MEDSENSUS: BENCHMARKING MEDICAL REASONING IN OMNI-MODAL MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Clinical reasoning rarely relies on a single source of evidence. Clinicians combine complementary observations to understand the same problem from different perspectives. Although modern models can now process heterogeneous inputs together, publicly accessible medical datasets with reliable clinical annotation remain scarce. Most existing benchmarks still focus on individual modalities or narrower pairings. We introduce **MedSensus**, a medical benchmark built from publicly available resources and reviewed by clinical experts. MedSensus comprises 9,539 evaluation instances organized into 1,865 evidence groups that share visual and acoustic evidence. Questions may have one or multiple correct answers. This structure supports question accuracy and group accuracy, with the latter requiring all related questions to be answered correctly. The benchmark also provides a factorial evaluation suite that selectively withholds the image, audio, or question stem, enabling estimation of the conditional ablation effect of each input source. Across fifteen representative models, the best model under direct prompting reaches 63.84% question accuracy and 17.04% group accuracy, compared with the clinician reference of 91.12% and 72.97%. These results show that accuracy on individual questions does not imply consistent reasoning across clinical evidence. By combining expert validation with controlled diagnostic evaluation, MedSensus provides a basis for studying how medical models integrate heterogeneous evidence.

1 INTRODUCTION

Clinical reasoning depends on the synthesis of complementary evidence whose relevance varies with the question and clinical context. Recent models can process heterogeneous inputs jointly [12, 16, 21, 26], yet the ability to accept several inputs does not establish whether a model selects and combines clinically relevant information. Rigorous evaluation therefore requires aligned evidence, verified clinical targets, and questions that reveal whether additional information improves or disrupts reasoning.

Existing benchmarks address these requirements separately. Medical visual question answering resources span diverse images and clinical tasks [3, 10, 11, 14], and medical audio datasets cover speech and physiological sounds [17, 23, 25]. Other work organizes observations by patient case or evaluates several modalities outside medicine [2, 13, 24, 28]. Few public medical resources align visual and acoustic evidence with a verified finding and several related questions. Consequently, they provide limited support for testing consistency across questions or identifying which evidence supports a prediction.

The construction of an aligned benchmark also presents annotation challenges. Source labels may disagree, recordings may repeat rather than complement the target finding, and question wording may reveal shortcuts. MedSensus therefore treats evidence alignment as an explicit annotation objective. Every case requires a verified visual anchor, every question requires a supported reference answer, and every recording requires a documented clinical relation to the case. The same structure permits controlled evaluation with irrelevant, confusing, or conflicting evidence.

We introduce **MedSensus**, a medical benchmark derived from publicly available resources under clinical expert review. Figure 1 presents representative examples. Construction starts from 499 medical images spanning 15 visual modalities. Clinical experts reconcile the source annotations and

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.

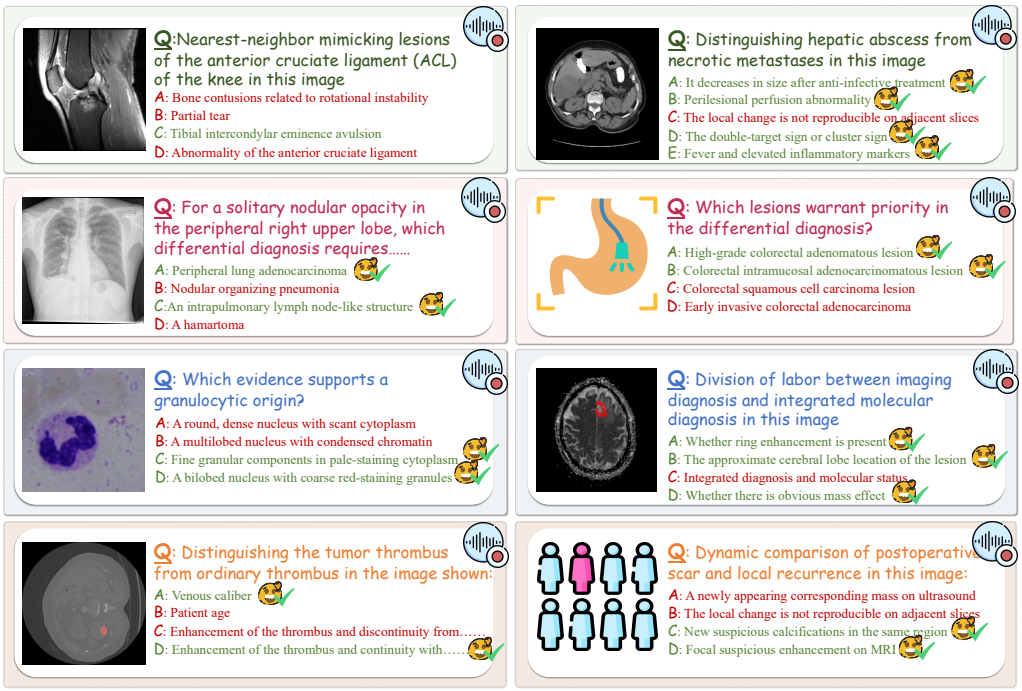


Figure 1: Representative MedSensus instances. Each panel pairs a medical image with a question authored by experts and its associated clinical or physiological recording. The examples illustrate variation in visual modality, recording type, reasoning task, and answer structure.

establish a verified visual finding for each case. Guided by these verified cases, experts create or revise 2,639 questions. The cases are then associated with 1,865 clinical or physiological recordings. Independent reviewers check the questions, reference answers, and the clinical validity of each recording association.

These components yield 9,539 evaluation instances in 1,865 evidence groups. Case groups organize all reviewed material from one clinical context, whereas each evidence group fixes one image and one recording across several related questions. This distinction supports separate measurement of individual answers and joint correctness under unchanged evidence. Questions may have several correct options, with exact matching penalizing both omissions and unsupported additions. Raw recordings are retained, while verified transcripts and expert descriptions serve only as controls for separating textual content from acoustic information.

MedSensus evaluates fifteen omni-modal models with question accuracy and group accuracy. The main experiments compare direct and reasoning prompts, isolate the contribution of each input source, estimate the contribution of raw recordings beyond textual representations, and control the relation between visual and acoustic evidence [1, 30]. Additional analyses cover responses without answer options, changes in option order, clinical subgroups, and common error patterns. Under direct prompting, the best model reaches 63.84% question accuracy and 17.04% group accuracy, compared with 91.12% and 72.97% for the clinician reference. The disparity indicates that correct individual answers do not necessarily translate into all related questions being answered correctly under the same clinical evidence.

Our contributions are:

- **A medical benchmark reviewed by clinical experts.** MedSensus provides 9,539 evaluation instances grounded in verified visual findings, reviewed questions and answers, and clinically reviewed associations between cases and recordings.
- **A benchmark for question accuracy, group-level correctness, and modality use.** Evidence groups support both question accuracy and group accuracy, while factorial controls measure the contribution of each input source.
- **An evaluation of fifteen current omni-modal models.** Experiments with fifteen models reveal a substantial gap between current systems and clinicians, particularly when all related questions must be answered correctly.

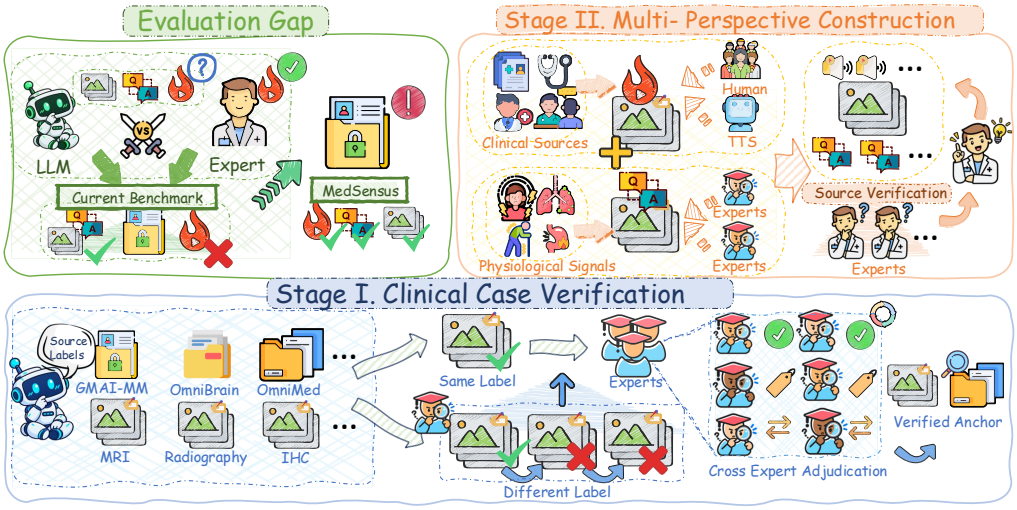


Figure 2: Construction of MedSensus. Experts reconcile source annotations to establish the visual anchor, author questions and answer sets, and review the clinical relation between each image and recording. Clinical speech and physiological signals provide complementary evidence, while controlled alternatives support evidence relation analysis.

2 The MedSensus Benchmark

2.1 Design Principles

MedSensus treats a clinical case, rather than an isolated question, as the basic unit of evaluation. Each case is anchored by a visual finding verified by experts, while recordings and questions define additional evidence and reasoning targets within the same clinical context. The design preserves which evidence is available for each prediction and repeats evaluation across questions that share that evidence. It can therefore measure both individual answer accuracy and joint correctness across questions under shared evidence, without assuming that every recording is diagnostic for every question.

The benchmark separates three nested units. A case group contains one image, all associated recordings, and all questions written for that context. An evidence group fixes one image and one recording while retaining the related questions. An evaluation instance selects one question from that evidence group. Formally,

$$C_i = (I_i, y_i^*, \mathcal{A}_i, \mathcal{Q}_i), \qquad x_{ikj} = (I_i, a_{ik}, q_{ij}, Y_{ikj}^*), \tag{1}$$

where y_i^* is the verified visual finding, $\mathcal{A}_i = \{a_{i1}, \ldots, a_{iK_i}\}$ is the set of recordings, $\mathcal{Q}_i = \{q_{i1}, \ldots, q_{iM_i}\}$ is the question set, and Y_{ikj}^* is the reference answer set for the corresponding image, recording, and question. A case group consequently produces K_i evidence groups and $K_i M_i$ evaluation instances. Every model call contains exactly one image, one recording, and one question; recordings are never concatenated.

2.2 Construction and Clinical Review

Figure 2 summarizes the construction process. The image finding, answer set, and relation of each recording to the case are treated as separate clinical judgments. They are reviewed before the components are combined into evaluation instances.

Visual anchors. Candidate images are collected from public medical datasets and repositories. Each retained case contains exactly one image. Nine candidate groups that required two to four images were excluded rather than converted into collages or reduced to a single frame. Repeated copies of an image are consolidated, with all source links preserved.

Source labels are treated as annotations to be examined, not as ground truth. Experts review each image together with its available labels and establish the visual finding that anchors the case. Agree-

ment across sources counts as corroboration only when the annotations were independently derived; inherited labels and duplicated records do not provide additional support. Conflicting interpretations are resolved against the image and source context rather than by automatic majority vote. Images available from only one source receive independent expert review.

Questions and answer sets. All questions and answer options are written by clinical experts. LLMs are used only for auxiliary checks of wording, answer consistency, and potential ambiguity; they do not construct questions or determine clinical targets, reference answers, or inclusion decisions. The questions probe findings, localization, morphology, diagnosis, management, evidence interpretation, and uncertainty. Because several statements can be simultaneously valid, the reference is a nonempty subset of the available options,

$$Y^*_{ikj} \subseteq \mathcal{V}_{ij}, \qquad |Y^*_{ikj}| \geq 1. \tag{2}$$

where $\mathcal{V}_{ij}$ is the option set. Reviewers who did not author the question assess its clinical validity, wording, options, and complete answer set. Correct options must follow from the reviewed case, while distractors must be plausible but unsupported by the intended evidence. Obvious cues in the wording or option structure are removed. This joint review is especially important when several options are correct, because either an omission or an unsupported addition changes the target answer.

Clinical and physiological audio. Each case is associated with one or more clinical or physiological recordings. Clinical speech includes human and synthesized recordings of patient history, clinician dictation, and clinical handoff. Physiological audio includes respiratory sounds and ECG sonifications with documented source signals or derivation procedures. Every image and recording pairing is reviewed for clinical validity. Assignments in the controlled relation study are further labeled concordant, noninformative, confusable, or discordant. Alternative physiological recordings remain separate interventions rather than replacements for the recording associated with the case.

For speech, semantic content is verified before delivery characteristics such as pauses, hesitation, or tempo are introduced. Human and synthesized recordings follow the same intended content, and their provenance and generation method are recorded. For physiological audio, the original signal or sonification procedure is retained. The raw recording remains the benchmark input; transcripts and expert descriptions are used only in controlled representation studies.

Final review and provenance. Visual verification, question validation, and audio alignment are assessed independently, so linguistic fluency and recording quality do not affect judgments of clinical validity. A case enters the benchmark only after its visual finding, complete answer sets, and recording assignments have passed review. Unresolved components are revised or excluded.

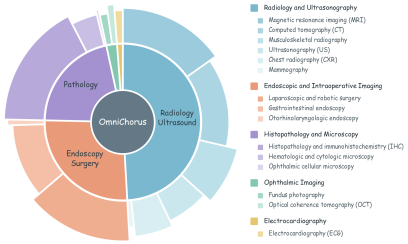


Figure 3: Visual composition of MedSensus. The 499 unique images span five imaging families and fifteen visual modalities.

The frozen manifest assigns stable identifiers to every case, image, recording, question, and evaluation instance. A separate provenance crosswalk retains source identifiers and file locations, while analysis metadata records task eligibility and evidence relation labels. Controlled assignments never overwrite the verified finding or the original recording relation.

2.3 Benchmark Composition

MedSensus contains 523 case groups, 499 image files, 1,865 audio recordings, and 2,639 unique questions. Pairing each question with every recording in its case gives $\sum_i K_i M_i = 9{,}539$ evaluation instances. Each image and recording pair defines one evidence group, producing 1,865 evidence groups in total. The images span 15 visual modalities and the cases cover 12 clinical domains.

The difference between 523 case groups and 499 image files is intentional. Thirty controlled ECG groups reuse six images, but define distinct image and audio relations and question contexts. They remain separate case groups while sharing the same underlying visual files.

A case group contains 2 to 21 questions and 1 to 11 recordings; 467 groups contain three or four recordings. Among the 2,639 questions, 967 have one correct option and 1,672 have several. This variation in group size motivates reporting question accuracy together with the stricter group accuracy.

Figure 3 summarizes five broad visual families: radiology and ultrasound, endoscopy and surgery, pathology and microscopy, ophthalmic imaging, and physiological signals. The benchmark covers findings, anatomical localization, quantitative assessment, diagnosis, clinical management, procedures, and uncertainty across the 12 represented clinical domains. Recording and question counts vary by case, so every experiment reports the size of its eligible subset. Additional examples appear in Appendix D.

2.4 Evaluation Tasks

The primary task presents one image, one recording, one question, and its answer options. MedSensus then derives controlled tasks from the same instances. To isolate the contribution of each input, let I, A, and Q denote the image, audio, and question stem, and let O denote the fixed options and response instruction. The factorial suite evaluates every subset of the three evidence inputs:

$$\mathcal{X}(S) = S \cup \{O\}, \qquad S \subseteq \{I, A, Q\}. \tag{3}$$

Its eight views include complete evidence, three conditions that omit one input, three conditions that retain one input, and a control containing only the options. An omitted input is never replaced by a placeholder or textual description.

The acoustic representation task compares clinical speech with its verified transcript and physiological signals with descriptions of their acoustic findings. In each case, the raw recording, text proxy, and their combination are evaluated on matched questions. A separate task removes answer options and scores free responses with rubrics written by experts.

The evidence relation task varies whether audio is concordant, noninformative, confusable, or discordant with the visual case. Questions are marked evidence invariant when their reference answer should remain unchanged, and evidence sensitive when clinically meaningful audio can change the supported answer. This distinction separates resistance to irrelevant or misleading evidence from appropriate revision when new evidence matters.

The benchmark manifest and analysis metadata link every instance to its case group, evidence group, media, answer set, task eligibility, and analysis strata. Comparisons therefore use matched examples and report the exact eligible sample count.

3 Evaluation and Findings

We evaluate question-level accuracy and evidence-group joint accuracy, the contribution of each input, and robustness across prompts, evidence representations, clinical strata, and answer formats. The main paper presents the factorial comparison, acoustic representation study, and visual and acoustic contributions. Appendix A reports confidence intervals, stratified results, robustness analyses, and error categories.

3.1 Evaluation Protocol

Models. We evaluate fifteen omni-modal models. The proprietary models, evaluated through provider APIs, are Qwen3.8-Omni-Flash [19], Qwen3.5-Omni-Plus [20], Gemini 3.8 Flash, Gemini 3.1 Pro, Gemini 3 Pro, Gemini 3 Flash, and Gemini 3.5 Flash [5, 6, 7, 8, 9], and Doubao-Seed-2.0-Lite [18]. The models with open weights are Qwen3-Omni-30B [27], MiniCPM-o 4.5 [4], Qwen2.5-Omni-7B and Qwen2.5-Omni-3B [26], OmniVinci-9B [29], OLA-7B [15], and Baichuan-Omni-1.5 [12]. The supplementary experimental log records the exact checkpoint or API version, access date, and decoding parameters for every model.

Clinician reference. Three practicing clinicians independently complete all 9,539 instances with full evidence and without model assistance. Each clinician evaluates every combination of image,

recording, and question as a separate instance, including repeated questions paired with different recordings. They are not tested under ablation or reasoning conditions. The main table reports their mean Acc_q and Acc_g; individual scores, agreement, and majority voting appear in the appendix.

Evaluation schedule. Complete evidence uses direct and reasoning prompts; all other factorial views use the direct prompt. Every condition preserves the same aligned instances and evidence groups. Audio conditions require one call per instance. Without audio, each of the 2,639 questions is answered once and replicated over matched recording identifiers only for scoring.

Prompts and decoding. The direct zero shot prompt requests only option letters; the reasoning prompt permits an explanation but requires the same structured final answer. Each call contains one question and its options. Inputs and prompt wording are identical across models within a condition. Decoding is deterministic when supported, and other sampling settings are recorded. A fixed retry policy is used; refusals, exhausted retries, and unparseable responses count as incorrect.

3.2 Metrics and Statistical Analysis

For an evaluation setting s, let G_s be the number of evidence groups, M_g the number of questions in group g, and $c_{gj}^{(s)} = \mathbb{1}[\widehat{Y}_{gj}^{(s)} = Y_{gj}^{(s)}]$ the indicator of an exact answer set match. With $N_s = \sum_g M_g$, the two primary metrics are

$$\text{Acc}_q^{(s)} = \frac{1}{N_s} \sum_{g=1}^{G_s} \sum_{j=1}^{M_g} c_{gj}^{(s)}, \qquad \text{Acc}_g^{(s)} = \frac{1}{G_s} \sum_{g=1}^{G_s} \prod_{j=1}^{M_g} c_{gj}^{(s)}. \tag{4}$$

Question accuracy scores individual predictions, whereas group accuracy requires every question sharing the same evidence to be correct. For conditions without audio, replicated predictions preserve matched weighting; no audio is presented to the model. We do not report accuracy over complete case groups as a primary metric because it depends mechanically on the number of recordings assigned to each case. Appendix B.1 therefore reports group accuracy by the number of questions.

For metric $m \in \{q, g\}$, the audio contribution is the paired decrease $\Delta_A^{(m)} = \text{Acc}_m^{\text{complete}} - \text{Acc}_m^{\text{without audio}}$; image and question contributions are defined analogously. Positive values indicate that retaining the input improves performance. Contributions are computed from unrounded accuracies and reported in percentage points.

Normalization ignores option order and repeated letters. Exact matching penalizes both omitted and unsupported options; option precision, recall, and set F1 are secondary measures. The chance baseline samples uniformly from all nonempty answer subsets and averages 4.22% across instances.

All aggregate results are computed from the frozen manifest. We obtain 95% confidence intervals by resampling groups and use paired bootstrap differences for conditions evaluated on the same groups. The prediction ledger stores the case, group, image, and audio identifiers; the input condition; the prompt and model versions; the reference and parsed answers; and the parsing status.

3.3 Overall Performance

Table 1 reports question accuracy and group accuracy under all eight cells of the factorial design, while retaining separate Direct and CoT columns for complete evidence. Table 2 compares original acoustic inputs with verified textual proxies and their combination. Figure 4 maps visual and acoustic contributions across models. Confidence intervals, clinical strata, and evaluation without answer options are reported in Appendix A.

Performance with complete evidence remains substantially lower than the clinician reference. Under direct prompting, Gemini 3.8 Flash obtains the highest question accuracy at 63.84% and the highest group accuracy at 17.04%, closely followed by Qwen3.8-Omni-Flash. The clinician reference reaches 91.12% and 72.97%, respectively. Models with open weights achieve lower accuracy than the leading proprietary models, but both groups exhibit the same pattern: gains on individual questions translate only weakly into gains on complete evidence groups. The clinician gap is therefore not explained solely by missing factual knowledge. Models answer many individual questions

	Complete evidence ($I+A+Q+O$)				Factorial controls (direct prompt)													
	Direct		CoT		Audio ablated $I+Q+O$		Image ablated $A+Q+O$		Question ablated $I+A+O$		Visual evidence $I+O$		Acoustic evidence $A+O$		Question stem $Q+O$		Option prior O	
Model	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g
Uniform subset random baseline	4.22	–	4.22	–	4.22	–	4.22	–	4.22	–	4.22	–	4.22	–	4.22	–	4.22	–
Clinician reference	91.12	72.97	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
Commercial and API-access models																		
Qwen3.8-Omni-Flash	63.26	16.35	64.83	17.96	56.91	9.12	58.92	11.85	44.38	5.20	38.92	1.82	43.56	4.50	54.16	7.61	32.88	0.91
Qwen3.5-Omni-Plus	59.41	12.76	61.21	14.32	54.08	7.35	55.76	9.49	42.31	4.40	36.06	1.34	39.47	3.11	51.98	6.43	32.49	0.86
Gemini 3.8 Flash	63.84	17.04	65.52	18.77	58.39	10.19	59.46	12.33	45.37	5.63	39.14	1.88	42.58	4.13	55.01	8.15	33.14	0.91
Gemini 3.1 Pro	61.57	14.69	63.22	16.25	56.58	8.90	57.75	10.94	43.82	4.99	37.72	1.61	41.11	3.65	53.12	7.02	33.06	0.91
Gemini 3 Pro	59.06	12.49	60.89	14.05	54.64	7.67	55.49	9.33	41.51	4.08	35.32	1.23	38.98	2.95	50.91	5.90	32.77	0.91
Gemini 3 Flash	56.73	10.67	58.38	11.85	52.70	6.60	53.31	7.94	39.92	3.54	33.84	1.02	37.48	2.57	49.29	5.15	33.68	1.02
Gemini 3.5 Flash	60.62	13.83	62.24	15.28	55.91	8.47	56.59	10.08	43.19	4.72	37.01	1.50	40.32	3.38	52.23	6.54	33.51	0.97
Doubao-Seed-2.0-Lite	57.95	11.58	60.23	13.41	53.27	6.86	54.57	8.69	39.58	3.43	33.53	0.97	37.69	2.63	50.39	5.63	33.43	0.97
Open-weight omni-modal models																		
Qwen3-Omni-30B	49.36	6.17	51.02	6.97	44.36	3.16	46.12	4.45	34.92	2.20	31.10	0.70	33.47	1.66	42.05	2.63	32.96	0.91
MiniCPM-o 4.5	45.28	4.34	47.06	5.04	40.91	2.25	42.18	3.11	34.31	2.04	30.02	0.59	32.48	1.50	38.44	1.82	33.27	0.97
Qwen2.5-Omni-7B	41.12	3.00	42.68	3.43	36.98	1.45	38.04	2.04	33.22	1.82	28.76	0.48	31.18	1.29	35.08	1.23	32.74	0.91
OmniVinci-9B	39.84	2.63	41.31	3.00	35.73	1.29	36.51	1.72	32.98	1.77	28.12	0.48	30.44	1.18	33.94	1.07	33.09	0.91
OLA-7B	38.06	2.20	39.62	2.57	34.08	1.02	34.92	1.45	32.42	1.66	27.16	0.38	29.51	1.07	32.78	0.91	32.61	0.86
Qwen2.5-Omni-3B	35.74	1.72	37.18	1.98	31.92	0.80	32.78	1.13	31.88	1.55	25.83	0.32	28.21	0.91	33.62	1.02	33.34	0.97
Baichuan-Omni-1.5	34.92	1.55	36.28	1.82	30.88	0.70	31.83	1.02	31.42	1.50	25.16	0.27	27.72	0.86	32.71	0.91	32.38	0.86

Table 1: Performance across complete and factorial input conditions. Acc_q is exact answer set accuracy, and Acc_g requires every question in an evidence group to be correct; all values are percentages. I, A, Q, and O denote image, audio, question stem, and answer options with the response instruction. Unparseable responses are incorrect. Clinicians are evaluated only with complete evidence under direct prompting.

correctly, while substantially fewer evidence groups have every related question answered correctly. Group accuracy captures this stricter all-correct requirement: a single incorrect answer makes the evidence group incorrect.

3.4 Contribution of Visual and Acoustic Evidence

Figure 4 separates complete input accuracy from sensitivity to visual and acoustic evidence. A point farther to the right or upward loses more accuracy when the corresponding input is removed. A point near the lower left is therefore not necessarily more robust, because marker area also reflects its complete input accuracy. Gemini 3.8 Flash and Qwen3.8-Omni-Flash combine the strongest overall performance with substantial use of both inputs.

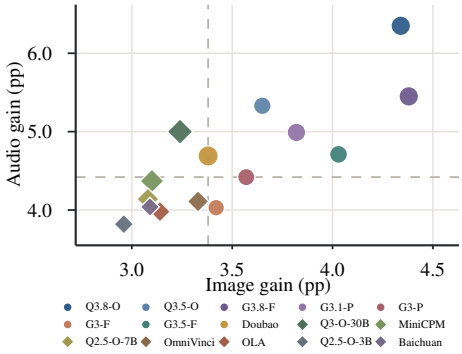


Figure 4: Conditional ablation effects of image and audio. Coordinates show the decrease in Acc_q when the corresponding modality is removed; marker area encodes complete input Acc_q.

Across the model suite, removing audio causes a larger accuracy decrease than removing the image when the question and options are held fixed. This comparison does not rank the intrinsic information content of the two modalities; each ablation measures the conditional contribution of one source in the presence of the remaining inputs. The result confirms that performance cannot be explained by visual question answering alone. Its consistency across proprietary models and models with open weights further shows that higher overall accuracy does not remove the same pattern of modality dependence.

The factorial conditions further separate evidence use from language priors. For most models, removing the question stem has the largest effect, while the question-only and options-only conditions remain substantially above the uniform-subset chance baseline. This indicates strong language and answer-structure priors in the benchmark. Image and audio nevertheless provide additional conditional gains over the question-and-options condition. Accordingly, complete-input accuracy should be interpreted together with the factorial controls rather than as evidence of clinical grounding by itself.

Table 2 tests whether the contribution of audio can be recovered from text alone. For speech, a transcript preserves lexical content but removes delivery cues carried by the waveform. For physiological signals, an expert description summarizes recognizable findings but cannot retain every property of the original signal. Across both settings, combining the recording with its textual representation performs better than either representation alone. Raw audio therefore contributes information that is lost during textualization, although the lower performance on physiological signals alone also shows that current models have difficulty representing acoustic patterns directly.

	Clinical speech								Physiological acoustic signals							
	Waveform		Verified transcript		Combined		Acoustic gain		Signal		Expert description		Combined		Acoustic gain	
Model	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g	Δ_q	Δ_g	Acc_q	Acc_g	Acc_q	Acc_g	Acc_q	Acc_g	Δ_q	Δ_g
Commercial and API-access models																
Qwen3.8-Omni-Flash	61.84	14.96	60.93	14.10	64.21	17.37	+3.28	+3.27	59.19	12.60	62.23	15.34	64.66	17.86	+2.43	+2.52
Qwen3.5-Omni-Plus	58.90	12.33	57.16	10.99	61.44	14.58	+4.28	+3.59	55.35	9.65	57.79	11.47	60.24	13.51	+2.45	+2.04
Gemini 3.8 Flash	62.76	15.87	61.29	14.48	65.57	18.87	+4.28	+4.40	60.40	13.62	62.25	15.39	64.94	18.18	+2.69	+2.79
Gemini 3.1 Pro	60.23	13.51	59.22	12.60	62.78	15.87	+3.57	+3.27	56.56	10.51	58.70	12.17	63.18	16.30	+4.48	+4.13
Gemini 3 Pro	57.33	11.10	56.82	10.72	61.14	14.32	+4.32	+3.59	54.58	9.12	56.16	10.24	60.40	13.62	+4.24	+3.38
Gemini 3 Flash	56.04	10.13	54.08	8.79	57.62	11.31	+3.54	+2.52	52.48	7.83	55.31	9.65	58.50	12.01	+3.19	+2.36
Gemini 3.5 Flash	58.96	12.39	59.10	12.49	62.58	15.71	+3.48	+3.22	56.49	10.46	58.44	11.96	62.23	15.34	+3.80	+3.38
Doubao-Seed-2.0-Lite	57.19	11.20	56.40	10.45	58.97	12.85	+2.57	+2.40	54.74	9.15	56.82	10.88	58.00	11.95	+1.18	+1.07
Open-weight omni-modal models																
Qwen3-Omni-30B	48.17	5.58	47.34	5.20	50.32	6.65	+2.98	+1.45	44.43	4.08	46.94	5.04	50.05	6.49	+3.12	+1.45
MiniCPM-o 4.5	43.36	3.70	43.41	3.70	46.46	4.83	+3.05	+1.13	39.54	2.57	42.95	3.54	46.71	4.93	+3.76	+1.39
Qwen2.5-Omni-7B	39.39	2.52	38.71	2.36	42.88	3.54	+4.17	+1.18	36.08	1.77	37.98	2.20	42.70	3.49	+4.71	+1.29
OmniVinci-9B	38.54	2.31	36.76	1.93	41.55	3.11	+4.79	+1.18	34.74	1.55	37.36	2.04	41.19	3.00	+3.83	+0.97
OLA-7B	36.15	1.77	36.25	1.82	38.89	2.41	+2.64	+0.59	33.24	1.29	36.18	1.82	38.68	2.36	+2.50	+0.54
Qwen2.5-Omni-3B	34.63	1.50	33.41	1.34	37.41	2.04	+3.99	+0.70	31.72	1.07	33.92	1.39	36.42	1.82	+2.50	+0.43
Baichuan-Omni-1.5	34.06	1.39	32.69	1.18	35.54	1.66	+2.85	+0.48	30.71	0.97	32.47	1.18	36.26	1.82	+3.80	+0.64

Table 2: Original acoustic inputs and textual representations. Acc_q and Acc_g are reported on matched questions. Acoustic gain is the combined representation score minus the corresponding textual representation score; all values are percentages or percentage points.

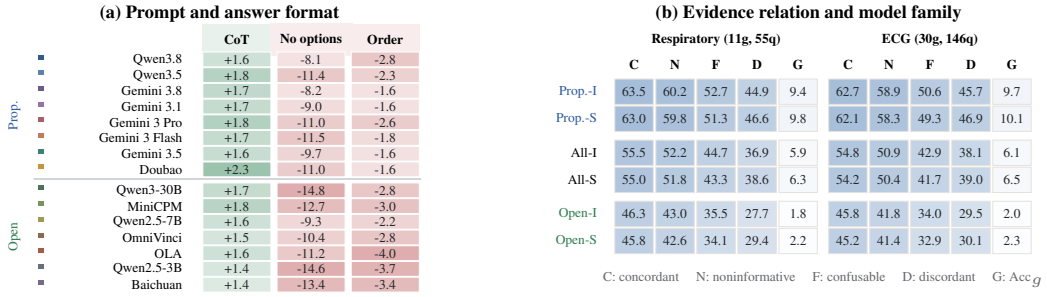


Figure 5: Sensitivity to prompting, answer format, and evidence relations. (a) Changes in Acc_q for each model under reasoning, option removal, and option permutation; color intensity is normalized within each column. (b) Mean performance by acoustic evidence relation and model group. I and S denote evidence invariant and evidence sensitive questions; values are percentages or percentage points.

3.5 Robustness and Diagnostic Analyses

Prompting and answer format. Figure 5 shows that a reasoning prompt yields only a modest improvement, whereas removing answer options causes a much larger loss (Tables 4 and 5). This limited gain indicates that explicit rationalization does not resolve the underlying errors. Candidate answers constrain the prediction space and provide semantic cues about the target decision. Option permutations reveal an additional weakness: average accuracy changes moderately, yet many individual questions are not answered correctly under every ordering (Table 6). Performance therefore depends on both option semantics and answer presentation.

Joint correctness across related questions. Question accuracy changes little as an evidence group contains more questions, but group accuracy declines sharply (Figure 8). The additional questions are therefore not uniformly harder; they create more opportunities for at least one incorrect answer within an evidence group. Part of the decline is mathematical because group accuracy is conjunctive, but this is also the clinical property being tested: a shared case representation should support mutually consistent decisions across questions. Most benchmark instances occur in groups of five or six questions, making this pattern representative of the benchmark distribution.

Clinical and structural variation. Performance differs across visual modalities and specialties (Figures 12 and 13). Answer structure also affects performance: questions with several correct answers are harder than those with one, and accuracy declines as the option set grows (Table 7). Exact set matching requires models to distinguish incomplete answers from unsupported additions, exposing calibration failures that conventional single answer accuracy does not capture. The relatively strong results on management questions and weaker results on finding identification further suggest that models use general clinical priors more effectively than precise perceptual evidence (Figure 9).

Acoustic evidence and conflict. Detailed breakdowns likewise show gains when raw recordings are added to their text proxies across speech and physiological signals (Figures 14 and 15). Accuracy deteriorates as respiratory and ECG evidence moves from concordant to noninformative, confusable, and discordant (Figure 16). Complete neglect of audio therefore cannot explain the observed behavior. Instead, performance depends on whether a model can determine when the recording should confirm, revise, or leave unchanged the conclusion supported by the visual case. The same trend in proprietary models and models with open weights indicates a shared difficulty in assessing the relevance of acoustic evidence. Manual error review supports this interpretation: visual shortcuts, acoustic shortcuts, integration failures, and incorrect responses to changing evidence dominate the error taxonomy (Table 8). These findings locate the principal source of error in clinical evidence integration rather than output formatting.

4 Related Work

Medical visual question answering and evaluation by clinical case. VQA-RAD and SLAKE established medical image question answering with clinically oriented questions [11, 14]. OmniMedVQA and GMAI-MMBench broadened coverage across imaging modalities and clinical tasks [3, 10]. These benchmarks primarily score questions independently. More recent benchmarks incorporate clinical context: LiveClin follows a clinical pathway, MedThinkVQA reasons across multiple images, and MEDSYN evaluates synthesis from multiple visual inputs [2, 24, 28]. They motivate evaluation at the case level, but to our knowledge do not jointly assess a medical image and raw clinical recording through several questions grounded in the same evidence.

Medical audio and omni-modal evaluation. CaReAQA and StethoLM evaluate reasoning over cardiac and respiratory sounds, while MedMosaic covers a broader range of medical recordings [17, 23, 25]. Audio is generally evaluated alone or with text, leaving its clinical relation to a matched image outside the task. In the general domain, OmniBench evaluates joint visual, acoustic, and textual inputs [13]. It evaluates general omni-modal performance, but is not designed to assess clinical alignment or consistency across questions. In MedSensus, the relation between an image and a recording is subject to expert review and remains fixed within each evidence group.

Interactions between modalities and diagnostic evaluation. MMMC studies disagreement between visual and textual inputs, MCR-Bench shows that audio and language models can favor text when it conflicts with audio, and misleading text can redirect medical image interpretation [1, 22, 30]. These studies isolate modality dominance, typically by perturbing one source for one prediction. More broadly, benchmark families differ not only in their modalities, but also in their evaluation unit and diagnostic scope. Independent questions measure local correctness, resources organized by case add clinical context, and conflict studies isolate a specific interaction. Few benchmarks combine clinically reviewed image and recording relations with repeated questions, multiple correct options, and exact answer set scoring. MedSensus holds an image and recording fixed across related questions and combines group accuracy with factorial ablations, acoustic representation controls, and controlled evidence relations.

5 Conclusion

MedSensus evaluates whether omni-modal models can integrate visual and acoustic clinical evidence while maintaining correctness across related questions. By organizing clinically reviewed evidence into groups, it compares individual accuracy with consistency under an unchanged clinical context. Current models benefit from both modalities and from acoustic information beyond text, but remain sensitive to answer presentation and conflicting evidence. Their central limitation is therefore not access to heterogeneous inputs, but reliable assessment of evidence relevance and reconciliation of clinical implications. The diagnostic evaluations distinguish limitations in modality representation from failures to select and reconcile relevant evidence. The remaining gap to clinicians shows that reliable case consistency has not yet been achieved. MedSensus provides a controlled framework for identifying these failures and guiding the development of more reliable medical reasoning systems.

AI USE STATEMENT

Generative AI tools were used to check question wording, identify potential ambiguity, and edit the manuscript. Clinical experts authored all questions and answer options and determined all clinical targets, reference answers, inclusion decisions, and compatibility judgments. The authors reviewed and verified all text produced with AI assistance and take full responsibility for the final content of this work.

ETHICS STATEMENT

MedSensus is constructed from publicly accessible medical datasets and repositories. Source provenance is retained for each image, recording, and derived benchmark item, and any release or redistribution will follow the licenses and terms of the original resources. The clinician reference was provided by three healthcare professionals on the research team, and only aggregate results are reported. MedSensus is intended for research evaluation, not for diagnosis, treatment selection, or other clinical decisions. Its case distribution and annotations may inherit limitations or biases from the underlying public sources. Synthetic speech and controlled evidence pairings are experimental inputs and should not be treated as records of real clinical encounters. Performance on this benchmark therefore does not establish that a model is safe or ready for clinical use.

REPRODUCIBILITY STATEMENT

Sections 2 and 3 specify benchmark construction, evaluation conditions, metrics, and inference procedures. The appendix reports confidence intervals, subgroup results, controlled evidence analyses, and the manual error review. We maintain a frozen evaluation manifest with stable case, image, audio, question, and evidence group identifiers. The manifest records provenance, experimental eligibility, reference answer sets, and the input assigned to every condition. The prediction ledger additionally stores the prompt setting, model or API version, decoding configuration, parsed answer, and scoring status. Subject to source licenses and platform terms, we will release the benchmark annotations, evaluation code, prompts, manifests, and model configuration records needed to reproduce and extend the evaluation.

REFERENCES

[1] Thomas A Buckley, James A Diao, Cam N Srivastava, Peter G Brodeur, Pranav Rajpurkar, Adam Rodman, and Arjun K Manrai. Multimodal foundation models exploit text to make medical image predictions. *Nature Communications*, 2026.

[2] Boqi Chen, Xudong Liu, Jiachuan Peng, Marianne Frey-Marti, Kyle Lam, Bang Zheng, Lin Li, and Jianing Qiu. Medsyn: Benchmarking multi-evidence synthesis in complex clinical cases for multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23631–23655, 2026.

[3] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

[4] Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

[5] Google DeepMind. Gemini 3 flash model card. Google DeepMind Model Card, December 2025. URL https://deepmind.google/models/model-cards/gemini-3-flash/. Accessed 2026-09-21.

[6] Google DeepMind. Gemini 3 pro model card. Google DeepMind Model Card, November 2025. URL https://deepmind.google/models/model-cards/gemini-3-pro/. Accessed 2026-09-21.

[7] Google DeepMind. Gemini 3.1 pro model card. Google DeepMind Model Card, February 2026. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed 2026-09-21.

[8] Google DeepMind. Gemini 3.5 flash model card. Google DeepMind Model Card, May 2026. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/. Accessed 2026-09-21.

[9] Google DeepMind. Gemini 3.8 flash model card. Google DeepMind Model Card, September 2026. URL https://deepmind.google/models/model-cards/gemini-3-8-flash/. Published 2 September 2026; accessed 26 September 2026.

[10] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183. IEEE, 2024.

[11] 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.

[12] Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025.

[13] Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Hangyu Guo, Yiming Liang, Jiaheng Liu, Noah Wang, Jian Yang, Siwei Wu, et al. Omnibench: Towards the future of universal omni-language models. *Advances in Neural Information Processing Systems*, 38, 2026.

[14] 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.

[15] Zuyan Liu, Yuhao Dong, Jiahui Wang, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Ola: Pushing the frontiers of omni-modal language model. *arXiv preprint arXiv:2502.04328*, 2025.

[16] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[17] 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. *arXiv preprint arXiv:2605.00969*, 2026.

[18] Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity, 2026. URL https://arxiv.org/abs/2607.00248.

[19] Qwen Team. Qwen3.8-omni: Towards native omni-modal agents, 2026. URL https://arxiv.org/abs/2609.25611.

[20] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.

[21] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

[22] Cheng Wang, Gelei Deng, Xianglin Yang, Han Qiu, and Tianwei Zhang. When audio and text disagree: Revealing text bias in large audio-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4878–4888, 2025.

[23] Tsai-Ning Wang, Lin-Lin Chen, Neil Zeghidour, and Aaqib Saeed. Careaqa: A cardiac and respiratory audio question answering model for open-ended diagnostic reasoning. *arXiv preprint arXiv:2505.01199*, 2025.

[24] Xidong Wang, Yue Shen, Junying Chen, Jian Wang, Jinjie Gu, Ping Zhang, Lei Liu, Wang Benyou, et al. Liveclin: A live clinical benchmark without leakage. In *International Conference on Learning Representations*, volume 2026, pp. 97981–98011, 2026.

[25] Yishan Wang, Tsai-Ning Wang, Mathias Funk, and Aaqib Saeed. Stetholm: Audio language model for cardiopulmonary analysis across clinical tasks. *arXiv preprint arXiv:2603.00355*, 2026.

[26] 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, 2025. URL https://arxiv.org/abs/2503.20215.

[27] 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*, 2025.

[28] Zonghai Yao, Benlu Wang, Yifan Zhang, Junda Wang, Iris Xia, Zhipeng Tang, Shuo Han, Zhichao Yang, Arman Cohan, and Hong Yu. Medical thinking with multiple images. In *International Conference on Learning Representations*, volume 2026, pp. 101924–101988, 2026.

[29] Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Zhen Wan, Jinchuan Tian, An-Chieh Cheng, Ligeng Zhu, Yuanhang Su, Yuming Lou, et al. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. In *International Conference on Learning Representations*, volume 2026, pp. 56101–56138, 2026.

[30] Zongmeng Zhang, Wengang Zhou, Jie Zhao, and Houqiang Li. Robust multimodal large language models against modality conflict. *arXiv preprint arXiv:2507.07151*, 2025.

A Core Results and Reliability Checks

This section expands the main benchmark results and checks whether they remain stable under confidence intervals, clinician comparison, prompt changes, open responses, and option reordering. Acc_q is question accuracy and requires the predicted answer set to match the reference answer exactly. Acc_g is group accuracy and is reported only when all questions in the original group remain together.

Detailed model tables use compact column labels to keep their text readable. The labels follow the model order used in the main benchmark table.

A.1 Complete Results and Confidence Intervals

The main table reports every model score, but its dense layout makes the overall pattern and the uncertainty around each result difficult to see. Figure 6 therefore summarizes all eight input conditions by model family. Figure 7 then shows the complete-input 95% confidence interval for every model. Together, the two figures preserve the detailed comparison without repeating another large model matrix.

Every model performs best with the complete input. The commercial and API access models are also ahead of the open models in almost every condition. Removing the question causes the largest drop, and removing the image or audio also lowers accuracy. Group accuracy falls much more because one wrong answer makes the whole evidence group incorrect. The result shows the main challenge in MedSensus: models can answer some questions from the evidence, but they do not stay correct across all related questions.

The ordering of the input conditions is also stable across model families. The question stem carries the largest single share of useful information, but the complete input remains best. This means that text cues alone do not explain the final score. Images and recordings add smaller gains on average, yet those gains matter more at the group level because one missed clue can break an otherwise correct set of answers. The confidence intervals support the same conclusion: the gap between individual success and complete group success is a benchmark wide pattern rather than the result of a few unusual groups.

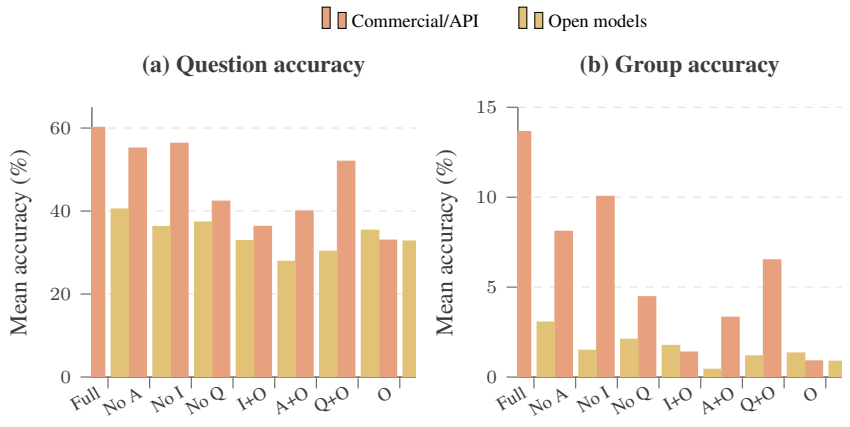


Figure 6: Mean performance across input conditions. Coral bars denote commercial and API-access models, and gold bars denote open-weight models. Values are family means under each input condition. I, A, Q, and O denote image, audio, question stem, and answer options with the fixed response instruction, respectively.

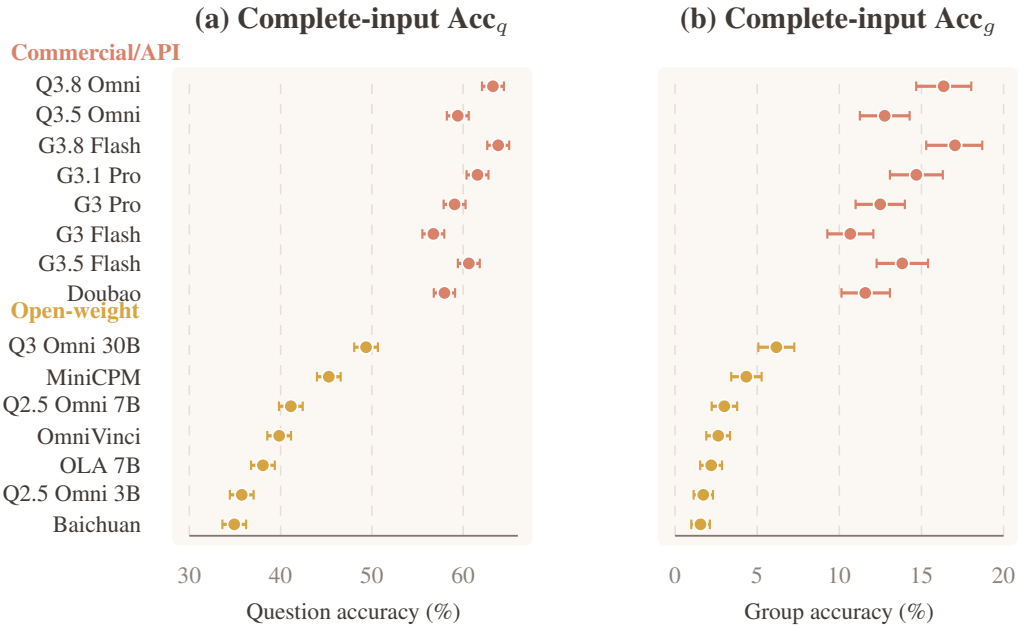


Figure 7: Complete-input accuracy with 95% confidence intervals. Points show model-level estimates under complete image–audio–question input, and horizontal lines show 95% confidence intervals obtained by evidence-group bootstrap resampling. Coral denotes commercial and API-access models, while gold denotes open-weight models.

A.2 Clinician Reference

A single clinician row in the main table does not show how the three evaluators performed. Table 3 adds their question accuracy and response time, then reports the clinician reference used in the main comparison. This makes the human reference easier to interpret.

All three clinicians reach high question accuracy, and their scores are close despite their different specialties. The clinician group accuracy is lower than question accuracy because every question in an evidence group must be correct, but it is still far above the best model result. The gap shows that the cases can be answered from the supplied evidence. The main difficulty for current models is staying correct across several questions about that evidence.

(a) Individual clinician results

Evaluator	Clinical role	Experience (years)	Completed instances	Acc_q 95% CI	Median time (s/query)
Clinician 1	Attending Radiologist	12	9,539	89.15% [87.2, 91.0]	18.4
Clinician 2	Senior Pulmonologist	8	9,539	92.35% [90.6, 94.1]	21.2
Clinician 3	Chief Pathologist	15	9,539	91.86% [90.0, 93.5]	24.5
Mean	Healthcare professionals	11.7	9,539	91.12% [89.3, 92.8]	21.4

(b) Clinician reference used in the main comparison

Reference	Acc_q	Acc_g	Fleiss' κ
Clinician reference	91.12%	72.97%	0.74

Table 3: Clinician performance on the complete benchmark. Panel (a) reports individual question accuracy and response time. Panel (b) reports the aggregate clinician reference used in the main table.

The agreement across three clinicians is important for interpreting this gap. They come from different specialties, yet all three perform well on the same set of cases. This makes it less likely that the result depends on one narrow clinical background. Their much higher group score also shows that the strict group metric is demanding but achievable. MedSensus is therefore measuring a real weakness in current models, not simply creating groups that no evaluator can complete.

A.3 Effect of Prompting

The main table includes direct and reasoning prompts, but the changes for each model are hard to compare in that wide layout. Table 4 puts the two prompts side by side while keeping the image, audio, question, and options unchanged. The table therefore measures the effect of the prompt rather than a change in evidence.

	Question accuracy (Acc_q)			Group accuracy (Acc_g)		
Model	**Direct**	**CoT**	Δ_q	**Direct**	**CoT**	Δ_g
Commercial and API-access models						
Qwen3.8-Omni-Flash	63.26	64.83	+1.57	16.35	17.96	+1.61
Qwen3.5-Omni-Plus	59.41	61.21	+1.80	12.76	14.32	+1.55
Gemini 3.8 Flash	63.84	65.52	+1.68	17.04	18.77	+1.73
Gemini 3.1 Pro	61.57	63.22	+1.65	14.69	16.25	+1.55
Gemini 3 Pro	59.06	60.89	+1.83	12.49	14.05	+1.55
Gemini 3 Flash	56.73	58.38	+1.65	10.67	11.85	+1.18
Gemini 3.5 Flash	60.62	62.24	+1.62	13.83	15.28	+1.45
Doubao-Seed-2.0-Lite	57.95	60.23	+2.28	11.58	13.41	+1.83
Open-weight omni-modal models						
Qwen3-Omni-30B	49.36	51.02	+1.66	6.17	6.97	+0.80
MiniCPM-o 4.5	45.28	47.06	+1.78	4.34	5.04	+0.70
Qwen2.5-Omni-7B	41.12	42.68	+1.56	3.00	3.43	+0.43
OmniVinci-9B	39.84	41.31	+1.47	2.63	3.00	+0.38
OLA-7B	38.06	39.62	+1.56	2.20	2.57	+0.38
Qwen2.5-Omni-3B	35.74	37.18	+1.44	1.72	1.98	+0.27
Baichuan-Omni-1.5	34.92	36.28	+1.36	1.55	1.82	+0.27

Table 4: Prompting sensitivity under complete image–audio–question evidence. Δ is CoT minus direct prompting in percentage points and is calculated before display rounding; only the final parsed answer set is scored.

Reasoning prompts give every model a small gain. The gain in question accuracy is similar across models, while stronger models gain more in group accuracy. Asking a model to reason may help it arrange the available information before answering. It does not solve errors in reading the image or audio. MedSensus therefore uses the same evidence under both prompts and treats prompting as a separate source of improvement.

The improvement is useful but modest compared with the loss caused by removing clinical evidence. Better prompting can help a model organize information that it has already read, but it cannot recover an image finding or an acoustic cue that the model failed to understand. The similar gains across systems also suggest that the main ranking is not an artifact of one prompt style. For this reason, the direct prompt remains the clearest basis for the controlled input comparisons in the main experiment.

A.4 Open Ended Validation

Multiple choice questions allow exact scoring, but the options can also help a model find the answer. Table 5 removes the options from a fixed subset while keeping the image, audio, and question. Expert written rules separate correct, partly correct, unsupported, and unscorable answers.

		Multiple choice	Open response result (%)				Accuracy decrease
Model	**Instances (n)**	**Acc$_q$**	**Correct**	**Partial**	**Incorrect / unsupported**	**Unscorable**	**(pp)**
Commercial and API-access models							
Qwen3.8-Omni-Flash	1,200	63.44	55.32	7.10	37.16	0.42	8.13
Qwen3.5-Omni-Plus	1,200	59.72	48.33	9.57	41.81	0.29	11.38
Gemini 3.8 Flash	1,200	63.74	55.59	6.87	36.83	0.71	8.15
Gemini 3.1 Pro	1,200	60.95	51.96	8.60	38.70	0.74	8.99
Gemini 3 Pro	1,200	58.70	47.75	9.24	42.81	0.21	10.95
Gemini 3 Flash	1,200	57.13	45.64	7.36	46.65	0.36	11.49
Gemini 3.5 Flash	1,200	61.21	51.53	6.37	41.80	0.30	9.68
Doubao-Seed-2.0-Lite	1,200	58.40	47.38	9.23	42.46	0.93	11.02
Open-weight omni-modal models							
Qwen3-Omni-30B	1,200	49.41	34.57	6.51	57.59	1.33	14.84
MiniCPM-o 4.5	1,200	45.71	33.00	8.45	57.19	1.37	12.71
Qwen2.5-Omni-7B	1,200	41.39	32.11	5.91	61.04	0.93	9.27
OmniVinci-9B	1,200	39.29	28.90	5.40	64.78	0.92	10.40
OLA-7B	1,200	38.24	27.05	6.48	65.66	0.81	11.19
Qwen2.5-Omni-3B	1,200	35.44	20.82	7.59	70.18	1.41	14.62
Baichuan-Omni-1.5	1,200	34.49	21.12	5.65	72.16	1.07	13.37

Table 5: Open ended validation on a fixed subset of 1,200 instances. The multiple choice Acc$_q$ is computed on the same subset and may differ from the full benchmark score. The image, audio, and question are retained while the options are removed. Experts score the responses using written answer rules.

Every model performs worse without the options. Stronger models keep more of their multiple choice performance, while several open models produce more partial or unsupported answers. Unscorable answers are rare, so most of the drop comes from answers that are not precise enough. The options clearly help, but they do not remove the need to understand the case. This is why MedSensus checks both exact option selection and answers written without options.

The pattern also helps separate recognition from answer production. Partial answers show that a model sometimes identifies part of the clinical finding but cannot state the complete supported answer. Incorrect and unsupported answers are more common than unscorable responses, so the decrease is not mainly a formatting problem. The stronger systems remain stronger after the options are removed, which means that the model ranking is not created only by multiple choice cues.

A.5 Answer Option Order Robustness

The open response test removes the options, but it does not show whether their order affects the answer. We repeat the complete input evaluation after changing the option order and map each prediction back to the original option. Table 6 reports accuracy and answer consistency across the different orders.

Model	Instances (n)	Original order Acc_q	Shuffled order mean Acc_q	Change (pp)	Answer consistency	Correct under all orders
Commercial and API-access models						
Qwen3.8-Omni-Flash	9,539	63.26	60.48	-2.78	92.84%	53.81%
Qwen3.5-Omni-Plus	9,539	59.41	57.11	-2.30	94.06%	49.79%
Gemini 3.8 Flash	9,539	63.84	62.27	-1.57	89.19%	56.33%
Gemini 3.1 Pro	9,539	61.57	59.94	-1.63	90.27%	52.11%
Gemini 3 Pro	9,539	59.06	56.46	-2.60	90.89%	49.49%
Gemini 3 Flash	9,539	56.73	54.90	-1.83	94.49%	48.52%
Gemini 3.5 Flash	9,539	60.62	59.00	-1.62	90.48%	52.31%
Doubao-Seed-2.0-Lite	9,539	57.95	56.33	-1.62	92.51%	48.64%
Open-weight omni-modal models						
Qwen3-Omni-30B	9,539	49.36	46.56	-2.80	85.54%	36.57%
MiniCPM-o 4.5	9,539	45.28	42.26	-3.02	84.64%	36.07%
Qwen2.5-Omni-7B	9,539	41.12	38.90	-2.22	88.39%	29.73%
OmniVinci-9B	9,539	39.84	37.00	-2.84	84.44%	29.47%
OLA-7B	9,539	38.06	34.07	-3.99	87.70%	24.18%
Qwen2.5-Omni-3B	9,539	35.74	32.02	-3.72	84.08%	23.14%
Baichuan-Omni-1.5	9,539	34.92	31.56	-3.36	87.76%	24.49%

Table 6: Results under three fixed option orders. Predicted letters are mapped back to the original options before scoring. Responses without a clear answer count as inconsistent.

Changing the option order lowers accuracy for every model. The drop is usually smaller for the commercial and API access models. Many predictions keep the same meaning across orders, but far fewer questions are answered correctly in every order. A reliable answer should follow the clinical meaning of an option, not its place in the list. This test shows that option order remains a source of error in MedSensus.

Answer consistency and accuracy capture different failures. A model can give the same wrong answer under every order, or change between options while remaining correct only once. Reporting both measures avoids treating stable errors as robust reasoning. The lower all order accuracy of the open models shows that their answers are more easily affected by surface changes that do not alter the clinical content.

B Benchmark Structure and Clinical Breakdowns

The main results pool all questions and evidence groups. This section opens that aggregate score and examines how performance changes with group size, reasoning task, recording type, disease category, visual modality, clinical specialty, and answer format.

B.1 Group Structure and Benchmark Composition

Group accuracy requires every question in an evidence group to be correct, so groups with more questions are harder to complete. The main score does not show this effect. Figure 8 places accuracy and group composition side by side, making the effect of group size visible without adding another dense table.

Question accuracy changes little across the larger group sizes, but group accuracy falls as more answers must be correct together. Most instances come from groups with five or six questions, and most case groups contain a moderate number of instances. The low group scores are partly caused by

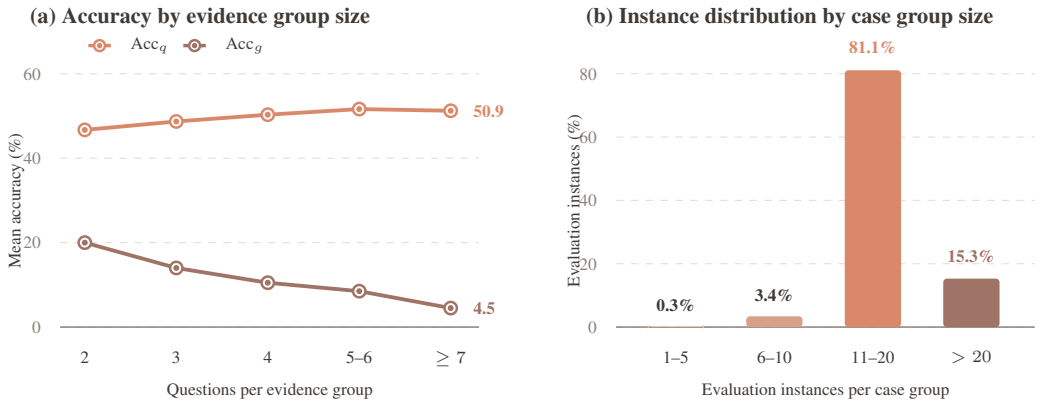


Figure 8: Group structure of MedSensus. Panel (a) shows mean question accuracy and group accuracy under complete input as a function of the number of questions in each evidence group. Panel (b) shows the distribution of evaluation instances across case-group sizes.

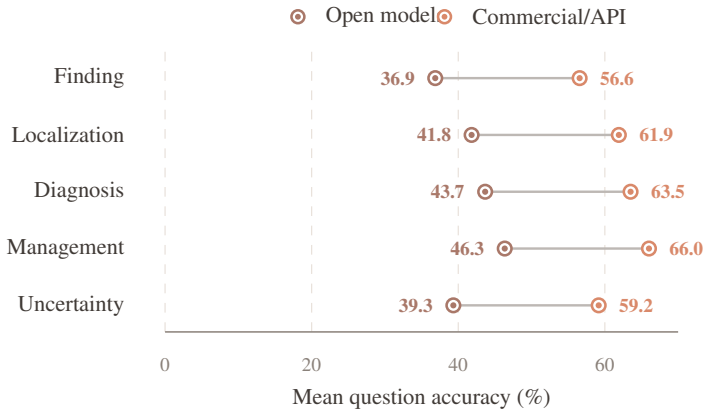


Figure 9: Mean complete-input question accuracy by clinical reasoning task. Each row compares commercial/API and open-weight model means for one clinical reasoning task. Task-level case-group counts are not additive because one group can contain more than one question type.

the strict scoring rule, but question accuracy does not fall in the same way. This is why MedSensus reports both measures: one tests individual answers, and the other tests whether those answers stay correct under shared evidence.

The two curves also rule out a simpler explanation for the low group score. If larger groups merely contained harder questions, question accuracy would fall sharply with group size. Instead, it remains broadly stable while group accuracy declines. The decline therefore reflects the increasingly strict all-correct requirement of group accuracy: a model may answer individual questions correctly while failing to answer every question in the group correctly.

B.2 Clinical Reasoning Task Categories

Overall accuracy can hide important differences between clinical tasks. A model may recognize a finding but fail to locate it, diagnose it, or judge whether the evidence is enough. Figure 9 separates these tasks and compares the two model families without another dense model matrix.

Finding identification is the hardest task for most models, while clinical management receives the highest scores. Localization, diagnosis, and judgments about whether the evidence is enough fall between these two ends. Finding questions often require a precise reading of the image or audio. Management questions can sometimes also use general clinical knowledge. The commercial and

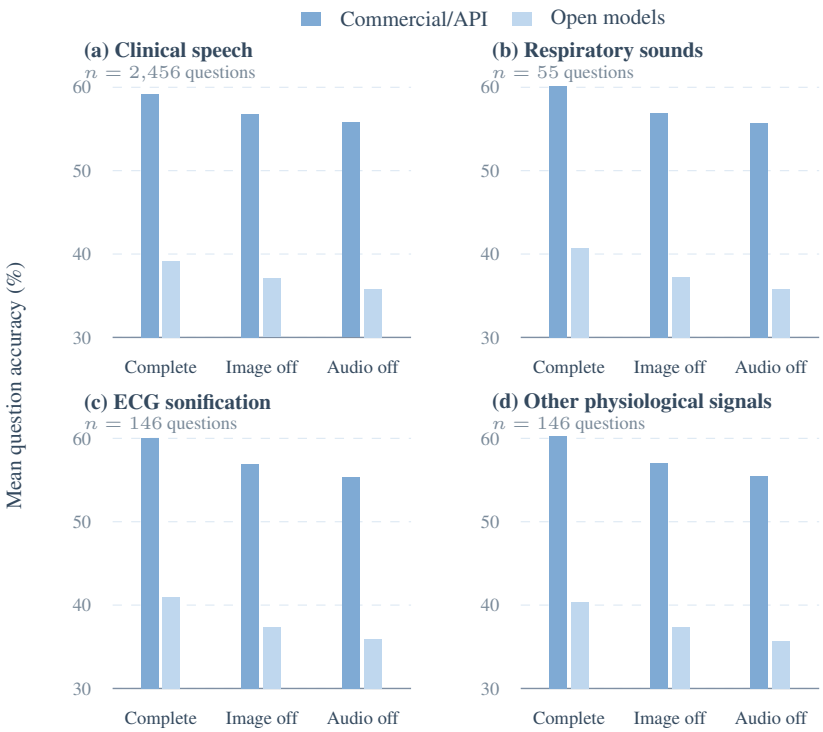


Figure 10: Image and audio removal by recording type. Each panel compares the commercial and API mean with the open model mean under complete input, image removal, and audio removal. Question counts can overlap because one text question may be paired with more than one recording type.

API access models remain ahead across all tasks, so the overall ranking is not caused by one type of question. MedSensus covers both the reading of clinical evidence and its use in later decisions.

The task ordering also shows why a single diagnosis score would be incomplete. A model can use familiar medical knowledge to choose a reasonable diagnosis or management step while still missing the exact finding that appears in the image or recording. MedSensus keeps these questions together because later clinical reasoning should be supported by the earlier evidence. A system that reaches the right decision for the wrong reason may look strong on one task but will not remain reliable across the full group.

B.3 Image and Audio Effects by Recording Type

The main ablation results show that image and audio both matter, but they place all recordings in one pool. Figure 10 repeats the test for clinical speech, respiratory sounds, electrocardiogram sonification, and other physiological signals. It shows whether the value of audio comes from one type of recording or appears across the benchmark. This separation also follows earlier medical audio evaluations, which show that speech, heart and lung sounds, and other recordings require different kinds of reasoning [17, 23, 25].

Complete input is best for every audio type and for both model groups. Removing audio usually causes a larger drop than removing the image, although the size of the drop changes across models and recordings. Speech gives the model words that it can process directly. Physiological sounds require the model to read a sound pattern, and the results vary more across systems. These differences show why MedSensus keeps several types of audio and tests each one with the same visual evidence.

The two removals answer different questions. The image removal test asks whether the recording can support the answer without its visual setting. The audio removal test asks whether the image and question already contain enough information. Their effects are not identical across recording types,

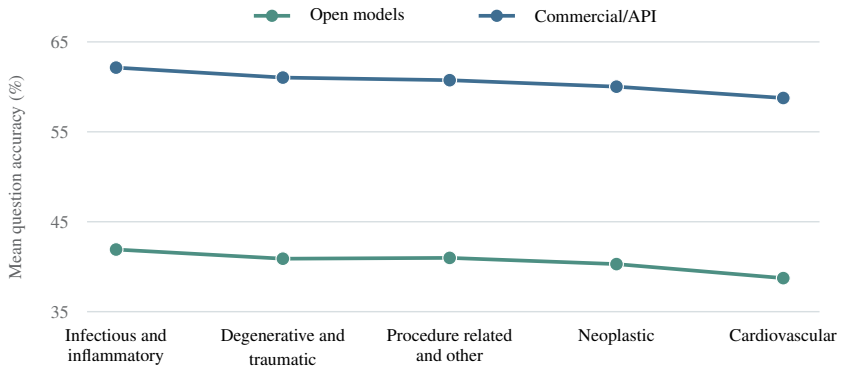


Figure 11: Mean question accuracy by clinical category under complete input. The two lines report the open model mean and the commercial and API mean on the same scale.

which shows that audio is not one uniform source. Speech often states clinical facts directly, while physiological sounds require pattern recognition. A benchmark that pooled them without separate analysis would hide this difference.

B.4 Clinical Category Breakdown

Pooling all cases can hide differences between clinical conditions. Figure 11 groups the complete input results by clinical category and compares the two model families directly. This view asks whether the overall difficulty comes from one group of conditions or appears across the benchmark.

Models perform best on infectious and inflammatory conditions and worst on cardiovascular disease. The other categories are closer together. Commercial and API access models remain ahead in every category. Cardiovascular cases in MedSensus include signals and image patterns that require careful reading, which may contribute to the lower scores. The category sizes are also different, so small gaps should not be overread. The broad pattern shows that the challenge is not limited to one disease group.

The distance between the two model families is also similar across the five categories. This suggests that the overall ranking is not created by one unusually easy category for the stronger systems. At the same time, the small differences between neighboring categories should be read as descriptive rather than as a clinical ordering of difficulty. The useful result is the shared pattern: no category removes the need to combine the supplied evidence.

B.5 Performance by Clinical Visual Modality

MedSensus contains radiology, microscopy, endoscopy, eye imaging, and electrocardiographic traces. One image ablation score cannot show whether a model reads all of these images equally well. Figure 12 places all visual modalities on one shared scale so that the broad pattern is visible without splitting the result across three tables.

Average performance is higher for pathology, optical coherence tomography, and magnetic resonance imaging. Electrocardiographic tracing is much harder for the model set as a whole. The same model leads on most image types, but a different model leads on electrocardiographic traces. Medical images use very different visual patterns, so skill on one type does not ensure skill on another. MedSensus pairs this range of images with audio and questions to test whether a model can use visual evidence across medical settings.

Most modalities form a fairly narrow middle range. This means that the overall visual result is not driven by a single easy image family. The clear exception is the electrocardiographic trace. Its lower average suggests that models have more difficulty with structured wave patterns than with many image based findings. This difference matters in MedSensus because the visual input is not just background context. It can change which part of the accompanying audio is relevant to the question.

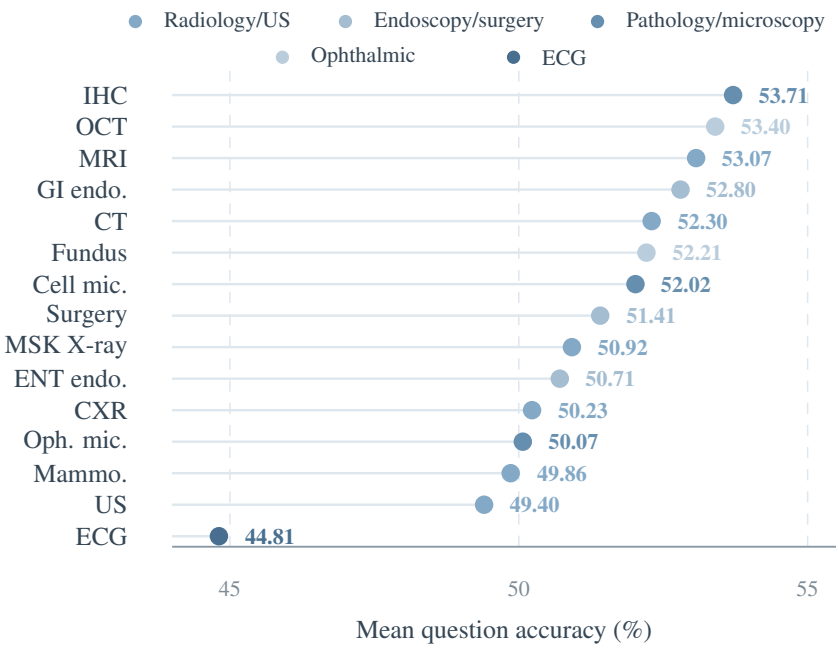


Figure 12: Mean question accuracy by visual modality. Modalities are ordered by complete-input question accuracy, and blue shades identify the broad visual family.

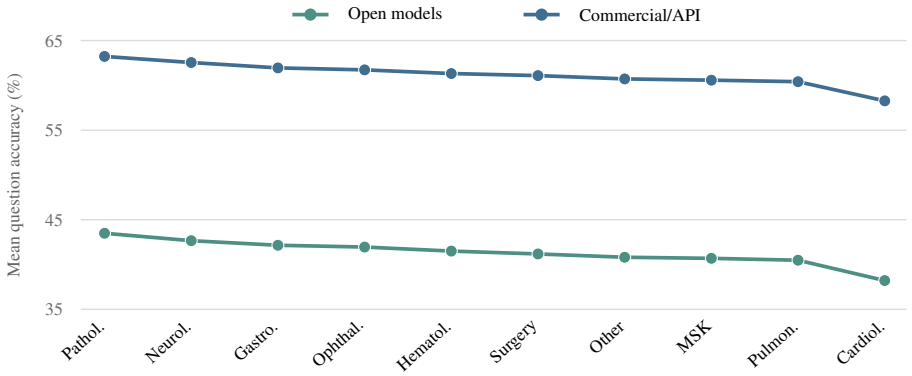


Figure 13: Mean question accuracy across clinical specialties. The two lines report the open model mean and the commercial and API mean for each specialty.

B.6 Clinical Specialty Breakdown

Visual modality and disease category describe the case in different ways. Clinical specialty gives another useful view because each specialty uses its own mix of images, sounds, and questions. Figure 13 compares complete input performance across these areas and makes the gap between model families easier to see.

Average performance is highest in pathology and neurology and lowest in cardiology. Commercial and API access models remain ahead of open models in every specialty. This agrees with the visual modality results, where electrocardiographic traces are also difficult. MedSensus covers many areas of medicine so that strong performance cannot come from one favored specialty.

The ranking is broadly shared by both model families. This agreement suggests that the lower cardiology result is linked to the evidence in those cases, rather than to one model family alone. Even so, the differences among most specialties are modest. The larger and more stable result is the gap

between commercial and open models, which remains visible in every specialty. A model therefore needs broad clinical evidence use, not strength in only one area, to perform well on MedSensus.

B.7 Performance by Answer Format

Exact answer set matching counts both a missing correct option and an extra wrong option as errors. The overall score does not show how this affects different question formats. Table 7 separates questions with one correct answer from questions with several correct answers. It also compares questions with different numbers of options.

(a) Number of correct answers

Answer format	n	All mean	Commerc mea	Open mean	Set F1	Best model	Best Acc_q
One correct answer	967	55.07	64.94	43.79	–	Gemini 3.8 Flash	69.15
Several correct answers	1,672	48.97	58.04	38.60	68.14	Gemini 3.8 Flash	60.88

(b) Number of answer options

Response options	n	All mean	Commercial/ mean	Open mean	Best model	Best Acc_q
3–4	1,178	53.91	63.70	42.71	Gemini 3.8 Flash	67.90
5	742	51.20	60.48	40.59	Gemini 3.8 Flash	64.16
6	561	48.69	57.74	38.34	Gemini 3.8 Flash	60.92
≥ 7	158	44.93	53.53	35.10	Gemini 3.8 Flash	56.84

Table 7: Performance by answer structure. Panel (a) compares questions with one versus several correct answers, and Panel (b) stratifies questions by the number of response options. Values under All, Commercial/API, and Open are mean question accuracy (Acc_q, %). Set F1 is additionally reported for questions with multiple correct answers.

Questions with one correct answer are easier for both model groups. Accuracy also falls as the number of options grows. When several answers are correct, the model must select every supported option and reject every wrong one. It may understand part of the case and still fail exact scoring. The higher set F1 shows that this happens often. MedSensus keeps these questions because clinical evidence can support more than one finding or action, and a reliable model must give the complete answer.

The gradual decline as more options are added is not only a chance effect. Larger option sets usually contain more clinically plausible alternatives, so the model must make finer distinctions between supported and unsupported choices. The gap between exact accuracy and set F1 shows that many errors are near misses rather than completely unrelated answers. Even so, a near miss can matter clinically when it adds an unsupported action or leaves out a required finding, which is why exact set accuracy remains the primary measure.

For physiological recordings, models usually perform better with an expert description than with the raw signal. Combining the signal and description is best. The description states the main finding in words, while the signal keeps details about timing, strength, and pattern. MedSensus includes both so that it can test language understanding, audio understanding, and their combined use.

The combined input improves every communication type and every physiological signal group. This is stronger evidence than a gain in only one audio subset. It suggests that text and sound are not simple replacements for one another: the text makes the main content explicit, while the recording retains details that are difficult to express fully in a short transcript or description. The remaining gap between raw signals and expert descriptions also shows that current models still benefit when a person first turns an acoustic pattern into clinical language.

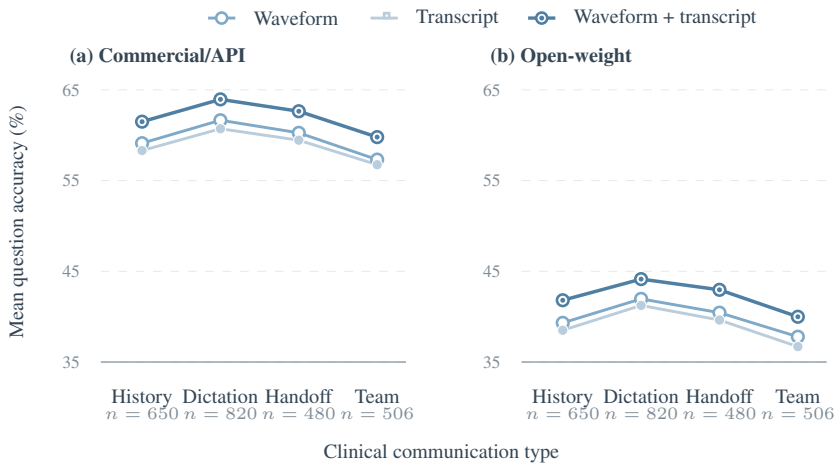


Figure 14: Clinical speech results by communication type. Each panel compares the original waveform, the verified transcript, and their combination. Lines report family-level mean question accuracy; counts show the number of eligible questions in each communication type.

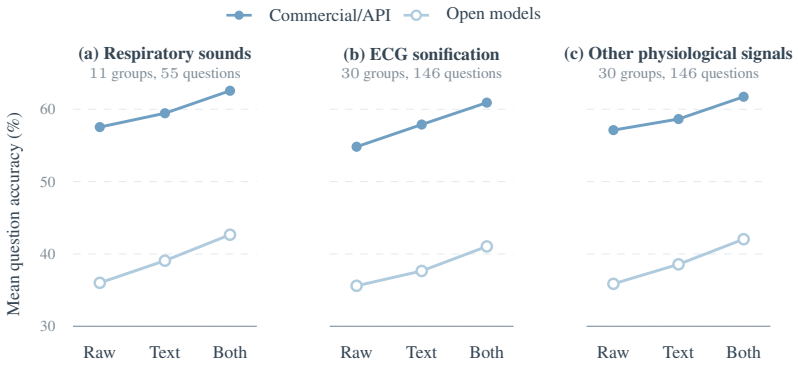


Figure 15: Physiological acoustic representation. Each panel compares the original signal, its expert description, and their combination. Lines report mean question accuracy for commercial/API and open-weight models.

B.8 Detailed Evidence Relation Analysis

An ablation shows whether a source helps, but it does not show whether the model knows when that source should be trusted. Figure 16 keeps the image fixed and changes how the audio relates to the case. Some questions should keep the same answer, while others should change when the audio changes. Prior studies show that conflicting modalities can shift a model toward the wrong source of evidence [1, 22, 30].

C Evidence Representation and Model Behavior

The previous section shows where performance changes across the benchmark. This section asks how models use the supplied evidence. It compares sound with text representations, changes the relation between image and audio, and links the resulting score changes to errors found by manual review.

C.1 Audio and Text Representation Results

The main text compares raw audio with text across the full benchmark. It does not show whether the result changes with the type of speech or physiological signal. Figures 14 and 15 repeat the comparison for each audio group.

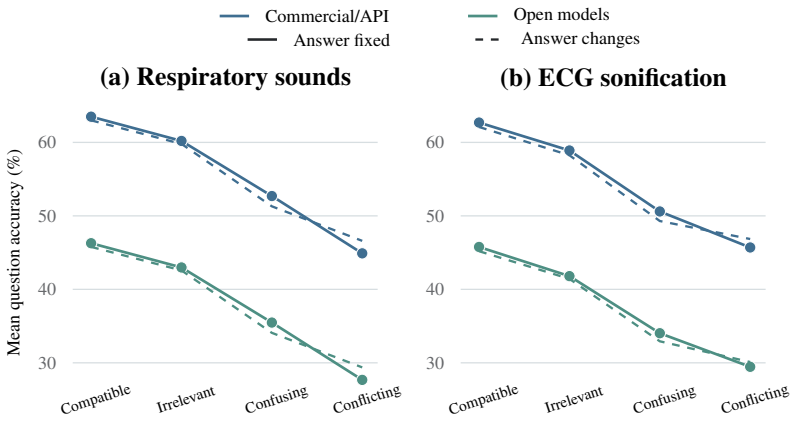


Figure 16: Evidence relation analysis for respiratory sounds and ECG sonification. Solid lines show questions whose answer should remain fixed; dashed lines show questions whose answer should change with the audio.

For clinical speech, the waveform and transcript together give the best result for every communication type. The waveform alone is usually slightly better than the transcript alone, which suggests that the way something is said can add useful information. Models handle clinician dictation best and have more difficulty with team discussion, where the content comes from several views.

The two audio types show the same pattern. Performance is best when the image and audio agree. It falls with unrelated audio, falls further with a plausible but confusing recording, and is lowest when the sources conflict. The same drop appears even for questions whose answer should not change. Models can therefore let a convincing recording override evidence that should remain more important. MedSensus tests this directly because more evidence is helpful only when the model knows which source matters for the question. Across all four relations, group accuracy remains below 11% for commercial and API models and below 3% for open models.

The answer behavior split makes the failure especially clear. For questions whose answer should stay fixed, a change after replacing the audio is an error in evidence selection. For questions whose answer should change, keeping the old answer means that the model ignored useful new evidence. Current models show both behaviors. The challenge is therefore not simply to trust or ignore audio, but to decide how much it matters for the question being asked.

C.2 Manual Error Analysis

The score tables show where performance changes, but they do not show what the errors look like. We review a fixed sample of wrong predictions under complete input and place each error in a clear category. Table 8 links the scores to common model behavior.

More than half of the reviewed errors come from using only the image or only the audio. Using only the image is more common. The next large group contains cases where the model reads both sources but combines them incorrectly. Models also follow conflicting audio or fail to change an answer when the audio is useful. Errors in the answer set and responses that cannot be scored are less common. Many failures therefore come from choosing or combining evidence, not only from failing to see or hear it. This is the main behavior that MedSensus is designed to measure.

These categories also connect the separate experiments. Visual and audio shortcuts explain why removing either source can change the answer. Failure to use relevant audio matches the weak results under changed evidence, while answer set errors explain part of the gap between exact accuracy and set F1. The manual review therefore does more than list mistakes. It shows that the same evidence use problems appear under several different tests, which makes the central result more credible.

Error category	Count	Share	Meaning
Visual shortcut	124	31.0%	The prediction follows the image and ignores audio needed for the answer.
Audio shortcut	86	21.5%	The prediction follows the audio and ignores the image needed for the answer.
Evidence integration error	65	16.2%	The model reads both sources but draws the wrong conclusion from them.
Overuse of conflicting audio	42	10.5%	Conflicting audio overrides evidence that is more relevant to the question.
Failure to use relevant audio	38	9.5%	The answer does not change when relevant audio should change it.
Answer set error	35	8.8%	At least one correct option is omitted or an unsupported option is added.
Unscorable output	10	2.5%	The response does not contain one clear answer set that can be scored.

Table 8: Manual review of a fixed sample of incorrect predictions under complete input.

C.3 What the Appendix Results Show Together

The main paper reports the central benchmark results, while the appendix asks whether the same conclusion remains visible under different views of the data. This final section brings those views together. It is useful because no single table can separate question answering skill from evidence use, consistency across related questions, and sensitivity to the way evidence is presented.

Three patterns appear throughout the appendix. First, a model can answer many individual questions correctly and still fail to answer every question linked to the same image and audio pair. The lower group accuracy is therefore not a second version of question accuracy. It measures whether the model can keep a sound clinical interpretation while the question changes. Second, the image, audio, question stem, and answer options all affect predictions. The question stem and options provide useful clues, but removing the image or audio still causes a clear loss. This means that complete input accuracy should be read together with the removal experiments. Third, the representation experiments show that text and sound support each other. A transcript or expert description makes some information easier to use, while the original sound can retain cues that the text does not fully express.

The category, modality, and specialty results also show that these patterns are not limited to one narrow part of medicine. Some areas are harder than others, especially when the visual signal is unusual or the answer depends on several sources at once, but the main model ordering remains broadly stable. The controlled evidence tests explain why: current models do not always know which source should guide a particular answer. They may follow a plausible but irrelevant clue, ignore a useful recording, or combine two correct observations in the wrong way. MedSensus is built around this distinction. Its purpose is not only to ask whether a model can accept images and audio, but whether it can select and combine the right evidence for each clinical question, repeatedly, within the same reviewed case.

D Illustrative Clinical Cases

The following pages show how one image and one recording support several questions from the same case. Each page covers four medical settings and reports example model answers.

Musculoskeletal, abdominal, thoracic, and endoscopic cases. The first page (Figure 17) covers knee MRI, abdominal imaging, thoracic imaging, and endoscopy. Its questions move from separating nearby mimics to choosing the most important lesion. They show why an image label alone is not enough: the model must use the clue that matters for the question.

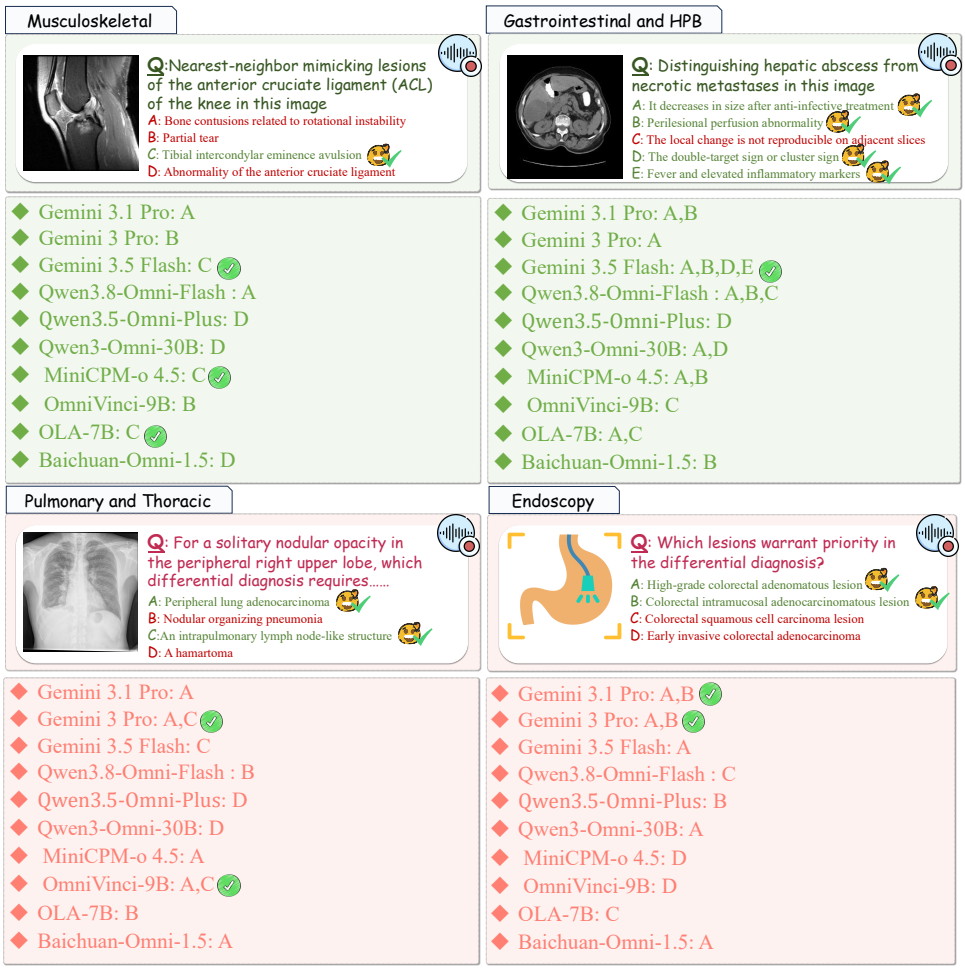


Figure 17: Cases from musculoskeletal, gastrointestinal and hepatopancreatobiliary, pulmonary and thoracic, and endoscopic settings. Each panel shows the image, question, answer choices, matched audio indicator, and model answers.

Surgical, tissue, genitourinary, and breast cases. The second page (Figure 18) focuses on evidence that changes with clinical context. The surgical example asks what action is taking place, the tissue example asks which lesion should be considered first, and the genitourinary case separates tumor thrombus from ordinary thrombus. The breast case compares postoperative change with local recurrence. These questions require the model to connect shape, location, continuity, and change over time with the information in the recording.

The four panels also move from direct observation to comparison. Instrument position and tissue continuity can support the surgical answer, but the other three cases require the model to reject alternatives that share part of the same appearance. The spread of model answers suggests that many systems notice a plausible clue without checking whether it is the clue that the question actually asks about.

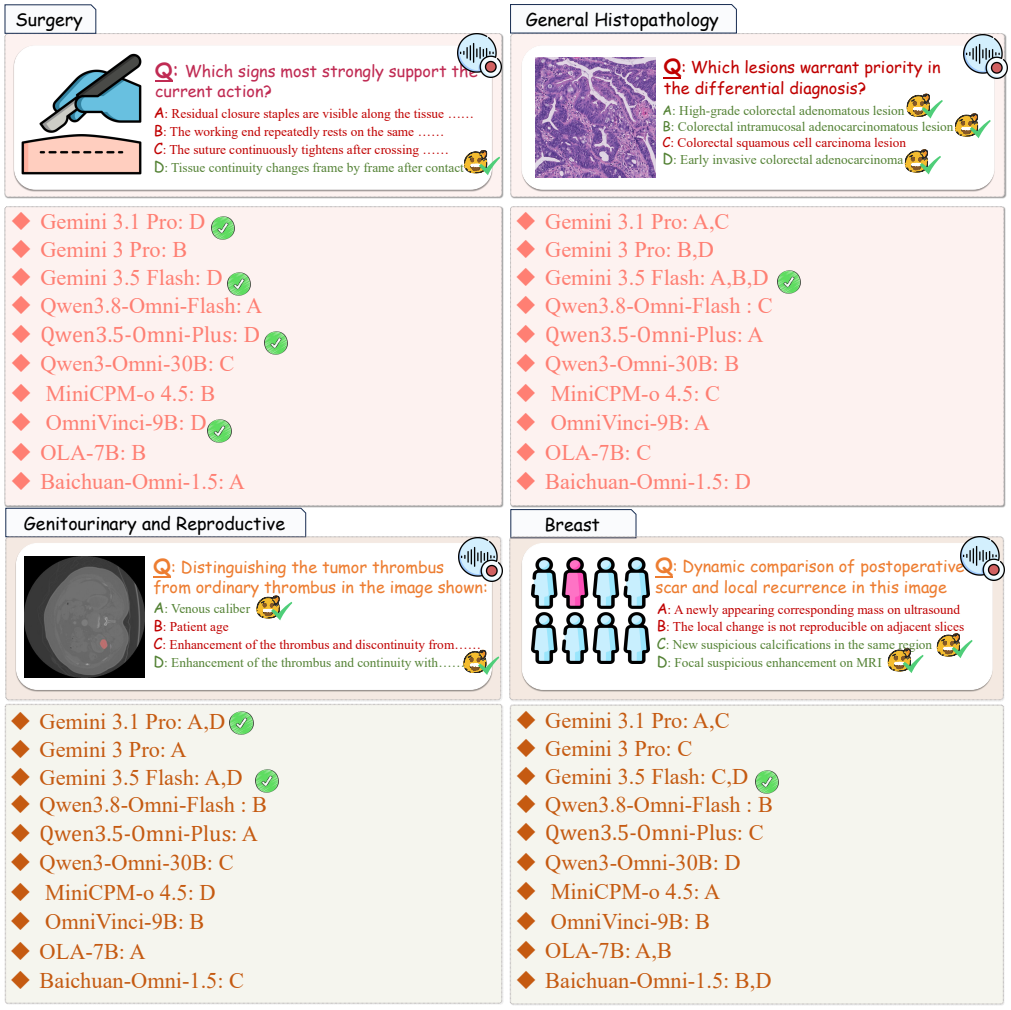


Figure 18: Cases from surgery, histopathology, genitourinary and reproductive medicine, and breast imaging. The examples move from recognizing an action to comparing clinically plausible alternatives.

Hematology, neurology, ophthalmology, and cardiovascular cases. The final page (Figure 19) shows that the useful source can differ from one question to the next. Cell morphology is central in hematology, while the neurology case separates what can be read from imaging from what requires an integrated diagnosis. The ophthalmology example adds a patient description, and the cardiovascular example combines a known myocardial infarction label with the current tracing. Together, these cases show the main MedSensus setting: the model must decide which part of the image, audio, and case history supports the answer.

These panels also make the evidence boundary visible. Imaging can locate a brain lesion, but it cannot by itself provide every molecular conclusion. A patient description can change how an ophthalmic image is read, while a known cardiac history must be checked against the current signal rather than copied into the answer. The disagreement across models therefore reflects more than visual recognition. It reflects whether each source is given the right role.

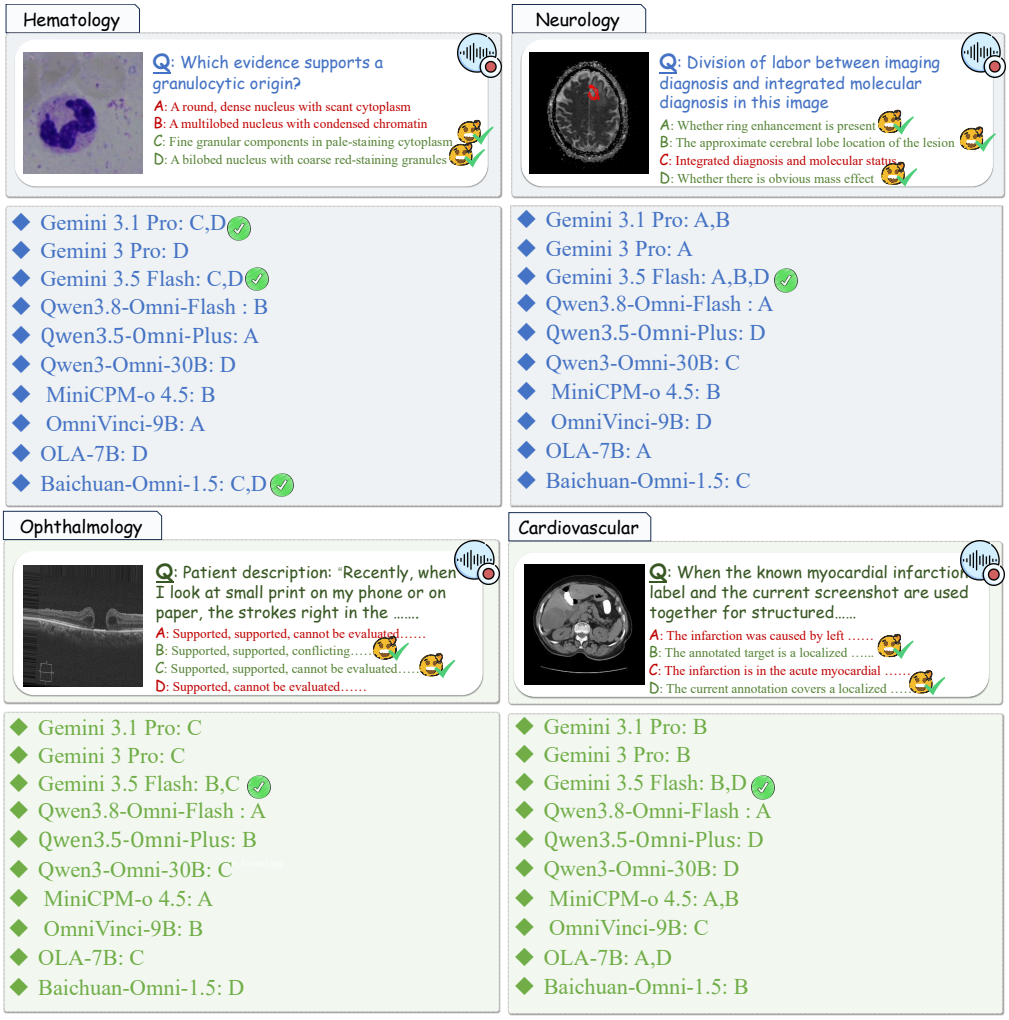


Figure 19: Cases from hematology, neurology, ophthalmology, and cardiovascular medicine. The panels illustrate how the evidence needed for the answer can shift across clinical fields.