CARE-MED: A CAPABILITY-RESOLVED BENCHMARK FOR MEDICAL PERCEPTION ACROSS PHOTOGRAPHIC AND PHYSIOLOGICAL-SIGNAL IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical multimodal large language models (MLLMs) are still compared mostly by a single aggregate accuracy, which reveals that a model errs but not which competency it lacks. Existing benchmarks also concentrate on sensor-reconstructed imaging (radiology, pathology, ultrasound, and microscopy), leaving camera-captured images and physiological-signal visualizations under-tested. We present CARE-Med, a 2,023-question Chinese-language benchmark over five under-tested domains: body fluids, in-vitro diagnostic (IVD) reagents, endoscopy and surgical fields, ex-vivo organs and tissues, and physiological-signal visualizations. Three of them (macroscopic body fluids, IVD-reagent readouts, and gross ex-vivo organs and tissues) have, to our knowledge, no dedicated vision–language evaluation, and a self-built four-level image taxonomy quantifies this coverage bias across seven prior benchmarks. Every question is mapped to leaves of a 34-leaf capability tree whose top level takes three knowledge types (factual, conceptual, procedural) from Bloom's revised taxonomy, and every question, answer, and label is verified by clinicians. Evaluating 14 frontier and medical MLLMs, we find that a single score, even when split by knowledge type, is misleading: within each frontier model the three knowledge types differ by only 1–3 points, yet its capability leaves span more than 30 (GPT-5.5 ranges from 53% to 89% and reaches only 71% overall). The three frontier models fail on largely the same capability leaves ($\rho \approx 0.8$), a shared weakness in concrete visual determination alongside strong explanation, and medical-specialized models redistribute competence across the tree rather than closing this gap. CARE-Med turns an opaque score into a diagnosis of where medical perception breaks.

1 INTRODUCTION

Multimodal large language models (MLLMs) are increasingly proposed for clinical use, and a growing suite of medical benchmarks tracks their progress (Chen et al., 2024; Hu et al., 2024; Zuo et al., 2025). Yet most benchmarks report a single aggregate accuracy, or at best tag questions by clinical task or imaging modality. Such a score tells us *that* a model errs; it rarely tells us which competency the model lacks, or on *which* kinds of images it breaks down. For a safety-critical domain, this coarse feedback is a poor guide to what is actually missing.

Two gaps compound the problem. **First, a coverage gap.** Existing medical MLLM benchmarks concentrate on sensor-reconstructed imaging—radiology, pathology, ultrasound, and microscopy (Hu et al., 2024; Zhang et al., 2023; Liu et al., 2021; He et al., 2020). The images a clinician photographs with an ordinary camera or endoscope (body fluids, in-vitro diagnostic reagent readouts, endoscopy and surgical fields, and gross ex-vivo organs and tissues) and the document-style images they read (physiological-signal visualizations such as electrocardiograms) are comparatively under-tested, despite being ubiquitous in practice. **Second, an attribution gap.** Cognitive competence has structure (recalling facts, explaining mechanisms, applying procedures), but a single score hides it, so failures cannot be traced to a specific skill.

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.

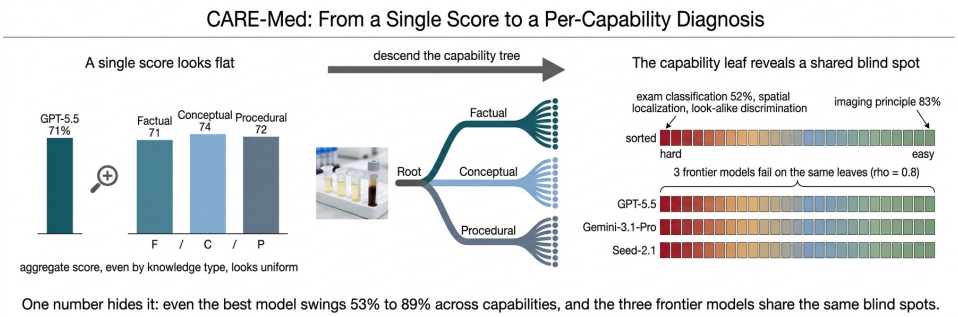


Figure 1: **From a single score to a per-capability diagnosis.** A single aggregate accuracy, even when split by the factual/conceptual/procedural knowledge types, makes models look uniformly mediocre; descending CARE-Med's capability tree to the leaf level exposes a large, structured spread and a systematic weakness that the three frontier models share. *(Illustrative; quantitative results in Section 5.)*

We introduce **CARE-Med**, a 2,023-question Chinese-language benchmark built to close both gaps and, in doing so, to answer a diagnostic question: *where do frontier medical MLLMs break beyond sensor-reconstructed imaging, and can we attribute those breaks to specific cognitive competencies?* Its questions cover five image domains that prior benchmarks leave sparsely tested, and every question is mapped to one or more nodes of a capability tree whose top level (factual, conceptual, and procedural knowledge) follows Bloom's revised *knowledge* dimension (Anderson & Krathwohl, 2001). To make questions demanding and image-dependent, items are posed as 6-to-8-option multiple-select and objective fill-in-the-blank, with distractors engineered against known MLLM visual-hallucination modes; ground truth is consolidated from three models' independent captions, and every question, answer key, and capability label is verified by clinicians.

Contributions.

- **Capability-resolved evaluation.** Every question is mapped to leaves of a four-level, 34-leaf capability tree whose top level takes three of the four knowledge types of Bloom's revised taxonomy (factual, conceptual, procedural); to our knowledge, CARE-Med is the first medical VQA benchmark to use this dimension for per-capability failure attribution.
- **Coverage of under-tested image domains.** 2,023 questions over five sparsely tested domains, three of which (*macroscopic* body fluids, IVD-reagent readouts, and *gross* ex-vivo organs and tissues) have to our knowledge no prior dedicated vision–language benchmark (claims scoped to taxonomy branches; Section 2).
- **A taxonomy-based coverage analysis and a difficulty-controlled construction protocol.** A four-level image taxonomy used, to our knowledge for the first time, to quantify the coverage bias of medical MLLM benchmarks and select under-tested domains; and an image-dependent construction protocol (multi-model caption consolidation, hallucination-targeted distractors, difficulty-targeted selection, clinician verification of every item).
- **A capability-resolved diagnosis of 14 frontier and medical MLLMs.** Three findings a leaderboard cannot give: (i) difficulty varies across capability leaves, not knowledge types: within each frontier model the three knowledge types differ by only 1–3 points, yet its capability leaves span more than 30 (and the strongest model reaches only 71%); (ii) the three frontier models fail on largely the *same* capability leaves ($\rho \approx 0.8$), a shared, systematic weakness in concrete visual determination (classification, localization, look-alike discrimination) even as they excel at explanation and reasoning; and (iii) medical-specialized models redistribute competence across the tree rather than closing the gap, with three of four medical-specialized open models falling below the two strongest general-purpose open models.

2 Related Work

Medical MLLM benchmarks. Medical VQA benchmarks have grown from radiology- and pathology-centric datasets (Liu et al., 2021; Lau et al., 2018; He et al., 2020; Zhang et al., 2023) to large multi-modality suites such as OmniMedVQA (Hu et al., 2024) and GMAI-MMBench (Chen et al., 2024), and to expert-level reasoning benchmarks such as MedXpertQA (Zuo et al., 2025), which attach flat modality labels or clinical-task tags. Chinese medical evaluation remains largely text-only (Liu et al., 2023b; Wang et al., 2024; Qiu et al., 2024; Ding et al., 2025), and its multimodal counterpart is sparse and again radiology- and pathology-centric (Liu et al., 2023a; Ding et al., 2025). A parallel line targets a single under-tested region: endoscopy (Liu et al., 2025), physiological signals (Liu et al., 2024), whole-slide pathology (Liang et al., 2025), and hematology (Wang et al., 2026; Malik et al., 2026). CARE-Med instead evaluates five such domains jointly and replaces modality tags with a knowledge-type capability tree. Our image taxonomy makes the scope of the resulting claims precise (Section 3.1): under it, macroscopic body fluids, IVD-reagent readouts, and gross ex-vivo organs and tissues have, to our knowledge, no prior dedicated vision–language benchmark,[1] whereas endoscopy and electrocardiograms are covered by prior work and we claim no domain novelty for them.

Image taxonomies and coverage analysis. Medical image type has long been organized through clinical ontologies (RadLex, SNOMED-CT, DICOM modality codes) built for reporting and records, and through appearance-based schemes for literature-figure retrieval such as the IRMA code (Lehmann et al., 2003), ImageCLEF modality hierarchies, and the image-type taxonomy of Müller et al. (2012). Recent MLLM benchmarks mostly use flat modality labels (Cai et al., 2025), with only shallow structure emerging (Jiang et al., 2026; Bao et al., 2026); coverage, where analyzed, is examined at the dataset level (Deng et al., 2026) or via lifecycle audits (Chen et al., 2025). The closest neighbor, MedBookVQA (Yip et al., 2025), organizes its textbook-derived data under a multi-tier taxonomy and compares prior benchmarks by coverage *counts*, but does not quantify their coverage *bias*. We do not claim the first medical image-type taxonomy; we claim its first use, to our knowledge, for quantifying the coverage bias of medical benchmarks and for selecting under-tested domains.

Cognitive taxonomies for evaluation. Bloom's revised taxonomy is increasingly used in AI evaluation, but almost always along its *cognitive-process* axis: Huber & Niklaus (2025) map LLM benchmarks onto process levels, and several visual benchmarks organize questions from remembering to creating (Gong et al., 2024; Burgess et al., 2025; Abootorabi et al., 2026). Medical work instead imports clinical competency frameworks or task-based capability trees (Gong et al., 2025), probes low-level perception as a blind spot distinct from reasoning (Bigverdi et al., 2025), or, for general MLLMs, builds hierarchical capability taxonomies (Li et al., 2024); psychometric critiques call for construct validity (Alaa et al., 2025). None organizes a medical VQA benchmark by the revised taxonomy's orthogonal *knowledge* dimension (Anderson & Krathwohl, 2001). Empirically, we also do not observe the monotonic difficulty gradient that process-level studies report: on CARE-Med the three knowledge types differ by only 1–3 points within a frontier model, and difficulty is governed at the leaf level (Section 5).

3 The CARE-Med Benchmark

CARE-Med contains 2,023 questions across five image domains. Four domains contribute 400 questions each (200 multiple-select and 200 fill-in-the-blank); the physiological-signal domain contributes 423 (260 multiple-select and 163 fill-in-the-blank), for 1,060 multiple-select and 963 fill-in-the-blank items in total. The questions draw on 1,430 unique images, and the five domains' image sets are disjoint. Every item is annotated with a domain (what is in the image) and one or more capability leaves (what skill is tested).

[1] IVD readout has been studied as image classification in point-of-care diagnostics, not as a vision–language task.

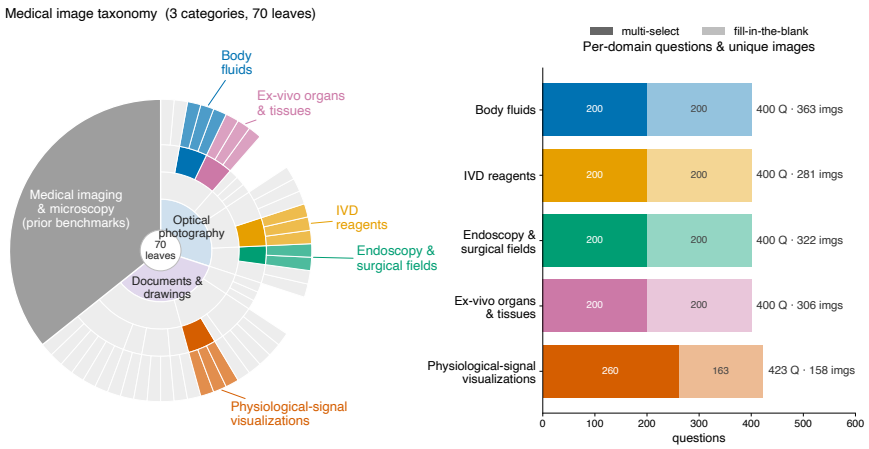


Figure 2: **CARE-Med targets under-tested image categories.** Left: the four-level, 70-leaf image taxonomy; CARE-Med's five domains (coloured) lie in category 1 (optical photography) and category 3 (documents and drawings), while category 2 (medical imaging and microscopy; grey), where prior benchmarks concentrate, is left out by design. Right: per-domain question counts by format and number of unique images.

3.1 Image Domains and Taxonomy

To reason about what benchmarks test, we build a medical image taxonomy with **four levels** and **three top-level categories**: (1) *optical photography* (ordinary camera or endoscope; macroscopically visible), (2) *medical imaging and microscopy* (sensor-reconstructed, not directly eye-visible), and (3) *documents and drawings*. A node is a leaf when it is terminal, at level 3 or level 4; the taxonomy has 70 leaves (Appendix D). It was curated through an iterative build–audit–expand loop in which a large model classifies crawled medical images, a medical team audits the assignments, and unclassifiable images trigger new categories.

CARE-Med targets five level-3 domains under categories 1 and 3: **body fluids** (the macroscopic appearance of urine, blood, sputum, and effusions, not their microscopic cytology), **in-vitro diagnostic (IVD) reagents** (pregnancy, glucose, and antigen test strips and cards), **endoscopy and surgical fields**, **ex-vivo organs and tissues** (gross specimens), and **physiological-signal visualizations** (electrocardiogram, electroencephalogram, and other waveform and curve charts). Four of the five sit in category 1; physiological-signal visualizations sit in category 3. None lies in category 2, the region on which prior benchmarks concentrate (Figure 2). The taxonomy also fixes the boundaries of these domains: blood *under a microscope* is cytology (category 2), whereas blood *in a tube or on a dressing* is macroscopic fluid appearance (category 1), so hematology and whole-slide benchmarks (Wang et al., 2026; Malik et al., 2026; Liang et al., 2025) occupy a different branch from our body-fluid domain, as microscopic histopathology does from our gross ex-vivo organs and tissues.

3.2 Taxonomy-Based Coverage Analysis

The taxonomy is also an instrument for measuring what existing benchmarks test. We classify the unique images of the publicly released evaluation sets of seven prior benchmarks (OmniMed-VQA, the GMAI-MMBench validation split, PMC-VQA, SLAKE, VQA-RAD, PathVQA, and EndoBench) into the 70 leaves with a single vision model (Seed-1.8, greedy decoding) that sees only the image, using a hierarchical prompt (category $\rightarrow$ group $\rightarrow$ leaf). Figure 3 shows the benchmark $\times$ leaf matrix; per-benchmark statistics are in Appendix Table 7.

Coverage is heavily skewed to category 2: five of the six imaging benchmarks place 82.5%–100% of their images there (mean $\approx 93\%$), concentrated in CT, MRI, X-ray, ultrasound, and H&E histopathology, and each populates only 5–42 of the 70 leaves. The two benchmarks with substan-

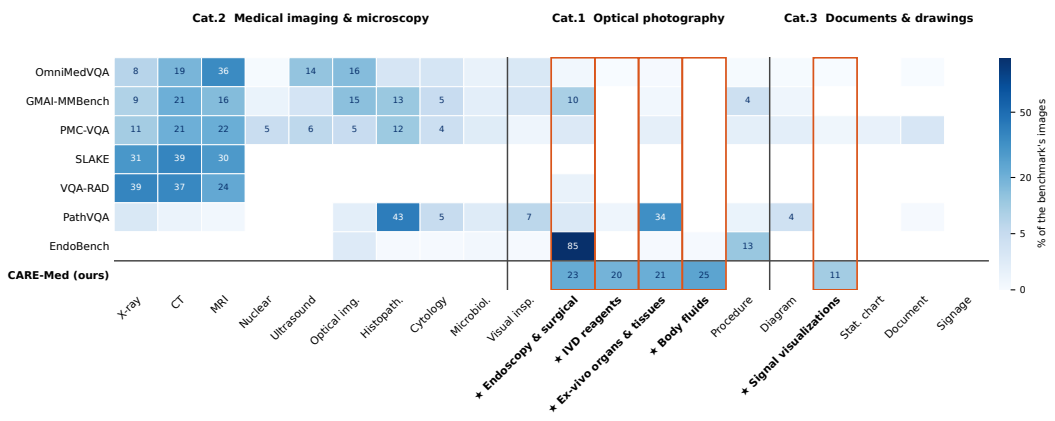


Figure 3: **Prior benchmarks crowd category 2; CARE-Med populates the neglected leaves.** Each cell is the share (%) of a benchmark's unique images in a taxonomy leaf group (white = no image; a curated 20-group view of the 70 leaves). Category-2 groups are populated by every general and radiology benchmark, whereas the five CARE-Med target regions (⋆) are nearly empty outside single specialized benchmarks (EndoBench for endoscopy, PathVQA for ex-vivo specimens, plus a 10% endoscopy share in GMAI-MMBench). The CARE-Med row shows the share of its unique images in each target domain.

tial category-1 mass are each concentrated in one domain: EndoBench (98% category 1) is almost entirely endoscopy, and PathVQA (44% category 1) draws its category-1 images mostly from ex-vivo pathology specimens. Category 3 never exceeds 3.7% of any benchmark. Pooling all seven benchmarks covers 54 of the 70 leaves: category 2 is saturated (25/25), category 3 reaches only 13/24, and 16 leaves are populated by no benchmark at all, among them physiological-signal curve charts, natural and puncture/drainage body fluids, and high-risk (Class III) IVD reagents, all CARE-Med targets. By domain, body fluids, IVD reagents, and physiological-signal visualizations receive essentially no images across the seven benchmarks (1, 12, and 3 of about 102k images), while endoscopy and ex-vivo specimens are populated mainly by a single specialized benchmark each (EndoBench and PathVQA; GMAI-MMBench adds a 10% endoscopy share).

To guard against a self-built taxonomy and classifier encoding our preferred domains: first, the classifier's "other/unclassifiable" rate is low (0–5.6% per benchmark), so the taxonomy absorbs heterogeneous medical images. Second, two medical annotators independently checked the classifier's assignment for a stratified sample of 480 images across the seven benchmarks: per-annotator accuracy was 96.3% and 95.6% (Wilson 95% CI $[94.2, 97.6]$ and $[93.4, 97.1]$), the assignment was judged correct by both annotators for 94.2% of images, and inter-annotator agreement was 96.5% (Gwet's AC1 = 0.96). Third, we will release the taxonomy definitions, the classification prompt, and the per-image assignments.

3.3 The Capability Tree

Complementary to what the image shows, CARE-Med annotates which competency each question tests. The capability tree has four levels. Its top level follows the *knowledge* dimension of Bloom's revised taxonomy (Anderson & Krathwohl, 2001), which distinguishes four knowledge types; we adopt the first three: **factual** (recall and distinguish entities, terms, quantitative readings, and classifications), **conceptual** (explain mechanisms, differentiate look-alike findings, and reason about imaging principles), and **procedural** (apply standard workflows, quality assessment, and guideline or care-pathway steps). Metacognitive knowledge, the fourth type, concerns a learner's awareness of their own cognition and is not elicited by single-turn image questions with a fixed answer key (Zhang et al., 2025); the three we use give the tree an axis orthogonal to the modality and clinical-task tags of prior benchmarks. These refine into competency areas and groups, ending in 34 leaf capabilities (13 factual, 10 conceptual, 11 procedural; Appendix C). Each question is mapped to one or more leaves (3.2 on average), because a single image typically supports several capabilities; labels are proposed at question generation and verified by clinicians during the per-question review

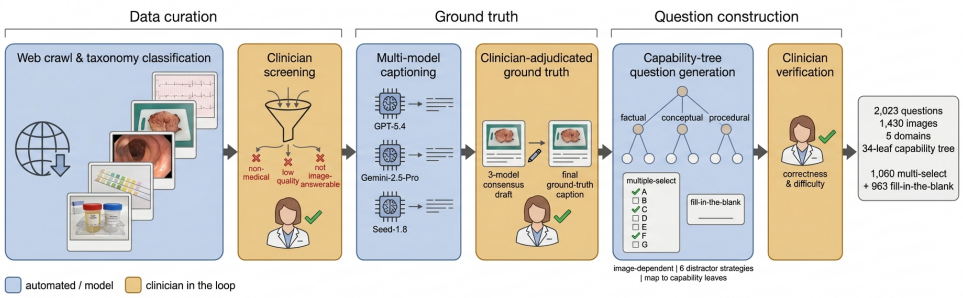


Figure 4: **CARE-Med construction pipeline** (schematic): automated/model stages (blue) and clinician-in-the-loop stages (orange); §3.4 details each stage.

(§3.4). The three leaves with fewer than 30 items are excluded from leaf-range and hardest/easiest statistics.

3.4 Benchmark Construction

CARE-Med was built in two phases that share one generation recipe (Figure 4); Appendix A lists every model used at each stage.

Shared recipe. (1) *Sourcing:* images are drawn from public web image–text collections and from figures in published medical books and literature, assigned to the taxonomy, and the five target domains are taken from the corresponding branches. (2) *Captioning:* three multimodal LLMs (GPT-5.4, Gemini-2.5-Pro, and Seed-1.8) independently caption each image from the image alone, without surrounding text. (3) *Ground truth:* the three captions are consolidated into a single ground-truth description anchored on the image. (4) *Question generation:* a structured prompt generates multiple-select and fill-in-the-blank items from the image and its ground truth under three constraints: an image-dependence principle (no text-only-answerable items; no diagnostic term leaked into the stem), strict grounding in the ground truth, and mapping to the most appropriate capability leaves. Distractors follow six strategies that target common MLLM visual-hallucination modes, such as entity confusion, visual traps, and precise-data errors. Multiple-select items have 6–8 options and a set-valued answer; fill-in-the-blank items ask for a short, objective answer (a number, organ name, or grade).

Phase 1 (967 items). The first 544 items in the four photographic domains, and the 423 physiological-signal items, were curated by clinicians throughout: clinicians screened images for medical relevance, quality, and whether a question could be answered from the image alone; resolved disagreements among the three captions against the image to form the ground truth; and reviewed and corrected every generated question.

Phase 2 (1,056 items). To scale the four photographic domains, Phase 2 automates the intermediate stages. An LLM scores image–text relevance to filter images; ground-truth consolidation decomposes the three captions into atomic claims, checks each claim against the image, and keeps image-supported claims; questions are generated by GPT-5.4; and an LLM screen removes items with an unsupported or non-unique answer key, a medical error, an unsafe recommendation, text-only answerability, or internal ambiguity. From the screened pool, items are selected to balance format, capability leaves, answer patterns, and image sources, preferring items that a probe model answers incorrectly; such *difficulty-targeted selection* is also used by recent frontier benchmarks (Phan et al., 2025). The probe, Seed-1.6, is not among the evaluated models. Finally, clinicians reviewed every selected item (stem, options, answer key, and capability labels) and corrected or removed problematic ones.

4 EVALUATION PROTOCOL

Models. We evaluate 14 MLLMs in closed-API and open-weight form. For three leading U.S. and Chinese frontier model families we evaluate both the current flagship (GPT-5.5, Gemini-3.1-Pro, Seed-2.1) and its immediately preceding version (GPT-5.4, Gemini-2.5-Pro, Seed-1.8) to capture within-family progression.[2] The open-weight models are four general-purpose models (Qwen3-VL-8B, InternVL3.5-8B, GLM-4.6V-Flash, MiniCPM-V-4.5) and four medical-specialized models (Lingshu-32B, MedGemma-27B, Citrus-V-8B, HuatuoGPT-V-34B). Appendix B.1 lists model versions.

Scoring. Each model answers every question from the image and a short instruction to output only its final answer (Appendix B.2); all 14 models returned a response for every item, so accuracies are computed over all 2,023 items. Multiple-select items are scored by exact set match between the predicted and reference option sets after normalizing letters, separators, and order; when the option letters cannot be parsed, the judge model below compares the response with the reference under a strict exact-match instruction. Fill-in-the-blank answers are graded by an LLM judge (Seed-1.8, temperature 0) under a conservative medical-equivalence rubric: the judge sees the question, the reference answer, and the model's answer, but neither the image nor the identity of the answering model (Li et al., 2025); synonyms, abbreviations, and unit variants are accepted, whereas a missing key entity, site, grade, negation, or procedural step is scored incorrect. We report accuracy overall and per format, with item-level bootstrap 95% confidence intervals (Appendix F).

Capability-resolved analysis. A model's accuracy at a leaf (or higher node) is computed over all items whose label set includes it, pooling both formats. Because the roster is heterogeneous, we report accuracy per model and by tier, never as a roster-wide mean, and measure whether models share a difficulty structure by the Spearman correlation of their leaf rankings.

Models that contributed to construction. Three evaluated models contributed captions to the ground truth (GPT-5.4, Gemini-2.5-Pro, Seed-1.8; marked †), GPT-5.4 also generated the Phase-2 questions, and Seed-1.8 serves as the fill-in judge. We keep these models on the leaderboard and flag them. Three observations bear on possible self-generation and self-preference effects: the two top-ranked models contributed to neither the ground truth nor the questions; Seed-1.8, the judge, has the lowest fill-in accuracy of the six closed models, with a fill-versus-multiple-select gap in the same range as the other closed models (Appendix F); and every final item was verified by clinicians, which limits one documented source of generator self-bias, namely wrong labels in the answer key (Yuan et al., 2025).

Image dependence and contamination. To test whether questions require vision rather than text priors, we re-run the three flagships with the image withheld, following work that measures image reliance by removing the image (Sadanandan & Behzadan, 2026) and that screens benchmarks for text-only answerability (Yue et al., 2025). Because images are web-collected, contamination is possible (Xu et al., 2026; Yan et al., 2026); the questions, answer keys, and distractors are newly written and unpublished, which limits memorization, and upon release we will keep a private held-out split and embed a canary string in the released files.

5 RESULTS AND ANALYSIS

Accuracy spans 71% to 17% across the 14 models, so we report it per model and by tier and never average across the roster; "frontier" denotes the three flagships (GPT-5.5, Gemini-3.1-Pro, Seed-2.1).

5.1 MAIN RESULTS

The benchmark is far from solved. The strongest model, GPT-5.5, reaches only 71.1% (Table 1), and the best open-weight model (Lingshu-32B, 45.4%) trails the weakest closed model (Seed-1.8,

[2]Claude was not evaluated because of regional access constraints; Med-CMR (Gong et al., 2025) and SNS-Bench-VL (Guo et al., 2025) likewise omit it.

Table 1: Main leaderboard (accuracy %, all 2,023 items; ranked by overall accuracy). MS = multiple-select (exact set match); Fill = fill-in-the-blank (LLM judge). † ground-truth caption contributor (flagged, kept on the board); ‡ medical-specialized. Item-level bootstrap 95% CIs are ±1.7–2.2 points overall (Appendix F).

Model	Overall	MS	Fill	Model	Overall	MS	Fill
GPT-5.5	**71.1**	77.4	64.3	Lingshu-32B‡	45.4	46.3	44.4
Seed-2.1	68.4	73.6	62.6	Qwen3-VL-8B	37.8	31.0	45.3
GPT-5.4†	63.8	67.2	60.1	InternVL3.5-8B	37.6	33.4	42.2
Gemini-3.1-Pro	63.3	67.5	58.6	GLM-4.6V-Flash	34.5	26.4	43.4
Gemini-2.5-Pro†	61.7	62.8	60.5	MedGemma-27B‡	29.5	29.2	29.8
Seed-1.8†	58.2	62.9	53.0	MiniCPM-V-4.5	22.7	7.4	39.6
				Citrus-V-8B‡	21.2	14.7	28.2
				HuatuoGPT-V-34B‡	17.4	8.1	27.5

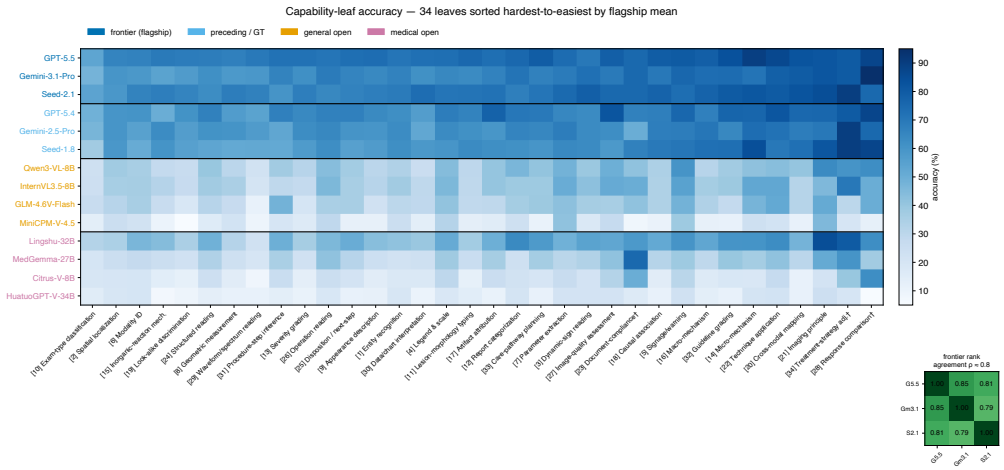


Figure 5: **Frontier models share a systematic blind spot, and medical specialization does not close it.** Per-leaf accuracy for all 14 models (rows, grouped by tier) across the 34 leaves (columns, sorted hardest to easiest by the flagship mean; † marks leaves with fewer than 30 items). The three flagships are weakest on largely the same left-hand leaves (exam-type classification, spatial localization, modality identification), and their leaf rankings agree at Spearman $\rho \approx 0.8$ (inset).

58.2%). In each closed family the current flagship scores above its predecessor (significantly for GPT and Seed, paired McNemar $p < 10^{-12}$; not significantly for Gemini, 63.3 vs. 61.7, $p = 0.19$), consistent with CARE-Med tracking capability rather than noise. The two leading models contributed neither ground truth nor questions, which bounds any self-generation advantage. Per domain (Appendix Table 10), body fluids are the easiest domain for 13 of the 14 models, IVD reagents are hardest for the frontier, and physiological-signal visualizations are hardest for most open models.

5.2 The Capability Profile

Difficulty lives at the capability leaf, not the knowledge type. Aggregating to the three knowledge types barely separates difficulty for the frontier: within each flagship, factual, conceptual, and procedural accuracy differ by only 1–3 points (GPT-5.5: 70.5/73.6/72.3; Gemini-3.1-Pro: 62.8/63.7/63.9; Seed-2.1: 68.0/70.6/68.3; Appendix Table 12). Descending to the leaves changes the picture: the capability leaves span more than 30 points within each flagship (GPT-5.5 ranges from 53% on exam-type classification to 89% on micro-mechanism), well beyond what random reassignment of items to leaves produces (size-preserving permutation test, $p < 0.05$ for each flagship; Appendix F).

The frontier models share a systematic blind spot. The three flagships agree strongly on which capabilities are hard: their leaf-difficulty rankings correlate at Spearman $\rho = 0.85$, 0.81, and 0.79 (Figure 5). The shared hardest leaves are exam-type classification (52% flagship mean), spatial localization (61%), modality identification, inorganic-reaction mechanism, look-alike discrimination, and structured reading (all $\approx 64\%$); the shared easiest are imaging-principle explanation (83%), cross-modal mapping (82%), technique-application reasoning (82%), and micro-mechanism explanation (81%). Apart from the reagent-chemistry leaf, the frontier's weakness is concrete visual determination (pinning down what an image is, where a structure lies, and which of two look-alikes it shows), while its strength is explanation and reasoning.

Medical-specialized models redistribute competence rather than closing the gap. Three of the four medical-specialized open models (HuatuoGPT-V 17.4%, Citrus-V 21.2%, MedGemma 29.5%) fall below the general-purpose Qwen3-VL and InternVL3.5 ($\approx 38\%$); only Lingshu-32B (45.4%) leads the open tier. Relative to their own averages, general open models are strongest on factual recognition, whereas the two strongest medical models (Lingshu-32B and MedGemma-27B) are strongest on procedural competencies; yet no open model closes the gap to the frontier, which is widest on conceptual leaves (frontier-minus-open gap of 42 points, versus 37 factual and 39 procedural). Medical specialization is thus associated with a shift in where competence lies rather than a uniform gain in perception.

Multiple-select accuracy exceeds fill-in accuracy, by a capability-dependent margin. Within a capability, the flagships score far higher on multiple-select than on fill-in items for classify-and-decide leaves (multiple-select minus fill: exam-type classification $+50$, procedure-step inference $+39$, modality identification $+38$), whereas the gap reverses for read-a-value leaves (parameter extraction -15, waveform reading -11; Appendix Figure 10).

5.3 Image Dependence and Robustness

Answers depend on the image. Withholding the image drops the three flagships by 21.5–23.3 points overall (GPT-5.5 $71.1 \rightarrow 49.6$, Gemini-3.1-Pro $63.3 \rightarrow 40.1$, Seed-2.1 $68.4 \rightarrow 45.0$), consistently across all five domains (Appendix Table 9 and Figure 10, right). The drop is largest for fill-in-the-blank, which falls by 33–35 points, while multiple-select falls by 9–14 points; this is why CARE-Med pairs the two formats. On the items that none of the flagships answers without the image, the ranking of all 14 models is essentially unchanged (Spearman 0.96; Appendix F).

The results are statistically reliable. Overall accuracies carry bootstrap 95% confidence intervals of ± 1.7–2.2 points; the tier gaps are large relative to them (e.g., the weakest closed model exceeds the strongest open model by 12.8 points, McNemar $p < 10^{-24}$); the leaf-level spread of each flagship exceeds a size-preserving permutation null; and the fill-in judge's own answers score lowest among the closed models (§4). Appendix F reports all intervals and tests.

6 Discussion and Conclusion

CARE-Med evaluates medical MLLMs on under-tested photographic and physiological-signal images and attributes their failures to competencies through a 34-leaf capability tree. Across 14 models, the aggregate score, and even the knowledge type, hides the structure of ability: accuracy varies by more than 30 points across capability leaves within each frontier model, the three frontier models fail on largely the same leaves, and medical-specialized models redistribute competence rather than closing the gap. Several considerations temper these claims. Images are web-collected, so contamination remains possible, although newly written questions mitigate it and a held-out split will be kept at release; we report no human baseline; CARE-Med is Chinese-language by design; and our domain-novelty claims are scoped to the macroscopic branch of the taxonomy. CARE-Med thus makes evaluation diagnostic: not only whether a model fails, but where.

AI use statement

Generative AI models are a component of the research itself and were also used as research assistants. *As research components:* multimodal LLMs produced image captions, consolidated ground-truth descriptions, generated candidate questions with answer keys and capability labels, screened

candidate questions, and served as the difficulty probe during item selection (Section 3.4); every question in the final benchmark, with its answer and capability labels, was then verified and corrected by clinicians. An LLM also serves as the judge for fill-in-the-blank answers (Section 4), and a vision model classifies prior-benchmark images for the coverage analysis, audited by two human annotators (Section 3.2). *As research assistants:* we used generative AI tools to give feedback on the methodology and evaluation design, to implement data-processing, evaluation, and analysis code, to help interpret results, to translate example items into English, to create figures, to search and summarize related literature, and to draft and edit parts of the paper. We have reviewed all AI-assisted work: code outputs were checked against the raw evaluation records, all reported numbers were recomputed from those records, and all citations were checked against the cited works. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

CARE-Med is built from medical images drawn from public web image–text collections and from figures in published medical books and literature. Because such images may carry licensing and privacy implications, we will release question text, answers, capability labels, and taxonomy annotations through an access-controlled process in which requesters state their intended use and agree to the terms of use, and we will distribute images only where their licence permits redistribution; otherwise we will provide source references. Released images are screened for personally identifying content, and we will maintain a takedown process. The benchmark is intended for evaluating and improving medical AI systems; it is not a diagnostic tool, and its questions and labels should not be used for clinical decision-making.

Reproducibility statement

Section 3 and Appendix A describe the construction pipeline, including every model used at each stage; Section 4 and Appendix B.2 give the answer and judging prompts and the scoring rules; Appendix B.1 lists the evaluated model versions; and Appendices E and F report per-domain, per-leaf, and robustness results with confidence intervals. The capability-tree and image-taxonomy definitions are given in full in Appendices C and D. We will release the benchmark questions with their capability labels, the taxonomy definitions and classification prompt, the per-image coverage assignments for prior benchmarks, and the evaluation code.

References

Mohammad Mahdi Abootorabi, Omid Ghahroodi, Anas Madkoor, Marzia Nouri, Doratossadat Dastgheib, Mohamed Hefeeda, and Ehsaneddin Asgari. Almieyar-Oryx-BloomBench: A Bilingual Multimodal Benchmark for Cognitively Informed Evaluation of Vision-Language Models. arXiv preprint arXiv:2606.05531, 2026.

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. Medical Large Language Model Benchmarks Should Prioritize Construct Validity. arXiv preprint arXiv:2503.10694, 2025.

Lorin W. Anderson and David R. Krathwohl. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, 2001.

Zhijie Bao, Fangke Chen, Licheng Bao, Chenhui Zhang, Wei Chen, Jiajie Peng, and Zhongyu Wei. MedRCube: A Multidimensional Framework for Fine-Grained and In-Depth Evaluation of MLLMs in Medical Imaging. arXiv preprint arXiv:2604.13756, 2026.

Mahtab Bigverdi, Wisdom O. Ikezogwo, Kevin Zhang, Hyewon Jeong, Mingyu Lu, Sungjae Cho, Linda G. Shapiro, and Ranjay Krishna. MedBLINK: Probing Basic Perception in Multimodal Language Models for Medicine. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1234–1244, 2025. doi: 10.1109/ICCVW69036.2025.00132.

James Burgess, Jeffrey J Nirschl, Laura Bravo-Sánchez, Alejandro Lozano, Sanket Rajan Gupte, Jesus G. Galaz-Montoya, Yuhui Zhang, Yuchang Su, Disha Bhowmik, Zachary Coman, Sarina M. Hasan, Alexandra Johannesson, William D. Leineweber, Malvika G Nair, Ridhi Yarlagadda, Connor Zuraski, Wah Chiu, Sarah Cohen, Jan N. Hansen, Manuel D Leonetti, Chad Liu, Emma Lundberg, and Serena Yeung-Levy. MicroVQA: A Multimodal Reasoning Benchmark for Microscopy-Based Scientific Research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19553–19564, 2025. doi: 10.1109/CVPR52734.2025.01821. arXiv:2503.13399.

Zhenyang Cai, Junying Chen, Rongsheng Wang, Weihong Wang, Yonglin Deng, Dingjie Song, Yize Chen, Zixu Zhang, and Benyou Wang. Exploring Compositional Generalization of Multimodal LLMs for Medical Imaging. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 13057–13079, 2025. doi: 10.18653/v1/2025.acl-long.639. arXiv:2412.20070.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Junjun He, and Yu Qiao. GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, pp. 94327–94427, 2024. doi: 10.52202/079017-2992. arXiv:2408.03361.

Wenting Chen, Guo Yu, Yiu-Fai Cheung, Meidan Ding, Jie Liu, Zizhan Ma, Wenxuan Wang, and Linlin Shen. Beyond the Leaderboard: Rethinking Medical Benchmarks for Large Language Models. arXiv preprint arXiv:2508.04325, 2025.

Zhongying Deng, Cheng Tang, Ziyan Huang, Jiashi Lin, Ying Chen, Junzhi Ning, Chenglong Ma, Jiyao Liu, Wei Li, Yinghao Zhu, Shujian Gao, Yanyan Huang, Sibo Ju, Yanzhou Su, Pengcheng Chen, Wenhao Tang, Tianbin Li, Haoyu Wang, Yuanfeng Ji, Hui Sun, Shaobo Min, Liang Peng, Feilong Tang, Haochen Xue, Rulin Zhou, Chaoyang Zhang, Wenjie Li, Shaohao Rui, Weijie Ma, Xingyue Zhao, Yibin Wang, Kun Yuan, Zhaohui Lu, Shujun Wang, Jinjie Wei, Lihao Liu, Dingkang Yang, Lin Wang, Yulong Li, Haolin Yang, Yiqing Shen, Lequan Yu, Xiaowei Hu, Yun Gu, Yicheng Wu, Benyou Wang, Minghui Zhang, Angelica I. Aviles-Rivero, Qi Gao, Hongming Shan, Xiaoyu Ren, Fang Yan, Hongyu Zhou, Haodong Duan, Maosong Cao, Shanshan Wang, Bin Fu, Xiaomeng Li, Zhi Hou, Chunfeng Song, Lei Bai, Yuan Cheng, Yuandong Pu, Xiang Li, Wenhai Wang, Hao Chen, Jiaxin Zhuang, Songyang Zhang, Huiguang He, Mengzhang Li, Bohan Zhuang, Zhian Bai, Rongshan Yu, Liansheng Wang, Yukun Zhou, Xiaosong Wang, Xin Guo, Guanbin Li, Xiangru Lin, Dakai Jin, Mianxin Liu, Wenlong Zhang, Qi Qin, Conghui He, Yuqiang Li, Ye Luo, Nanqing Dong, Jie Xu, Wenqi Shao, Bo Zhang, Qiujuan Yan, Yihao Liu, Jun Ma, Zhi Lu, Yuewen Cao, Zongwei Zhou, Jianming Liang, Shixiang Tang, Qi Duan, Dongzhan Zhou, Chen Jiang, Yuyin Zhou, Yanwu Xu, Jiancheng Yang, Shaoting Zhang, Xiaohong Liu, Siqi Luo, Yi Xin, Chaoyu Liu, Haochen Wen, Xin Chen, Alejandro Lozano, Min Woo Sun, Yuhui Zhang, Yue Yao, Xiaoxiao Sun, Serena Yeung-Levy, Xia Li, Jing Ke, Chunhui Zhang, Zongyuan Ge, Ming Hu, Jin Ye, Zhifeng Li, Yirong Chen, Yu Qiao, and Junjun He. Project Imaging-X: A Survey of 1000+ Open-Access Medical Imaging Datasets for Foundation Model Development. arXiv preprint arXiv:2603.27460, 2026.

Jinru Ding, Lu Lu, Chao Ding, Mouxiao Bian, Jiayuan Chen, Wenrao Pang, Ruiyao Chen, Xinwei Peng, Renjie Lu, Sijie Ren, Guanxu Zhu, Xiaoqin Wu, Zhiqiang Liu, Rongzhao Zhang, Luyi Jiang, Bing Han, Yunqiu Wang, and Jie Xu. MedBench v4: A Robust and Scalable Benchmark for Evaluating Chinese Medical Language Models, Multimodal Models, and Intelligent Agents. arXiv preprint arXiv:2511.14439, 2025.

Haozhen Gong, Xiaozhong Ji, Yuansen Liu, Wenbin Wu, Xiaoxiao Yan, Jingjing Liu, Kai Wu, Jiazhen Pan, Bailiang Jian, Jiangning Zhang, Xiaobin Hu, and Hongwei Bran Li. Med-CMR: A Fine-Grained Benchmark Integrating Visual Evidence and Clinical Logic for Medical Complex Multimodal Reasoning. arXiv preprint arXiv:2512.00818, 2025.

Yunye Gong, Robik Shrestha, Jared Claypoole, Michael Cogswell, Arijit Ray, Christopher Kanan, and Ajay Divakaran. BloomVQA: Assessing Hierarchical Multi-modal Comprehension. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14905–14918, 2024. doi: 10.18653/v1/2024.findings-acl.885. arXiv:2312.12716.

Hongcheng Guo, Zheyong Xie, Shaosheng Cao, Boyang Wang, Weiting Liu, Anjie Le, Lei Li, and Zhoujun Li. SNS-Bench-VL: Benchmarking Multimodal Large Language Models in Social Networking Services. arXiv preprint arXiv:2505.23065, 2025.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. PathVQA: 30000+ Questions for Medical Visual Question Answering. arXiv preprint arXiv:2003.10286, 2020.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. OmniMed-VQA: A New Large-Scale Comprehensive Evaluation Benchmark for Medical LVLM. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22170–22183, 2024. doi: 10.1109/CVPR52733.2024.02093. arXiv:2402.09181.

Thomas Huber and Christina Niklaus. LLMs Meet Bloom's Taxonomy: A Cognitive View on Large Language Model Evaluations. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 5211–5246, 2025. URL https://aclanthology.org/2025.coling-main.350/.

Juntao Jiang, Jiangning Zhang, Yali Bi, Jinsheng Bai, Weixuan Liu, Weiwei Jin, Zhucun Xue, Yong Liu, Xiaobin Hu, and Shuicheng Yan. M3CoTBench: Benchmark Chain-of-Thought of MLLMs in Medical Image Understanding. arXiv preprint arXiv:2601.08758, 2026.

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: 180251, 2018. doi: 10.1038/sdata.2018.251.

Thomas M. Lehmann, Henning Schubert, Daniel Keysers, Michael Kohnen, and Berthold B. Wein. The IRMA Code for Unique Classification of Medical Images. In *Medical Imaging 2003: PACS and Integrated Medical Information Systems*, volume 5033 of *Proc. SPIE*, pp. 440–451, 2003. doi: 10.1117/12.480677.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, 2024. doi: 10.1109/CVPR52733.2024.01263.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2757–2791, 2025. doi: 10.18653/v1/2025.emnlp-main.138. arXiv:2411.16594.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, and Linlin Shen. WSI-LLaVA: A Multimodal Large Language Model for Whole Slide Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22718–22727, 2025. doi: 10.1109/ICCV51701.2025.02109. arXiv:2412.02141.

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 *IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pp. 1650–1654, 2021. doi: 10.1109/ISBI48211.2021.9434010. arXiv:2102.09542.

Junling Liu, Ziming Wang, Qichen Ye, Dading Chong, Peilin Zhou, and Yining Hua. Qilin-Med-VL: Towards Chinese Large Vision-Language Model for General Healthcare. arXiv preprint arXiv:2310.17956, 2023a.

Junling Liu, Peilin Zhou, Yining Hua, Dading Chong, Zhongyu Tian, Andrew Liu, Helin Wang, Chenyu You, Zhenhua Guo, Lei Zhu, and Michael Lingzhi Li. Benchmarking Large Language Models on CMExam – A Comprehensive Chinese Medical Exam Dataset. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 52430–52452, 2023b. doi: 10.52202/075280-2283. arXiv:2306.03030.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teach Multimodal LLMs to Comprehend Electrocardiographic Images. arXiv preprint arXiv:2410.19008, 2024.

Shengyuan Liu, Boyun Zheng, Wenting Chen, Zhihao Peng, Zhenfei Yin, Jing Shao, Jiancong Hu, and Yixuan Yuan. EndoBench: A Comprehensive Evaluation of Multi-Modal Large Language Models for Endoscopy Analysis. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 2350–2397, 2025. doi: 10.52202/085713-0076. arXiv:2505.23601.

Hajra Malik, Hafiza Tooba Aftab, Abdul Rehman, Mohsen Ali, and Waqas Sultani. Multilingual Hematology Visual Question Answering Dataset. arXiv preprint arXiv:2606.25246, 2026.

Henning Müller, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, and Sameer Antani. Creating a Classification of Image Types in the Medical Literature for Visual Categorization. In *Medical Imaging 2012: Advanced PACS-based Imaging Informatics and Therapeutic Applications*, volume 8319 of *Proc. SPIE*, 2012. doi: 10.1117/12.911186.

Long Phan, Alice Gatti, Ziwen Han, et al. Humanity's Last Exam. *arXiv preprint arXiv:2501.14249*, 2025.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards Building Multilingual Language Model for Medicine. *Nature Communications*, 15, 2024. doi: 10.1038/s41467-024-52417-z. arXiv:2402.13963.

Binesh Sadanandan and Vahid Behzadan. Consistent but Dangerous: Per-Sample Safety Classification Reveals False Reliability in Medical Vision-Language Models. arXiv preprint arXiv:2603.20985, 2026.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. CMB: A Comprehensive Medical Benchmark in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 6184–6205, 2024. doi: 10.18653/v1/2024.naacl-long.343. arXiv:2308.08833.

Yuanlong Wang, Weichi Chen, Adrian Rajab, Wenfang Liu, Yulan Jin, Andrew Srisuwananukorn, and Ping Zhang. PBSBench: A Multi-Level Vision-Language Framework and Benchmark for Hematopathology Whole Slide Image Interpretation. arXiv preprint arXiv:2604.17570, 2026.

Bruce Changlong Xu, Lan Wu, and Alexander Ryu. A Controlled Audit of Pretraining Contamination in Public Medical Vision-Language Benchmarks. arXiv preprint arXiv:2606.10066, 2026.

Zhiling Yan, Dingjie Song, Zhe Fang, Yisheng Ji, Xiang Li, Quanzheng Li, and Lichao Sun. LiveMedBench: A Contamination-Free Medical Benchmark for LLMs with Automated Rubric Evaluation. arXiv preprint arXiv:2602.10367, 2026.

Sau Lai Yip, Sunan He, Yuxiang Nie, Shu Pui Chan, Yilin Ye, Sum Ying Lam, and Hao Chen. MedBookVQA: A Systematic and Comprehensive Medical Benchmark Derived from Open-Access Book. arXiv preprint arXiv:2506.00855, 2025.

Peiwen Yuan, Yiwei Li, Shaoxiong Feng, Xinglin Wang, Yueqi Zhang, Jiayi Shi, Chuyi Tan, Boyuan Pan, Yao Hu, and Kan Li. Silencer: From Discovery to Mitigation of Self-Bias in LLM-as-Benchmark-Generator. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 143773–143795, 2025. doi: 10.52202/085713-4317. arXiv:2505.20738.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, et al. MMMU-Pro: A More Robust Multi-discipline Multimodal Understanding Benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15134–15186, 2025. doi: 10.18653/v1/2025.acl-long.736.

Geng Zhang, Yizhou Ying, Sihang Jiang, Jiaqing Liang, Guanglei Yue, Yifei Fu, Hailin Hu, and Yanghua Xiao. From Remembering to Metacognition: Do Existing Benchmarks Accurately Evaluate LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13440–13457, 2025. doi: 10.18653/v1/2025.findings-emnlp.724. URL https://aclanthology.org/2025.findings-emnlp.724/.

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. arXiv preprint arXiv:2305.10415, 2023.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding. In *Proceedings of the 42nd International Conference on Machine Learning (ICML), PMLR 267*, pp. 80961–80990, 2025. arXiv:2501.18362.

A Construction Details

Table 2 lists, for each stage, who or what performs it in the two construction phases, and Table 3 gives the per-domain item counts. Both phases use the same three caption models and the same question-generation constraints (Section 3.4).

Table 2: Stage-by-stage construction in the two phases. Phase 1 covers the first 544 items of the four photographic domains and all 423 physiological-signal items; Phase 2 covers the remaining 1,056 photographic-domain items.

Stage	Phase 1 (967 items)	Phase 2 (1,056 items)
Sourcing	Web image–text collections and figures from medical books and literature; taxonomy classification; an LLM removes non-medical images and checks image–text consistency	Same sources; an LLM scores image–text relevance and only fully relevant pairs are kept
Image screening	Clinicians: medical relevance, quality, answerability from the image alone	LLM relevance filter (above)
Captioning	GPT-5.4, Gemini-2.5-Pro, Seed-1.8, image only	Same
Ground truth	Content on which the three captions agree is kept; clinicians check and correct every disagreement against the image	LLM claim-level consolidation: decompose captions into atomic claims, verify each against the image, keep image-supported claims
Question generation	LLM with the structured prompt (image-dependence, grounding, leaf mapping, hallucination-targeted distractors)	GPT-5.4 with the same prompt, which requires at least one procedural leaf among the questions generated for each image
Question screening	Clinicians	LLM screen; items are removed for an unsupported or non-unique key, a medical error, an unsafe recommendation, text-only answerability, or internal ambiguity
Selection	Clinician curation	Greedy balancing of format, leaves, answer patterns, and image sources, preferring items the probe (Seed-1.6, not evaluated) answers incorrectly
Final verification	Clinicians review every item: stem, options, answer key, capability labels	Same

Table 3: Items per domain and construction phase. Only Phase-2 items were selected with the difficulty probe.

Domain	Phase 1	Phase 2	Total
Body fluids	64	336	400
IVD reagents	190	210	400
Endoscopy & surgical fields	130	270	400
Ex-vivo organs & tissues	160	240	400
Physiological-signal visualizations	423	0	423
Total	967	1,056	2,023

Phase-2 question screen. The LLM screen receives the image, the ground-truth description, and the generated item (stem, options, key, capability mapping, and the generator's rationale). It checks core validity (image dependence, evidence for the key, uniqueness of the answer, safety of procedural recommendations), distractor fairness (each wrong option clearly wrong and not overlapping the key), and protects hard items that require integrating visual facts with procedural knowledge from being removed merely for difficulty. Items with any hard failure are removed; items with minor issues (wording, distractor overlap, option parallelism, leaf-mapping mismatch) are retained for the clinician review, which corrects them.

Difficulty-targeted selection. For each domain, a greedy selector first covers capability leaves not yet represented, then brings each available leaf to a minimum support, and finally fills the quota; its score combines format quotas, a target share of probe-incorrect items, the screen's difficulty estimate, long-tail leaves, per-image and per-source caps, and caps on repeated answer patterns. Phase-2 selection keeps at most one new question per image in the body-fluid domain and caps questions per image elsewhere.

B Evaluation Details

B.1 Evaluated Models

Table 4: Evaluated models. *Access*: closed (API only) or open weights. *Tier*: for each closed vendor we evaluate the current flagship and its immediately preceding version; open-weight models are general-purpose or medical-specialized. *Identifier*: the API model name (with snapshot date where the provider exposes one) or the Hugging Face repository id; —: no public snapshot identifier. † Preceding-version model that also contributed ground-truth captions. ‡ Medical-specialized model. Params: nominal size from the public release; n/d = not disclosed.

Model	Access	Tier	Identifier	Params
GPT-5.5	Closed	Current flagship	`gpt-5.5-2026-04-24`	n/d
GPT-5.4†	Closed	Preceding version	`gpt-5.4-2026-03-05`	n/d
Gemini-3.1-Pro	Closed	Current flagship	`gemini-3.1-pro-preview`	n/d
Gemini-2.5-Pro†	Closed	Preceding version	`gemini-2.5-pro`	n/d
Seed-2.1	Closed	Current flagship	`---`	n/d
Seed-1.8†	Closed	Preceding version	`---`	n/d
Lingshu-32B‡	Open	Medical open	`lingshu-medical-mllm/Lingshu-32B`	32B
MedGemma-27B‡	Open	Medical open	`google/medgemma-27b-it`	27B
Citrus-V-8B‡	Open	Medical open	`jdh-algo/Citrus-V-8B-v1.0`	8B
HuatuoGPT-V-34B‡	Open	Medical open	`FreedomIntelligence/HuatuoGPT-Vision-34B-hf`	34B
Qwen3-VL-8B	Open	General open	`Qwen/Qwen3-VL-8B-Instruct`	8B
InternVL3.5-8B	Open	General open	`OpenGVLab/InternVL3_5-8B`	8B
GLM-4.6V-Flash	Open	General open	`zai-org/GLM-4.6V-Flash`	9B
MiniCPM-V-4.5	Open	General open	`openbmb/MiniCPM-V-4_5`	8B

B.2 Answer Instructions and Judge Prompts

Answer generation. Each item is sent to the model under test as a single user turn (no system prompt) that contains the image and a text prompt made of the question stem, a format instruction and, for multiple-select items, the lettered options. The two format instructions are shown in Figure 6 (English translations of the Chinese originals).

Scoring and the judge. Multiple-select answers are scored deterministically: the predicted and reference answers are upper-cased, separators and punctuation are removed, and the resulting sets of option letters must be identical. Only when the option letters cannot be parsed (e.g., because the answer contains prose) is an LLM judge called with the fallback prompt (Figure 7); across the

```
[Multiple-select]
{question stem}
Output only the letter(s) of the correct option(s), in uppercase;
concatenate multiple letters directly, without any separator, space,
punctuation or explanation.
{options A, B, C, ..., one per line}

[Fill-in-the-blank]
{question stem}
Output only the final answer text (a phrase or a single sentence is
enough); no explanation, no reasoning process, no numbering, no extra
punctuation.
```

Figure 6: Answer-generation prompt. The format instruction is appended to every item; the image is attached to the same user turn.

14 models this occurred for 10.1% of multiple-select model–item pairs (1,502 of 14,840). Fill-in-the-blank answers are always scored by the LLM judge with the conservative open-answer prompt (Figure 8). The judge sees only the question text (plus the option list in the multiple-select fallback), the reference answer and the model's answer; it sees neither the image nor the identity of the answering model. It must output `1` (correct) or `0` (incorrect), with deterministic decoding (temperature 0, top-p 1, at most 8 output tokens); responses without a parseable verdict are re-queried. The prompts were issued in Chinese; faithful English renderings are given in Figures 7 and 8.

```
You are a strict grader of medical examinations.
You will see [Question], [Options], [Reference answer] and [Model
answer].
Your task is not to solve the question again, but to judge whether the
model answer gives a final answer that is exactly identical to the
reference answer.

Grading rules:
1.  Consider only the option letters A-I.
2.  Ignore order, spaces and all separators/punctuation.
3.  If the model answer does not clearly state final options (e.g., it
only analyses, restates the question, discusses reasons, expresses
uncertainty, or gives mutually contradictory options), score 0.
4.  Missing options, extra options or superfluous options all score 0.
5.  [Question] and [Options] serve only to help identify whether the
model gave an explicit answer; do not use them to infer, on the model's
behalf, the answer it should have chosen.
6.  Even if the model's reasoning looks sound, score 0 unless its final
options are exactly identical to the reference answer.

[Question]
{question}
[Options]
{options}
[Reference answer]
{reference}
[Model answer]
{prediction}

Output only 1 or 0:  output 1 if exactly identical, otherwise output 0.
```

Figure 7: Judge prompt A: multiple-select fallback, used only when exact letter-set matching cannot parse the option letters.

```
You are a strict grader of medical examinations.
You will see [Question], [Reference answer] and [Model answer].
Your task is not to answer the question again, but to judge whether the
model answer is medically equivalent to the reference answer in the
context of this question.

Core principles:
1.  The reference answer takes precedence.  [Question] is used only to
understand the context and resolve references; do not use the question
stem to fill in information the model omitted.
2.  The model answer must cover the necessary information in the
reference answer.  If it omits a key entity, site, nature, degree,
quantity, condition, negation, boundary restriction, next action,
examination method, order of management or risk-control requirement,
score 0.
3.  Be strict for medical questions:  if the model answer gives only a
more generic direction or boilerplate such as "correlate clinically /
further examination / continue evaluation" without stating the key
medical point of the reference answer, score 0.
4.  If the reference answer contains several necessary points, missing
any one of them scores 0.
5.  Allow synonymous rephrasing, abbreviations vs. full names, common
acronyms, interchange of Chinese and English terms, and differences in
units or symbols.  If the reference answer explicitly accepts any value
within a range, a valid value within that range may score 1.
6.  If the model answer adds information that conflicts with the
reference answer, is self-contradictory, or changes the medical meaning,
score 0.
7.  Be especially strict for procedural-knowledge questions:  if the
reference answer contains specific actions, examination names or an
order of steps (such as "re-check, monitor, send for testing, culture,
biochemistry, imaging confirmation, consultation, stop/continue,
follow-up"), and the model answer generalizes them, omits steps, changes
their order or keeps only vague advice, score 0.

[Question]
{question}
[Reference answer]
{reference}
[Model answer]
{prediction}

Output only 1 or 0:  output 1 if medically equivalent, otherwise output
0.
```

Figure 8: Judge prompt B: fill-in-the-blank items (conservative medical-equivalence judge).

C CAPABILITY TREE

Table 5: The full CARE-Med capability tree: 4 levels, 3 knowledge types, 9 areas, 20 groups and 34 leaves (13 factual, 10 conceptual, 11 procedural). *Type*: F = factual, C = conceptual, P = procedural (level 1). *Area ▷ Group*: level-2 and level-3 nodes. Leaf names are those used in Table 11; definitions are translated from the capability-tree specification given to the question generator. n: number of items labelled with the leaf. Items may carry several leaves (mean 3.16 leaves per item; 6,398 item–leaf labels over 2,023 items), so n sums to more than the number of items. Leaves with $n < 30$ (IDs 23, 28, 34) are reported but excluded from leaf-range statistics.

ID	Leaf	Type	Area ▷ Group	Definition	n
Factual knowledge — correctly recall and distinguish the terms and definitions tied to a medical image (13 leaves)					
1	Entity recognition	F	Basic definitions & terminology ▷ Entities & structures	Accurately identify medical-device components, specific organs, or other typical entities in the image.	393
2	Spatial localization	F	Basic definitions & terminology ▷ Entities & structures	Identify the position of, and spatial relations among, organs/devices, vessels and nerves (e.g., liver segments, lung lobes and segments, spinal levels; anatomical orientation).	161
3	Dynamic-sign reading	F	Basic definitions & terminology ▷ Processes & states	Recognize a dynamic physiological or pathological moment captured in a static image (e.g., active bleeding, peristalsis).	33
4	Legend & scale reading	F	Basic definitions & terminology ▷ Abstract concepts	Interpret the measurement marks carried by the image itself, such as dashed measurement lines, crosshair cursors and asterisks.	380
5	Signage/warning reading	F	Basic definitions & terminology ▷ Abstract concepts	Recognize special symbols, colour bands and hazard/warning signs in the medical environment.	75
6	Modality identification	F	Basic definitions & terminology ▷ Abstract concepts	Identify the imaging principle, sequence or stain of the image (e.g., contrast-enhanced CT, FLAIR, H&E staining).	191
7	Parameter extraction	F	Data & facts ▷ Quantitative data	Extract machine-printed parameters at the image margin (e.g., slice thickness, window width/level) or graduations on tubes and measuring cups.	407
8	Geometric measurement	F	Data & facts ▷ Quantitative data	Use an in-image ruler or grid to measure objective quantities such as lesion long-axis diameter, cardiothoracic ratio or ECG rate.	326
9	Appearance description	F	Data & facts ▷ Qualitative facts	Accurately describe qualitative sensory facts that cannot be precisely quantified, such as turbidity, colour, or a moist/dry appearance.	405
10	Exam-type classification	F	Classification & standards ▷ Typological classification	Classify the image at the macroscopic physical level (e.g., nuclear medicine, ultrasound, gross pathology, endoscopic photograph, IVD reagent) by imaging principle, body site or clinical use.	230
11	Lesion-morphology typing	F	Classification & standards ▷ Typological classification	Apply a classical morphological classification from appearance such as solid vs. cystic or fracture displacement (e.g., Garden type).	205
12	Report categorization	F	Classification & standards ▷ Typological classification	Classify a report by report type, content, terminology and quality-control grade.	114

continued on next page

Table 5 (continued)

ID	Leaf	Type	Area ▷ Group	Definition	n
13	Severity grading	F	Classification & standards ▷ Hierarchical grading	Map the depth or extent of a lesion to a severity grade (mild/moderate/severe) or a TNM stage.	118
Conceptual knowledge — explain the internal logic and mechanisms contained in the image (10 leaves)					
14	Micro-mechanism	C	Mechanism understanding ▷ Molecular/cellular mechanisms	Explain physiological or pathological activity at the cellular or molecular level (e.g., proliferation driven by a gene mutation).	38
15	Inorganic-reaction mechanism	C	Mechanism understanding ▷ Molecular/cellular mechanisms	Explain the reaction mechanism of non-living matter (e.g., dye staining, calcium deposition, a reagent colour change).	68
16	Macro-mechanism	C	Mechanism understanding ▷ Organ/system mechanisms	Explain responses at the organ or system level (e.g., splenomegaly and varices caused by portal hypertension).	149
17	Artifact attribution	C	Knowledge association & differentiation ▷ Causes/risk factors	Recognize metal or motion artifacts, over-exposure or staining distortion, and trace them to the underlying physical or operational cause.	71
18	Causal association	C	Knowledge association & differentiation ▷ Causes/risk factors	From an imaging feature, infer the most likely causal or associated factor (e.g., extensive consolidation $\rightarrow$ pneumonia).	443
19	Look-alike discrimination	C	Knowledge association & differentiation ▷ Comparison & differentiation	Differentiate highly similar-looking lesions (same disease, different appearance; different diseases, same appearance) from subtle differences in margin, enhancement or colour.	558
20	Cross-modal mapping	C	Knowledge association & differentiation ▷ Comparison & differentiation	From a gross or CT image, infer how the lesion would appear under another modality (e.g., microscopy, MRI, PET).	69
21	Imaging principle	C	Principle explanation ▷ Basic scientific principles	Explain why an examination produces the observed appearance (e.g., high density $\leftrightarrow$ calcification; ultrasound reflection).	33
22	Technique application	C	Principle explanation ▷ Basic scientific principles	Explain why a given examination is preferred for a condition (e.g., X-ray first for fractures; ultrasound in children to avoid radiation).	110
23	Document-compliance check	C	Principle explanation ▷ Ethics/communication/methodology	Check whether signatures, alterations and layout of prescriptions, consent forms and reports comply with regulations.	12
Procedural knowledge — interpret the standard workflows, methods and protocols contained in the image (11 leaves)					
24	Structured reading	P	Workflows & steps ▷ Diagnostic & assessment workflow	Extract positive findings strictly in the standard reading order (e.g., "outside-in"; "locate–characterize–quantify").	325
25	Disposition / next-step	P	Workflows & steps ▷ Treatment & management workflow	Recognize an emergency or pathological situation in the image and decide the specific clinical next step.	504
26	Operation-technique reading	P	Methods & techniques ▷ Operating steps & methods	Identify the operating gesture, probe placement or patient positioning shown in the image and its significance.	48

continued on next page

Table 5 (continued)

ID	Leaf	Type	Area ▷ Group	Definition	n
27	Image-quality assessment	P	Methods & techniques ▷ Assessment methods & criteria	Assess whether noise, artifacts or occlusion push the image below the threshold required for clinical diagnosis.	117
28	Response comparison	P	Methods & techniques ▷ Assessment methods & criteria	Compare images before and after treatment using RECIST or device displacement to assess response or regression.	8
29	Waveform/ spectrum reading	P	Methods & techniques ▷ Data interpretation & analysis	Read amplitude, baseline shift and frequency variation of ECG/EEG traces on a grid.	105
30	Data/chart interpretation	P	Methods & techniques ▷ Data interpretation & analysis	Interpret the core medical meaning of the data in a chart or table.	179
31	Procedure-step inference	P	Standard protocols & guidelines ▷ Standard operating procedures	Infer the next standard step from the visual logic of a flowchart or algorithm diagram.	85
32	Guideline grading	P	Standard protocols & guidelines ▷ Clinical practice guidelines	Apply a guideline to lesion features (e.g., BI-RADS, TI-RADS category) to assess risk and recommend an intervention.	85
33	Care-pathway planning	P	Standard protocols & guidelines ▷ Clinical practice guidelines	Combine anatomy, stage and severity to derive a complete, guideline-concordant diagnostic and treatment pathway.	343
34	Treatment-strategy adjustment	P	Standard protocols & guidelines ▷ Clinical practice guidelines	From the dynamic change of a lesion, decide the guideline-required path (continue follow-up vs. change regimen).	10

D IMAGE TAXONOMY, COVERAGE, AND POSITIONING

Table 6: The self-built medical-image taxonomy used to define CARE-Med's domains and to measure the coverage of prior benchmarks: 4 levels, 3 top-level categories, 10 level-2 groups, 30 level-3 subclasses and 70 leaves (21 in category 1, 25 in category 2, 24 in category 3). Each row is a level-3 subclass; its level-4 leaves are listed with their codes. A subclass without level-4 children is itself a leaf (*terminal*). Parenthesized text gives representative contents. ★ marks CARE-Med's five target level-3 nodes: 1.1.2 (device-assisted view = *Endoscopy & surgical fields*), 1.2.1 (*in-vitro diagnostic (IVD) reagents*), 1.4.1 (*Ex-vivo organs & tissues*), 1.4.2 (*Body fluids*) and 3.2.1 (*Physiological-signal visualizations*). #: number of leaves under the subclass.

Code	Group (L2)	Subclass (L3)	Leaves (L4)	#
Category 1: Optical photography — ordinary optical camera or endoscope; macroscopically visible (21 leaves)				
1.1.1	Physiological structures	Visible on inspection	1.1.1.1 abnormal lesions visible on inspection (skin, mucosal and limb lesions on the body surface); 1.1.1.2 normal structures visible on inspection (normal surface anatomy)	2
1.1.2	Physiological structures	★ Device-assisted view (*Endoscopy & surgical fields*)	1.1.2.1 device-visible abnormal lesions (lesions seen through endoscopes or in the surgical field); 1.1.2.2 device-visible normal structures (ENT endoscopy, bronchoscopy, GI endoscopy, capsule endoscopy, laparoscopy, hysteroscopy, arthroscopy, surgical field, etc.)	2

continued on next page

Table 6 (continued)

Code	Group (L2)	Subclass (L3)	Leaves (L4)	#
1.2.1	Drugs & devices	★ Medical devices: IVD reagents only (*IVD reagents*)	1.2.1.1 Class I (low-risk reagents such as diluents and stains); 1.2.1.2 Class II (moderate-risk reagents such as pregnancy tests, blood-glucose strips, *H. pylori* test cards); 1.2.1.3 Class III (high-risk reagents such as COVID-19 antigen tests, HIV self-tests, tumour-marker assays)	3
1.2.2	Drugs & devices	Drugs	1.2.2.1 biologics (blood products, vaccines, cell and gene therapies, recombinant proteins); 1.2.2.2 chemical drugs (oral, injectable, topical, aerosol forms); 1.2.2.3 traditional Chinese medicines	3
1.3.1	Medical scenes & activities	Clinical interaction (history taking, palpation, auscultation)	*(terminal: level-3 leaf)*	1
1.3.2	Medical scenes & activities	Procedures (surgical overview, nursing procedures, rehabilitation training)	*(terminal: level-3 leaf)*	1
1.3.3	Medical scenes & activities	Care settings (consulting room, ward, operating room)	*(terminal: level-3 leaf)*	1
1.4.1	Ex-vivo biological samples	★ Organs & tissues (*Ex-vivo organs & tissues*)	1.4.1.1 whole organs (ex-vivo organs with an intact capsule or boundary: parenchymal, hollow, glandular, supporting/circulatory); 1.4.1.2 anatomical regions (organ parts with identifiable regional landmarks); 1.4.1.3 resected tissue (tissue lacking an organ outline: masses and nodules, curettings, biopsy cores, fragments)	3
1.4.2	Ex-vivo biological samples	★ Body fluids (*Body fluids*)	1.4.2.1 natural physiological fluids (urine, blood, sputum, saliva, semen, breast milk, etc.); 1.4.2.2 puncture, drainage and lavage fluids (pleural and peritoneal effusions, cerebrospinal fluid, lavage fluid, spent dialysate, etc.); 1.4.2.3 secretions (cervical/vaginal discharge, wound exudate, pus, etc.)	3
1.4.3	Ex-vivo biological samples	Animal samples	1.4.3.1 parasites; 1.4.3.2 laboratory animals	2
Category 2: Medical imaging & microscopy — sensor-reconstructed; not directly visible to the eye (25 leaves)				
2.1.1	Radiology	Projection imaging	2.1.1.1 plain radiography (chest, bone and joint, mammography, dental panoramic); 2.1.1.2 X-ray contrast studies (DSA, coronary angiography, barium meal, urography, hysterosalpingography, angiography)	2
2.1.2	Radiology	Tomography	2.1.2.1 non-contrast CT (incl. CBCT); 2.1.2.2 contrast-enhanced CT (incl. CT perfusion); 2.1.2.3 CTA; 2.1.2.4 3D reconstruction	4
2.1.3	Radiology	Magnetic resonance imaging	2.1.3.1 MRI (DWI, MRS, MRCP, CMR, etc.); 2.1.3.2 contrast-enhanced MRI; 2.1.3.3 MRA	3
2.1.4	Radiology	Nuclear medicine	2.1.4.1 PET-CT; 2.1.4.2 SPECT; 2.1.4.3 PET-MRI	3
2.2.1	Acoustic, optical & electrical imaging	Ultrasound	2.2.1.1 B-mode; 2.2.1.2 colour Doppler; 2.2.1.3 3D ultrasound; 2.2.1.4 4D ultrasound	4
2.2.2	Acoustic, optical & electrical imaging	Optical imaging	2.2.2.1 optical-technique imaging (OCT, fluorescence); 2.2.2.2 optical magnification (slit lamp, ophthalmoscope, dermoscope)	2

continued on next page

Table 6 (continued)

Code	Group (L2)	Subclass (L3)	Leaves (L4)	#
2.2.3	Acoustic, optical & electrical imaging	Other modalities	2.2.3.1 thermography (infrared thermal imaging); 2.2.3.2 electrical impedance imaging	2
2.3.1	Tissues & cells	Histopathology	2.3.1.1 routine staining (H&E, intra-operative frozen section); 2.3.1.2 special staining (immunohistochemistry, immunofluorescence, Masson, PAS, Congo red, etc.)	2
2.3.2	Tissues & cells	Cytology	2.3.2.1 routine cytology (blood smear, bone-marrow smear, exfoliative cytology, liquid-based cytology (TCT/TCL), etc.); 2.3.2.2 special cytology (ICC/IF, in-situ hybridization, multiplex immunofluorescence, imaging flow cytometry, etc.)	2
2.3.3	Tissues & cells	Microbiological microscopy (bacteria, fungi, spirochetes, parasites, viruses, etc.)	*(terminal: level-3 leaf)*	1
Category 3: Documents & drawings — manually processed, drawn or transcribed; not directly photographed (24 leaves)				
3.1.1	Medical drawings & models	Physiological structures	3.1.1.1 organ schematics; 3.1.1.2 tissue schematics; 3.1.1.3 cell schematics; 3.1.1.4 microorganism schematics; 3.1.1.5 protein/molecule schematics; 3.1.1.6 mechanism diagrams	6
3.1.2	Medical drawings & models	Drugs & devices	3.1.2.1 medical devices; 3.1.2.2 drugs	2
3.1.3	Medical drawings & models	Medical scenes & activities	3.1.3.1 clinical interaction; 3.1.3.2 procedures; 3.1.3.3 care settings	3
3.1.4	Medical drawings & models	Ex-vivo biological samples	3.1.4.1 body fluids; 3.1.4.2 animal samples	2
3.2.1	Medical records & data	★ Physiological-signal visualization (*Physiological-signal visualizations*)	3.2.1.1 waveforms (ECG, EEG, EMG, fetal heart-rate monitoring, pulse-oximetry waveform, etc.); 3.2.1.2 curves (pulmonary function, audiogram, ambulatory blood pressure, uroflowmetry, etc.); 3.2.1.3 maps/topographs (EEG topography, corneal topography, body-composition analysis, etc.)	3
3.2.2	Medical records & data	Medical statistical charts (epidemiological and demographic charts; line, scatter and histogram plots, etc.)	*(terminal: level-3 leaf)*	1
3.2.3	Medical records & data	Medical forms & documents	3.2.3.1 tabular data (laboratory reports, duty rosters); 3.2.3.2 semi-structured forms (prescriptions, package inserts, payment slips); 3.2.3.3 unstructured text (medical records, informed-consent forms); 3.2.3.4 mixed (charts combined with text)	4
3.3.1	Medical signage	Environmental signs (e.g., sterile zone)	*(terminal: level-3 leaf)*	1
3.3.2	Medical signage	Warning signs (e.g., radiation warning signs)	*(terminal: level-3 leaf)*	1

continued on next page

Table 6 (continued)

Code	Group (L2)	Subclass (L3)	Leaves (L4)	#
3.3.3	Medical signage	Identification markers (e.g., inpatient wristbands)	*(terminal: level-3 leaf)*	1
Total (3 categories, 10 groups, 30 subclasses)				**70**

Table 7 gives the per-benchmark statistics behind Figure 3, and Table 8 positions CARE-Med against representative medical MLLM benchmarks.

Table 7: Taxonomy-based coverage of prior medical benchmarks (Seed-1.8 classifier, image-only input). Columns: unique images; image share per top-level category (Cat.1 optical photography, Cat.2 medical imaging & microscopy, Cat.3 documents & drawings); non-medical/other share; number of the 70 taxonomy leaves populated; and the three most concentrated leaves (% of the benchmark's images). Category 2 dominates every general/radiology benchmark; the seven together leave 16 leaves—including CARE-Med targets—uncovered.

Benchmark	#Img	Cat.1 (%)	Cat.2 (%)	Cat.3 (%)	Other (%)	Leaves /70	Top-3 concentrated leaves (% of images)
OmniMedVQA	82,059	2.1	97.8	0.0	0.1	41	MRI 36, CT 18, ultrasound 13
GMAI-MMBench	4,550	16.6	82.5	0.4	0.5	32	CT 16, MRI 16, magnified-optical 11
PMC-VQA	1,440	3.9	86.9	3.7	5.6	42	MRI 20, CT 13, X-ray 9
SLAKE	180	0.0	100.0	0.0	0.0	5	CT 34, X-ray 31, MRI 29
VQA-RAD	451	0.4	99.6	0.0	0.0	6	X-ray 39, CT 27, MRI 24
PathVQA	6,719	44.0	52.2	3.6	0.2	30	H&E 36, resected tissue 21, whole organ 12
EndoBench	6,303	98.4	1.6	0.0	0.0	11	endoscopy-abn. 59, endoscopy-normal 26, procedure 13
Union (7)	—	—	—	—	—	**54**	Cat.2 saturated (25/25); Cat.3 only 13/24; 16 leaves uncovered

Table 8: Positioning of CARE-Med against representative medical MLLM benchmarks. *F/C/P attribution*: every question carries one or more leaves of a factual/conceptual/procedural knowledge-type capability tree. *Fluids/IVD/ex-vivo*: dedicated coverage of macroscopic body-fluid appearance, in-vitro diagnostic reagent readouts, and gross ex-vivo organs and tissues. ✓ yes; ∼ partial or incidental; × no.

Benchmark	Lang.	Organizing axis	F/C/P attribution	Fluids / IVD / ex-vivo (gross)	Signal visualiz.	Fill-in-the-blank
OmniMedVQA (Hu et al., 2024)	EN	Modality (flat)	×	×	×	×
GMAI-MMBench (Chen et al., 2024)	EN	Clinical task & modality	×	×	×	×
MedXpertQA (Zuo et al., 2025)	EN	Body system, reasoning	×	×	∼	×
EndoBench (Liu et al., 2025)	EN	Endoscopy tasks (1 domain)	×	×	×	×
MedBench v4 (Ding et al., 2025)	ZH	Clinical task (rad./path.)	×	×	×	∼
MedBookVQA (Yip et al., 2025)	EN	Modality / anatomy / dept.	×	×	∼	×
CARE-Med (ours)	ZH	**Knowledge-type tree (F/C/P)**	✓	✓	✓	✓

E Additional Results

This section reports the image-dependence control (Table 9), per-domain accuracy (Table 10), the knowledge-type versus leaf-range summary (Figure 9), the format comparison and no-image drop (Figure 10), and the full per-leaf matrix (Table 11).

F Robustness Analyses

All analyses use the per-item evaluation records of the 14 models (2,023 items) and the three no-image runs; bootstrap intervals use 10,000 item-level resamples (seed 0).

Confidence intervals and pairwise tests. Table 13 gives 95% bootstrap intervals for every model and format, and Table 14 gives paired exact McNemar tests for each flagship against its predecessor and for adjacent leaderboard positions.

Table 9: Image-dependence control. Withholding the image drops the three flagships by 21.5–23.3 points overall; the drop is concentrated in the fill-in-the-blank format (-33 to -35 points), while multiple-select retains more text-answerability.

Flagship	With image	No image	Δ
GPT-5.5	71.1	49.6	-21.5
Gemini-3.1-Pro	63.3	40.1	-23.1
Seed-2.1	68.4	45.0	-23.3
Fill-in-the-blank only (with image → no image):			
GPT-5.5	64.3	29.0	-35.3
Gemini-3.1-Pro	58.6	24.9	-33.6
Seed-2.1	62.6	29.5	-33.1

Table 10: Per-domain accuracy (%). Fluids = body fluids; IVD = in-vitro diagnostic reagents; Endo. = endoscopy & surgical fields; Ex-vivo = ex-vivo organs & tissues; Signal = physiological-signal visualizations. † ground-truth-contributing model. Body fluids are the easiest domain for 13 of the 14 models; IVD reagents are hardest for the frontier, and physiological-signal visualizations are hardest for most open models.

Model	Fluids	IVD	Endo.	Ex-vivo	Signal	Overall
GPT-5.5	77.8	65.2	69.0	69.8	73.8	71.1
Seed-2.1	74.8	62.0	65.5	64.2	74.9	68.4
GPT-5.4†	74.0	59.2	63.0	63.0	60.0	63.8
Gemini-3.1-Pro	74.5	57.5	59.8	60.8	63.8	63.3
Gemini-2.5-Pro†	68.0	55.5	60.0	60.8	64.3	61.7
Seed-1.8†	71.2	50.5	52.8	56.0	60.3	58.2
Lingshu-32B	64.8	45.2	41.5	45.2	31.2	45.4
Qwen3-VL-8B	53.5	38.2	32.0	38.5	27.4	37.8
InternVL3.5-8B	52.8	33.5	31.8	36.2	33.8	37.6
GLM-4.6V-Flash	46.5	31.5	33.5	36.0	25.5	34.5
MedGemma-27B	43.8	25.8	25.5	31.0	22.0	29.5
MiniCPM-V-4.5	32.8	23.8	19.8	21.2	16.3	22.7
Citrus-V-8B	35.2	18.5	18.5	21.5	12.5	21.2
HuatuoGPT-V-34B	26.0	16.8	17.2	16.0	11.1	17.4

Leaf-level structure. Leaf accuracies are noisier than aggregates, so we test whether the flagships' leaf spreads exceed what chance would produce. Table 15 compares each flagship's observed max–min leaf accuracy (leaves with $n \geq 30$) with a permutation null that reassigns item outcomes across the label sets while preserving every leaf's size and the co-occurrence of labels; the right block permutes within answer format. All three spreads exceed the null's 95th percentile. Table 16 reports the flagships' pairwise leaf-rank correlations with bootstrap intervals; because resampling adds noise to every leaf accuracy, which attenuates rank correlations, the percentile intervals sit below the point estimates.

Image-necessary items. Table 17 re-scores all models on the items that none of the three flagships answers correctly when the image is withheld. The ranking of the 14 models is essentially unchanged (Spearman $\rho = 0.965$).

Judge self-preference. Seed-1.8 grades fill-in answers and is also evaluated. Table 18 lists each model's multiple-select and fill-in accuracy: Seed-1.8 has the lowest fill-in accuracy of the six closed models, and its fill-minus-multiple-select gap (-10.0) lies within the range of the other closed models (-2.3 to -13.1), which does not suggest a large self-preference effect.

Table 11: Full capability-point profile: accuracy (%) on each of the 34 leaves (F/C/P = factual/conceptual/procedural) for all 14 models; n = items carrying the leaf is given in Table 5; leaves 23, 28, and 34 have fewer than 30 items and are excluded from leaf-range statistics. Column codes: G5.5=GPT-5.5, S2.1=Seed-2.1, G5.4=GPT-5.4, Gm3.1=Gemini-3.1-Pro, Gm2.5=Gemini-2.5-Pro, S1.8=Seed-1.8, Lin=Lingshu-32B, Qwn=Qwen3-VL-8B, Ivl=InternVL3.5-8B, GLM=GLM-4.6V-Flash, MdG=MedGemma-27B, MCP=MiniCPM-V-4.5, Ctr=Citrus-V-8B, Hua=HuatuoGPT-V-34B.

#	Capability leaf (type)	G5.5	S2.1	G5.4	Gm3.1	Gm2.5	S1.8	Lin	Qwn	Ivl	GLM	MdG	MCP	Ctr	Hua
1	Entity recognition (F)	71	69	62	66	63	62	42	34	37	30	22	25	18	19
2	Spatial localization (F)	66	59	60	60	60	59	35	38	36	32	25	25	20	19
3	Dynamic-sign reading (F)	70	79	64	73	64	64	55	39	42	36	33	18	21	21
4	Legend & scale reading (F)	73	71	68	64	64	59	51	44	46	41	38	32	30	19
5	Signage/warning reading (F)	80	73	76	76	68	71	60	60	55	48	39	39	31	23
6	Modality identification (F)	66	66	60	59	55	51	47	36	37	36	28	23	20	20
7	Parameter extraction (F)	74	75	65	69	68	64	49	48	51	42	31	42	26	21
8	Geometric measurement (F)	67	68	57	58	60	52	32	30	35	27	24	22	15	12
9	Appearance description (F)	74	68	64	64	60	57	43	31	31	23	24	12	15	12
10	Exam-type classification (F)	53	54	49	48	47	37	33	24	24	27	20	14	19	15
11	Lesion-morphology typing (F)	76	71	66	63	61	63	37	30	29	29	26	14	13	15
12	Report categorization (F)	75	68	70	71	66	61	64	46	39	37	24	24	22	21
13	Severity grading (F)	72	68	64	62	63	58	36	31	28	20	23	17	14	11
14	Micro-mechanism (C)	89	82	76	71	74	84	61	42	53	47	37	24	18	21
15	Inorganic-reaction mechanism (C)	69	68	50	54	63	60	44	28	32	26	25	10	15	7
16	Macro-mechanism (C)	79	77	67	74	70	72	53	32	37	34	30	12	17	14
17	Artifact attribution (C)	73	72	77	66	59	62	49	41	41	30	37	15	25	11
18	Causal association (C)	79	77	67	70	68	70	52	34	34	34	32	11	15	12
19	Look-alike discrimination (C)	68	66	61	57	57	55	35	27	27	19	23	5	13	9
20	Cross-modal mapping (C)	83	83	75	80	68	72	48	41	32	32	30	12	12	12
21	Imaging principle (C)	82	85	76	82	70	82	85	64	45	52	52	45	21	24
22	Technique application (C)	86	80	82	78	72	70	63	44	53	52	42	22	27	22
23	Document-compliance check (C)	75	75	67	75	50	67	58	42	50	42	75	8	50	17
24	Structured reading (P)	68	63	67	62	61	54	49	40	38	32	34	16	24	14
25	Disposition / next-step (P)	73	66	67	66	60	55	50	37	36	36	32	12	21	12
26	Operation-technique reading (P)	71	65	67	69	65	62	46	44	46	35	42	29	25	19
27	Image-quality assessment (P)	79	75	82	69	62	58	53	50	51	36	42	27	32	21
28	Response comparison (P)	88	75	88	100	75	88	62	62	50	50	38	12	62	0
29	Waveform/spectrum reading (P)	70	67	54	59	57	52	23	22	21	11	20	10	9	11
30	Data/chart interpretation (P)	73	74	56	61	53	59	40	26	30	29	23	14	18	11
31	Procedure-step inference (P)	72	60	67	66	53	52	49	38	29	48	35	20	26	14
32	Guideline grading (P)	84	80	76	74	65	73	60	38	39	28	40	12	19	12
33	Care-pathway planning (P)	79	71	69	67	65	60	58	41	38	39	37	11	20	11
34	Treatment-strategy adjustment (P)	80	90	80	80	90	90	80	60	70	30	60	20	40	20

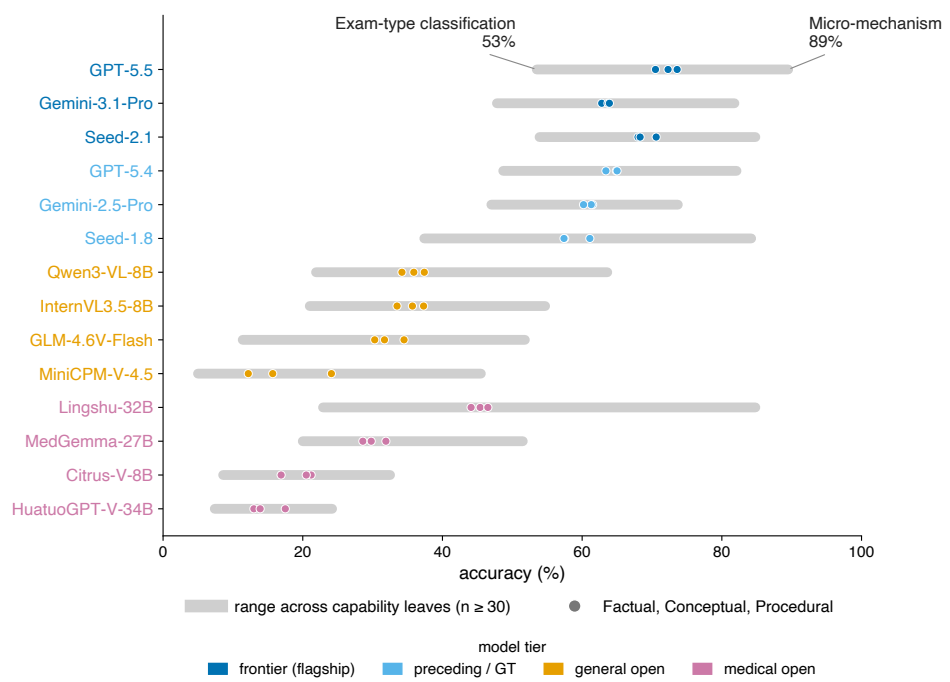


Figure 9: **Knowledge type barely separates difficulty; the capability leaf does.** For each model, the three knowledge-type accuracies (factual, conceptual, procedural; dots) cluster within a few points for most models, whereas accuracy across capability leaves with $n \geq 30$ (bar) spans far more—GPT-5.5, for instance, ranges from 53% (exam-type classification) to 89% (micro-mechanism). Colour denotes model tier.

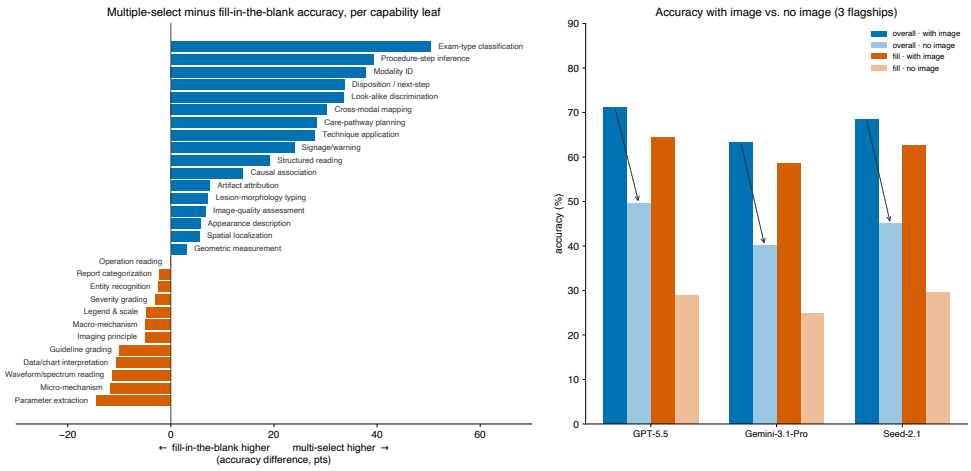


Figure 10: **Format gap by capability, and answers depend on the image.** Left: per capability, multiple-select minus fill-in-the-blank accuracy (flagship average)—multiple-select accuracy is far higher for classify-and-decide competencies, but the gap reverses for read-a-value competencies. Right: withholding the image drops the three flagships by 21.5–23.3 points overall and 33–35 points on fill-in-the-blank, confirming genuine image dependence (see also Table 9).

G Example Items

Figures 11–15 show one item per domain, three in the multiple-select and two in the fill-in-the-blank format. Items are posed to models in Chinese; the English translations are faithful to the original

Table 12: Knowledge-type accuracy vs. capability-leaf range (leaves with $n \geq 30$). For the three flagships the knowledge types differ by only 1–3 points while their leaves span more than 30. Selected models; full matrix in Table 11.

Model	Factual	Conceptual	Procedural	Leaf range ($n \geq 30$)
GPT-5.5	70.5	73.6	72.3	53–89
Seed-2.1	68.0	70.6	68.3	54–85
Gemini-3.1-Pro	62.8	63.7	63.9	48–82
Lingshu-32B	44.1	45.4	46.5	23–85
Qwen3-VL-8B	37.4	34.2	35.9	22–64
HuatuoGPT-V-34B	17.5	13.0	13.9	7–24

Table 13: Accuracy (%) with 95% bootstrap confidence intervals (multiple-select and fill-in intervals resample within format).

Model	Overall	Multi-select	Fill-in
GPT-5.5	71.1 [69.2, 73.1]	77.4 [74.8, 79.8]	64.3 [61.2, 67.3]
Seed-2.1	68.4 [66.3, 70.3]	73.6 [70.8, 76.2]	62.6 [59.6, 65.7]
GPT-5.4	63.8 [61.7, 65.9]	67.2 [64.3, 70.0]	60.1 [57.0, 63.2]
Gemini-3.1-Pro	63.3 [61.1, 65.4]	67.5 [64.8, 70.3]	58.6 [55.6, 61.7]
Gemini-2.5-Pro	61.7 [59.6, 63.8]	62.8 [60.0, 65.7]	60.5 [57.4, 63.7]
Seed-1.8	58.2 [56.0, 60.4]	62.9 [60.0, 65.8]	53.0 [49.7, 56.1]
Lingshu-32B	45.4 [43.3, 47.6]	46.3 [43.3, 49.3]	44.4 [41.3, 47.7]
Qwen3-VL-8B	37.8 [35.7, 40.0]	31.0 [28.3, 33.8]	45.3 [42.2, 48.4]
InternVL3.5-8B	37.6 [35.5, 39.7]	33.4 [30.6, 36.2]	42.2 [39.0, 45.3]
GLM-4.6V-Flash	34.5 [32.5, 36.6]	26.4 [23.8, 29.1]	43.4 [40.3, 46.5]
MedGemma-27B	29.5 [27.6, 31.5]	29.2 [26.5, 32.0]	29.8 [26.9, 32.7]
MiniCPM-V-4.5	22.7 [20.9, 24.6]	7.4 [5.8, 9.0]	39.6 [36.4, 42.6]
Citrus-V-8B	21.2 [19.4, 23.0]	14.7 [12.5, 16.9]	28.2 [25.3, 31.2]
HuatuoGPT-V-34B	17.4 [15.7, 19.0]	8.1 [6.5, 9.7]	27.5 [24.6, 30.4]

wording. Because the source images carry licensing constraints on redistribution (see the Ethics statement), they are not reproduced here; each box instead gives a neutral one-line summary of the image content and lists the capability leaves the item is mapped to (IDs and names as in Table 5; F/C/P = factual/conceptual/procedural).

Image (not reproduced): two test tubes of liquid placed side by side against a blue background; no text, scale bar or other markers.

Question. Observe the liquid samples in these two test tubes. Regarding their visual features and possible clinical significance, which of the following statements are correct?

A. The liquid in the left tube is uniformly red and turbid, suggesting that intravascular haemolysis may have occurred.

B. The liquid in the right tube is milky-white and turbid, which can be seen in chylothorax or chylous ascites.

C. The left tube contains an anticoagulated blood sample, and the red blood cells have settled evenly to the bottom.

D. The liquid in the right tube is typical purulent cerebrospinal fluid, for which bacterial culture is usually positive.

E. The bright red colour of the liquid in the left tube is due to improper arterial blood collection.

F. The right tube contains a sample of lipid-emulsion injection used for parenteral nutrition support.

G. Both tubes contain urine samples, indicating haematuria and chyluria, respectively.

Reference answer. AB

Capability leaves. [9] Appearance description (F); [18] Causal association (C); [19] Look-alike discrimination (C).

Figure 11: Example item from the *Body fluids* domain (multiple-select).

Image (not reproduced): a hand-held, white cassette-type rapid test with a round sample well (S) and an oval result window marked C and T, photographed in dim light.

Question. Observe the result of this in-vitro diagnostic reagent. Which of the following descriptions are correct?

A. The control zone (C) is coloured and the test zone (T) is not, indicating a negative result.

B. A faint line appears in the test zone (T), suggesting a possible very early pregnancy; retesting is recommended.

C. This reagent is used for the rapid detection of specific drug metabolites in urine.

D. A clear line in the control zone (C) indicates that this use of the test is valid.

E. Only the control zone (C) is coloured, indicating that the reagent has expired or was used improperly, so the result is invalid.

F. The test is based on the polymerase chain reaction (PCR) principle and is highly sensitive.

G. The result is negative, which completely rules out pregnancy; no further examination is needed.

Reference answer. AD

Capability leaves. [1] Entity recognition (F); [4] Legend & scale reading (F); [18] Causal association (C); [25] Disposition / next-step (P).

Figure 12: Example item from the *in-vitro diagnostic (IVD) reagents* domain (multiple-select).

Image (not reproduced): white-light laryngoscopic comparison images; panel (a) before treatment and panel (b) after treatment.

Question. In image (a), the large, smooth mass that obstructs the airway is located mainly in the left _____ region.

Reference answer. Aryepiglottic fold

Capability leaves. [1] Entity recognition (F); [2] Spatial localization (F).

Figure 13: Example item from the *Endoscopy & surgical fields* domain (fill-in-the-blank).

Image (not reproduced): gross specimen of a coiled, resected bowel segment on a green background, with a centimetre scale along the top edge.

Question. Using the scale at the top of the image, the maximum transverse diameter of the main part of this ex-vivo bowel specimen (the distended bowel segment on the right) is approximately _____ cm. (Give the answer as an integer.)

Reference answer. 9

Capability leaves. [4] Legend & scale reading (F); [8] Geometric measurement (F).

Figure 14: Example item from the *Ex-vivo organs & tissues* domain (fill-in-the-blank).

Image (not reproduced): a cropped black-and-white ECG rhythm strip on standard grid paper; the lead name, paper speed and voltage calibration are not shown.

Question. Carefully observe this ECG rhythm strip. Which of the following descriptions accurately reflect the electrophysiological phenomena shown?

A. The entire strip shows a sustained, regular tachycardia.

B. An abrupt transition from a slow rate/pause to a rapid rate is visible.

C. During the rapid phase, the QRS complexes are wide and bizarre.

D. During the rapid phase, the QRS complexes are narrow.

E. During the rapid phase, the RR intervals are absolutely irregular.

F. Before the middle of the strip, one markedly prolonged inter-beat interval is visible.

G. Definite ventricular-fibrillation waves are visible.

H. Several P waves are followed by a single QRS complex, i.e., conduction at a fixed ratio.

Reference answer. BDF

Capability leaves. [9] Appearance description (F); [11] Lesion-morphology typing (F); [19] Look-alike discrimination (C); [29] Waveform/spectrum reading (P).

Figure 15: Example item from the *Physiological-signal visualizations* domain (multiple-select).

Table 14: Paired exact McNemar tests on all 2,023 items. First block: flagship versus predecessor within each family; second block: adjacent leaderboard positions. n_{10}/n_{01}: items answered correctly only by A / only by B; p_{Holm} adjusts over all 15 tests.

Pair (A vs B)	Δ (pt)	n_{10}	n_{01}	p	p_{Holm}
GPT-5.5 vs GPT-5.4	+7.3	284	136	4.2×10^{-13}	5.5×10^{-12}
Seed-2.1 vs Seed-1.8	+10.2	325	119	3.6×10^{-23}	5.0×10^{-22}
Gemini-3.1-Pro vs Gemini-2.5-Pro	+1.5	280	249	0.192	0.616
GPT-5.5 vs Seed-2.1	+2.8	196	140	0.0026	0.016
Seed-2.1 vs GPT-5.4	+4.5	287	195	3.2×10^{-5}	0.0003
GPT-5.4 vs Gemini-3.1-Pro	+0.5	268	257	0.663	1.000
Gemini-2.5-Pro vs Seed-1.8	+3.6	299	227	0.0019	0.014
Seed-1.8 vs Lingshu-32B	+12.8	449	191	6.9×10^{-25}	1.0×10^{-23}
Lingshu-32B vs Qwen3-VL-8B	+7.6	356	202	7.1×10^{-11}	8.6×10^{-10}
Qwen3-VL-8B vs InternVL3.5-8B	+0.2	291	286	0.868	1.000
InternVL3.5-8B vs GLM-4.6V-Flash	+3.1	338	276	0.014	0.069
GLM-4.6V-Flash vs MedGemma-27B	+5.0	354	253	4.8×10^{-5}	0.0004
MedGemma-27B vs MiniCPM-V-4.5	+6.8	363	225	1.4×10^{-8}	1.5×10^{-7}
MiniCPM-V-4.5 vs Citrus-V-8B	+1.5	237	206	0.154	0.616
Citrus-V-8B vs HuatuoGPT-V-34B	+3.8	247	170	0.0002	0.0015

Table 15: Observed leaf spread (points) versus a size-preserving permutation null (1,000 permutations).

Model	Observed	Null mean	Null p95	p	Null mean	Null p95	p
GPT-5.5	36.0	20.1	28.8	0.0090	20.6	28.9	0.0080
Gemini-3.1-Pro	34.0	21.4	30.2	0.014	21.4	30.7	0.022
Seed-2.1	30.9	20.5	28.5	0.023	21.1	30.3	0.043

Table 16: Pairwise Spearman ρ between flagship leaf-accuracy profiles with 95% item-bootstrap percentile intervals.

Pair	ρ ($n \geq 30$, 31 leaves)	ρ (all 34)
GPT-5.5 vs Gemini-3.1-Pro	0.835 [0.550, 0.889]	0.847 [0.452, 0.867]
GPT-5.5 vs Seed-2.1	0.813 [0.565, 0.887]	0.813 [0.504, 0.882]
Gemini-3.1-Pro vs Seed-2.1	0.763 [0.490, 0.848]	0.787 [0.410, 0.844]

Table 17: Accuracy (%) on the image-necessary items (none of GPT-5.5, Gemini-3.1-Pro, Seed-2.1 is correct without the image), with 95% bootstrap intervals, and each model's rank on the full benchmark and on this subset.

Model	Full	Subset	Subset MS	Subset fill	Rank
GPT-5.5	71.1	48.3 [44.9, 51.7]	39.0	52.2	1→1
Seed-2.1	68.4	47.1 [43.7, 50.5]	39.0	50.4	2→2
GPT-5.4	63.8	39.7 [36.4, 42.9]	21.6	47.2	3→5
Gemini-3.1-Pro	63.3	42.9 [39.7, 46.4]	27.8	49.2	4→4
Gemini-2.5-Pro	61.7	43.1 [39.7, 46.4]	23.2	51.3	5→3
Seed-1.8	58.2	38.3 [35.0, 41.6]	30.3	41.7	6→6
Lingshu-32B	45.4	27.1 [24.1, 30.3]	15.4	32.0	7→9
Qwen3-VL-8B	37.8	28.7 [25.7, 31.9]	10.0	36.5	8→7
InternVL3.5-8B	37.6	28.0 [24.9, 31.1]	14.9	33.4	9→8
GLM-4.6V-Flash	34.5	25.2 [22.3, 28.1]	4.1	33.9	10→10
MedGemma-27B	29.5	15.9 [13.5, 18.5]	6.2	20.0	11→12
MiniCPM-V-4.5	22.7	23.8 [20.9, 26.9]	2.5	32.7	12→11
Citrus-V-8B	21.2	15.7 [13.3, 18.2]	2.9	21.0	13→13
HuatuoGPT-V-34B	17.4	13.7 [11.4, 16.1]	2.1	18.6	14→14

Table 18: Multiple-select (exact set match) versus fill-in (LLM-judged) accuracy (%), with 95% bootstrap intervals for the difference.

Model	MS	Fill	Fill−MS
GPT-5.5	77.4	64.3	-13.1 [-17.0, -9.1]
Seed-2.1	73.6	62.6	-11.0 [-15.0, -6.9]
GPT-5.4	67.2	60.1	-7.0 [-11.3, -2.8]
Gemini-3.1-Pro	67.5	58.6	-9.0 [-13.1, -4.7]
Gemini-2.5-Pro	62.8	60.5	-2.3 [-6.4, 2.0]
Seed-1.8	62.9	53.0	-10.0 [-14.2, -5.7]
Lingshu-32B	46.3	44.4	-1.9 [-6.3, 2.5]
Qwen3-VL-8B	31.0	45.3	14.2 [10.0, 18.4]
InternVL3.5-8B	33.4	42.2	8.8 [4.5, 12.9]
GLM-4.6V-Flash	26.4	43.4	17.0 [13.0, 21.2]
MedGemma-27B	29.2	29.8	0.6 [-3.4, 4.6]
MiniCPM-V-4.5	7.4	39.6	32.2 [28.7, 35.6]
Citrus-V-8B	14.7	28.2	13.5 [10.0, 17.1]
HuatuoGPT-V-34B	8.1	27.5	19.4 [16.1, 22.8]