. Shah, Sami Lachgar, Philip Andrew Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Agüera y Arcas, Nenad Tomašev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle K. Barral, Dale R. Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950, January 2025. ISSN 1546-170X. doi: 10.1038/s41591-024-03423-7. URL http://dx.doi.org/10.1038/s41591-024-03423-7.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild, 2012. URL https://arxiv.org/abs/1212.0402.

Ivan Sviridov, Amina Miftakhova, Artemiy Tereshchenko, Galina Zubkova, Pavel Blinov, and Andrey Savchenko. 3MDBench: Medical multimodal multi-agent dialogue benchmark. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 26614–26654, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1353. URL https://aclanthology.org/2025.emnlp-main.1353/.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. LVBench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967, October 2025.

Jack Webb. Synthetic Australian medical documents - sample, 2026. URL https://huggingface.co/datasets/RootCauseAnalytics/synthetic-australian-medical-documents-sample.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report, 2025a. URL https://arxiv.org/abs/2503.20215.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. DermaVQA: A multilingual visual question answering dataset for dermatology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Lecture Notes in Computer Science, pp. 209–219. Springer Nature Switzerland, 2024a. ISBN 9783031720864. doi: 10.1007/978-3-031-72086-4_20. URL http://dx.doi.org/10.1007/978-3-031-72086-4_20.

Wen-wai Yim, Yujuan Fu, Zhaoyi Sun, Asma Ben Abacha, Meliha Yetisgen, and Fei Xia. DermaVQA: A multilingual visual question answering dataset for dermatology. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Lecture Notes in Computer Science, pp. 209–219. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-72086-4_20.

Wen-wai Yim, Asma Ben Abacha, Robert Doerning, Chia-Yu Chen, Jiaying Xu, Anita Subbarao, Zixuan Yu, Fei Xia, M. Kennedy Hall, and Meliha Yetisgen. WoundcareVQA: A multilingual visual question answering benchmark dataset for wound care. *Journal of Biomedical Informatics*, 170:104888, October 2025a. ISSN 1532-0464. doi: 10.1016/j.jbi.2025.104888. URL http://dx.doi.org/10.1016/j.jbi.2025.104888.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Meliha Yetisgen, Noel Codella, Roberto Andres Novoa, and Josep Malvehy. DermaVQA-DAS: Dermatology assessment schema (DAS) & datasets for closed-ended question answering & segmentation in patient-generated dermatology images, 2025b. URL https://arxiv.org/abs/2512.24340.

TaeKeun Yoo. External eye photographs for screening referable blepharitis, 2025. URL https://data.mendeley.com/datasets/bp37vw4z8d/1.

Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. PMC-VQA: Visual instruction tuning for medical visual question answering, 2023. URL https://arxiv.org/abs/2305.10415.