MMBU: A Massive Multi-modal Biomedical Understanding Benchmark to Probe the Perception Capabilities of Vision-Language Models

Ryan D'Cunha[1,⋆], Alejandro Lozano[1,*], Xiaoxiao Sun[1,*], Daniel Vela Jarquin[3], Min Woo Sun[1], Josiah Aklilu[1], James Burgess[1], Yuhui Zhang[1], Paola Avila Robayo[1], Jin Ye[4], Ming Hu[4], Zhongying Deng[5], Junjun He[6], Xin Chen[7], Yue Yao[7], Robert Tibshirani[1], Jeffrey J. Nirschl[2], and Serena Yeung-Levy[1]

[1] Stanford University, Stanford, CA, USA
[2] University of Wisconsin–Madison, Madison, WI, USA
[3] Instituto Tecnológico de Monterrey, Mexico
[4] Monash University, Melbourne, VIC, Australia
[5] University of Cambridge, Cambridge, United Kingdom
[6] Shanghai Jiao Tong University, Shanghai, China
[7] Shandong University, Shandong, China
{rdcunha, lozanoe, xxsun}@stanford.edu

🤗 https://huggingface.co/datasets/ryandcunha/mmbu

Abstract. Vision and language models (VLMs) hold immense promise to transform biomedical imaging workflows, from detecting lesions in chest X-rays to profiling cellular features in microscopy. Realizing this potential, however, requires robust and fine-grained visual perception. Models need to correctly interpret subtle features in images, and they must do so across diverse biomedical modalities, scales, and contexts. Nevertheless, current benchmarks remain limited. To address these gaps, we introduce the **M**assive **M**ultimodal **B**iomedical **U**nderstanding (MMBU) benchmark. It is the largest biomedical vision and language benchmark to date, covering 35 submodalities with rich structured metadata. It includes both open and closed versions of ungrounded classification, grounded classification, and object detection, enabling systematic evaluation of model performance across biological scales, clinical settings, and imaging modalities. Evaluating 15 open-weight and 2 frontier VLMs, we find that while medical adaptation provides measurable gains for some models, the high accuracy often reported on established benchmarks can mask deficiencies in visual perception and domain generalization.

Keywords: Biomedical Vision Language Models, Multimodal Biomedical Imaging, Domain Adaptation, Visual Understanding, Benchmarking, Object Detection, Classification

⋆ Equal contribution.

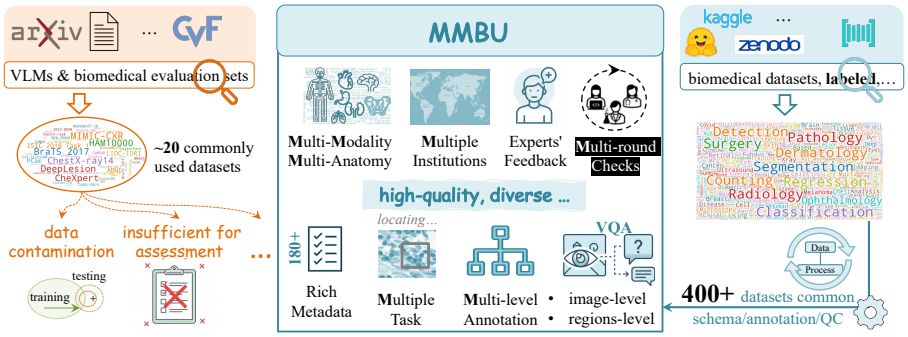

Fig. 1: The data landscape of MMBU. Current biomedical VLM evaluation relies on roughly 20 commonly used datasets. However, as the training data for large models expands, this evaluation becomes inadequate due to issues such as data pollution and a lack of diversity. We introduce MMBU to address this issue.

1 Introduction

Biomedical vision-language models (VLMs) are increasingly explored for a wide range of biomedical applications [9,31], from clinical decision support [32] to scientific discovery [15]. To realize this potential, VLMs must first exhibit robust visual perception – a major failure mode of current VLMs [1] – as they need to capture both fine- and coarse-grained features and interpret them robustly across heterogeneous patient populations, biomedical modalities, biological scales, and acquisition conditions [24, 29]. However, existing evaluations [4, 7, 11] are often limited to a small set of commonly used datasets and do not provide sufficient metadata to stratify performance to inspect model behavior in detail. Moreover, the training sets underlying these benchmarks are frequently incorporated into the pretraining or instruction-tuning pipelines of modern biomedical VLMs, leading to substantial distributional overlap between training and evaluation data. As a result, strong performance on established benchmarks may overestimate real-world perceptual capabilities and mask important failure modes.

Thus, to rigorously assess visual perception in this setting, benchmarks must therefore go beyond narrow domain-specific evaluation and support systematic analysis across tasks, modalities, specimens, and metadata axes. Motivated by this need, we introduce **MMBU** (**M**assive **M**ulti-modal **B**iomedical **U**nderstanding), a large-scale biomedical benchmark designed to evaluate perception capabilities in a broader and more fine-grained way.

Unlike existing single-domain biomedical benchmarks that focus on specific tasks or modalities, such as PathVQA [4], VQA-RAD [7], and SLAKE [11], **MMBU** aims to provide a more reliable evaluation of biomedical VLMs through broader coverage across visual tasks, imaging modalities, biological specimens, and acquisition conditions. Such a broader evaluation is necessary, as existing benchmarks cover only a limited portion of the biomedical imaging landscape

and recent studies have revealed a gap between benchmark performance and real-world capability. For example, state-of-the-art biomedical VLMs remain highly sensitive to prompt variation and batch effects—systematic shifts arising from differences in scanners, acquisition protocols, or institutions—and often fail to consistently outperform their general-domain counterparts after biomedical adaptation [3, 6, 13]. Although these findings are based on relatively small-scale evaluations, they raise important questions about the robustness of current biomedical VLMs and the effectiveness of existing adaptation strategies. By enabling systematic analysis across diverse biomedical modalities, domains, and acquisition conditions, **MMBU** provides a more comprehensive assessment of biomedical visual perception and helps characterize the limitations of current models.

These findings highlight the need for a benchmark that can systematically evaluate biomedical VLMs across a broader range of visual tasks, domains, topics, and modalities, and thereby more accurately assess when and how domain adaptation improves visual perception. Such an evaluation requires carefully annotated, expert-curated data and a unified task taxonomy that supports comprehensive and comparable analysis. To address this need, we assembled a multidisciplinary team of clinicians, bioinformaticians, statisticians, and computer scientists to define a taxonomy of biomedical visual perception tasks and the corresponding domains across biology and medicine. Guided by expert input, we curated data, conducted quality assurance, and iteratively refined a unified benchmark.

1. **The MMBU Benchmark:** Figure 1 summarizes the strengths of our contribution. We introduce the largest biomedical VLM benchmark to date (Table 1), covering 11 modalities (across 35 submodalities) from 20 specimens and 95 unique regions of interest within those specimens, spanning 410 datasets. Each dataset is converted into closed- and open-ended visual question answering (VQA) tasks and manually annotated by a group of experts. MMBU evaluates core perception tasks, including classification and detection, and provides rich, structured annotations for each datapoint, covering image provenance, dataset name, domain, modality, submodality, stain (when applicable), specimen, specimen subregion, topic, original task description, context, institution of acquisition, and URL.
2. **Systematic evaluation of medically adapted VLMs:** We leverage MMBU to characterize biomedical VLMs (and their base counterparts). Despite steady progress, we find high error rates across all models and tasks, limited but measurable improvement from medical adaptation, substantial weaknesses in spatially grounded tasks such as object detection, and inconsistent generalization beyond commonly used biomedical benchmarks and MMBU.

Given the significant room for improvement, we release MMBU to enable systematic, large-scale evaluation of biomedical VLMs and to support the development of more reliable models for biomedical perception tasks.

Table 1: Current landscape of biomedical vision–language benchmarks. We compare MMBU with existing benchmarks across biomedicine. While prior benchmarks offer valuable but narrow coverage of selected modalities or tasks, MMBU fills a critical gap by unifying a substantially broader range of biomedical domains, submodalities, and metadata sources. *If a dataset does not provide the annotation required by a given column, our experts re-annotate the missing fields using the same criteria as MMBU.

Benchmark	Submodalities	Size	Topics	Metadata	Source
PathVQA [4]	1	6K	×	×	Textbook, PEIR
Cholec80-VQA	1	9K	×	×	Cholec80
VQA-RAD [7]	3	3K	×	×	MedPix teaching cases
SLAKE [11]	3	2K	×	×	MSD, ChestX-ray8, CHAOS
RadBench [26]	6	137K	×	×	13 paired image–text datasets
MMMU (H & M) [30]	6	2K	×	×	Exams, quizzes, textbooks
OmniMedVQA [5]	12	128K	×	×	73 classification datasets
GMAI-MMBench [2]*	32	26K	321	2	284 public and hospital datasets
MMBU	**35**	**78K**	**458**	13	410 curated datasets

2 Related Works

Biomedical VLMs Recent advances in general-purpose VLMs have rapidly expanded their visual reasoning, language grounding, and instruction-following capabilities. These developments have motivated parallel efforts to probe their biomedical strengths [20, 33] and to adapt them to medical workflows through instruction tuning or reinforcement learning, leading to models such as LLaVA-Med [10], MedGemma [21], Lingshu [27], and OctoMed [18]. This accelerating model development underscores the need for rigorous benchmarks that can capture improvements across biomedical domains, modalities, and tasks.

Domain-specific biomedical VQA benchmark Several specialized benchmarks have been developed to evaluate the capabilities of VLMs within specific biomedical imaging modalities—for example, VQA-RAD [7], U2-Bench [8], and SLAKE for radiology; PathVQA [4] for pathology; and MicroBench [14], and MicroVQA [1] for microscopy. Although these benchmarks provide detailed, domain-specific assessments and have catalyzed progress within their respective areas, their narrow scope limits their ability to measure how models generalize across the full depth and breadth of biomedical tasks.

Multi-domain biomedical VQA benchmarks To support broader evaluation, several recent medical benchmarks span multiple domains, modalities, and tasks. Notable examples include OmniMedVQA [5], MMMU (Health & Medicine) [30], and GMAI-MMBench [2]. Among these, GMAI-MMBench is one of the most comprehensive, drawing on 284 datasets across 18 clinical tasks. Despite their scale, these benchmarks still have several key limitations: Existing large-scale benchmarks focus primarily on diagnostic imaging (e.g., radiology, pathology, endoscopy) and provide limited coverage of fundamental biological imaging modalities (e.g., fluorescence microscopy, electron microscopy) that reveal underlying cellular and molecular processes driving disease. These benchmarks are largely human-centric, often overlook additional specimens relevant

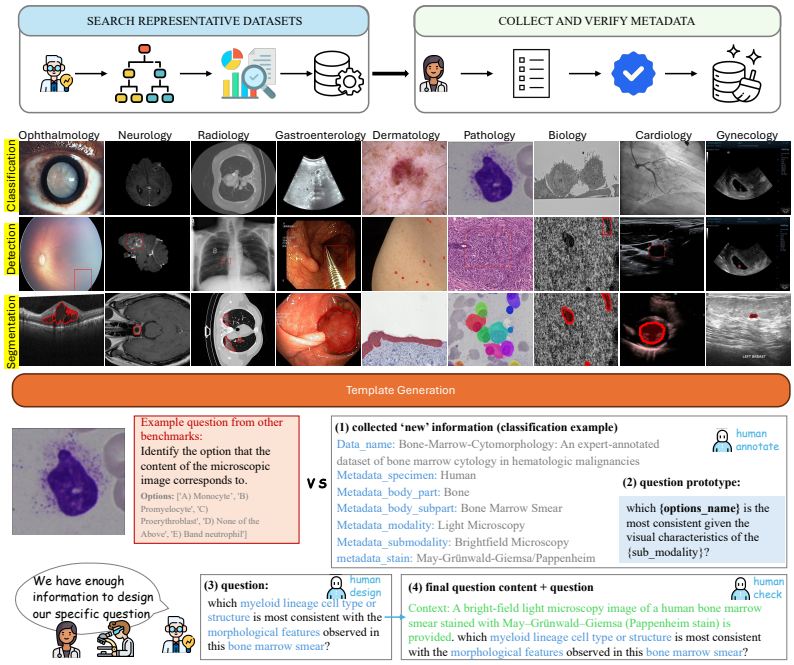

Fig. 2: Multi-task visual examples and metadata-driven question construction in MMBU. **TOP:** Data collection metadata extraction and standardization. **Middle:** Representative samples from diverse medical domains and modalities across three task types, including classification, detection, and segmentation. **Bottom:** Example of question construction in MMBU using newly collected metadata. An example benchmark question is first shown (left), then reformulated through a four-step pipeline (right): (1) collect additional fine-grained metadata from the image and source context (e.g., specimen, body part/subpart, modality, submodality, and stain), (2) define a question prototype conditioned on metadata fields, (3) instantiate a draft domain-specific question targeting clinically meaningful attributes, and (4) produce a finalized question after human design and verification.

to human health, and are predominantly classification-focused, offering few or no auxiliary tasks.

Furthermore, they generally lack the contextual information and metadata necessary to stratify performance across meaningful axes; critical details such as provenance, institution, specimen type, topic, and context are frequently absent. For instance, while GMAI-MMBench provides categorization through its lexical tree, MMBU offers more granular and refined annotations, with up to 13 structured attributes per sample encompassing acquisition parameters, task context, and technical factors (e.g., stain). This rich metadata is essential for analyzing model robustness and performance degradation due to batch effects or domain shifts; analyses that are central to MMBU's evaluation objectives but challenging to conduct systematically with existing benchmarks.

3 Dataset collection methodology

Our benchmark creation workflow is illustrated in Figure 2 and consists of four stages: (1) presenting representative examples spanning classification, detection, and segmentation to demonstrate task and domain diversity in MMBU; (2) collecting fine-grained metadata from images and source context; (3) constructing metadata-conditioned question templates and instantiating draft questions; and (4) performing human-in-the-loop refinement and validation to produce final questions. This metadata-driven, human-in-the-loop pipeline yields more specific and clinically grounded prompts than directly reusing generic benchmark questions.

3.1 Data Discovery and Quality Filtering

We began by designing a multi-granularity dataset taxonomy with input from a multidisciplinary committee of two clinicians, two wet-lab specialists, and two bioinformaticians (provided in the supplementary materials). The taxonomy was refined through four iterations and covers medical and scientific specialties, image modalities and submodalities, anatomical entities, and biological contexts. Once finalized, we solicited from co-authors representative datasets for each taxonomized item, using the following inclusion criteria: (i) publicly accessible datasets with clear licensing, (ii) high-quality images, defined as images that are not blurry, where the subject is clearly visible, and without preprocessing artifacts, and (iii) human-generated labels, excluding datasets annotated by LLMs. Additionally, datasets were required to cover core perception tasks (classification, detection, and segmentation). To avoid data contamination, we excluded datasets frequently used to train medical VLMs (e.g., literature-derived datasets). Using these criteria, experts identified 122 relevant datasets in the first round.

In the second round, we developed a discovery agent to search for datasets to increase coverage of underrepresented categories (as identified by our taxonomy) across major public repositories, including Zenodo, Kaggle, Papers with Code, and Hugging Face. After retrieval, four annotators verified the metadata by cross-checking the original dataset descriptions (and subsequently dropping any dataset not complying with the inclusion criteria); this round identified an additional 1.2K candidate datasets. In the third round, five experts reviewed the aggregated candidate pool and removed datasets with missing annotations, corrupted files, unclear licensing, or preprocessing or formatting errors. This yielded 440 candidate datasets, from which we selected 410 datasets.

3.2 Metadata Collection and Standardization

The originally collected datasets had varying organizational structures, file formats, and often very limited structured metadata. To address this, we developed a standardized pipeline to harmonize all datasets into a common format. This pipeline handles dataset downloading, integrates annotators to verify and correct extracted metadata, and implements image processing across multiple

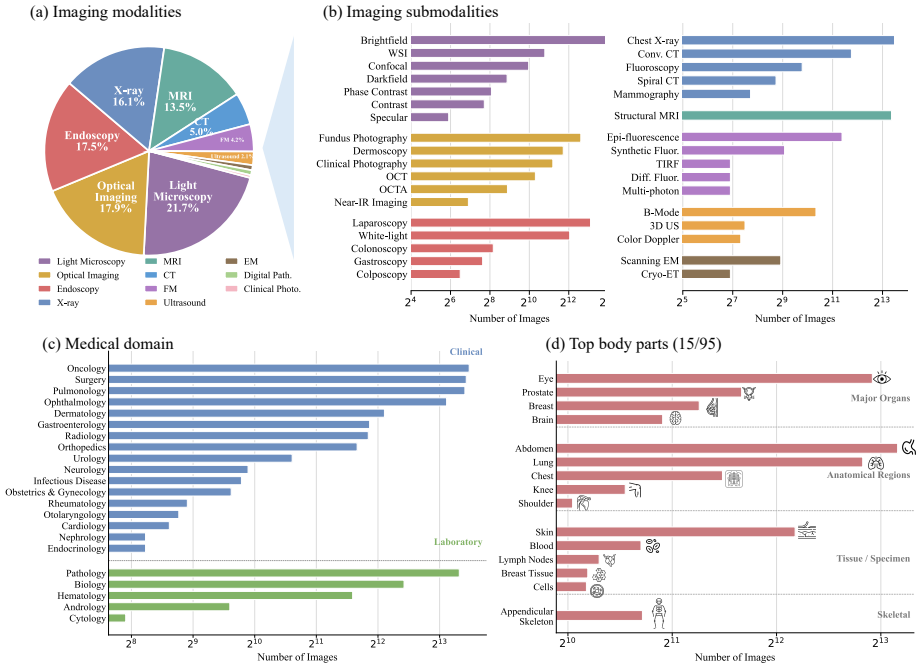

Fig. 3: Overview of MMBU dataset composition across modalities, submodalities, medical domains, and body parts. (a) Distribution of top-level imaging modalities. (b) Distribution of imaging submodalities, shown in two panels for readability, with counts on a log2 scale. (c) Distribution of medical domains grouped into clinical and laboratory categories. (d) Top body parts (15 out of 95 total body-part categories), further organized into major organs, anatomical regions, tissue/specimen, and skeletal groups. Across all panels, bar lengths indicate image counts.

formats—standardizing all of them to PNG (e.g., TIFF, JPEG, DICOM, NIfTI, SVS, NDPI, MRXS, OME-TIFF).

3.3 Question Construction, Validation and Refinement

We provide three versions of templates: two automatic and one manually designed. **No Context:** a standard template shared across all tasks, regardless of dataset. **Modality:** the same template structure, but augmented with the image modality as context in the question. **Full Context:** Annotators used the collected metadata for each dataset to populate a predetermined template. They were asked to include key attributes such as specimen type (e.g., human tissue, human fluid, human smear), imaging modality (e.g., light microscopy, computed tomography, magnetic resonance imaging), stain or contrast when applicable (e.g., hematoxylin and eosin, T1-weighted contrast), preparation method (e.g., smear, section, biopsy), and anatomical site or compartment (e.g., leg, liver,

lung). Annotators were further instructed to ensure that all template instructions are grounded in pixel-level visual content, to avoid disease names or clinical outcomes, to avoid any terms suggesting malignancy, and to be clear by not using acronyms. Additional details are provided in the Supplementary Materials.
Validation and Refinement. Lastly, to ensure data quality, all standardized datasets, templates, and metadata are re-evaluated and corrected (if needed) by a team of 3 experts.

3.4 Dataset Description

Figure 3 shows the main MMBU statistics. In summary, MMBU comprises 410 biomedical datasets, including 250 for classification, 84 for segmentation, and 76 for object detection. These datasets span 11 unique modalities (across 35 submodalities) and cover 27 medical and scientific domains. Collectively, they include 20 distinct specimens and 95 regions of interest within those specimens. After standardization, datasets are converted into four categories: ungrounded classification, grounded classification (using segmentation masks), grounded classification (using bounding boxes), and object detection: each having two setups (open and closed), resulting in a total of eight configurations, each with three different templates (24 setups in total).

4 Benchmark Experiments

4.1 Task Definition

Ungrounded Classification. In ungrounded classification, the model predicts a label or answers a question based on the entire image, without any region of interest highlighted or annotated.
Grounded Classification. In grounded classification, the model predicts a label for the content within a region of interest, specified by either a segmentation mask or a bounding box. For both classification tasks, the closed setup requires the model to select the correct answer from a candidate pool, whereas in the open setup, the model must generate the answer directly.
Object Detection. In object detection, the model is tasked with predicting a bounding box given an image and a target class. In the closed setup, the model chooses the correct bounding box from a candidate pool; in the open setup, it must predict the bounding box directly. Further mathematical setup is described in the Supplementary Materials.

4.2 Experimental Setup

Model Selection. To capture the current VLM landscape, we selected 15 open and 2 frontier autoregressive VLMs spanning different sizes, instruction-tuning strategies, reasoning capabilities, and medically fine-tuned variants. For each medically adapted model, we include its corresponding base counterpart to directly assess the impact of biomedical fine-tuning on model performance. We

further verified that all medically adapted models have no dataset overlap with our collected datasets, as reported in the original publications. Models (color-coded by family throughout the rest of this work) include:

- **GPT-5.4-mini** [17] and **GPT-4.1-mini** [16], representing different generations of frontier models to create a comparison to open-source models.
- **Lingshu-7B** and **Lingshu-32B** [27], a family of generalist biomedical VLMs trained in a multi-stage paradigm that progressively infuses medical knowledge, instruction-tuned on over 7.1M multimodal samples. We also evaluate **Qwen2.5-VL-7B-Instruct** and **Qwen2.5-VL-32B-Instruct** [19], the corresponding base models.
- **OctoMed-7B** [18], a recent medical VLM with multimodal reasoning capabilities, built on **Qwen2.5-VL-7B-Instruct**.
- **MedGemma-4B-it** and **MedGemma-1.5-4B-it** [21], Google's open models for medical text and image comprehension, and the model they were built on, **Gemma-3-4B-it** [23].
- **LLaVA-Med-v1.5-7B** [10], one of the first VLMs trained on a large-scale, broad-coverage biomedical figure-caption dataset, and its corresponding base model, **LLaVA-1.5-7B** [12].
- One of the strongest open-source general-purpose VLMs, **InternVL3.5-8B** [25], which uses a cascaded reinforcement learning approach that enhances reasoning ability via a coarse-to-fine training strategy.
- Additional models from the Qwen series, including **Qwen2.5-VL-3B-Instruct** [19] and the latest **Qwen3-VL-4B-Instruct**, **Qwen3-VL-8B-Instruct**, and **Qwen3-VL-32B-Instruct** [28].

This set of models is non-exhaustive; it is designed to capture representative trends across model size, architecture, reasoning ability, and domain specialization. We employ model checkpoints from the respective official Hugging Face repositories and conduct all experiments with NVIDIA L40S and A100 GPUs.

4.3 Evaluation

Metrics. We report micro-averaged F1-score with 95% confidence intervals obtained via 1,000-iteration bootstrap resampling over datapoints. For classification tasks (ungrounded and grounded), correctness is determined by exact string match after normalization. For object detection, a prediction is considered correct when its Intersection-over-Union (IoU) with the ground-truth bounding box meets or exceeds 0.5.

Closed-ended VQA answer extraction and scoring For closed-ended VQA, we use a fully deterministic extraction and scoring protocol rather than LLM-as-a-judge evaluation [22]. A hierarchical parser scans model outputs for explicit answer fields, formatting tags, ordered lists, and valid option labels, and then compares the extracted label against the ground-truth option set. Responses with no parseable answer or with labels outside the valid option set are counted as incorrect.

Table 2: Evaluated VLMs on MMBU tasks. Micro-averaged F1 scores (with 95% bootstrap confidence intervals) of 17 representative VLMs and a random baseline on *classification* and *detection* tasks in MMBU. Sub-columns report Closed VQA, Open VQA, and their difference (Δ). **Bold** indicates the best score per column.

Model	Ungrounded Classification			Grounded Classification (from Detection)			Grounded Classification (from Segmentation)			Object Detection		
	Closed	Open	Δ	Closed	Open	Δ	Closed	Open	Δ	Closed	Open	Δ
GPT-5.4-mini	0.533 ± 0.006	0.104 ± 0.004	0.430 ± 0.007	0.410 ± 0.010	0.132 ± 0.010	0.278 ± 0.014	0.472 ± 0.005	0.126 ± 0.008	0.346 ± 0.009	0.100 ± 0.004	**0.053** ± 0.007	0.046 ± 0.008
GPT-4.1-mini	**0.539** ± 0.006	0.070 ± 0.002	0.468 ± 0.006	0.430 ± 0.010	0.087 ± 0.007	0.344 ± 0.012	0.501 ± 0.011	0.065 ± 0.004	0.436 ± 0.012	0.094 ± 0.003	0.046 ± 0.006	0.048 ± 0.007
Gemma-3-4B	0.303 ± 0.006	0.022 ± 0.002	0.281 ± 0.007	0.283 ± 0.008	0.026 ± 0.003	0.257 ± 0.009	0.389 ± 0.008	0.009 ± 0.002	0.381 ± 0.008	0.083 ± 0.004	0.024 ± 0.005	0.058 ± 0.006
MedGemma-4B	0.439 ± 0.009	0.043 ± 0.001	0.396 ± 0.009	0.371 ± 0.012	0.066 ± 0.005	0.305 ± 0.012	0.310 ± 0.008	0.042 ± 0.005	0.268 ± 0.009	0.076 ± 0.003	0.018 ± 0.004	0.058 ± 0.005
MedGemma-1.5-4B	0.403 ± 0.007	0.041 ± 0.002	0.362 ± 0.007	0.335 ± 0.011	0.045 ± 0.002	0.290 ± 0.011	0.428 ± 0.008	0.042 ± 0.003	0.385 ± 0.009	0.071 ± 0.002	0.000 ± 0.000	0.071 ± 0.002
InternVL3.5-8B	0.517 ± 0.006	0.111 ± 0.003	0.406 ± 0.007	**0.439** ± 0.014	0.122 ± 0.006	0.318 ± 0.015	0.514 ± 0.006	0.088 ± 0.005	0.426 ± 0.008	0.061 ± 0.004	0.019 ± 0.004	0.042 ± 0.006
Lingshu-7B	0.418 ± 0.008	0.054 ± 0.002	0.364 ± 0.008	0.360 ± 0.014	0.047 ± 0.006	0.313 ± 0.015	0.258 ± 0.008	0.049 ± 0.006	0.209 ± 0.010	0.012 ± 0.002	0.001 ± 0.001	0.011 ± 0.002
Lingshu-32B	0.469 ± 0.006	0.073 ± 0.002	0.396 ± 0.006	0.368 ± 0.010	**0.150** ± 0.005	0.218 ± 0.012	0.310 ± 0.006	0.053 ± 0.002	0.257 ± 0.007	0.067 ± 0.005	0.004 ± 0.002	0.063 ± 0.005
Qwen2.5-VL-3B	0.381 ± 0.008	0.029 ± 0.002	0.352 ± 0.008	0.336 ± 0.016	0.027 ± 0.004	0.308 ± 0.017	0.240 ± 0.008	0.026 ± 0.004	0.214 ± 0.009	0.014 ± 0.002	0.016 ± 0.004	-0.002 ± 0.004
Qwen2.5-VL-7B	0.361 ± 0.006	0.031 ± 0.002	0.330 ± 0.006	0.320 ± 0.008	0.030 ± 0.005	0.289 ± 0.009	0.357 ± 0.005	0.030 ± 0.004	0.326 ± 0.006	0.029 ± 0.002	0.012 ± 0.003	0.017 ± 0.004
Qwen2.5-VL-32B	0.526 ± 0.006	0.088 ± 0.002	0.437 ± 0.007	0.401 ± 0.013	0.079 ± 0.004	0.322 ± 0.014	**0.693** ± 0.005	0.115 ± 0.007	0.577 ± 0.008	0.092 ± 0.006	0.047 ± 0.006	0.045 ± 0.009
Qwen3-VL-4B	0.293 ± 0.006	0.079 ± 0.002	0.214 ± 0.006	0.363 ± 0.008	0.093 ± 0.008	0.270 ± 0.011	0.540 ± 0.009	0.065 ± 0.002	0.475 ± 0.009	0.040 ± 0.005	0.001 ± 0.001	0.039 ± 0.005
Qwen3-VL-8B	0.472 ± 0.009	0.086 ± 0.003	0.386 ± 0.009	0.391 ± 0.014	0.098 ± 0.007	0.292 ± 0.016	0.526 ± 0.007	0.074 ± 0.007	0.452 ± 0.010	0.006 ± 0.001	0.001 ± 0.001	0.005 ± 0.001
Qwen3-VL-32B	0.530 ± 0.008	**0.137** ± 0.006	0.393 ± 0.010	0.433 ± 0.012	0.145 ± 0.008	0.288 ± 0.014	0.630 ± 0.005	**0.182** ± 0.011	0.447 ± 0.012	0.028 ± 0.003	0.003 ± 0.002	0.024 ± 0.003
LLaVA-v1.5-7B	0.399 ± 0.006	0.026 ± 0.002	0.374 ± 0.006	0.310 ± 0.008	0.020 ± 0.004	0.289 ± 0.009	0.299 ± 0.010	0.024 ± 0.001	0.275 ± 0.010	0.031 ± 0.003	0.000 ± 0.000	0.031 ± 0.003
LLaVA-Med-7B	0.395 ± 0.006	0.033 ± 0.001	0.362 ± 0.006	0.340 ± 0.010	0.038 ± 0.004	0.302 ± 0.011	0.080 ± 0.004	0.035 ± 0.004	0.045 ± 0.005	0.033 ± 0.003	0.002 ± 0.002	0.030 ± 0.004
OctoMed-7B	0.494 ± 0.007	0.070 ± 0.002	0.424 ± 0.008	0.346 ± 0.011	0.059 ± 0.008	0.287 ± 0.014	0.573 ± 0.010	0.056 ± 0.002	0.517 ± 0.010	0.033 ± 0.003	0.009 ± 0.003	0.024 ± 0.004
Random	0.199 ± 0.004	— —	— —	0.162 ± 0.003	— —	— —	0.183 ± 0.004	— —	— —	**0.172** ± 0.004	— —	— —

Open-ended VQA answer extraction and scoring For open-ended VQA, exact string matching is insufficient because correct answers may be expressed using synonymous biomedical terminology. We therefore use Qwen3-32B as a semantic equivalence judge for free-form responses. The judge receives only the model response and ground-truth answer and is instructed to apply strict biomedical equivalence, including explicit checks for negation, laterality, severity, uncertainty, and diagnostic specificity. For example, responses such as "pneumonia" and "no pneumonia" are treated as non-equivalent.

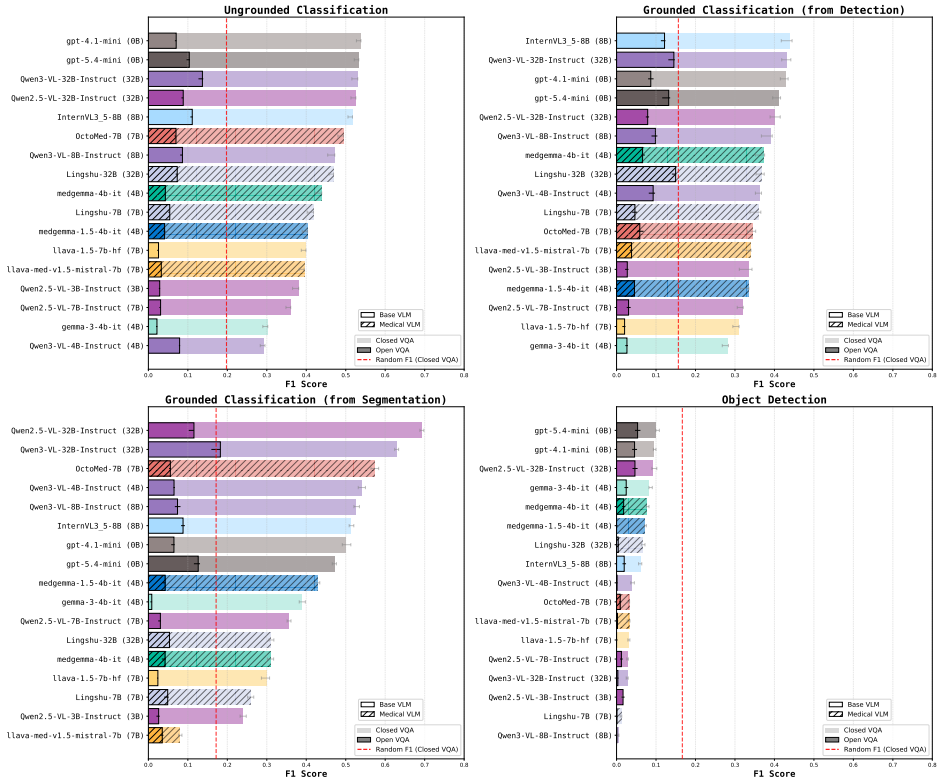

Fig. 4: Aggregate performance on MMBU. Performance of a representative set of VLMs on the *classification* and *detection* tasks in the benchmark. Solid outlines denote open-format results, while boxes without outlines denote closed-format results. Models are ranked by their closed-format performance. Dashes indicate models adapted to medical data. Similar colors indicate models of the same family.

5 Results

5.1 Overall Performance

Table 2 summarizes the F1 scores of 17 VLMs and a random baseline across four core perception tasks on MMBU: ungrounded classification, grounded classification derived from detection annotations, grounded classification derived from segmentation annotations, and object detection. Each task is evaluated under both closed-ended (multiple-choice) and open-ended (free-form) VQA formats; the gap between the two (Δ) quantifies how much performance degrades when the answer space is unconstrained.

Three high-level trends emerge. First, absolute F1 scores remain low overall: the best closed-format result reaches 0.693 (Qwen2.5-VL-32B on grounded classification from segmentation), but this is an exception, with most scores falling

below the 0.5 threshold for adequate performance. Second, the closed-to-open gap is consistently large, with an average Δ of 0.26 for classification tasks, indicating that models rely heavily on answer-choice cues rather than genuine visual understanding. Third, object detection remains a critical failure mode: no VLM surpasses the random baseline (F1 = 0.172) in the closed setting, and performance in the open setting is near zero for most models, revealing fundamental limitations in spatial reasoning.

Across model families, larger variants generally outperform their smaller counterparts (e.g. Qwen2.5-VL-32B vs. 3B and 7B; Lingshu-32B vs. 7B), yet the gains are modest and task-dependent. Notably, no single model ranks first across all four tasks: GPT-4.1-mini leads ungrounded classification (0.539), Qwen2.5-VL-32B leads grounded classification from segmentation (0.693), and InternVL3.5-8B leads grounded classification from detection (0.439) as well as open-ended tasks overall. Meanwhile, medically adapted models do not uniformly outperform their base counterparts; for instance, LLaVA-Med-7B underperforms LLaVA-v1.5-7B across tasks, while MedGemma-1.5-4B shows clear gains over Gemma-3-4B. These patterns are analyzed in detail in the following subsections.

5.2 Key Observations

High Error Rates Indicate Significant Room for Improvement. Figure 4 shows the average F1 scores aggregated across all domains and broken down by task type. Each subplot shows both the open-format version of the task (solid outline) and the closed-format version (no outline). Notably, there is a substantial F1 performance gap between closed and open format across models: 0.38 for ungrounded classification, 0.29 for grounded classification from detection, 0.36 for grounded classification from segmentation, and 0.04 for object detection (Supplementary Materials). Overall, regardless of task category, model choice, or adaptation strategy, models mostly achieve low F1 scores, often only slightly above random performance (e.g. 0.2 F1 in closed classification). These observations align with prior small-scale [1,13] and large-scale benchmarking [2]. Furthermore, while a few models frequently appear among the top performers (e.g., InternVL3.5-8B, MedGemma-4B, GPT-4.1-mini, and Qwen2.5-VL-32B), no model consistently ranks first across the aggregation of all tasks or across all domains (as shown in Figure 4), indicating there is no single "generalist" biomedical VLM. Given that VLMs consistently fail in the open format version of each task, the rest of the analysis emphasizes the closed format unless specified otherwise.

VLMs Perform Above Random Performance on Classification-Related Tasks. In general, VLMs perform best on closed classification-related tasks, with average F1 scores across models of 0.44 for ungrounded classification, 0.37 for grounded classification from detection, and 0.42 for grounded classification from segmentation. Indeed, closed classification tasks are heavily used to train and benchmark VLMs. In ungrounded closed classification, MedGemma-4B performs above the random baseline by +0.24, achieving an F1 of 0.439. However, this advantage does not translate into fine-grained grounded classification performance:

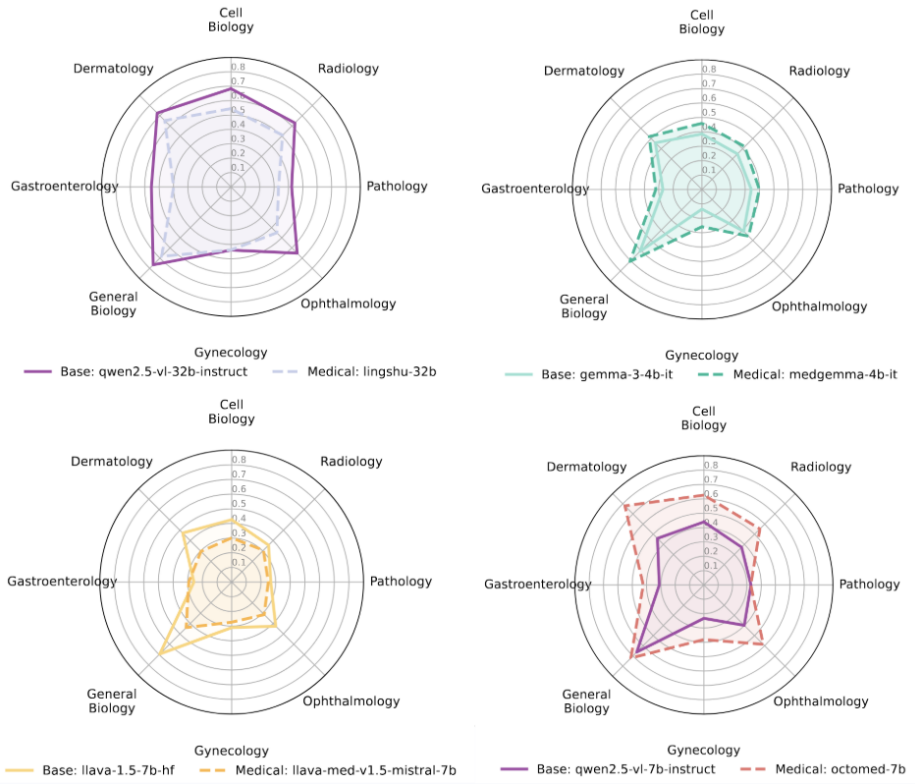

Fig. 5: Comparison of model performance across aggregated biomedical domains. The radial plots compare base (solid line) models and their medically adapted (shown by dashes) counterparts across aggregated domains in MMBU.

InternVL3.5-8B outperforms MedGemma-4B on both detection-derived classification (0.439 vs. 0.371) and segmentation-derived classification (0.514 vs. 0.310). These results suggest medical specialization improves classification performance but does not guarantee superior fine-grained grounded reasoning.

VLMs Exhibit Poor Object Detection Capabilities. Notably, all VLMs fail at object detection in both open and closed settings, even models that explicitly support it (e.g. Qwen2.5-VL family). The best performer, Qwen2.5-VL-32B, reaches only 0.10 F1 on the closed format version, where the task is to select the correct bounding box from multiple options. This highlights a core limitation: models can classify objects inside bounding boxes but cannot reliably locate the boxes themselves (lacking spatial awareness). Although many VLMs are not trained for referral segmentation, this gap remains significant given that segmentation and detection are foundational in biomedical imaging.

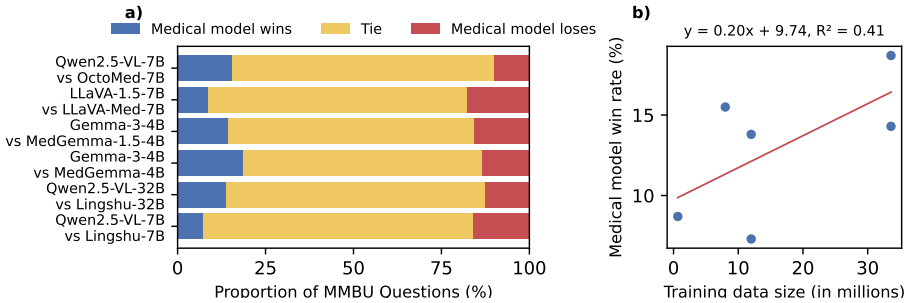

Fig. 6: Medical models vs. base models. a) Head-to-head comparison of general-purpose and medical vision–language models on MMBU, showing the proportion of questions where the medical model wins, ties, or loses. b) Relationship between medical model training data size (in millions of examples) and win rate on MMBU.

5.3 Additional Findings and Discussion

Beyond the aggregate results in Table 2, we investigate two questions central to the current biomedical VLM landscape: (i) does medical adaptation consistently improve performance over general-purpose base models, and (ii) do gains observed on established benchmarks transfer to the broader evaluation provided by MMBU?

In conjunction, Medically Adapted Models Outperform Non-Medical VLMs. While no single model consistently outperforms all others, stratifying performance by metadata shows that medically adapted models generally have a modest advantage. For example, when evaluating performance by submodality, medical models outperform non-medical models 62.28% of the time (Supplementary Materials) with an average F1 lead of only 0.08. We see the best performance from medical models on Near-Infrared Imaging, Phase Contrast Microscopy, and Specular Microscopy.

Medical adaptation yields limited but measurable gains. Given that medical models perform slightly better on MMBU, we further investigate their performance relative to their corresponding base models. Figure 6 (a) compares the closed classification performance of base models and their medically adapted counterparts. Across all adaptation methods, models tie in at least 70% of cases. Notably, only about half of the adapted models outperform their base counterparts, with the strongest gains observed in MedGemma-4B (18.7% wins / 13.4% losses), OctoMed-7B (15.5% wins / 9.9% losses), and Lingshu-32B (13.8% wins / 12.4% losses). Figure 5 further stratifies performance by a reduced set of domains of interest. Indeed, only OctoMed-7B and MedGemma-4B consistently exhibit improvements with respect to their base counterparts. We further find that pretraining data size plays a more important role in adaptation success than model size (Figure 6 (b)), with an estimated +0.20% increase in win rate per additional 1M training datapoints ($R^2 = 0.41$). These results highlight that medical adaptation is feasible, although substantial gains remain limited when

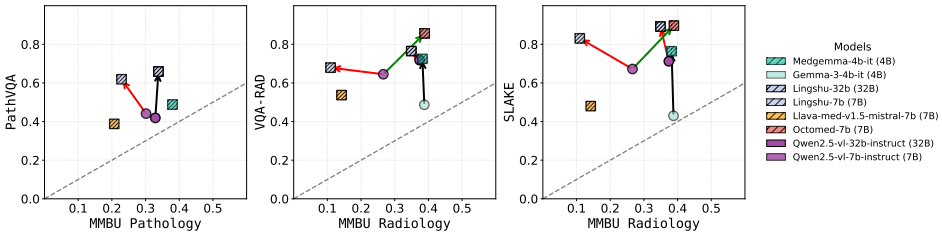

Fig. 7: Performance comparison across MMBU subsets (x-axis) and legacy benchmarks (y-axis). Arrows indicate changes after medical adaptation. Most specialized models excel on legacy datasets (PathVQA, VQA-RAD, SLAKE) but generalize poorly to MMBU, while MedGemma and OctoMed improve on both, indicating better overall generalization.

evaluated broadly, partially aligning with recent findings showing that adapted models often underperform their base counterparts more frequently than they outperform them [6].

Cross-Dataset Generalization of Medically Adapted VLMs. Figure 7 compares base and adapted VLMs on legacy benchmarks (PathVQA, VQA-RAD, SLAKE) and the corresponding MMBU subsets. While all adapted models improve on legacy benchmarks, gains do not consistently transfer to MMBU: OctoMed improves on both; MedGemma improves on legacy benchmarks but remains flat on MMBU; and Lingshu models improve on legacy benchmarks but degrade on MMBU. These results suggest that adaptation primarily optimizes for widely used benchmarks, likely because their training sets are commonly used to fine-tune [10] or pretrain [27] medical VLMs, with limited generalization to broader biomedical tasks.

6 Conclusion

MMBU exposes pervasive weaknesses in biomedical VLMs' perceptual abilities: low ungrounded and grounded classification accuracy, poor detection, and inconsistent cross-dataset generalization, despite increasing model size or adaptation. While no benchmark can replace expert evaluation, MMBU provides a representative proxy for measuring progress. By consolidating diverse perceptual tasks and rich metadata, it highlights clear directions for future work: stronger spatial modeling, improved perception, and adaptation methods that generalize beyond currently established datasets.

References

1. Burgess, J., Nirschl, J.J., Bravo-Sánchez, L., Lozano, A., Gupte, S.R., Galaz-Montoya, J.G., Zhang, Y., Su, Y., Bhowmik, D., Coman, Z., Hasan, S.M., Johannesson, A., Leineweber, W.D., Nair, M.G., Yarlagadda, R., Zuraski, C., Chiu,

W., Cohen, S., Hansen, J.N., Leonetti, M.D., Liu, C., Lundberg, E., Yeung-Levy, S.: Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research (2025), https://arxiv.org/abs/2503.13399

2. Chen, P., Ye, J., Wang, G., Li, Y., Deng, Z., Li, W., Li, T., Duan, H., Huang, Z., Su, Y., Wang, B., Zhang, S., Fu, B., Cai, J., Zhuang, B., Seibel, E.J., He, J., Qiao, Y.: Gmai-mmbench: A comprehensive multimodal evaluation benchmark towards general medical ai (2024), https://arxiv.org/abs/2408.03361
3. Gu, Y., Fu, J., Liu, X., Valanarasu, J.M.J., Codella, N., Tan, R., Liu, Q., Jin, Y., Zhang, S., Wang, J., et al.: The illusion of readiness: Stress testing large frontier models on multimodal medical benchmarks. arXiv preprint arXiv:2509.18234 (2025)
4. He, X., Zhang, Y., Mou, L., Xing, E., Xie, P.: Pathvqa: 30000+ questions for medical visual question answering (2020), https://arxiv.org/abs/2003.10286
5. Hu, Y., Li, T., Lu, Q., Shao, W., He, J., Qiao, Y., Luo, P.: Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm (2024), https://arxiv.org/abs/2402.09181
6. Jeong, D.P., Mani, P., Garg, S., Lipton, Z.C., Oberst, M.: The limited impact of medical adaptation of large language and vision-language models. arXiv preprint arXiv:2411.08870 (2024)
7. Lau, J.J., Gayen, S., Ben Abacha, A., Demner-Fushman, D.: A dataset of clinically generated visual questions and answers about radiology images. Scientific Data **5** (2018). https://doi.org/10.1038/sdata.2018.251, https://www.nature.com/articles/sdata2018251
8. Le, A., Liu, H., Wang, Y., Liu, Z., Zhu, R., Weng, T., Yu, J., Wang, B., Wu, Y., Yan, K., Sun, Q., Jiang, M., Pei, J., Liu, S., Zheng, H., Li, Z., Noble, A., Souquet, J., Guo, X., Lin, M., Guo, H.: U2-bench: Benchmarking large vision-language models on ultrasound understanding (2025), https://arxiv.org/abs/2505.17779
9. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36**, 28541–28564 (2023)
10. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Advances in Neural Information Processing Systems **36** (2024)
11. Liu, B., Zhan, L.M., Xu, L., Ma, L., Yang, Y., Wu, X.M.: Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering (2021), https://arxiv.org/abs/2102.09542
12. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2024), https://arxiv.org/abs/2310.03744
13. Lozano, A., Nirschl, J., Burgess, J., Gupte, S.R., Zhang, Y., Unell, A., Yeung-Levy, S.: {\mu}-bench: A vision-language benchmark for microscopy understanding. arXiv preprint arXiv:2407.01791 (2024)
14. Lozano, A., Nirschl, J., Burgess, J., Gupte, S.R., Zhang, Y., Unell, A., Yeung-Levy, S.: Micro-bench: a vision-language benchmark for microscopy understanding. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. pp. 30670–30685 (2024)
15. Maruf, M., Daw, A., Mehrab, K.S., Manogaran, H.B., Neog, A., Sawhney, M., Khurana, M., Balhoff, J.P., Bakış, Y., Altintas, B., et al.: Vlm4bio: A benchmark

dataset to evaluate pretrained vision-language models for trait discovery from biological images. Advances in Neural Information Processing Systems **37**, 131035–131071 (2024)

16. OpenAI: GPT-4.1 mini. `https://developers.openai.com/api/docs/models/gpt-4.1-mini` (2025), openAI API model documentation. Model ID: gpt-4.1-mini; snapshot: gpt-4.1-mini-2025-04-14. Accessed: 2026-05-27
17. OpenAI: GPT-5.4 mini. `https://developers.openai.com/api/docs/models/gpt-5.4-mini` (2026), openAI API model documentation. Model ID: gpt-5.4-mini; snapshot: gpt-5.4-mini-2026-03-17. Accessed: 2026-05-27
18. Ossowski, T., Zhang, S., Liu, Q., Qin, G., Tan, R., Naumann, T., Hu, J., Poon, H.: Octomed: Data recipes for state-of-the-art multimodal medical reasoning. arXiv preprint arXiv:2511.23269 (2025)
19. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025), `https://arxiv.org/abs/2412.15115`
20. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
21. Sellergren, A., Kazemzadeh, S., Jaroensri, T., Kiraly, A., Traverse, M., Kohlberger, T., Xu, S., Jamil, F., Hughes, C., Lau, C., et al.: Medgemma technical report. arXiv preprint arXiv:2507.05201 (2025)
22. Shi, L., Ma, C., Liang, W., Diao, X., Ma, W., Vosoughi, S.: Judging the judges: A systematic study of position bias in llm-as-a-judge. In: Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics. pp. 292–314 (2025)
23. Team, G.: Gemma 3 technical report (2025), `https://arxiv.org/abs/2503.19786`
24. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine **5**(1), 48 (2022)
25. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., Wang, Z., Chen, Z., Zhang, H., Yang, G., Wang, H., Wei, Q., Yin, J., Li, W., Cui, E., Chen, G., Ding, Z., Tian, C., Wu, Z., Xie, J., Li, Z., Yang, B., Duan, Y., Wang, X., Hou, Z., Hao, H., Zhang, T., Li, S., Zhao, X., Duan, H., Deng, N., Fu, B., He, Y., Wang, Y., He, C., Shi, B., He, J., Xiong, Y., Lv, H., Wu, L., Shao, W., Zhang, K., Deng, H., Qi, B., Ge, J., Guo, Q., Zhang, W., Zhang, S., Cao, M., Lin, J., Tang, K., Gao, J., Huang, H., Gu, Y., Lyu, C., Tang, H., Wang, R., Lv, H., Ouyang, W., Wang, L., Dou, M., Zhu, X., Lu, T., Lin, D., Dai, J., Su, W., Zhou, B., Chen, K., Qiao, Y., Wang, W., Luo, G.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency (2025), `https://arxiv.org/abs/2508.18265`
26. Wright, C., Reeves, P.: Radbench: benchmarking image interpretation skills. Radiography **22**(2), e131–e136 (2016)
27. Xu, W., Chan, H.P., Li, L., Aljunied, M., Yuan, R., Wang, J., Xiao, C., Chen, G., Liu, C., Li, Z., et al.: Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning. arXiv preprint arXiv:2506.07044 (2025)

28. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388
29. Yang, Y., Zhang, H., Gichoya, J.W., Katabi, D., Ghassemi, M.: The limits of fair medical imaging ai in real-world generalization. Nature medicine **30**(10), 2838–2848 (2024)
30. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., Wei, C., Yu, B., Yuan, R., Sun, R., Yin, M., Zheng, B., Yang, Z., Liu, Y., Huang, W., Sun, H., Su, Y., Chen, W.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi (2024), https://arxiv.org/abs/2311.16502
31. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
32. Zhou, J., Li, H., Chen, S., Chen, Z., Han, Z., Gao, X.: Large language models in biomedicine and healthcare. npj Artificial Intelligence **1**(1), 44 (2025)
33. Zhou, Y., Ong, H., Kennedy, P., Wu, C.C., Kazam, J., Hentel, K., Flanders, A., Shih, G., Peng, Y.: Evaluating gpt-4v (gpt-4 with vision) on detection of radiologic findings on chest radiographs. Radiology **311**(2) (2024). https://doi.org/10.1148/radiol.233270, https://pubs.rsna.org/doi/full/10.1148/radiol.233270, pMID: 38712869