

MicroVQA: A Multimodal Reasoning Benchmark for Microscopy-Based Scientific Research

James Burgess*¹ Jeffrey J Nirschl*¹ Laura Bravo-Sánchez^{1,*}
 Alejandro Lozano¹ Sanket Rajan Gupte¹ Jesus G. Galaz-Montoya¹ Yuhui Zhang¹
 Yuchang Su² Disha Bhowmik³ Zachary Coman³ Sarina M. Hasan⁴
 Alexandra Johannesson⁵ William D. Leineweber¹ Malvika G Nair³ Ridhi Yarlagadda³
 Connor Zuraski¹ Wah Chiu¹ Sarah Cohen³ Jan N. Hansen¹ Manuel D Leonetti⁶
 Chad Liu⁶ Emma Lundberg^{1,5} Serena Yeung-Levy¹

¹Stanford University, ²Tsinghua University, ³University of North Carolina at Chapel Hill,
⁴Princeton University, ⁵KTH Royal Institute of Technology, ⁶Chan Zuckerberg Biohub Network

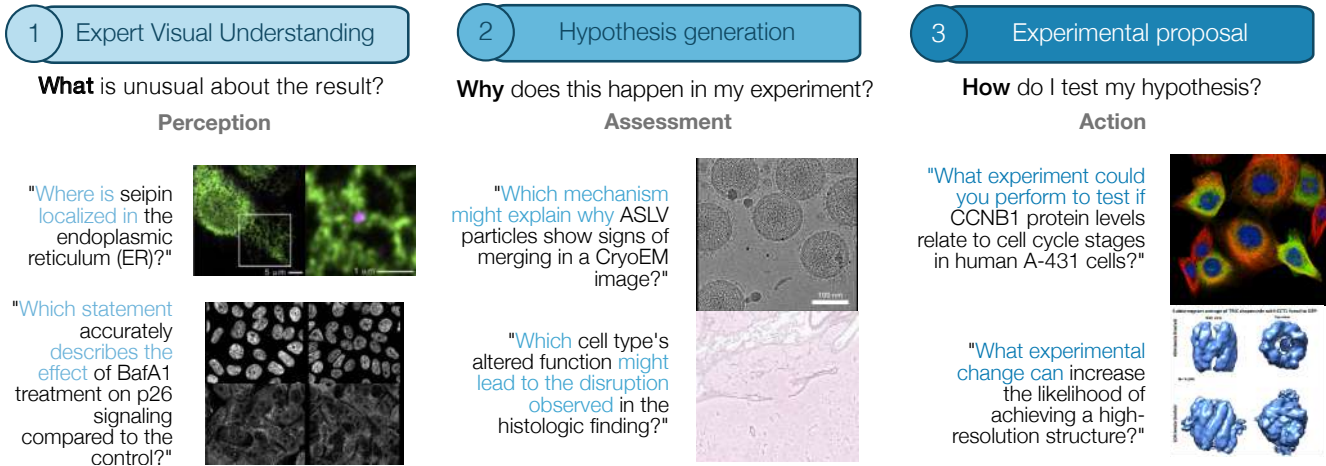


Figure 1. A scientific experimentation cycle that drives discovery: researchers analyze experiments, develop hypotheses, and design further experiments to test their ideas - generating new data that sparks the next cycle. We release MicroVQA, a scientist-curated benchmark for visual question answering (VQA) to test these reasoning capabilities. MicroVQA evaluates scientific reasoning in the context of biological microscopy, with questions spanning expert visual understanding, hypothesis generation, and experimental proposal.

Abstract

Scientific research demands sophisticated reasoning over multimodal data, a challenge especially prevalent in biology. Despite recent advances in multimodal large language models (MLLMs) for AI-assisted research, existing multimodal reasoning benchmarks target up to college-level difficulty, while research-level benchmarks emphasize lower-level perception, falling short of the complex multimodal reasoning needed for scientific discovery. To bridge this gap, we introduce MicroVQA, a visual-question answering (VQA) benchmark designed to assess three reasoning capabilities vital in research workflows: expert image understanding, hypothesis generation, and experiment pro-

posal. MicroVQA consists of 1,061 multiple-choice questions (MCQs) curated by biological experts across diverse microscopy modalities, ensuring VQA samples represent real scientific practice. We find that standard MCQ creation methods do not properly test our targeted reasoning capabilities, motivating a new two-stage pipeline: an optimized LLM prompt structures question-answer pairs into MCQs; then, an agent-based 'RefineBot' generates more challenging distractors. Benchmarking on state-of-the-art MLLMs reveal a peak performance of 43%; models with smaller LLMs only slightly underperform top models, suggesting that language-based reasoning is less challenging than multimodal reasoning; and tuning with scientific articles enhances performance. Expert analysis of chain-of-thought

reasoning failures indicates that multimodal reasoning errors are frequent, followed by knowledge errors and overgeneralization. These insights highlight the challenges in multimodal scientific reasoning, showing MicroVQA is a valuable resource advancing AI-driven biomedical research. MicroVQA is available [here](#).

1. Introduction

The transformative potential of AI in scientific discovery lies in its ability to interpret and reason over complex, multimodal data while integrating specialist domain knowledge. Recently, multimodal large language models (MLLMs) have demonstrated notable capabilities in data comprehension, recall, and multi-step reasoning across various modalities on exam-style questions [1, 3, 6, 40, 57, 67]. This progress has generated enthusiasm for scientific AI applications, ranging from multimodal chat AI assistants [25, 58, 69, 78] to AI agents capable of orchestrating complex experiments [10, 13, 21, 45]. However, while MLLMs perform well on general exam tasks, they often lack the specialized, contextual reasoning needed for real scientific challenges, where tasks require advanced image-based reasoning, analysis, and hypothesis-driven experimentation.

In recent years, efforts to apply scientific MLLMs across a range of domains have laid an essential groundwork. Advances have been achieved in image understanding [32, 34, 38, 41, 47], literature analysis [32, 38, 55, 63], and problem-solving at school and early college levels [43, 71]. In applied settings, AI has facilitated scientific tasks in chemistry for compound design [29], code-based analysis [52, 60], and diagnostic pathology and radiology [42]. However, there has been less progress on tasks requiring *higher-order reasoning* in *research-level* settings. The gap is due in part to a lack of multimodal benchmarks for scientific reasoning – unlike college-level tasks where existing exams can form benchmarks, research tasks require expert curation.

Microscopy provides a compelling use case for developing such a benchmark. It requires more than pattern recognition or factual recall – it demands the ability to synthesize visual information with experimental context, formulate hypotheses, and propose experimental follow-ups. To evaluate these advanced capabilities, we introduce MicroVQA, a visual-question answering (VQA) benchmark tailored for multimodal scientific reasoning. Grounded in the real-world scientific practice of biological microscopy, MicroVQA incorporates over 1,061 multiple-choice questions (MCQs) manually created by biological researchers across diverse disciplines, with each question taking more than 30 minutes to generate. Careful expert curation ensures the benchmark is relevant to real scientific applications.

MicroVQA is designed to evaluate three key reasoning tasks crucial to scientific inquiry: (1) *expert image understanding* – the ability to recognize scientifically relevant

features and integrate contextual information about sample preparation; (2) *hypothesis generation* – the formulation of scientifically plausible explanations from experimental data; and (3) *experiment proposal* – the ability to suggest further experiments that could validate or refute hypotheses. These capabilities align with the iterative cycles of scientific research and demand both abductive reasoning – inferring the best explanation from multiple possibilities – and deductive reasoning – moving from general principles to specific conclusions [67].

We found that generating MCQs from open-ended scientific VQA samples posed unique challenges – standard methods create distractors that are far too easy for MLLMs to eliminate without truly testing multimodal knowledge or reasoning. To ensure question quality and difficulty, we developed a two-stage pipeline for MCQ generation: first, an optimized LLM prompt structures question-answer pairs into well-formed MCQs; then, a novel agent-based ‘RefineBot’ system increases question difficulty by generating more challenging distractors. This ensures that MicroVQA’s MCQs better test true scientific capabilities, rather than simple pattern recognition.

We benchmarked frontier MLLMs on MicroVQA, finding that the best performers achieved a modest peak of 43%, which shows a significant gap between current models and expert-level scientific reasoning. The variation between different MLLMs was small; notably, smaller LLMs only slightly underperform their larger counterparts, suggesting that the key challenge for our MCQs is not language-only reasoning, but multimodal reasoning or knowledge. We also find that finetuning MLLMs on scientific publications improves MicroVQA performance. Finally, a group of biomedical researchers perform a detailed qualitative analysis into MLLM failure modes, uncovering three major failure modes: Expert perception is the most common, followed by knowledge errors, and then overgeneralization errors. This further supports that visual reasoning is a key challenge in MicroVQA.

In summary, our contributions are as follows:

- We introduce MicroVQA, a benchmark for multimodal reasoning in biological microscopy. We defined three key tasks for scientific research – expert image understanding, hypothesis generation, and experiment proposal. Then, expert researchers curated 1,060 diverse VQA samples.
- We develop a novel two-stage MCQ generation pipeline to address the challenge of creating rigorous and challenging MCQs from long-format scientific questions.
- We provide quantitative and qualitative benchmarking of frontier MLLMs, highlighting areas for improvement.

2. Related work

MLLM reasoning benchmarks While *perception* focuses on identifying an image’s visual features, *visual rea-*

soning further integrates prior knowledge to derive new conclusions using logical inference strategies like deduction, induction, abduction, and analogical reasoning [67]. Reasoning in MLLMs has focused on image-language VQA [20, 35, 43, 67, 71, 72]. The most prominent examples are MMMU [71, 72], ScienceQA, [43] and MathVista [44]. While they do test reasoning, they are derived from exams up to the college level, while our benchmark emphasizes research-level difficulty. Other popular multimodal benchmarks test commonsense inference [59, 74], spatial or relational awareness [27, 75] and integrating prior knowledge [46]. Another line of work is visual abductive reasoning [7, 24], or finding the most likely explanation for an observation from incomplete information; which relates closely to the hypothesis generation in this work.

MLLMs in science The most related work considers multimodal VQA in scientific or medical domains like microscopy images [41], biological images [47], medical images [23, 26], and chemical structure drawings [34]. Still, they mostly test more straightforward perception and recognition. Figure comprehension for scientific articles does require more reasoning with multimodal data [32, 36–38, 50]; however, the images here are curated for publication, and the emphasis is on understanding content, rather than performing practical tasks. On the other hand, our benchmark includes realistic experiment images and tests more advanced reasoning like hypothesis generation and experiment planning. Scientific tasks have received much more attention in language-only LLMs [12, 17, 28, 32, 48, 51, 52, 56, 60, 66], with GPQA being a notable work targeting PhD-level multi-step reasoning [51], but these obviously neglect multimodal reasoning, which is important to many scientific areas [21]. In terms of models, most MLLMs in the biomedical domain are targeted at clinical tasks [26, 33, 49, 54, 62, 76].

Multiple-choice question construction A challenge in creating exams for education research and benchmarks in machine learning is generating multiple-choice questions with quality distractors [4, 22], commonly using LLM generation. While zero-shot distractor prompting works in some cases [26], others perform few-shot prompting or LLM-tuning with high-quality examples [8, 18, 19]. A recent result highlighted by Cambrian-1 showed that many recent VQA benchmarks are exploitable by language shortcuts, with MLLMs having similar evaluation scores without having access to the image [61]. A specific language-only strategy known in VQA is the ‘language bias’ – models ignore the image, and make guesses corresponding to which MCQ option is statistically more likely under the pretraining set. In this work, we find that zero-shot MCQ generation is inadequate, likely due to language shortcuts; we therefore

Table 1. MicroVQA benchmark attributes.

Dataset feature	Value
Total questions	1,061
Multi-image questions	423
Avg. MCQ question length	66
Avg. MCQ answer length	15
Avg. raw question length	158
Avg. raw answer length	52
Unique image sets	255
Image Modalities	Light, Fluoro, Electron
Image Scales	Tissue, Cell, Subcell, Atomic
Organisms	31
Research areas	33
Expert question creators	12
Time to create 1 question	30-40 mins
Time to quality check 1 MCQ	5 mins

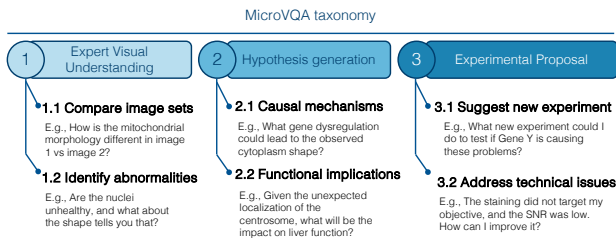


Figure 2. MicroVQA taxonomy of sub-tasks.

develop a novel method to improve MCQ generation.

3. The MicroVQA Benchmark

Here, we present the MicroVQA benchmark for visual-question-answering (VQA) which advances multimodal scientific reasoning by tackling two major evaluation challenges. First, defining tasks is challenging because scientific research encompasses many complex activities [25] – we therefore define three important tasks important to biomedical discovery. Second, VQA samples cannot be readily sourced from existing exams or classification datasets – we therefore assembled a team of biologists to curate challenging reasoning problems. In Sec. 3.1, we provide an overview of the MicroVQA benchmark; Sec. 3.2 details the key tasks; and finally, Sec. 3 demonstrates that MicroVQA is at the forefront of measuring research-level reasoning in MLLMs. In the later Sec. 4, we describe our approach for generating multiple-choice questions.

3.1. Overview of MicroVQA

MicroVQA is a comprehensive dataset of 1,061 VQA triplets, manually curated by a team of expert biological researchers from diverse fields, and key attributes are summarized in Tab. 1. These expert-crafted questions test reasoning across three key tasks: expert understanding, hypothesis generation, and experiment proposal, and we additionally provide a taxonomy of sub-tasks in Sec. 3. The

questions cover a broad spectrum, ranging from biological issues, such as “What is unusual about the cell shape?” – to technical imaging challenges like “Is this feature due to low signal-to-noise ratio (SNR) or is it real?”. We provide the ‘raw’ questions and answers written by experts, which are often long and nuanced, along with converted multiple-choice questions (MCQs) suitable for MLLM evaluation.

The images encompass the most common microscopy modalities used in human biology: brightfield, fluorescence, and electron microscopy. The sample types span the full range of microscopic scales – tissue, cellular, sub-cellular, and atomic – and emphasize organisms relevant to human biology and medically motivated tasks, namely human and mouse. Over 60% of the samples have multiple images because comparison is essential in microscopy research, and because multichannel fluorescence images cannot be represented in standard RGB. All these attributes are provided as metadata tags to facilitate deeper error analysis, thereby enhancing MicroVQA’s value as a resource for advancing AI capabilities in scientific research.

3.2. MicroVQA scientific reasoning tasks

To define specific reasoning tasks aimed at advancing biomedical research, we established the following criteria: (1) the tasks should use image-text inputs and text outputs suitable for MLLMs, (2) they should require higher-order reasoning beyond simple image processing, and (3) they should emphasize core experimental activities such as experimentation, analysis, and action [25], rather than auxiliary tasks like literature review or writing. The tasks were developed through interviews with nine co-author PIs and postdocs specializing in microscopy research (Appendix D). We identified three essential capabilities that we now introduce, with examples shown in Sec. 3.

Expert visual understanding After performing an experiment, scientists must interpret data within its experimental context. In biological microscopy, this involves identifying patterns like protein distributions or cell morphologies, while also assessing technical aspects such as artifacts and noise levels. The task requires MLLMs to perform anomaly detection and image comparison, going beyond simple perception in two key ways: analysis must consider sample preparation context, and expert knowledge is needed to evaluate biological features and technical artifacts.

Hypothesis generation The next step is proposing mechanistic hypotheses to explain experimental data. For example, when a drug treatment causes mitochondrial fragmentation in cells, a hypothesis might be that the drug disrupted calcium homeostasis, activating the DRP1 protein through a multi-step pathway, leading to fragmentation. This requires *abductive reasoning* [7, 15, 24, 67] as one must se-

lect from many possible hypotheses given incomplete information. The process demands the MLLM to integrate relevant domain knowledge with experimental context and visual features to reason about which hypothesis best explains the observations.

Experiment proposal The final step is designing experiments to validate hypotheses. In microscopy, this often involves selecting appropriate assays and controls, requiring knowledge of suitable protocols and *deductive reasoning* about whether proposed experiments will provide evidence for or against the hypothesis. This task also may involve adjusting a particular experiment to address technical issues, like how to ensure a good signal-to-noise ratio, and this requires both image understanding and knowledge of protocols.

Table 2. Comparing scientific multimodal benchmarks close to MicroVQA for in terms of reasoning level or difficulty beyond college level. We show difficulty level, scientific domain, dataset source, and size. Compared to others, MicroVQA either has higher difficulty level, or it has higher reasoning level (which is established in Fig. 3). Compared to others at the same reasoning level, namely MMMU-Pro, it has similar size.

Benchmark	Difficulty level	Domain	Source	Size
MicroVQA	research	microscopy	expert-curated	1,061
LabBench [32]	research	biology	webQA	181
PathVQA [23]	graduate	pathology	textbooks	16.3k
OmnimedVQA* [26]	graduate	medical	classification dataset	127.9k
Microbench [41]	graduate	microscopy	classification dataset	17.2k
MMMU [71]	undergraduate	general	textbooks, webQA	11k
MMMU Pro [72]	undergraduate	general	MCQ dataset	1,730
Science QA [43]	grade-school	science	exams	16.8k

3.3. Analysis of MicroVQA benchmark

Scientific discovery requires complex reasoning beyond basic perception and knowledge, and it requires expertise beyond school-level education. Here we show that MicroVQA addresses a gap that prior multimodal scientific benchmarks do not test *high level reasoning* at *research-level* difficulty.

To assess reasoning levels in MLLM benchmarks, we use Bloom’s taxonomy [9], which hierarchically classifies cognitive difficulty from recall to evaluation. We fine-tune GPT-4o to classify Bloom’s taxonomy on related MLLM benchmarks with results in Fig. 3; in Tab. 2 we additionally show key attributes of the same benchmarks. Fig. 3 shows that benchmarks having research-level and graduate-level difficulty are dominated by lower reasoning questions (levels 1-2), while our MicroVQA has higher level reasoning (levels 3-4). This is not surprising since OmnimedVQA

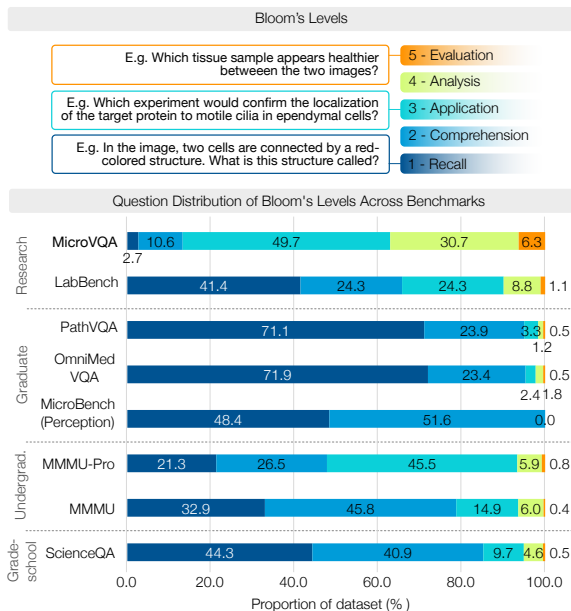


Figure 3. Composition of scientific MLLM benchmarks regarding estimated Bloom’s taxonomy [9]. Higher levels are more cognitively challenging. MicroVQA has more questions at higher levels compared to other benchmarks, for example, MMMU [71] and ScienceQA [43], while perception-driven medical benchmarks like OmniMedVQA are at lower levels.

and MicroBench derive questions from prior classification datasets, while our task definition and data collection approach does reflect high-level reasoning.

On the other hand, some benchmarks have higher Bloom’s reasoning level closer to MicroVQA – MMMU, MMMU-Pro, and Science-QA – but their difficulty level is lower – undergraduate or grade-school. The most comparable benchmark for reasoning is the undergraduate-level MMMU-Pro, and its dataset size is 1,730 compared to our 1,061: this shows that for reasoning-intensive benchmarks, sizes beyond 1,000 are very high.

4. MCQ generation process

Here we describe the complete MicroVQA curation process for creating ‘raw’ VQA samples and mapping them to challenging multiple choice questions (MCQs), which is summarized in Fig. 4.

4.1. Creation of raw VQA samples by experts

Expert Generation of Raw VQA Samples Creating VQA triples suitable for research-level reasoning tasks is time-consuming and requires expertise. We recruited 12 human experts to each create approximately 90 VQA samples, taking about 30–40 minutes per sample. Each submission included an image set, a question, a detailed response, and contextual information like experimental context and mo-

tivation. They are raw VQA triples, (v_0, q_0, a_0) – ‘raw’ means the question and answer can be much longer and more detailed than typical VQA benchmarks. This allowed expert annotators to include enough detail to match what is a valuable input and response for an MLLM. Appendix E shows materials that prompt question creators. The images were sourced from the contributor’s own experiments, image databases, or research papers published since 2024, and all image licenses permit redistribution.

Quality Control We had three quality control criteria: questions should be challenging and not solvable with undergraduate-level knowledge; they should be aligned with the defined tasks; and there should be diversity in image content. Questions were reviewed against these criteria for each contributor, and feedback was provided for further VQA samples. These review rounds occurred after submitting the first 10 questions, and the first 40 questions.

4.2. MCQ generation from raw VQA samples

To benchmark MLLMs, we transform raw expert-created VQA samples to multiple-choice questions (MCQs). We propose new methods for MCQ generation, which we first motivate.

Motivation: naive MCQ generation is inadequate The standard approach for VQA benchmarks is to zero-shot prompt an LLM with raw VQA samples [4], but this has two problems. The first is that generated MCQs do not follow established educational MCQ design principles (Appendix E.2), likely because the raw VQA questions and answers are often long and variable. Our stage 1 ‘exam alignment’, which we describe in the next section, addresses this challenge. The second challenge is that generated MCQs do not properly test MLLM capabilities. Evaluating GPT-4o on these MCQs scores 93% even *without the image*, despite most of the dataset questions clearly requiring image analysis, and we investigate this by inspecting the chain-of-thought outputs in Fig. 4 and Appendix E.2. Many distractors can be easily eliminated based on general biology knowledge, or because they are too vague compared to the correct option. We hypothesize that MCQ and distractor generation is challenging when the target answer string is long, and where the subject matter is specialized, suggesting that other scientific benchmarks may encounter similar issues. Overall, this motivates a more involved approach to distractor generation.

Stage 1: Exam-style MCQ Generation Here we ensured that multiple-choice questions (MCQs) conformed to established design principles for biomedical exams (Fig. 4). This effort was led by a physician-scientist co-author trained

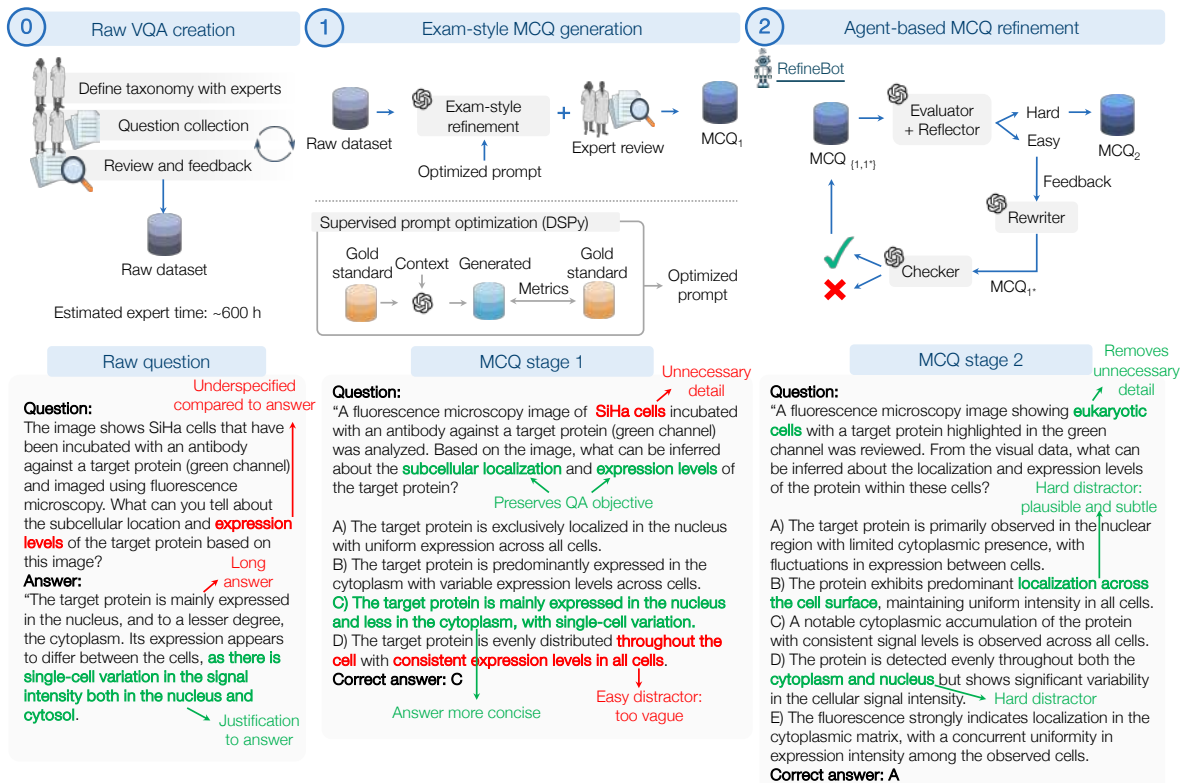


Figure 4. Constructing the MicroVQA multiple choice questions. (0) We defined tasks with domain biological scientists and created 1,061 raw VQA samples. (1) The raw samples were aligned to an exam-style MCQ by manually transforming a small set and optimizing an LLM prompt to match that alignment. (2) MCQs are further improved using RefineBot, a new iterative method to make MCQs more challenging. The lower panel shows an example MCQ from raw VQA to final: the annotations highlight key changes that we further explore in Appendix E.2, where red indicates issues, and green indicates good attributes.

in medical board exam item writing, who reviewed educational literature on MCQ design [53] and Bloom’s Taxonomy for assessing cognitive skills in biomedical exams [5, 14, 73]. We began by manually transforming 50 user-submitted raw question-answer pairs, (q_0, a_0) , into one-best-answer MCQs with k distractors, (q_1, a_1, \mathbf{d}_1) . These ‘Gold Standard’ MCQs retained the original questions’ meanings while strictly adhering to NBME standards [53] and minimizing cues that enable test-wise guessing. We used these gold-standard MCQs in a supervised learning setup to develop a general LLM prompt, p , that maps raw question-answer pairs to exam-style MCQs. We leverage the DSPy framework [30, 31], which optimizes p to match the gold-standard MCQ outputs. Additionally, the optimization process incorporated supervision from LLM-based quality metrics: content similarity, NBME-aligned formatting, and the absence of extraneous clues. Further details are in Appendix E.2.

Stage 2: RefineBot question refinement Although Stage 1 produces exam-aligned MCQs, many remain easily solvable due to eliminable distractors (see Appendix D). To enhance difficulty, we introduce *RefineBot*, a method that

increases MCQ complexity (Fig. 4). The key idea is that weaknesses in MCQ construction are revealed by the chain-of-thought (CoT) answers. The first LLM agent, the *evaluator/reflector*, answers the MCQ with CoT and then reflects on the strategies used. For example, a question about processes inside a virus might have a distractor referring to surface processes – this can be eliminated without referencing the image. The LLM summarizes its solution strategy into a reflection text: in this case, that the distractor was implausible based on knowledge that the stated process does not occur inside the virus, and so does not match the question. It passes the reflection to the *rewriter* LLM agent. The rewriter revises the question stem and generates new distractors to invalidate the identified strategies: in this example it may create distractors referring to processes that do occur inside a virus. To prevent significant changes to the question-answer pair over iterations, we employ an LLM *checker* to ensure semantic consistency with the original pair. The revised MCQ is returned to the evaluator; if it still answers correctly, the loop continues. The process stops if the rewriter fails the *checker* or after n iterations. If it fails, RefineBot can be rerun with a different seed, often succeeding—a form of inference-time scaling [11]. All agents are GPT-4o-0806, but to mitigate potential

Table 3. VQA accuracy on MicroVQA by task: expert visual understanding (V), hypothesis generation (H), experiment proposal (E). Models marked * were used in MCQ generation, which may affect comparative performance, which we investigate in Sec. 5.1. We therefore additionally include models marked **, which are ‘weaker’ models from the same provider. The model † is the base LLaVA for LLaVA-Med. Best values per model subset in bold.

	Model	V	H	E	Overall
closed source	*GPT-4o [1]	40.3	32.7	32.0	35.4
	**GPT-4o-mini	41.3	38.5	36.0	39.0
	*Claude 3.5 Sonnet [6]	45.3	38.2	34.7	39.0
	**Claude-3-Haiku	46.5	38.2	45.3	42.8
	Gemini-1.5-Pro [57]	48.3	41.0	43.1	44.2
	Gemini-1.5-Flash-8B	48.5	37.3	40.4	42.2
open source	VILA1.5-40B [39]	46.5	39.2	39.6	42.0
	VILA1.5-13B	41.8	35.4	41.3	39.1
	Qwen2-VL-72B [65]	41.3	41.9	42.4	41.8
	Qwen2-VL-7B	44.3	39.2	40.9	41.5
	Llama 3.2 90B [16]	39.1	35.3	39.1	37.5
	Llama 3.2 11B	30.1	28.1	28.4	28.9
	Pixtral 12B [2]	39.8	33.6	32.0	35.6
medical	LLaVA-Med-Mistral-7B [33]	43.0	37.3	47.1	41.6
	†LLaVA-Mistral-7B [40]	39.8	31.6	43.1	37.1
	Random	22.0	21.9	21.8	21.9

Table 4. Ablation of MicroVQA accuracy scores during the MCQ generation (process shown in Fig. 4). Accuracy is high because MCQs have shortcuts (Sec. 4.1) after ‘Stage 1’ exam alignment for a challenging benchmark, but is lower after ‘Stage 2’. Final column is the relative decrease in accuracy. Models with * were used in Stage 2 generation and have the biggest accuracy drops (**bolded**). They are grouped with different models from the same provider in **.

	Stage 1	Stage 2	Relative drop
*GPT-4o	79.2	33.8	-57.3
**GPT-4o-mini	75.7	39.0	-48.4
*Claude-3.5-Sonnet	82.0	39.8	-51.5
**Claude-3-Haiku	76.8	42.8	-44.2
Gemini-1.5-Pro	79.5	44.1	-44.5
Gemini-1.5-Flash	74.5	42.2	-43.3
Qwen2-VL-72B	72.8	41.9	-42.5
Qwen2-VL-7B	68.7	41.5	-39.6

bias against 4o in final evaluation, we also use Claude-3.5-Sonnet-0620 as the *evaluator* agent; we assess biases in the experiments section.

5. Experiments

5.1. Benchmarking MLLMs with MicroVQA

We evaluate state-of-the-art multimodal large language models (MLLMs) on the MicroVQA benchmark. To ensure a comprehensive assessment, we include multiple model sizes spanning three categories: closed source, open source, and specialized medical models – medical is the closest domain to microscopy research with specialist MLLMs. We

utilize standard chain-of-thought prompting [72], but for specialist models that struggle with formatting predictions, we employ an additional parsing method (Appendix F). We report mean accuracy in multiple-choice VQA tasks for each model over three random seeds in Tab. 3.

MicroVQA is challenging for all MLLMs Our evaluation reveals a substantial performance gap between current MLLMs and expert-level microscopy analysis in MicroVQA. The leading model, Gemini Pro, achieves only 44.2 overall, and the highest performing task – expert visual understanding – scores only 48.5. There is surprisingly little variation in performance between models: the open source VILA-40B and Qwen2-VL-72B score 42.0 and 41.8, with Llama-90B only 4.5 points lower. GPT-4o and Claude-Sonnet were used in MCQ construction, which causes a small bias that we discuss below. However other models from the same providers – GPT-4o-mini and Claude-Haiku – also perform similarly to the top-performing Gemini at 39.0 and 42.8. Next, we consider performance across the 3 tasks. The ‘hypothesis generation’ task stands out as the hardest for all models, and the gap is strongest for smaller models. For example hypothesis generation is 4.9 below overall score in Gemini Flash, and 3.7 worse on VILA-13b. On the other hand, the visual understanding task has higher scores in most models, for example by 4.1 points on the top-performing Gemini Pro. Overall though, no frontier model stands out as significantly stronger. We hypothesize that the hardest questions in MicroVQA emphasize microscopy image understanding, and that no generalist models have good microscopy image understanding.

Smaller models are surprisingly competitive For all models with a lower parameter version – Gemini, QwenVL, VILA, and Llama – the drop in performance due to size is less than 3 for all except Llama, where the drop is 8.6. Moreover, these drops are smaller than on other multimodal reasoning benchmarks like MMMU, where for example, Gemini drops by 9.6 between Pro and Flash-8B [71]. Typically, smaller models have the same vision encoder, but a smaller LLM [65]. Since the size of the language model has a small impact, we hypothesize that for the ‘solved’ part of MicroVQA, the language-only reasoning is relatively simple compared to domains like math that require multi-step logic. This suggests that other aspects of MLLM reasoning are more challenging, like multimodal reasoning. If true, this suggests that future work could focus on stronger image representations, and this idea is supported by the qualitative error analysis in Sec. 5.2.

Specialist biomedical training does improve performance While no specialist MLLMs have been developed

for microscopy research in particular, LLaVA-Med is fine-tuned on scientific articles from PubMed that include data types overlapping with MicroVQA – especially (tissue) pathology. In Tab. 3, we compare LLaVA-Med against its base model, LLaVA-Mistral-7B and find overall stronger performance by 4.5 points. This suggests two clear opportunities for improving MicroVQA performance. The first is to instruct-tune with a dataset that better aligns with the MicroVQA data domain: not only tissue pathology, but also fluorescence and electron microscopy, which is available in scientific articles [77]. The second is to simply begin training from a stronger base model like VILA. Beyond that, there is more to explore in MLLM specialization, such as tuning the base image encoder for microscopy.

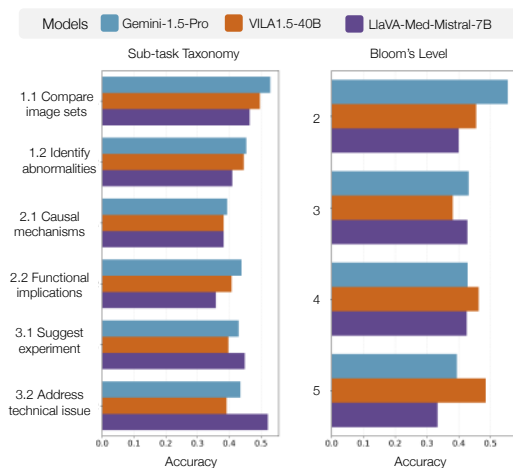


Figure 5. Performance by sub-task and Bloom’s level for best models: Gemini-1.5-Pro (closed source), VILA1.5-40B (open-source), and LLaVA-Med-Mistral-7B (medical).

MCQ generation is very effective but can disadvantage RefineBot models

We introduced a new MCQ generation method via our agent-based system, RefineBot, to make MCQs more challenging. Tab. 4 shows that between stage 1 and stage 2 (before and after RefineBot), the relative accuracy for all models drops by between 39% and 57%. However, the most significant drops are for models used in the refining process (GPT-4o and Claude-3.5-Sonnet), suggesting that the adversarial process introduced a small bias against these models. Further evidence is that their weaker related models – GPT-4o-mini and Claude-3-Haiku – had smaller accuracy drops. Despite making fair evaluation challenging, our use of frontier LLMs in the refinement process is well-motivated, as these models are best positioned to generate challenging, high-quality MCQs that probe the boundaries of current capabilities. Moreover, the bias is relatively modest – less than 10 points between GPT-4o and Gemini – while the benefit is a huge increase in difficulty across all test MLLMs, as seen in Tab. 4.

Metadata-based analysis supports deeper error analysis

Fig. 5 shows error rates based on attribute tag. In Appendix F we explore results further – namely that multi-image reasoning has surprisingly good performance, higher level Bloom’s questions are harder, and (not shown) rare image modalities perform worse.

5.2. Error analysis on MicroVQA

To understand MicroVQA’s performance, three biomedical experts reviewed the Chain-of-Thought (CoT) responses of 30 samples on Claude 3.5 Sonnet, with each review requiring at least 45 mins of careful analysis. Errors were classified into a major category and minor contributing errors were also noted. Expert perception accounted for 50% of errors, while 30% were knowledge errors, and 13% were due to overgeneralization or oversimplification errors. Additionally, we identified that the most common categories were perception error, knowledge, and overgeneralization. The remaining errors were due to reasoning or hallucination. Since the dominant error type is perception, we show an example in Appendix F.

Expert perception errors occurred when the model misinterpreted visual features, leading to incorrect inferences. As shown in the Appendix F ‘perception’ example, the response correctly perceives small electron-dense objects but incorrectly classifies them as ribosomes. In contrast, ribosomes tend to be larger and have a more regular shape and lower contrast. Language associations may have contributed to the preference for ribosomes since they have been studied by EM since the 1950s. Overall, perception errors were the dominant error in our analysis, which suggests that future methods should improve vision representations in MLLMs for microscopy.

Knowledge errors highlight gaps in nuanced biomedical knowledge. One example from this category in Appendix F deals with understanding how protein localization changes during cell signaling. An experiment gives information about a transmembrane signaling protein, and the question is asked to interpret the image and explain the function in these cells. The model perceives a punctate distribution and assumes the role of coordinating intracellular vesicle trafficking. This represents a shallow interpretation that could have been mitigated with specialized biomedical knowledge. These errors could be mitigated by giving MLLMs access to biology databases at inference time.

Overgeneralization errors reflect the model’s tendency to apply broad scientific principles without regard for specific context. In one example, the question asks to compare the risk of malignancy of a tumor, schwannoma, to other

nerve sheath tumors. During reasoning, the model proposes a simplified question—what is true about schwannoma—which proceeds to answers while ignoring the comparison.

6. Conclusion

MLLMs hold great potential for advancing scientific discovery. We aim to inspire the development of broadly applicable scientific AI systems. Future directions include training models for more robust image representations, integrating knowledge bases, exploring open evaluation methods, and studying LLMs’ reasoning over hypotheses. Beyond microscopy, we hope our benchmark serves as a blueprint for AI evaluation in expert-driven domains, fostering collaboration across fields like biology, chemistry, and materials science.

References

- [1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 7
- [2] Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Devendra Chaplot, Jessica Chudnovsky, Saurabh Garg, Theophile Gervet, Soham Ghosh, Amélie Héliou, Paul Jacob, et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 7
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 2
- [4] Elaf Alhazmi, Quan Z Sheng, Wei Emma Zhang, Munazza Zaib, and Ahoud Alhazmi. Distractor generation for multiple-choice questions: A survey of methods, datasets, and evaluation. *arXiv preprint arXiv:2402.01512*, 2024. 3, 5
- [5] Lorin W. Anderson and David R. Krathwohl, editors. *A Taxonomy for Learning, Teaching, and Assessing. A Revision of Bloom’s Taxonomy of Educational Objectives*. Allyn & Bacon, New York, 2 edition, 2001. 6, 8
- [6] Anthropic. The claude 3 model family: Opus, sonnet, haiku. Model card, Anthropic, 2024. 2, 7
- [7] Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen-tau Yih, and Yejin Choi. Abductive commonsense reasoning. *arXiv preprint arXiv:1908.05739*, 2019. 3, 4
- [8] Semere Kiros Bitew, Johannes Deleu, Chris Develder, and Thomas Demeester. Distractor generation for multiple-choice questions with predictive prompting and large language models. *arXiv preprint arXiv:2307.16338*, 2023. 3
- [9] Benjamin S. Bloom. *Taxonomy of Educational Objectives: The Classification of Educational Goals. Handbook I: The Cognitive Domain*. David McKay Co Inc, New York, 1956. 4, 5
- [10] Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. 2
- [11] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024. 6
- [12] Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Yongge Li, Mujie Lin, Shuwen Yang, et al. Sciassess: Benchmarking llm proficiency in scientific literature analysis. *arXiv preprint arXiv:2403.01976*, 2024. 3
- [13] Anne E Carpenter, Beth A Cimini, and Kevin W Eliceiri. Smart microscopes of the future. *nature methods*, 20(7):962–964, 2023. 2
- [14] Alison Crowe, Clarissa Dirks, and Mary Pat Wenderoth. Biology in bloom: implementing bloom’s taxonomy to enhance student learning in biology. *CBE Life Sci. Educ.*, 7(4):368–381, 2008. 6, 8
- [15] Igor Douven. Abduction, 2011. 4
- [16] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7
- [17] Kehua Feng, Keyan Ding, Weijie Wang, Xiang Zhuang, Zeyuan Wang, Ming Qin, Yu Zhao, Jianhua Yao, Qiang Zhang, and Huajun Chen. Sciknoweval: Evaluating multi-level scientific knowledge of large language models. *arXiv preprint arXiv:2406.09098*, 2024. 3
- [18] Wanyong Feng, Jaewook Lee, Hunter McNichols, Alexander Scarlatos, Digory Smith, Simon Woodhead, Nancy Otero Ornelas, and Andrew Lan. Exploring automated distractor generation for math multiple-choice questions via large language models. *arXiv preprint arXiv:2404.02124*, 2024. 3
- [19] Nigel Fernandez, Alexander Scarlatos, Simon Woodhead, and Andrew Lan. Divert: Distractor generation with variational errors represented as text for math multiple-choice questions. *arXiv preprint arXiv:2406.19356*, 2024. 3
- [20] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 3
- [21] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. Empowering biomedical discovery with ai agents. *arXiv preprint arXiv:2404.02831*, 2024. 2, 3
- [22] Mark J Gierl, Okan Bulut, Qi Guo, and Xinxin Zhang. Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of educational research*, 87(6):1082–1116, 2017. 3, 7
- [23] 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. 3, 4

- [24] Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. The abduction of sherlock holmes: A dataset for visual abductive reasoning. In *European Conference on Computer Vision*, pages 558–575. Springer, 2022. 3, 4
- [25] Tom Hope, Doug Downey, Daniel S Weld, Oren Etzioni, and Eric Horvitz. A computational inflection for scientific discovery. *Communications of the ACM*, 66(8):62–73, 2023. 2, 3, 4, 6
- [26] Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183, 2024. 3, 4, 2, 7
- [27] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 3
- [28] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. Pubmedvqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*, 2019. 3
- [29] Yeonghun Kang and Jihan Kim. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. *Nat. Commun.*, 15(1):4705, 2024. 2
- [30] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive NLP. *arXiv preprint arXiv:2212.14024*, 2022. 6
- [31] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023. 6
- [32] Jon M Laurent, Joseph D Janizek, Michael Ruzo, Michaela M Hinks, Michael J Hammerling, Siddharth Narayanan, Manvitha Ponnampati, Andrew D White, and Samuel G Rodrigues. Lab-bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024. 2, 3, 4
- [33] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36, 2024. 3, 7
- [34] Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, et al. Chemvlm: Exploring the power of multimodal large language models in chemistry area. *CoRR*, 2024. 2, 3
- [35] Lin Li, Guikun Chen, Hanrong Shi, Jun Xiao, and Long Chen. A survey on multimodal benchmarks: In the era of large ai models. *arXiv preprint arXiv:2409.18142*, 2024. 3
- [36] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*, 2024. 3
- [37] Shengzhi Li and Nima Tajbakhsh. Scigraphqa: A large-scale synthetic multi-turn question-answering dataset for scientific graphs. *arXiv preprint arXiv:2308.03349*, 2023.
- [38] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, et al. Mmsci: A multimodal multi-discipline dataset for phd-level scientific comprehension. *arXiv preprint arXiv:2407.04903*, 2024. 2, 3
- [39] Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26689–26699, 2024. 7
- [40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 7
- [41] Alejandro Lozano, Jeffrey Nirschl, James Burgess, Sanket Rajan Gupte, Yuhui Zhang, Alyssa Unell, and Serena Yeung-Levy. μ -bench: A vision-language benchmark for microscopy understanding. *arXiv preprint arXiv:2407.01791*, 2024. 2, 3, 4
- [42] Ming Y ”Lu, Bowen Chen, Drew F K Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V Parwani, Luca L Weishaupt, and Faisal” Mahmood. ”a multimodal generative AI copilot for human pathology”. *Nature*, 634(8033):”466–473”, 2024. 2
- [43] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521, 2022. 2, 3, 4, 5
- [44] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3
- [45] Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, pages 1–11, 2024. 2
- [46] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 3
- [47] M Maruf, Arka Daw, Kazi Sajeed Mehrab, Harish Babu Manogaran, Abhilash Neog, Medha Sawhney, Mridul Khurana, James P Balhoff, Yasin Bakis, Bahadir Altintas, et al. Vlm4bio: A benchmark dataset to evaluate pretrained vision-

- language models for trait discovery from biological images. *arXiv preprint arXiv:2408.16176*, 2024. 2, 3
- [48] Adrian Mirza, Nawaf Alampara, Sreekanth Kunchapu, Benedict Emoekabu, Aswanth Krishnan, Mara Wilhelmi, Macjonathan Okereke, Juliane Eberhardt, Amir Mohammad Elahi, Maximilian Greiner, et al. Are large language models superhuman chemists? *arXiv preprint arXiv:2404.01475*, 2024. 3
- [49] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023. 3
- [50] Shraman Pramanick, Rama Chellappa, and Subhashini Venugopalan. Spiqqa: A dataset for multimodal question answering on scientific papers. *arXiv preprint arXiv:2407.09413*, 2024. 3
- [51] David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023. 3
- [52] Loïc A Royer. Omega—harnessing the power of large language models for bioimage analysis. *Nature Methods*, pages 1–3, 2024. 2, 3
- [53] Billings Melissa S, DeRuchie Kristine, Hussie Kieran, Kulesher Allison, Merrell Jacquelyn, Morales Amy, Paniagua Miguel A, Sherlock Jennifer, Swygert Kimberly, and Tyson Julie. Nbm item-writing guide: Constructing written test questions for the health sciences, 2021. 6, 8
- [54] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, et al. Capabilities of gemini models in medicine. *arXiv preprint arXiv:2404.18416*, 2024. 3
- [55] Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnampati, Samuel G Rodrigues, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024. 2
- [56] Liangtai Sun, Yang Han, Zihan Zhao, Da Ma, Zhennan Shen, Baocai Chen, Lu Chen, and Kai Yu. Scieval: A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 19053–19061, 2024. 3
- [57] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2, 7
- [58] Surendrabikram Thapa and Surabhi Adhikari. Chatgpt, bard, and large language models for biomedical research: opportunities and pitfalls. *Annals of biomedical engineering*, 51(12):2647–2651, 2023. 2
- [59] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visiolinguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 3
- [60] Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kit-tithat Krongchon, Yao Li, et al. Scicode: A research coding benchmark curated by scientists. *arXiv preprint arXiv:2407.13168*, 2024. 2, 3
- [61] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 3
- [62] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 1(3):AIoa2300138, 2024. 3
- [63] David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, et al. Sciriff: A resource to enhance language model instruction-following over scientific literature. *arXiv preprint arXiv:2406.07835*, 2024. 2
- [64] Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. *arXiv preprint arXiv:2406.09411*, 2024. 15
- [65] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 7
- [66] Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *arXiv preprint arXiv:2307.10635*, 2023. 3
- [67] Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*, 2024. 2, 3, 4
- [68] Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhramil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, et al. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*, 2024. 7
- [69] Tong Xie, Yuwei Wan, Wei Huang, Yufei Zhou, Yixuan Liu, Qingyuan Linghu, Shaozhou Wang, Chunyu Kit, Clara Grazian, Wenjie Zhang, et al. Large language models as master key: unlocking the secrets of materials science with gpt. *arXiv preprint arXiv:2304.02213*, 2023. 2
- [70] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. Mmt-bench: A comprehensive multimodal bench-

- mark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006*, 2024. [2](#)
- [71] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. [2](#), [3](#), [4](#), [5](#), [7](#)
- [72] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Ming Yin, Botao Yu, Ge Zhang, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*, 2024. [3](#), [4](#), [7](#), [14](#)
- [73] Nikki B Zaidi, Charles Hwang, Sara Scott, Stefanie Stallard, Joel Purkiss, and Michael Hortsch. Climbing bloom’s taxonomy pyramid: Lessons from a graduate histology course. *Anat. Sci. Educ.*, 10(5):456–464, 2017. [6](#), [8](#)
- [74] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019. [3](#)
- [75] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5317–5327, 2019. [3](#)
- [76] Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, pages 1–13, 2024. [3](#)
- [77] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Hanwen Xu, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv preprint arXiv:2303.00915*, 2023. [8](#), [2](#), [15](#)
- [78] Zhiling Zheng, Oufan Zhang, Ha L Nguyen, Nakul Rampal, Ali H Alawadhi, Zichao Rong, Teresa Head-Gordon, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. Chatgpt research group for optimizing the crystallinity of mofs and cofs. *ACS Central Science*, 9(11):2161–2170, 2023. [2](#)

MicroVQA: A Multimodal Reasoning Benchmark for Microscopy-Based Scientific Research

Supplementary Material

Appendices A-C include acknowledgments, author contributions, ethics, and discussion of limitations and future work. The remaining sections correspond to sections in the paper. Appendix D is benchmark details from the main Sec. 3. Appendix E is MCQ generation details from the main Sec. 4. Appendix F is the experiments details from the main Sec. 5.

We release an anonymous codebase called `microVQA`, which has the code used for dataset construction and evaluation. They are attached as supplementary material, and we will release all code to the public upon publication.

Table of Contents

A Acknowledgments and author contributions	1
B Ethics	1
C Limitations and future work	2
D Benchmark details	3
D.1 Accessing MicroVQA Benchmark	3
D.2 Dataset Schema and Structure	3
D.3 Training contamination mitigation	3
D.4 Benchmark attribute details	3
D.5 Related benchmarks	4
D.6 Bloom’s taxonomy	4
D.7 Methodology for proposing tasks	5
E MCQ generation details	6
E.1 Question collection process details	6
E.2 Experiment on naive MCQ generation	7
E.3 Stage 1 MCQ generation (exam alignment)	7
E.4 Stage 2 MCQ generation (RefineBot) details	8
E.5 Final MCQ quality check	12
E.6 MCQ generation results: stage 1 exam alignment	12
E.7 MCQ generation results: stage 2 RefineBot	13
F. Experiments details	14
F.1. Evaluation prompts	14
F.2. Model details	14
F.3. More experiments on attribute-level information	15
F.4. Error analysis	16

A. Acknowledgments and author contributions

B. Ethics

Ethical use of biomedical data: MicroVQA was developed with a commitment to ethical practices in handling biomedical research data. All microscopy images included in the dataset were either original, unpublished data or sourced from open-access articles (e.g., CC-BY licensed) published after January 2024. The dataset does not include patient-identifiable

information, and MicroVQA adheres to applicable privacy and research ethics guidelines. The dataset does not constitute human subjects research.

Bias and data diversity: Recognizing the potential biases in AI models trained on biomedical data, we have prioritized diversity in curating MicroVQA. The dataset includes microscopy images spanning multiple organisms, imaging modalities, and experimental conditions to provide a diverse benchmark for performance evaluation. However, we acknowledge that certain imaging types, such as brightfield and fluorescence microscopy, may be overrepresented due to their prevalence in the field. Future iterations of MicroVQA will seek to address such imbalances to improve dataset representativeness.

Potential societal impacts: Biomedical AI models, including those trained on MicroVQA, have the potential to influence research and healthcare significantly, both positively and negatively. While these models can accelerate scientific discovery, they may also perpetuate biases in training data or result in uneven performance across demographic or biological groups. To mitigate these risks, MicroVQA includes metadata annotations to support the analysis of potential biases and emphasizes diverse data sourcing. We will engage with the research community to address emerging ethical concerns.

Data licensing and usage: MicroVQA is distributed under the Creative Commons Attribution-ShareAlike 4.0 International (<https://creativecommons.org/licenses/by-sa/4.0/>) license to promote transparency and collaboration in the research community. This license reflects our efforts to balance open access with the ethical use of data, enabling both academic and commercial applications of the dataset while adhering to original license requirements. Full licensing details are available at <https://creativecommons.org/licenses/by-sa/4.0/>.

C. Limitations and future work

While MicroVQA advances research-level reasoning benchmark significantly, we can identify a number of limitations and avenues for future work.

Open vs closed evaluation Like most VQA benchmarks, we used multiple-choice (MCQ) evaluation. However downstream applications will obviously operate in an open setting, without options and requiring more detail. MLLM research does have some open benchmarks [70], though impartial evaluation is a significant challenge. Our early experiments with open prompting suggest that models like GPT-4o tend to give very vague answers, and tend to depend strongly on the text part of the input prompt.

Dataset scale Our final dataset had 1061 samples, which is smaller than some other VQA benchmarks like MicroBench [41] and OmniMedVQA [26]. MicroVQA’s scale sufficient to compare overall scores between models, however statistical power starts to become limiting when trying to compare different models on subsets of the dataset. How does this compare to other benchmarks? As we can see from Tab. 2 and Fig. 3, the larger datasets tend to be those with lower-level reasoning according to Bloom’s taxonomy. Their samples tend to be derived automatically from metadata tags or classification labels. Benchmarks with comparably high-level reasoning, like MMMU-Pro, have a similar scale to ours (1,700). In the language-only domain, the research-level reasoning benchmark GPQA has 448 samples. This trend reflects a clear, practical limitation: it is very time-consuming for humans to create difficult reasoning questions, and the pool of capable experts is small.

Breadth of dataset attributes in microscopy We aimed for MicroVQA to cover a broad range of the microscopy discipline; for example, we have samples from all four relevant scales – tissues, cell, subcell, atomic. However each human annotator is an expert in a subset of microscopy and biology, so there are practical limits on dataset coverage. We, therefore, chose to focus on the most common imaging modalities – brightfield, fluorescence, electron – while skipping less common modalities like Raman spectroscopy. In terms of sample types, we emphasized those relevant to humans – human and mouse – while having only a few samples from rare organisms like bottlenose dolphin.

Coverage of scientific reasoning Looking even more broadly, the motivation of this work is to advance reasoning in scientific research generally. However, we use microscopy in particular. This is because of a clear trade-off in breadth vs depth. Our dataset strikes a good balance, given that microscopy represents a large portion of visual data in biomedicine and biology. Future work can reuse our framework to generate reasoning VQA benchmarks in other fields of visual biomedicine and biology, such as in medicine [26] or ecology [77].

Resolving MCQ shortcuts A key contribution in this paper is the RefineBot for generating multiple choice questions (MCQs) that are too challenging to solve using text-only shortcuts. Our approach used the outputs of existing MLLMs as signal to make distractors harder. However, this cannot guarantee the removal of shortcuts, and more generally, harder distractors may exist. Therefore, the next generation of LLMs may be able to identify and exploit new shortcuts. If that occurs, we can simply run RefineBot again using the new LLM, creating MicroVQA-2. This strategy could be used for other VQA evaluations.

D. Benchmark details

D.1. Accessing MicroVQA Benchmark

MicroVQA is an expert-curated benchmark for multimodal scientific reasoning, specifically in microscopy-based biological research. The dataset comprises 1,061 multiple-choice questions (MCQs) created by biological researchers to represent authentic scientific practice. Each MCQ is paired with an associated microscopy image and metadata describing the experimental context. A demonstration subset of 28 instances is available for initial exploration.

The MicroVQA dataset is publicly accessible on HuggingFace at <https://huggingface.co/datasets/microvqa/microvqa>. It is distributed under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license, allowing reuse and share-alike redistribution with attribution. The dataset is available in Apache Arrow and Parquet formats to ensure efficient data handling. These formats support fast serialization, memory mapping, and streaming from the web, facilitating seamless integration into machine learning pipelines.

D.2. Dataset Schema and Structure

The dataset schema defines the fields included for each data instance. Figure 6 provides a schema summary, while a sample data instance is detailed below 7.

```
{
  "image_id": datasets.Value("string"),
  "image": datasets.Image(decode=True, id=None),
  "label": datasets.ClassLabel(),
  "label_name": datasets.Value("string"),
  "domain": datasets.Value("string"),
  "subdomain": datasets.Value("string"),
  "modality": datasets.Value("string"),
  "submodality": datasets.Value("string"),
  "stain": datasets.Value("string"),
  "microns_per_pixel": datasets.Value("string"),
  "questions": datasets.Value("string"),
}
```

Figure 6. Example data schema.

D.3. Training contamination mitigation

To mitigate contamination, all data in MicroVQA were sourced from original, unpublished microscopy images or open-access articles (e.g., CC-BY-licensed) published after January 2024. Additionally, unique canary strings were embedded in the dataset to detect potential contamination during fine-tuning or model evaluation.

D.4. Benchmark attribute details

The MicroVQA benchmark is designed to evaluate multimodal reasoning capabilities in biomedical research contexts. The dataset incorporates diverse attributes such as organism, research subject, imaging modalities, and biological length scales (image scales) to ensure a comprehensive assessment¹. These metadata not only enhance the scientific relevance but also enable analysis of reasoning across different biological and microscopy contexts. This section describes the methodology used to define, compute, and annotate these key attributes.

The benchmark comprises 1,061 total questions, including 423 questions containing multiple images combined into panels. A collage of sample images is shown in 24. The average MCQ token lengths were computed using the `o200k_base` tokenizer, with separate calculations for the raw input and formatted multiple-choice questions. Imaging modalities are

```

{
  "image_id": "d54bb7ec-284f-4218-a47d-af87bb371de5",
  "image": datasets.Image(decode=True, id=None),
  "label": datasets.ClassLabel(),
  "label_name": "pathology",
  "domain": "pathology",
  "subdomain": "gastrointestinal pathology",
  "modality": "light microscopy",
  "submodality": "brightfield microscopy",
  "stain": "H&E",
  "microns_per_pixel": "Not provided",
  "questions": "Question:
  ``A hematoxylin and eosin (H&E)-stained micrograph shows cellular
  formations within a gastrointestinal tissue sample with unique
  organization and cellular patterns. What pattern and morphology do these
  cells most closely align with?

  A) Bundles of elongated cells with pale pink cytoplasm and uniform oval
  nuclei in a crisscross pattern
  B) Circular groups of cells with mucin in central spaces and small,
  dark nuclei
  C) Dense arrays of polygonal cells with small nucleoli and vacuolated
  cytoplasm
  D) Rows of cells with conspicuous cytoplasm and rounded nuclei
  E) Diffused arrangement of cells with transparent cytoplasm
  and irregularly clustered chromatin

  Correct Answer: A) Bundles of elongated cells with pale pink cytoplasm
  and uniform oval nuclei in a crisscross pattern``"
}

```

Figure 7. Example data instance.

broadly categorized into light, fluorescence, and electron microscopy. Biological length scales span four levels: tissue (millimeter scale), cellular (micrometer scale), subcellular (nanometer scale), and atomic (angstrom scale), enabling reasoning across diverse structural hierarchies.

To annotate attributes such as organism, specimen, and research subject, we manually reviewed random samples from the dataset and created a taxonomy to cover the most likely categories. A biomedical domain expert manually labeled these categories for 50 raw input question-answer pairs and performed initial prompt engineering. We used DSPy to optimize the prompt and few-shot examples on the labeled dataset. The input context included the taxonomy of organisms⁸ and research subjects⁹.

These annotations were the gold standard to guide DSPy prompt engineering for automated attribute labeling on the full dataset. Prompt instructions included fallback rules (e.g., tagging "None" for indeterminate attributes) to ensure accuracy and minimize hallucination. The optimized pipeline annotated all 1,061 questions while ensuring the validity of the dataset attributes.

D.5. Related benchmarks

In section 3.3 and Tab. 2 we compare MicroVQA to related scientific multimodal benchmarks, focusing on its intersection of high-difficulty and advanced reasoning questions. Beyond these key attributes, we examine the sources of the questions. Most benchmarks rely on existing question repositories, such as exams, web-based QA datasets, textbooks, or other MCQ datasets. This approach bypasses the need for MCQ generation methods, as the questions and options are pre-tested and curated, making these benchmarks more representative of established tasks. Another common strategy involves reformatting classification datasets into MCQs using question templates, which increases task diversity but often limits scenario uniqueness and language variability. In contrast, MicroVQA is built by directly querying experts for challenging, domain-specific research questions. By starting with long form questions and answers, our approach requires an innovative MCQ generation method, ensuring uniquely tailored questions that push the boundaries of domain-specific reasoning and knowledge assessment.

D.6. Bloom’s taxonomy

Bloom’s Taxonomy is a model for evaluating cognitive skill levels in educational assessments. It has six hierarchical levels of learning: recall, understand/comprehend, apply, analyze, evaluate, and create. Multiple-choice questions can test five of


```

Taxonomy of organisms in biology | # Taxonomy of Organisms in Biology
and Biomedical Research:
Apis mellifera (Honeybee)
  Social insect with complex behavior.
  Keywords: honeybee, social behavior, ecology, pollination.

Arabidopsis thaliana (Thale Cress)
  Model plant organism.
  Keywords: Arabidopsis, plant genetics, photosynthesis.

Caenorhabditis elegans (C. elegans)
  Transparent nematode worm.
  Keywords: worm, development, aging, programmed cell death.

Danio rerio (Zebrafish)
  Fish with transparent embryos.
  Keywords: zebrafish, vertebrate development, embryology.
...

Drosophila melanogaster (Fruit Fly)
  Insect with rapid life cycle.
  Keywords: fruit fly, genetics, development.

Escherichia coli (E. coli)
  Bacterium used in genetics and molecular biology.
  Keywords: bacteria, gene expression, plasmids.
...

Homo sapiens (Human Cells)
  Cultured human cells (e.g., HeLa, HEK293, A431, U2OS).
  Keywords: human cell lines, immortalized cell lines

In vitro (None)
  Purified molecules and reactions in controlled environments.
  No living cells or organisms.
  Keywords: purified proteins, purified molecules, cell-free systems
...

```

Figure 8. Examples of taxonomy classes used as context to LLM to assign an organism to a question. A YAML file with the full taxonomy will be released with the code.

Bloom’s levels but not the sixth level, “create.” These levels provide a structured framework for assessing lower-order and higher-order thinking skills, which are essential for designing robust benchmarks like MicroVQA.

Initial attempts to use LLMs to classify Bloom’s levels for biomedical visual question answering (VQA) benchmarks revealed a systematic overestimation of cognitive skill levels. This discrepancy limits the naive use LLMs for automated Bloom’s classification, perhaps due to the nuances of applying Bloom’s taxonomy in domain-specific contexts. To ensure accurate classification of Bloom’s level, a physician-scientist with NBME training in item writing and familiarity with Bloom’s classification manually labeled Bloom’s level for 200 examples from public MCQ questions, MicroVQA, MicroBench, and ScienceQA. These examples were used to fine-tune GPT-4o-mini (gpt-4o-mini-2024-07-18) to classify Bloom’s level and align with expert judgments.

The fine-tuned Bloom’s classifier was applied to MicroVQA and numerous public science/biomedical benchmarks to assess the cognitive skills evaluated by each question. Each question was mapped to one of Bloom’s levels, providing insights into the distribution of cognitive tasks across the benchmark. This analysis revealed that while existing benchmarks predominantly tested lower-order skills (e.g., recall and comprehension), MicroVQA has a greater emphasis on questions that require higher-order reasoning (e.g., analyzing, evaluating), filling a critical gap in multimodal scientific reasoning assessments.

D.7. Methodology for proposing tasks

Here, we expand on the task proposal methodology introduced in Sec. 3.2 To define specific reasoning tasks aimed at advancing biomedical research, we conducted a comprehensive expert consultation process. Through structured interviews with nine Principal Investigators and postdocs specializing in microscopy research, including participants from both academia and industry, we developed a framework for tasks that would meaningfully integrate multimodal large language models (MLLMs) into microscopy research workflows.

```

Taxonomy of research subjects| #Taxonomy of Research Subjects in Biology:
Anatomy
  Study of the structure of organisms and their parts.
  Keywords: organs, tissues, morphology, physical organization.

Biochemistry
  Chemical processes within living organisms.
  Keywords: proteins, enzymes, nucleic acids, metabolic pathways.
  ...

Cell and Molecular Biology
  Structure and function of cells and their molecules.
  Keywords: cellular processes, gene expression, molecular pathways.
  ...

Developmental Biology
  Process by which organisms grow and develop.
  Keywords: embryology, morphogenesis, genetic regulation.
  ...

Genetics
  Study of heredity and gene function.
  Keywords: DNA, inheritance, genetic variation, gene expression.

Immunology
  Immune system in health and disease.
  Keywords: antibodies, immune response, pathogens, vaccines.

Microbiology
  Study of microorganisms.
  Keywords: bacteria, viruses, fungi, parasites, microbial ecology.

Neurobiology
  Structure and function of the nervous system.
  Keywords: neurons, brain, neural circuits, cognition.
  ...

Structural Biology
  Molecular structure of biological macromolecules.
  Keywords: protein folding, nucleic acid structures, crystallography.
  ...

```

Figure 9. Examples of taxonomy classes used as context to LLM to assign a research subject to a question. A YAML file with the full taxonomy will be released with the code.

The task selection process followed specific criteria: tasks must utilize image-text inputs and text outputs suitable for MLLMs; they should require higher-order reasoning beyond simple image processing; and they should emphasize core experimental activities (experimentation, analysis, and action) rather than auxiliary tasks like literature review or writing [25].

Our methodology involved a two-phase interview process. The first phase consisted of initial one-hour individual sessions with eight research groups featuring open-ended, unstructured discussions about key challenges in microscopy experiments, potential applications of MLLMs, current workflows and unmet needs, and possible AI system integration points. We then conducted follow-up interviews one week later to consolidate findings and build consensus.

To avoid anchoring biases, we intentionally withheld predefined categories during initial interviews. This approach enabled unbiased task identification and allowed us to capture a broad spectrum of potential use cases. Through this iterative process, we identified three essential capabilities that align with the scientific life cycle for microscopy experiments: advanced image interpretation, hypothesis generation, and experimental design/hypothesis testing.

E. MCQ generation details

E.1. Question collection process details

Data collection used a structured submission form shared with collaborators. The form required users to provide the following:

- Image or Image Set: Users uploaded microscopy images, optionally including multiple images for comparative questions.
- Experiment Details: A brief description of the experimental setup, imaging modalities, and any relevant biological context.
- Question Text: Each question was aligned with one of three predefined use cases: identifying interesting features, explaining mechanisms, or proposing next steps.

- Answer and Distractors: Contributors provided a correct answer and biologically plausible incorrect answers.

Quality control was performed in multiple rounds. To familiarize users with the task and data requirements, each user submitted an initial form with 4-6 questions for manual quality control and feedback. Feedback was provided to users on improving question specificity, ensuring the questions required image interpretation rather than relying solely on general knowledge. Contributors were encouraged to refine their submissions based on the feedback to better align with the tasks and goals of the MicroVQA benchmark.

Users were encouraged to submit data from multiple sources, imaging modalities (e.g., fluorescence, electron microscopy), and biological scales (e.g., tissue, cellular, subcellular) to promote diversity. Contributors could upload microscopy images from their experiments or use CC-BY publicly available images, such as those from preprints or publications released after January 2024. This ensured that all images were either original or sourced from open-license publications, enabling their use in an open-source benchmark.

E.2. Experiment on naive MCQ generation

In Sec. 4.2, we discuss how the naive approach to MCQ generation – simply zero-shot prompting GPT – leads to MCQs that are very easy for frontier models to solve. Which ultimately motivated our two-stage MCQ generation approach. Here, we discuss those experiments in more detail. We create a naive version of MicroVQA using a standard prompt [26, 68]:

```
You are an expert in molecular and cell biology, and in microscopy.

I will give you an original biology-related question and its answer, your task is to rephrase an equivalent question with identical answer. The question related to an image, and we don't show the image. Meanwhile, I want to transfer this QA-pair into a multi-choice question. Please generate 5 incorrect options to construct the candidate options.

{{QUESTION}}

{{ANSWER}}
```

Then, we evaluated this naive MicroVQA on the top closed source models with the image, and to study the over-reliance on language shortcuts, we evaluated a text-only version. In the text-only setting, we add the following to the prompt: “If an image is mentioned, ignore this information and try your best to answer the question.”. Tab. 5 shows the results of this experiment.

Table 5. Performance on MicroVQA MCQs with naive MCQ generation. We report overall VQA and VQA without the image.

	VQA	VQA-no-image
GPT-4o	85.1	82.7
Claude-3.5-Sonnet	91.4	88.4
Gemini-1.5-Pro	88.5	82.4

All models score over 80% on questions that are from a specialized domain. We hypothesize that this reflects a problem with the non-trivial task of MCQ generation from raw VQAs for several reasons. First, the high ‘VQA-no-image’ scores across all models suggest that selecting the correct option does not require image understanding, despite the raw VQA samples being designed to incorporate image-based reasoning. An alternative explanation is that the questions themselves may be too easy for advanced models. Second, we test this alternate explanation by performing open evaluation on a subset of the raw VQA queries— that is, we queried GPT-4o with the questions but without the options. Here, the models rarely gave good responses, even after careful prompting. This suggests that models are leveraging test-taking strategies to eliminate distractors [22]. Third, we examined the chain-of-thought responses in the VQA-no-image case, which suggested the models often eliminate distractors based on information from the text question. Therefore, we conclude that there are issues with the improper generation of the MCQs using the naive strategy.

What exactly are the problems with MCQ generation? Qualitatively, the types of errors are the same as those we discuss in Appendix E.7, so we refer readers to that section to get a deeper understanding.

E.3. Stage 1 MCQ generation (exam alignment)

The first stage transformed raw, open-ended questions and long-form answers into well-structured MCQs aligned with established principles of biomedical exam design. A physician-scientist trained in medical board exam item writing initiated

the process by reviewing biomedical educational literature, NBME guidelines [53], and Bloom’s taxonomy for assessing cognitive skills [5, 14, 73]. The expert manually converted 50 raw question-answer pairs into one-best-answer MCQs with biologically relevant distractors using the literature review and their prior training. This manual effort ensured that the MCQs preserved the original question’s intent and scientific validity while adhering to educational standards. Distractors were crafted to require reasoning and prevent reliance on superficial cues like word association or general knowledge, thus maintaining the depth and complexity of the original content.

To scale this process, the expert designed an initial LLM prompt for automated Stage 1 MCQ generation. This prompt and the manually curated training data were input for DSPy-based prompt optimization. Using the MIPROV2 optimizer, the prompt was refined within a Chain of Thought (CoT) framework combined with retrieval-augmented generation (RAG) pipelines powered by o1-mini. Optimization was guided by a custom metric evaluated by a separate LLM judge (o1-mini). The judge assessed each revised MCQ across three criteria: (1) semantic similarity to the original question and answer, (2) adherence to NBME formatting standards, and (3) minimization of extraneous information that could provide unintended clues. The optimization metric was weighted to prioritize preserving the original meaning and scientific validity, with lesser emphasis on formatting and extraneous text. Additional implementation details are provided in the source code.

E.4. Stage 2 MCQ generation (RefineBot) details

The goal of RefineBot is to take MCQs that are too easy for MLLMs – especially where they are leveraging language-only shortcuts – and to rewrite them to make them more challenging. It is the third panel in Fig. 4. For more details on what types of improvements are achieved, see Appendix E.7. The complete code for RefineBot is released.

A single RefineBot iteration Below, we show the pseudo-code for a single-question revision. The input is a question string, a list of options, and the index of the correct answer in the options. It returns a tuple (code, MCQ) where code is a string on whether the run was successful, and MCQ is either the revised multiple choice question if the rewrite was successful or otherwise None.

```

1 def revise_mcq(question: str, choices: list, correct_idx: int, max_iters: int = 5):
2     # Store original version
3     original = {'question': question, 'choices': choices, 'correct_idx': correct_idx}
4     reflections = []
5
6     for iter in range(max_iters):
7         # Try answering without image
8         result = evaluate_without_image(question, choices, correct_idx)
9
10        if not result.can_answer:
11            code = 'SUCCESS_NO_CHANGE' if iter == 0 else 'SUCCESS_REWRITE'
12            return code, (question, choices, correct_idx)
13
14        if iter == max_iters - 1:
15            return 'FAIL_ITERATIONS', None
16
17        # Reflect on strategies used to solve
18        reflection = analyze_why_answerable(result)
19        reflections += reflection
20
21        # Rewrite based on all reflections from all iterations
22        new_version = rewrite_question(reflections, original)
23
24        # Verify the rewrite didn't change the meaning meaning preserved
25        if not check_meaning_preserved(original, new_version):
26            return 'FAIL_REWRITE', None
27
28        # Update current version
29        question = new_version['question']
30        choices = new_version['choices']
31        correct_idx = new_version['correct_idx']
32
33    return 'FAIL_UNKNOWN', None

```

This process is also described in the main text. One detail to note: the `rewrite_question` function receives the reflections from all prior iterations, not just the current one. In this way, the rewriting function can consider all possible strategies that have been identified. All LLM calls are to `gpt-4o-2024-08-06`.

To better illustrate the method, we step through a single example. This way, we can show the prompts and understand responses.

1. The evaluation of a question with the `evaluate_without_image` prompt and response:

```
Prompt
-----
The following question is supposed to be paired with an image. We will not provide the image, so answer to the best of your ability.
In an electron microscopy image of a mouse primary embryonic cortical neuron expressing full-length human mHTT (BACHD), "eye-ball" shaped features are observed in the top left corner that appear anomalous. What is the most likely cause of these features?
Think step by step and then output the answer in the format of "The answer is (X)" at the end.

(a) Aggregation of human mHTT protein within the neuron
(b) Ice contamination from liquid nitrogen during sample preparation
(c) Formation of intracellular vesicles during neuron development
(d) Artifacts arising from excessive heavy metal staining
-----
Response (target answer is b)
-----
To determine the cause of the "eye-ball" shaped features observed in the electron microscopy image, we can analyze each option step by step:

(a) Aggregation of human mHTT protein within the neuron: Aggregations of proteins can indeed appear as anomalous structures under electron microscopy. However, these typically have irregular shapes and may not consistently resemble "eye-ball" shapes.

(b) Ice contamination from liquid nitrogen during sample preparation: Ice contamination can occur during sample preparation, particularly if samples are not handled properly when flash-frozen. This can lead to anomalous features appearing in microscopy images. These features might appear as circular or irregular shapes, potentially looking like "eye-balls."

(c) Formation of intracellular vesicles during neuron development: Intracellular vesicles are normal structures within neurons and other cells. Although they can appear during development, they are typically well characterized and documented, and are less likely to be described as anomalous in the context given.

(d) Artifacts arising from excessive heavy metal staining: Over staining can indeed produce artifacts in electron microscopy images. While staining artifacts are plausible, they might not specifically result in "eye-ball" shapes unless the staining process consistently targets specific round structures.

Considering the nature of the descriptors ("eye-ball" shaped, anomalous), option (b) ice contamination is often a source of unexpected features in microscopy images that could fit the appearance described, as ice can form circular shapes.

Therefore, the most likely cause of these features is (b) Ice contamination from liquid nitrogen during sample preparation.

The answer is (b).
```

2. If the answer is incorrect, then the question is already hard enough, and we can exit. Otherwise, we call the `analyze_why_answerable` function, which reflects on how this answer was achieved. It continues in the same conversation:

```
-----
User
-----
That is correct.

Explain how you were able to answer the question without access to the image - only the question_stem and choices.
What strategies did you use?

Then, briefly summarize these "language-only strategies" into a short list.

Finally, say whether your reasoning was closer to:
- 'deduction': you answered with high confidence because you had enough information.
- 'abduction': you answered with lower confidence you made an informed guess.

Be concise in your final response.
-----
Assistant
```

To answer the question without the image, I relied entirely on analyzing the given information and the plausible characteristics of each option. Here's a breakdown of my strategy:

1. **Keyword and Context Analysis**: I focused on key descriptors such as "eye-ball" shaped features and their being "anomalous" in an electron microscopy image of neurons expressing human mHTT. This guided me to consider characteristics commonly seen under electron microscopy.
2. **Knowledge of the Subject Matter**: I applied general knowledge about electron microscopy artifacts and protein behavior. This helped me weigh the likelihood of each option based on what typically causes specific image features.
3. **Process of Elimination**: By understanding the nature of each option including potential artifacts or structures (such as vesicles or ice) in microscopy-I systematically assessed the validity of each scenario in relation to the described features.
4. **Logical Inference**: I inferred which scenario most commonly results in anomalies resembling "eye-ball" features, favoring the simplest explanation.

Summary of "language only strategies":

- Analyze keywords and context
- Apply relevant subject matter knowledge
- Use process of elimination
- Make logical inferences

The reasoning was closer to 'abduction' as I made an informed guess based on patterns and typical causes, given that I lacked direct visual confirmation.

3. Next, we call `rewrite_question`. Here, the templated variable `{{conversations}}` has the full conversations produced by the function `analyze_why_answerable` in the previous call and from all other iterations. The `{{n_chat}}` just indicates how many conversations there are included:

Below, I will display `{{n_chat}}` chat conversations between a 'user' and an LLM 'assistant'.

In each conversation

- a user asks the assistant to answer a multichoice VQA question, however they do not provide the image. They only get the `question_stem` and choices.
- the `question_stem` and correct answer are similar to other conversations, however the distractors are different
- the assistant then answers correctly.
- the user then asks the assistant to explain how it answered the question with only the text.
- the assistant then summarizes what strategy they used to answer the question.

Altogether, these conversations give examples of language shortcuts that need to be avoided when constructing the `question_stem` and choices.

Here are the conversations:

`{{conversations}}`

Your task is to rewrite the `question_stem` and choices so that a different LLM 'assistant' cannot use the language-only strategies that were identified in these past conversations.

Include an 'explanation' about why your new set of distractors are better.

Your revised choices should include the correct answer at the 'correct_index'.

Your revised `question_stem` and choices should not significantly change the meaning of the question and correct answer.

You are free to change the distractors a lot to achieve this task.

Include `{{n_choices}}` choices.

The choices should not be made too different in length and detail to the original question stem and choices.

Note that the question stem may contain important cues that cannot be removed. E.g. if a question asks about the "green stain" in an image, you cannot change it to "stain" because it introduces ambiguity.

The choices should not include the letters in front like (a), (b), ...

Return a json

The output of this function is a JSON and is:

```
{
  "question_stem": "In an electron microscopy image of a neuron expressing human mHTT (BACHD) protein, distinct, spherical features are prominent and seem to be atypical. What is the most probable explanation for these features?",
  "choices": [
    "Concentration of cellular waste products leading to distinct morphological anomalies",
    "Formation of air bubbles introduced during sample immobilization processes",
    "Ice contamination from liquid nitrogen during sample preparation",
    "Inclusion bodies formed by aggregated proteins expressing during stress response",
    "Cellular morphological changes during mitotic division"
  ],
}
```

```

"correct_index": 2,
"explanation": "The revised options introduce plausible scientific phenomena that could also result in spherical or distinct anomalies. This set of distractors is more varied and grounded in legitimate alternative explanations without leading the responder towards the correct answer through language patterns. For example, air bubbles and protein inclusion bodies might mimic sphere-like structures. Additionally, by adding an unrelated phenomenon like cellular changes during division, the set confuses the language pattern strategy previously used. Thus, the correct choice of \"Ice contamination from liquid nitrogen\" stands out less and necessitates either visual confirmation or real understanding of sample preparation outcomes to identify correctly."
}

```

Notice that the response includes an explanation of why it made the changes, and the changes relate to what was in the reflection, adding interpretability.

4. Finally, we check that the rewrite has not changed the semantic meaning of the question and answer.

```

Prompt
*****
Below are two question-answer pairs.
The question-answer pairs are part of VQA triplets, and both pairs use the same image.

Are these question-answer pairs semantically equivalent?
Or are they significantly different?
Give a true/false and also an explanation.

QUESTION 1:
In an electron microscopy image of a mouse primary embryonic cortical neuron expressing full-length human mHTT (BACHD), "eye-ball" shaped features are observed in the top left corner that appear anomalous. What is the most likely cause of these features?

ANSWER 1:
Ice contamination from liquid nitrogen during sample preparation

QUESTION 2:
In an electron microscopy image of a neuron expressing human mHTT (BACHD) protein, distinct, spherical features are prominent and seem to be atypical. What is the most probable explanation for these features?

ANSWER 2:
Ice contamination from liquid nitrogen during sample preparation

*****
Response, is_equivalent: True
*****
is_equivalent: Both question-answer pairs revolve around electron microscopy images depicting neurons expressing the human mHTT (BACHD) protein. The main focus in each question is on the presence of unusual spherical (or 'eye-ball' shaped) features. In both cases, the identical explanation is provided: ice contamination from liquid nitrogen during sample preparation.

While the wording differs slightly, the core components, contexts, and the ultimate explanation remain consistent between the two pairs. Thus, they can be considered semantically equivalent as they describe the same phenomenon with the same reasoning for the observed anomalies.

```

5. If the meaning check step fails, then we exit with a fail as we notice that once the meaning deviates it will become more distorted with further iterations. Otherwise, we update the MCQ, and start the loop again. Where the first function call in the new loop will check whether this revised MCQ is now too hard or needs to be rewritten.

Calibrating the preservation of meaning Since there is a risk that the rewriter will change the semantic meaning of the question and answer, it is crucial that the `check_meaning_preserved` is effective. For topics requiring domain expertise, this requires prompt engineering with a domain expert providing feedback. This is one of the key human bottlenecks in tuning the prompts in this system.

Other sources of MCQ invalidation While the `check_meaning_preserved` function checks that the question and answer is correct, an additional risk is that the rewrite introduces distractors that are more correct than the target answer. In our manual review evaluations (we reviewed all 1,061 MCQs), this was very rare. However, if this were a significant problem in future systems, it may be addressable by tuning the rewriter bot, with an expert human evaluator in the prompt engineering loop.

Extending RefineBot beyond language shortcuts We constructed RefineBot specifically to address the issue of language shortcuts – where the question is answerable without using the image. This is appropriate because we work with vision-centric VQA, meaning the questions should require image interpretation. However, the framework should work in other

settings – namely language-only QA. This would require adjusting the prompts slightly. For example, the current reflection prompt asks how the question was successfully evaluated without using language; also the current rewriter prompt directs the LLM to address language shortcuts in its rewrite. Therefore, adapting RefineBot to language-only MCQ should make these prompts more general.

On the other hand, the method could also be updated for the more general VQA setting – to improve questions that do not have language shortcuts, but are easy for some other reason. Here, the evaluator and reflector agents, which are currently LLMs, could be switched for an MLLM, which also ingests the image (though this would be expensive for current state-of-the-art models). We hypothesize that current MLLMs are so weak at image perception, that this is not necessary for MicroVQA – that simply removing language shortcuts will suffice for creating a challenging benchmark. As MLLMs improve at perceiving microscopy images, this general VQA RefineBot could become more useful for ensuring that the benchmark remains challenging.

Repeated RefineBot iterations A single run of RefineBot may exit as a FAIL, either because the rewrite changed the meaning of the question, or because the maximum number of iterations was reached. However, RefineBot can give very different outputs when changing the random seed, and we find that simply running again with a different random seed can lead to a successful run. Therefore, if the bot fails, we rerun it.

If an MCQ exited RefineBot successfully, then the evaluator LLM was fooled by the question (got the answer incorrect). Yet, the question could still be suboptimal in detectable ways: the LLM could be capable of narrowing the MCQ down to two options, but have guessed the wrong option; moreover, this question could be easier for different models. To ensure it is difficult enough overall, we require that the MCQ is answered incorrectly for two random seeds from two different models: `gpt-4o-2024-08-06` and `claude-3-5-sonnet-20241022`. If not, then it fails, and we run the RefineBot again with a different random seed.

We discuss in the experiments Sec. 5.1 how this process leads to a small evaluation bias against the models used as evaluators. That section shows that despite the bias, this process makes the MCQs generally harder for all models. Future versions of RefineBot could mitigate this by including a larger family of evaluator models.

Finally, some MCQs fail to pass the RefineBot check. In these cases, we simply keep the stage 1 MCQ.

E.5. Final MCQ quality check

Since the MCQ generation process involves LLM generation, we need to verify the correctness of the final answer. Each MCQ is manually reviewed by the same expert who created that particular question (between 80 and 120 per person). For any question with issues, that expert makes minimal changes to the question to make it correct.

E.6. MCQ generation results: stage 1 exam alignment

The Stage 1 pipeline produced one-best-answer MCQs that preserved the content, testing focus, and validity of the expert-generated question-answer pairs. It generated an initial set of distractors following best practices outlined in educational and NBME guidelines. A key design goal of Stage 1 was to preserve the original question content without altering or introducing new information. Secondary objectives included ensuring proper formatting according to NBME guidelines and minimizing information that might provide clues to the correct answer without reference to the image.

While the pipeline ensured alignment with these goals, preserving the original content occasionally resulted in MCQs containing extra information, such as image descriptions or specific biological terms (i.e., gene or protein name), that could inadvertently help narrow the options. For example, stage 1 question stems sometimes included protein names, disease descriptions, or descriptions of image features, which could reduce the challenge. Figure 10 illustrates Stage 1 outputs that are appropriately formatted but may preserve details from the original input, making questions less challenging. While the distractors were biologically plausible, some options were overly easy to eliminate, reducing the challenge. These limitations are addressed and refined in Stage 2.

The `o1-mini-2024-09-12` LLM judge, employed during DSPy optimization, compared the Stage 1 MCQs to the original inputs to evaluate (1) semantic similarity and (2) adherence to MCQ formatting. The results showed that 96% of questions preserved the original meaning, and 100% adhered to NBME formatting guidelines. Manual analysis of the 4% labeled as at least partially dissimilar revealed that these cases typically involved open-ended original questions being reformulated into closed-ended MCQs. This transformation, a necessary step for creating one-best-answer MCQs, was not a failure but rather an expected outcome of Stage 1. In rare instances where the original submission included multiple correct answers, the output of Stage 1 retained one correct answer and excluded others to comply with the one-best-answer format. The LLM judge flagged these cases as partially dissimilar but these changes were essential for adhering to the MCQ structure.

A common issue in MCQ design is that correct answers tend to be longer and more detailed than distractors. To assess this, we computed the ratio of tokens in incorrect options to those in the correct answer, where values ≤ 1 indicate longer correct answers, and values ≥ 1 suggest shorter correct answers. Stage 1 MCQs showed a ratio of 0.87, indicating that correct answers were only slightly longer than distractors. By comparison, naive MCQ generation yielded a ratio of 0.46, meaning correct answers were nearly twice as long as distractors, making them easier to identify. Additionally, naive MCQ generation preserved the original meaning in only 89% of cases and met NBME formatting guidelines in just 70%, highlighting the efficacy of the Stage 1 process at preserving information and ensuring proper MCQ formatting.

E.7. MCQ generation results: stage 2 RefineBot

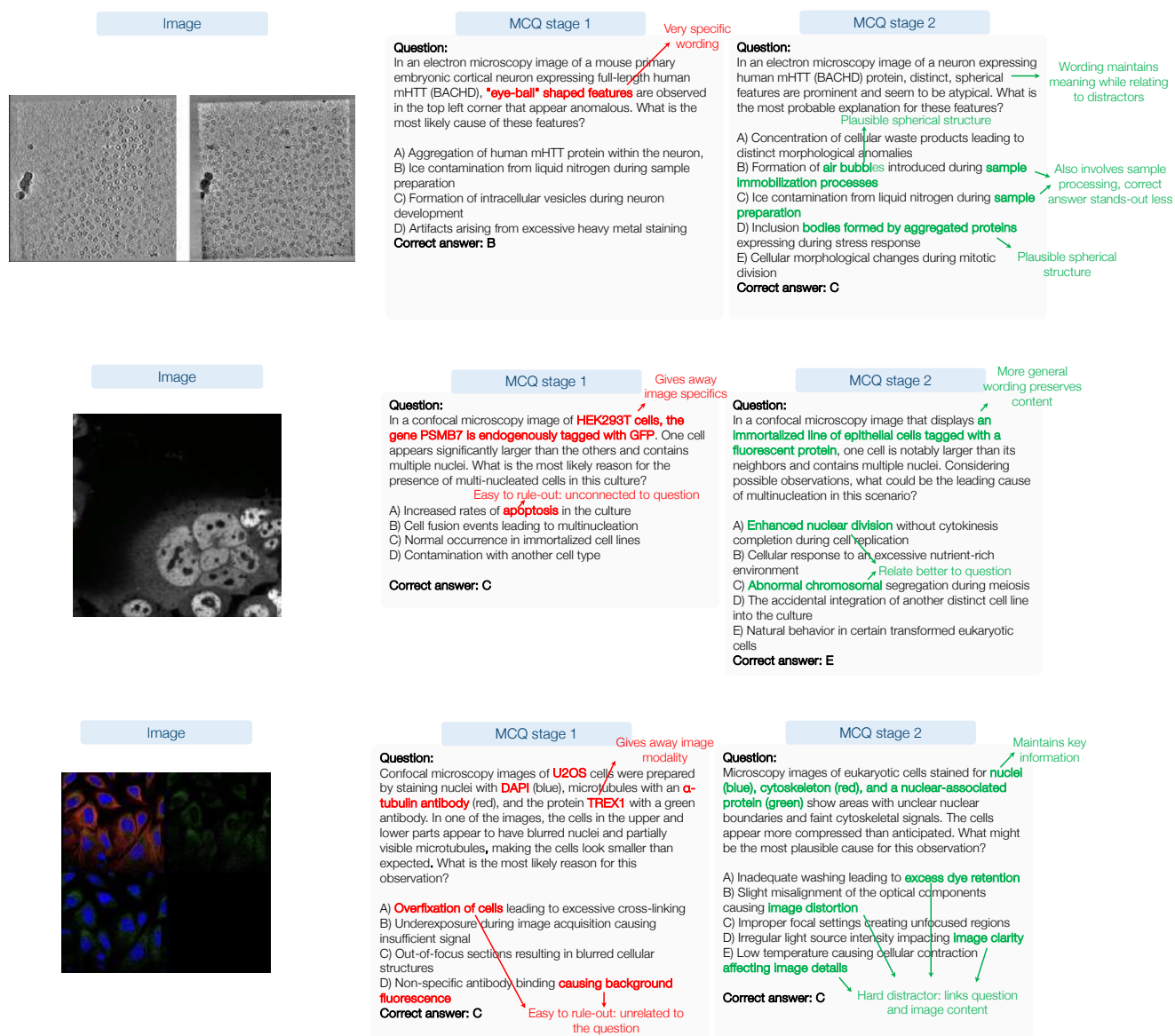


Figure 10. Examples of changes to questions and options between stage 1 and stage 2 (RefineBot) of our MCQ generation method. In red elements that need to be improved and in green improvements.

The second stage of our MCQ generation method focuses on increasing the difficulty of the questions and options. In Fig. 4, we illustrate the transformation of a question at each stage, highlighting issues and corresponding improvements. While stage 1 produces exam-style questions that preserve the QA objective, the wording of the questions and distractors

often enable MLLMs to exploit shortcuts rather than demonstrate knowledge. To address this, stage 2 (RefineBot) employs a chain-of-thought process. We analyze additional examples of improvements achieved in stage 2, showing how RefineBot improves both the style and content. Key strategies include:

1. Overly precise wording in the question stem: Precise phrasing can make it easy to identify the correct answer through pattern matching. For example, Fig. 10 (top row), changing ‘eye-ball shaped’ to ‘spherical’ retains the question’s intent while creating more challenging distractors referencing plausible alternative processes.
2. Over-description of the image eliminates the need for image interpretation. For example, in Fig. 10 (middle row), stage 1 includes details that reveal the specifics of the image, allowing the model to propose common issues without analyzing the image. Stage 2 replaces these specifics with general terms in the question stem.
3. Distractors unrelated to the specifics of the scenario described by the question. In Fig. 10 (bottom row), stage 2 increases distractor complexity by incorporating elements tied to the question’s context, making them less predictable.

F. Experiments details

F.1. Evaluation prompts

All prompts except for LLaVA and LLaVA-Med used the chain-of-thought prompt template from the MMMU-Pro code [72]:

```
The following is a multiple choice question (with answers).
Think step by step and then output the answer in the format of \"The answer is (X)\" at the end.

{{QUESTION}}

Options:
{{CHOICES}}
```

An example complete question is:

```
The following is a multiple choice question (with answers).
Think step by step and then output the answer in the format of "The answer is (X)" at the end.

A fluorescence microscopy image displays cultured mouse primary neurons stained with a specific marker localized to mitochondria. The fluorescence signal exhibits an uneven distribution across the neuronal cell, and there is an absence of signal within the nucleus. What is the most probable cause for the uneven distribution of the mitochondrial signal?

Options:
(1): Mitochondrial transport along axonal and dendritic processes ensures uneven distribution based on cellular energy demands.
(2): Differential expression of mitochondrial proteins leads to variable fluorescence signal across the neuron.
(3): Mitochondrial biogenesis occurs at specific neuronal regions, causing localized fluorescence accumulation.
(4): Mitochondrial anchoring to the cytoskeleton restricts their movement, resulting in uneven distribution.
(5): Mitochondrial dynamics, including fission and fusion, allow mitochondria to move and distribute throughout the neuron.
```

The answer is then extracted with the regex pattern: `answer is \((?[0-9])\)?`.

For LLaVA-Med, we added an extra parsing strategy. This was because we observed that while LLaVA-Med often gave reasonable responses to questions, it failed to follow instructions well. First, it did not follow the CoT prompt consistently to think step by step. It also didn’t use the right answer formatting: "The answer is (X)". Since we were more interested in biomedical reasoning adhering to VQA formatting, we decided to handle this model differently.

Specifically, to encourage CoT reasoning, we fixed the prefix of the response to:

```
Response: Let’s think through this step by step and evaluate each option..
```

Then we added a suffix to the end:

```
After carefully considering all the options, the most likely answer is (,
```

Then we continued generation, and did regular regex parsing to get the final answer.

F.2. Model details

Our main benchmarking experiments in Tab. 3 refer to models that have multiple versions. The exact model versions used for inference are shown in Tab. 6

Model Name	API Endpoint	Source & Details
GPT-4o	gpt-4o-2024-08-06	OpenAI API
GPT-4o-mini	gpt-4o-mini-2024-07-18	OpenAI API
Claude-3.5-sonnet	anthropic/claude-3-5-sonnet	OpenRouter API (Nov 12) Equiv: claude-3-5-sonnet-20241022 on Anthropic API
Claude-3-Haiku	anthropic/claude-3-haiku	OpenRouter API (Nov 12) Equiv: claude-3-haiku-20240307 on Anthropic API
Gemini-1.5-Pro	google/gemini-pro-1.5	OpenRouter API (Nov 12)
Gemini-1.5-Flash-8B	google/gemini-flash-1.5-8b	OpenRouter API (Nov 12)
VILA1.5-40B	Efficient-Large-Model/VILA1.5-40b	HuggingFace (Nov 12), Local inference
VILA1.5-13B	Efficient-Large-Model/VILA1.5-13b	HuggingFace (Nov 12), Local inference
Qwen2-VL-72B	Qwen/Qwen2-VL-72B-Instruct	Hyperbolic API (Nov 12)
Qwen2-VL-7B	Qwen/Qwen2-VL-7B-Instruct	Hyperbolic API (Nov 12)
Llama 3.2 90B	meta-llama/llama-3.2-90b-vision-instruct	OpenRouter API (Nov 12)
Llama 3.2 11B	meta-llama/llama-3.2-11b-vision-instruct	OpenRouter API (Nov 12)
Pixtral 12B	mistralai/pixtral-12b	OpenRouter API (Nov 12)
LLaVA-Med-Mistral-7B	microsoft/llava-med-v1.5-mistral-7b	HuggingFace (Nov 12), Local inference
LLaVA-Mistral-7B	liuhaotian/llava-v1.6-mistral-7b	HuggingFace (Nov 12), Local inference

Table 6. AI model API endpoints and their sources

For LLaVA and LLaVA-med, although the version number 1.6 is different from LLaVA-Med with 1.5, we do believe that LLaVA-Med was either trained from this LLaVA version, or from one that is very similar. They have identical architecture, and the version is very close; it is possible the LLaVA-Med copied the number incorrectly. The LLaVA-Med codebase or github issues does not clarify this at the time of writing.

Note that for OpenRouter function calls to Anthropic, the date is not included in the string. We believe they use timestamps for older model versions, but not for the latest version. So reproducing it may require checking if the model has been updated.

F.3. More experiments on attribute-level information

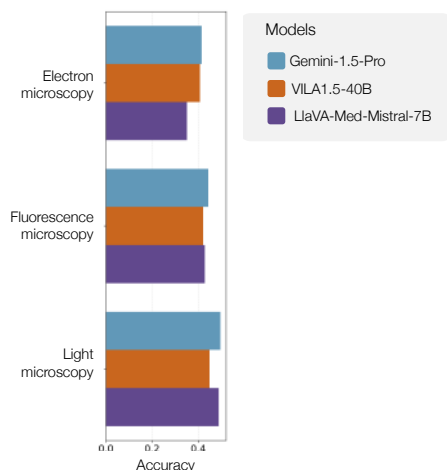


Figure 11. Performance by image modality type for the best models: Gemini-1.5-Pro (closed source), VILA1.5-40B (open-source), and LLaVA-Med-Mistral-7B (medical).

We described our metadata features in Sec. 3, and we can now leverage them to benchmark models on different attributes. Fig. 5 (left) shows all sub-tasks, and shows that ‘compare image sets’, which is common in biology to compare ‘treatment’ and ‘control’ groups. All models perform relatively well, despite multi-image reasoning being understudied in MLLMs [64]. Fig. 5 (right) shows that for Gemini and LLaVa-Med, higher Bloom’s levels – corresponding to more advanced reasoning – leads to worse performance. This supports the intuitive result that higher-level reasoning is more challenging – although VILA does not follow this trend. A final result in Fig. 11 shows that all models performed worse on cryo-electron microscopy images than brightfield or fluorescence microscopy. This may be due to cryo-EM’s challenging features (low contrast, low signal-to-noise ratio) and their under-representation in pretraining datasets [77].

F.4. Error analysis

F.4.1 Perception errors

Subject: Cell and molecular biology

Bloom's: Level 4 (Analysis)

Specimen: mammalian cells

A cryo-electron tomography (cryoET) slice of mammalian cells was acquired at a resolution of 1.3 Å and binned by 4 (resulting in 5.3 Å per pixel). Within the tomogram, dark circular like structures are observed in an organelle. What is the most likely identity of these structures?

Stress granules

✗ Ribosomes

AI: Let me think about this step by step:

...

2) Let's consider their size and appearance:

- dense, small, roughly circular objects
- characteristic electron density pattern of ribosomes

...

My prediction is Ribosomes

Error type: Perception

Error reason: These are granules characterized by heterogeneous, electron-dense aggregates. Ribosomes would be larger, more regular in shape, and not have as high contrast.

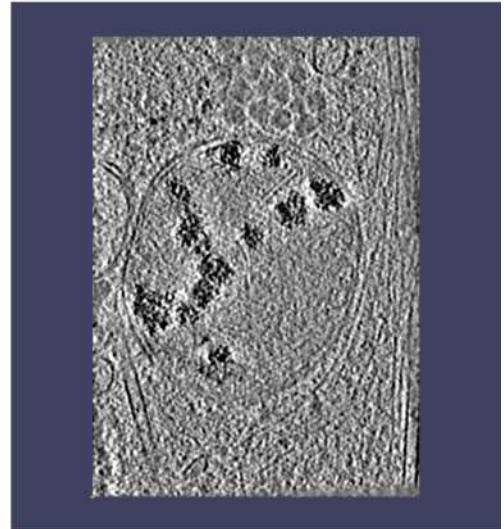


Figure 12

Subject: Anatomy
Bloom's: Level 4 (Analysis)
Specimen: gastrointestinal lesion biopsy

A hematoxylin and eosin (H&E)-stained micrograph shows cellular formations within a gastrointestinal tissue sample with unique organization and cellular patterns. What pattern and morphology do these cells most closely align with?

Bundles of elongated cells with pale pink cytoplasm and uniform oval nuclei in a crisscross pattern

✗ Circular groups of cells with mucin in central spaces and small, dark nuclei

AI: Let me analyze this step by step:

- 1) Looking at the H&E-stained micrograph:
 - It shows a circular cross-section of gastrointestinal tissue
 - The center shows normal pink-stained connective tissue
- 2) Focusing on the cellular arrangement:
 - The cells are arranged in circular groups
 - These groups form glands with lumens

...
4) The image shows typical intestinal glands (crypts) with:

- Circular arrangement with central lumens containing ...

...
My prediction is "Circular groups of cells with mucin in ..."

Error type: Perception

Error reason: No. The intestinal glands are not the lesion. There is a pink spindle cell lesion with a fascicular arrangement of cells.

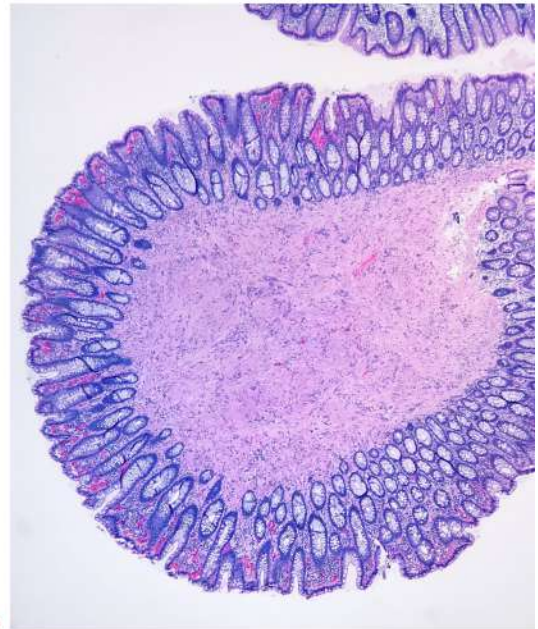


Figure 13

Subject: Pathology
Bloom's: Level 4 (Analysis)
Specimen: schwannoma tissue

An H&E-stained image of a nerve sheath tumor displays unique cell arrangements and marked changes within its structure. Given these observations, what is the most likely reason for the observed morphological alterations?

Outcome of a chronic process affecting the tumor

✗ Indication of rapid cellular proliferation

AI: Let me analyze this step by step:

1) Looking at the H&E-stained images, I can observe:

- Dense cellular proliferation
- Distinctive "herringbone" or swirling pattern

2) Key characteristics visible:

- The cells are arranged in fascicles
- There's a high nuclear to cytoplasmic ratio
- The pattern suggests active tumor growth

...

4) The morphological features strongly suggest active tumor growth with:

- High cellular density
- Signs of active proliferation

...

My prediction is Indication of rapid cellular proliferation

Error type: Perception

Error reason: No. One image shows cystic change, which can be seen in chronic, longstanding schwannomas. The other shows Verocay bodies and fascicular growth, which do not indicate rapid cellular proliferation.

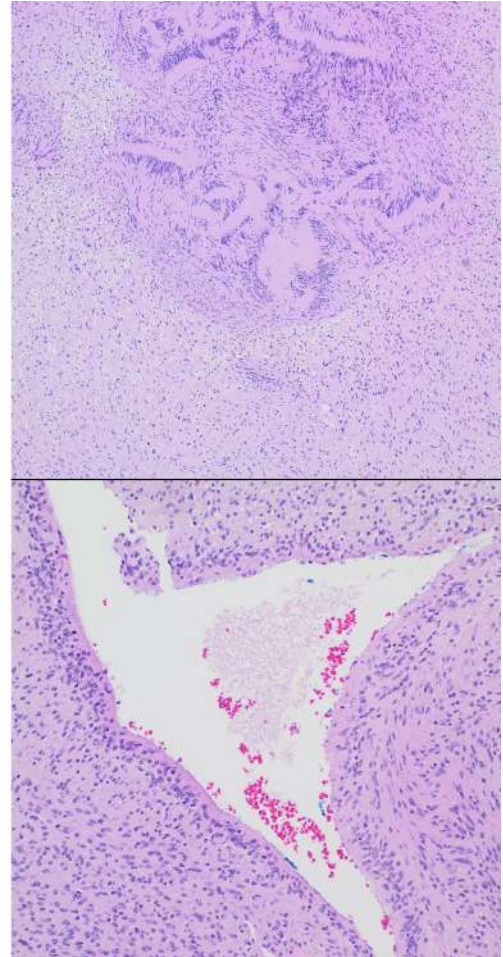


Figure 14

Subject: Physiology

Bloom's: Level 2 (Comprehension)

Specimen: mouse liver tissues (wounded and unwounded)

In a fluorescence microscopy image comparing wounded and unwounded liver tissues stained to visualize markers of immune response - neutrophil-derived migrasomes (Ly6G, yellow) and platelets (CD41, magenta), what is most likely observed about the distribution of these markers in unwounded tissue?

Slight overlap without any notable biological consequences

✗ No overlap, indicating complete segregation of cellular functions

AI: Let me think about this step by step:

1) The image shows two types of staining:

- Yellow (Ly6G) for neutrophil-derived migrasomes
- Magenta (CD41) for platelets
- The markers rarely overlap

...

My prediction is: "No overlap, indicating complete segregation of cellular functions"

Error type: Perception

Error reason: The markers rarely overlap, but there is very focal colocalization in some puncta. The selected answer argues for "complete segregation of cellular functions", which is a too extreme or absolute of a conclusion based on the image.

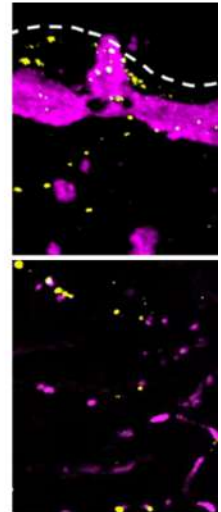


Figure 15

Subject: Neurobiology
Bloom's: Level 4 (Analysis)
Specimen: brain parenchyma

A reticulin stain of cerebral tissue, including its vasculature, reveals disorganized patterns around fiber structures in a patient with hemimegalencephaly. Which glial cell is likely responsible for this abnormal reticular fiber pattern as indicated by the stain?

Subpial astrocytes

✗ Vascular astrocytes

AI: Let's solve this step by step:

1) First, let's understand what we're looking at:

- The image shows a reticulin stain of cerebral tissue
- We can see disorganized patterns around fiber structures
- This is in a patient with hemimegalencephaly (enlarged brain hemisphere)

2) What does a reticulin stain show?

- Reticulin stains show type III collagen fibers
- In the brain, these are typically associated with blood vessels

...

5) In hemimegalencephaly:

- There's abnormal brain development
- Vascular patterns can be disrupted
- The supporting glial architecture is often abnormal

The answer is vascular astrocytes

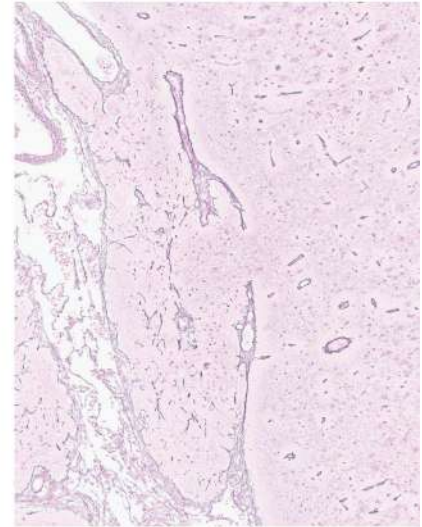


Figure 16

Error type: Perception

Error reason: It failed to recognize that the subpial region, not vasculature, is the primary site of abnormality in this context.

F.4.2 Knowledge errors

Subject: Neurobiology
Bloom's: Level 3 (Application)

A double immunofluorescence experiment was conducted on nerve cells, highlighting Sema4B (red marker) and Sox9, an astrocyte marker (green marker). Sema4B was found primarily surrounding the nuclear area and dispersed in the cytoplasm. Based on this pattern, what potential role might Sema4B play in these cells?

Sema4B acts as a signaling molecule interacting with cell surface receptors.

✗ Sema4B aids in transporting proteins across cell compartments.

AI: Let me solve this step by step:

1) Looking at the immunofluorescence images:

- Sox9 (green) marks astrocytes and Sema4B (red) shows a particular distribution

2) Analyzing Sema4B's distribution:

- Red punctate staining, primarily around the nucleus

- Shows dispersed pattern in cytoplasm

- Not concentrated in nucleus itself

3) This distribution pattern suggests:

- The protein is involved in cytoplasmic processes

- Vesicular-like pattern suggesting role in trafficking

...

5) Reasoning:

- The punctate pattern and distribution is characteristic of proteins involved in cellular trafficking

The answer is (4)

Error type: Misconception

Error reason: Sema4B is a transmembrane proteins that plays a role in cellular signalling. TM proteins can be internalized via endocytosis and does not mean that its main function is intracellular trafficking.

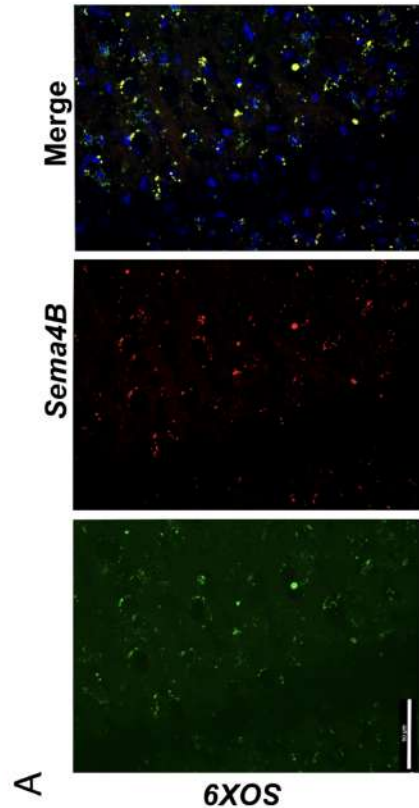


Figure 17

Subject: Neurobiology
Bloom's: Level 4 (Analysis)
Specimen: fibroblasts from an Alzheimer's patient

A Cryo-Electron Tomography image of fibroblasts from a 70-year-old patient with Alzheimer's disease shows three mitochondria gathered together. What is the most likely reason for this mitochondrial clustering?

Mitochondrial fission due to cellular stress

✗ Impaired mitophagy causing accumulation of damaged mitochondria

AI: Let's solve this step by step:

...

2) Analyzing each option: ...

- Fission: Would show smaller, fragmented mitochondria
- Impaired mitophagy: Would show accumulated, damaged mitochondria

...

My prediction is impaired mitophagy.

Error type: Misconception

Error reason: Impaired mitophagy may result in larger overall mitochondria, not an increase in small microchondrial aggregates. Fission is a better answer in this case.

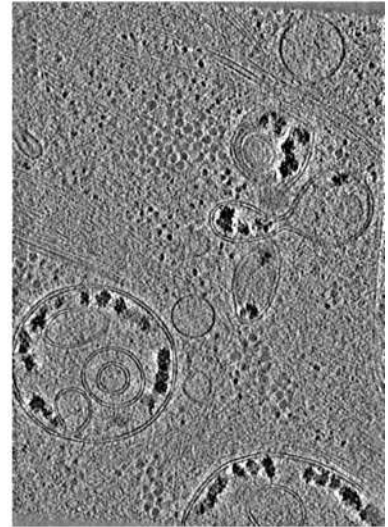


Figure 18

Subject: Neurobiology
Bloom's: Level 3 (Application)
Specimen: resection specimen for epilepsy

A microscopic analysis of tissue from an individual suffering from recurrent seizures involves using hematoxylin & eosin stain alongside a vimentin stain. Cells with strong vimentin expression are significantly larger than surrounding entities. Which identity is most probable for these prominent vimentin-expressing cells?

Aberrant large cells characteristic of cortical malformation

✗ Enlarged reactive astrocytes near injury sites

AI: Let me solve this step by step:

1) First, let's analyze what we're seeing in the images:

- The H&E stain (top image) shows tissue with scattered cells and some enlarged cells
- The vimentin stain (bottom image) shows strong brown staining of certain larger cells

2) Key information from the question:

- Patient has recurrent seizures
- Vimentin-positive cells are significantly larger
- Looking for most probable identity of these cells

...

4) In the context of seizures:

- Seizures can cause brain injury
- Brain injury triggers astrogliosis

...

My prediction is Enlarged reactive astrocytes near injury sites

Error type: Misconception

Error reason: The AI is able to interpret the question and image, but lacks the knowledge to recognize the link between recurrent seizures, cortical dysplasia, and large ballooned cells.

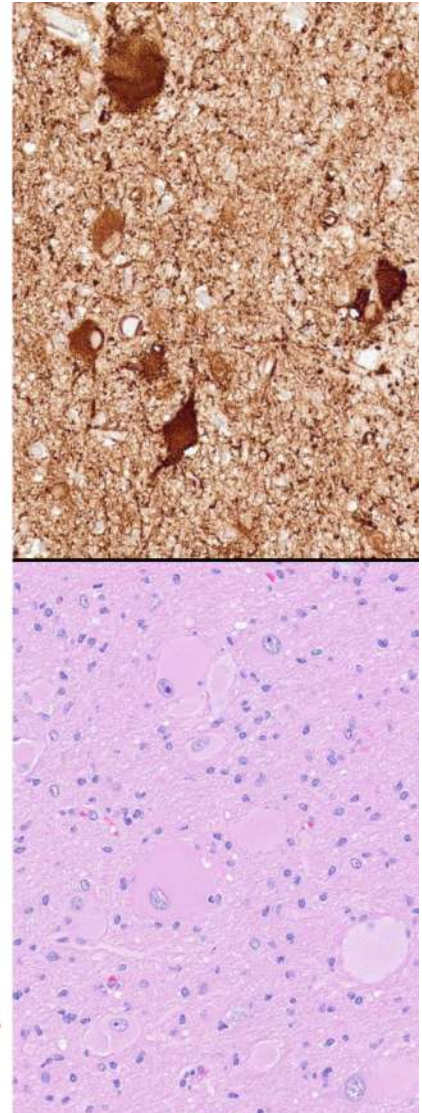


Figure 19

Subject: Neurobiology
Bloom's: Level 4 (Analysis)
Specimen: brain parenchyma

A certain specialized stain reveals unusual cell movement crossing boundaries typical of neuronal migration in a tissue resection from a case of structural brain anomaly. Which condition is primarily associated with this type of cellular pattern?

Type II lissencephaly

✗ Focal cortical dysplasia

AI: Let me solve this step by step:

- 1) The image shows a histological section with specialized staining that reveals abnormal cellular organization and migration patterns.
- 2) The question asks about unusual cell movement crossing typical neuronal migration boundaries.
- 3) Looking at the options:
 - Focal cortical dysplasia involves disorganized neurons in localized areas

...

- Type II lissencephaly shows abnormal neuronal migration and positioning

- 4) Key characteristics of focal cortical dysplasia include:

....

- 5) The histological image shows features consistent with focal cortical dysplasia, including:

...

The answer is (1)

Error type: Misconception

Error reason: Misconception about the scope and extent of neuronal migration patterns in focal cortical dysplasia (FCD) versus Type II lissencephaly. FCD will have abnormal cortical lamination, but the neurons are still within the expected gray/white matter of normal neuronal migration. They do not extend beyond and cross established neuronal boundaries such as the glial limitans.

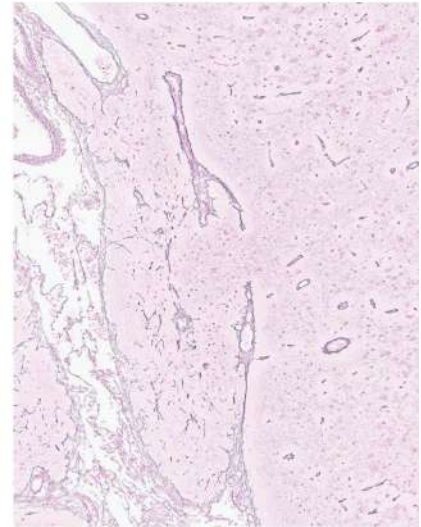


Figure 20

F.4.3 Overgeneralization or oversimplification errors

Subject: Pathology
Bloom's: Level 2 (Comprehension)
Specimen: schwannoma tissue

In a histological examination, an H&E-stained image shows Verocay bodies and cystic degeneration with pseudoepithelial metaplasia, indicative of schwannoma. How does the risk of malignant transformation in schwannoma compare to other nerve sheath tumors?

Schwannomas exhibit a lower risk of malignant transformation compared to other peripheral nerve sheath tumors.

✗ Schwannomas rarely undergo malignant transformation, regardless of tumor location or size.

AI: Let me think about this step by step:

1) The images show typical histological features of schwannoma with:

- Verocay bodies (palisading nuclei)

...

2) Regarding malignant transformation in schwannomas:

- Schwannomas are benign and rarely become malignant

...

My prediction is Schwannomas rarely undergo malignant transformation...

Error type: Overgeneralization

Error reason: The selected choice does not address the question asked "compared to other peripheral nerve sheath tumors?". The "best answer" for the question should compare to other tumors as specified in the question. The AI has "oversimplified" the question to ask a simpler question than the one asked.

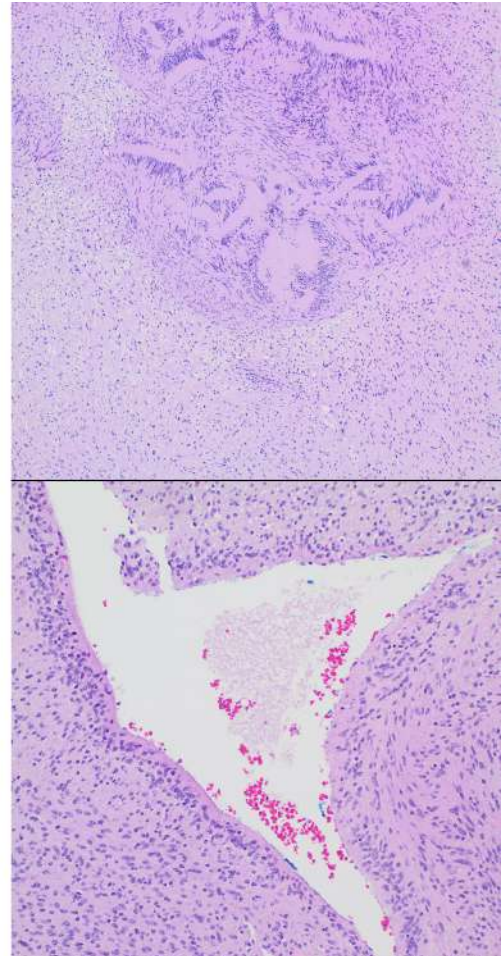


Figure 21

Subject: Cell and molecular biology
Bloom's: Level 4 (Analysis)
Specimen: HEK293 cells (wildtype and PEX19 knockdown)

Confocal microscopy images depict live HEK293 cells subjected to CRISPR/Cas9-mediated knockdown of a specific gene, alongside wildtype cells. The cells are stained with BODIPY-C12 (cyan) to highlight lipid droplets and Hoechst to stain the nuclei. In wildtype cells, distinct organellar structures are visible, whereas these structures are absent in the gene knockdown cells. What is the most likely cellular outcome resulting from the knockdown of this gene in HEK293 cells?

Impaired peroxisome biogenesis

✗ Altered lipid metabolism pathways

AI: Let me solve this step by step:

1) First, let's analyze what we're seeing in the images:

- Live HEK293 cells are shown
- BODIPY-C12 staining (cyan) shows lipid droplets

...

2) Key observations:

- The cyan staining (BODIPY-C12) indicates the presence of lipid droplets
- There's a difference between wildtype and knockdown cells in terms of organellar structures
- The staining specifically targets lipids

...

4) Reasoning:

- Since BODIPY-C12 specifically stains lipids
- The visible differences are in lipid-containing structures
- The most direct connection would be to lipid metabolism
- Changes in lipid metabolism would directly affect lipid droplet formation and distribution

The answer is (3)

Error type: Overgeneralization

Error reason: Assumes that the changes in lipid droplets are directly linked to altered lipid metabolism pathways, without fully integrating the implications of peroxisome biogenesis in lipid metabolism and organelle dynamics.

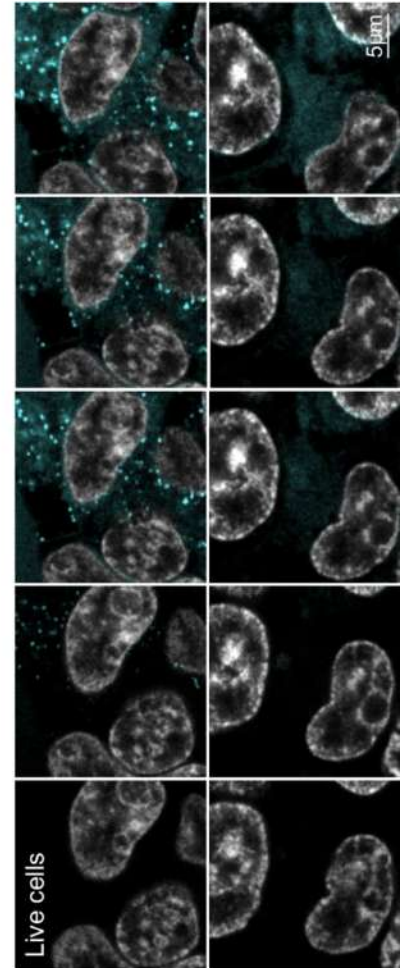


Figure 22

F.4.4 Hallucination or errors due to added details

Subject: Cell and molecular biology

Bloom's: Level 4 (Analysis)

Specimen: cellulose fibers in vitro

A CryoET tomographic slice image and its corresponding 3D annotation of cellulose fibers in vitro show differing amounts of cellulose fibers. Beyond the 2D nature of the slice versus the 3D nature of the annotation, which mechanism most likely accounts for the observed differences?

Variability in manual annotation leading to subjective bias.

✘ Overlapping fibers being misinterpreted as single fibers due to resolution limits.

AI: Let me solve this step by step:

1) First, let's analyze what we're seeing:

- There appears to be a difference in how many fibers are visible/annotated

2) Consider the limitations of CryoET imaging:

- Resolution is limited

- Multiple fibers can appear to overlap in 2D views

...

3) Evaluate the options:

- Resolution limits could cause overlapping fibers to appear as one in the 2D slice

- Manual annotation variability wouldn't explain systematic differences

...

- The resolution limits of the microscope make it difficult to distinguish overlapping fibers in the 2D view

My prediction is Overlapping fibers being misinterpreted ...

Error type: Hallucination

Error reason: The question does not state there are systematic differences. This detail is added during the AI CoT reasoning. A better answer would rule out easy or obvious causes of differences (random or systematic). Although fibers can overlap in Cryo-EM, it is generally considered to have a reasonably high resolution.

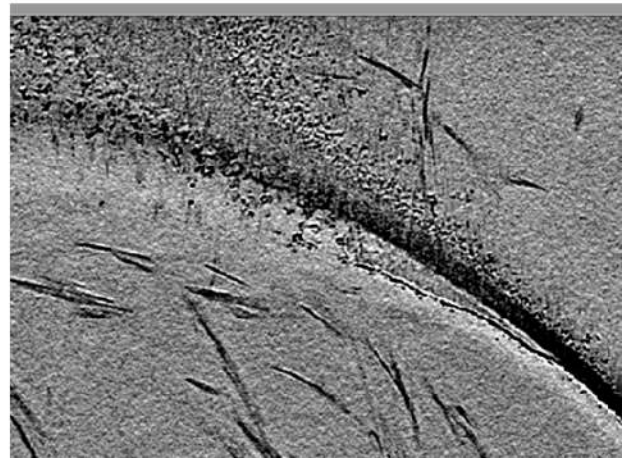
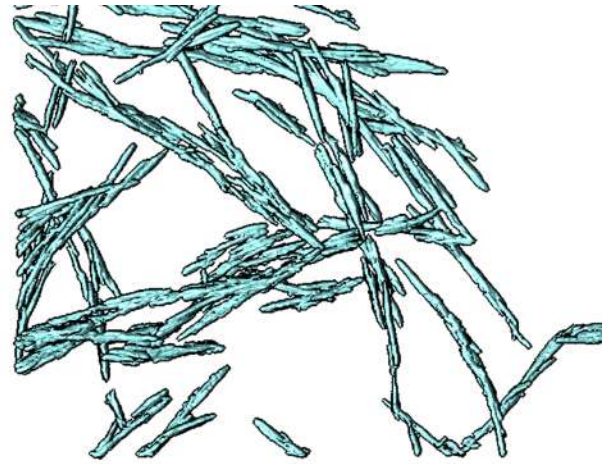


Figure 23

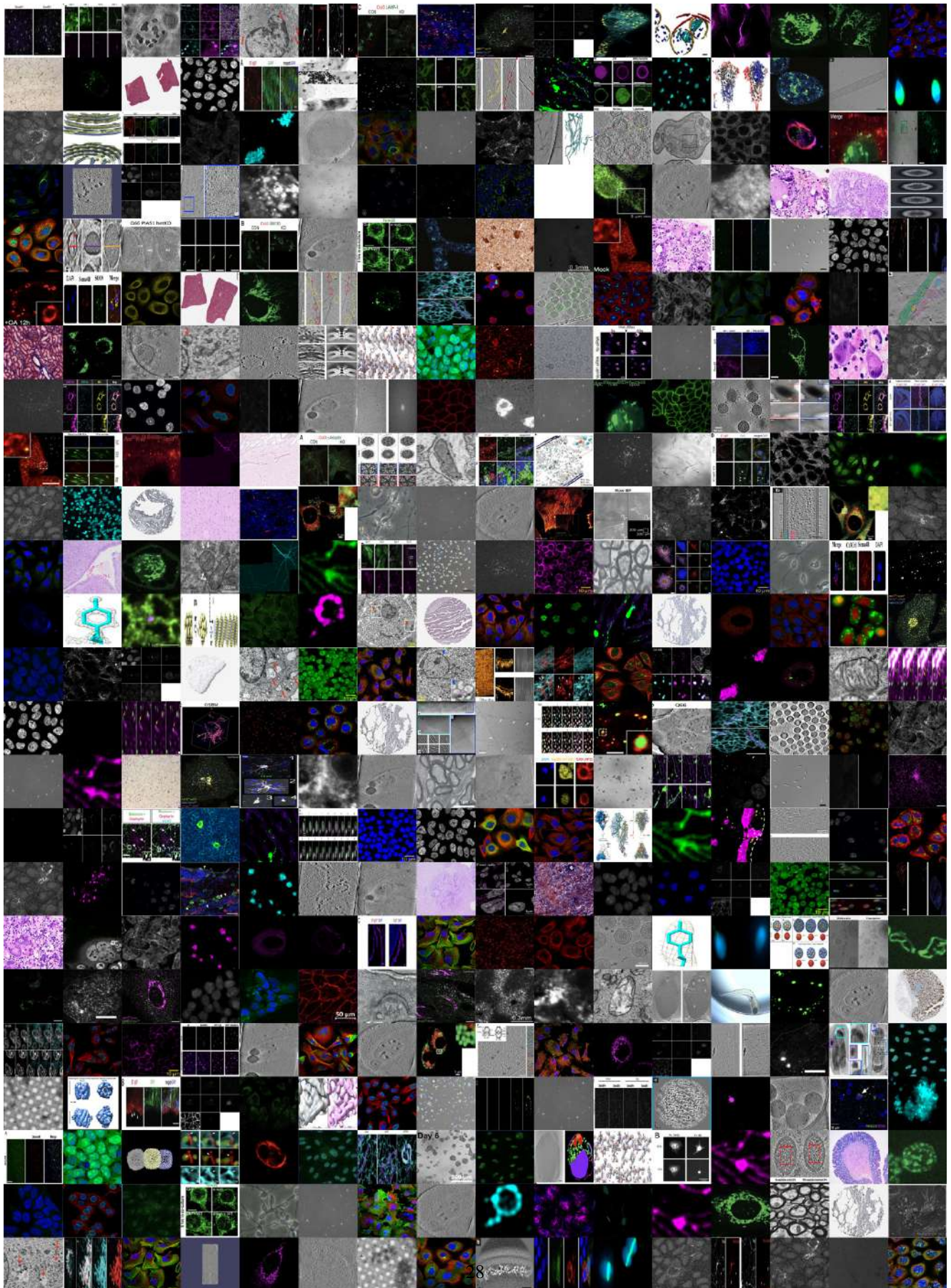


Figure 24. Collage of images from MicroVQA.