BIOREASONBENCH: A DIAGNOSTIC BENCHMARK FOR MULTIMODAL REASONING WITH STRUCTURED VISUAL EVIDENCE IN BIOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) can describe scientific images fluently, but it remains unclear whether they can use biological images as structured evidence across a complete reasoning chain. Existing evaluations are often text-only, recognition-oriented, or summarized by a single accuracy that obscures where reasoning fails. We introduce **BioReasonBench**, a diagnostic benchmark of 1,027 biology questions, spanning three question types: single-select, multiple-select, and open-ended. Of these, 894 (87.0%) are grounded in images of pedigrees, pathways, experimental setups, quantitative curves, and other biological content. Unlike the single-answer choice and single-target open responses common in multimodal benchmarks, 193/194 multiple-select items have two or three correct options, and all 298 open-ended items are multi-blank (2,238 blanks in total). Each item is organized along five diagnostic axes and source/year splits, and is evaluated with question-type-aware exact, partial-credit, and LLM-judge protocols. Across 13 MLLMs, mean accuracy falls from 61.2% on single-select to 28.3% on exact multiple-select and 11.6% on fully correct open-ended questions. Relation graphs, tables, and genetics are especially difficult. Partial option sets and partially correct blanks are common, revealing failures of correspondence, state tracking, calibration, and causal completeness rather than uniform factual ignorance. On `set0`, the strongest model scores 71.0/100, matching the lowest of 10 Grade-12 students, who average 88.1. We hope BioReasonBench provides actionable insights into fine-grained visual-reasoning mechanics and guides the development of next-generation MLLMs capable of trustworthy scientific inquiry.

1 INTRODUCTION

Multimodal large language models (MLLMs) already describe charts, diagrams, and documents with impressive fluency (Radford et al., 2021; Liu et al., 2023; OpenAI, 2023). In biology, however, the decisive question is not whether a model can name what is in an image, but whether it can use that image as evidence for a multi-step conclusion. A pedigree is useful only if symbols remain bound to individuals across generations; a pathway only if arrows preserve direction and regulation; a table or curve only if rows, columns, and conditions stay aligned with the claim being made. We call such content *structured visual evidence*: visual relations, symbol bindings, and state constraints that must be preserved while deriving the answer. The same figure can therefore be recognized correctly at the object level and still support an incorrect biological conclusion.

This gap is hard to see under recognition-oriented evaluation. Many biology questions demand more than local identification: models must coordinate diagram structure with latent mechanisms such as inheritance, regulation, transport, or energy flow, update intermediate states, and emit an answer that satisfies every required constraint. Failures then look like partially correct option sets, early state-tracking mistakes that cascade through later steps, or open responses that recover relevant vocabulary yet omit a required mediator, polarity, or conclusion. In short, fluency and partial knowledge can coexist with incomplete reasoning. The implication is direct. Symbols, arrows, rows, and states in a biology figure constitute part of the evidence, and a model uses that evidence only when the corresponding bindings remain valid in the final answer.

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

Table 1: Comparison with representative science benchmarks. #Bio counts biology-labeled items. Types: SS = single-select, MS = multiple-select, Open = open-ended. Axes lists the main reporting slices. Open describes exam-style open responses. BioReasonBench uses a five-axis taxonomy and multi-blank open items (2,238 blanks).

Benchmark	Modality	#Bio	Types	Axes	Open
GPQA	Text	105	SS	domain	n/a
MMLU / CMMLU	Text	454 / 169	SS	subject	n/a
MMMU	Text+Image	380	SS/Open	subject	short
AI2D	Image	n/a	SS	n/a	n/a
C-Eval / AGIEval	Text	417 / 210	SS	subject	limited
BioReasonBench	Text+Image	1,027	SS/MS/Open	5-axis	multi-blank

Existing science benchmarks do not isolate this failure mode (Table 1). Text-only suites such as GPQA, MMLU, and C-Eval stress domain knowledge without visual grounding (Rein et al., 2024; Hendrycks et al., 2021; Huang et al., 2023). Broad multimodal suites such as ScienceQA and MMMU include biology, but typically mix it with many disciplines and favor single-answer choice or short open responses (Lu et al., 2022; Yue et al., 2024). Diagram and chart benchmarks such as AI2D and ChartQA emphasize recognition or extraction more than exact option-set completion and multi-blank scientific answers (Kembhavi et al., 2016; Masry et al., 2022). Across these settings, a single accuracy number further collapses where reasoning breaks (by field, representation, or answer completion) into one leaderboard score. What is needed is a biology-focused benchmark that makes structured visual grounding, complete answering, and slice-level diagnosis first-class evaluation goals.

We introduce **BioReasonBench**, a diagnostic benchmark of 1,027 biology questions (894 image-grounded) spanning single-select, genuine multiple-select, and multi-blank open-ended formats (2,238 blanks). Each item is annotated along five axes (organization level, organism group, mechanism field, reasoning mode, and visual type) and scored with exact, partial-credit, and blank-/question-level protocols so that local correctness can be separated from answer completion (Figure 1). Our contributions are:

1. **A comprehensive multimodal biology benchmark.** BioReasonBench jointly evaluates single-select, option-set multiple-select, and multi-blank open-ended items grounded in pedigrees, pathways, experiments, tables, and related biological displays.
2. **A five-axis diagnostic taxonomy.** Organization, taxon, mechanism field, reasoning mode, and visual type support field- and representation-specific analysis beyond aggregate accuracy.
3. **An empirical diagnosis of current MLLMs.** Experiments localize the dominant deficits to relational grounding, state tracking, option-set calibration, and causal completeness, and show that Grade-12 students remain ahead on the same paper.

2 Related Work

Scientific and biomedical text reasoning. MMLU (Hendrycks et al., 2021), GPQA (Rein et al., 2024), PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), and BioASQ (Tsatsaronis et al., 2015) establish strong baselines for factual recall, domain knowledge, and multi-step scientific reasoning in text. They show that modern models can answer many difficult science and biomedical questions when the required evidence is already verbalized. However, their predominantly textual inputs cannot reveal whether a model can align symbols with entities in a pedigree, preserve row/column constraints in a biological table, or track state changes across a diagram. BioReasonBench complements this literature by keeping the scientific domain fixed while making visual evidence indispensable to the answer.

Multimodal science and image understanding. ScienceQA (Lu et al., 2022), MMMU (Yue et al., 2024), MathVista (Lu et al., 2024), AI2D (Kembhavi et al., 2016), ChartQA (Masry et al., 2022), and SEED-Bench (Li et al., 2024) evaluate multimodal scientific or image-based skills ranging from diagram understanding to chart reading and general academic problem solving. Collectively, these

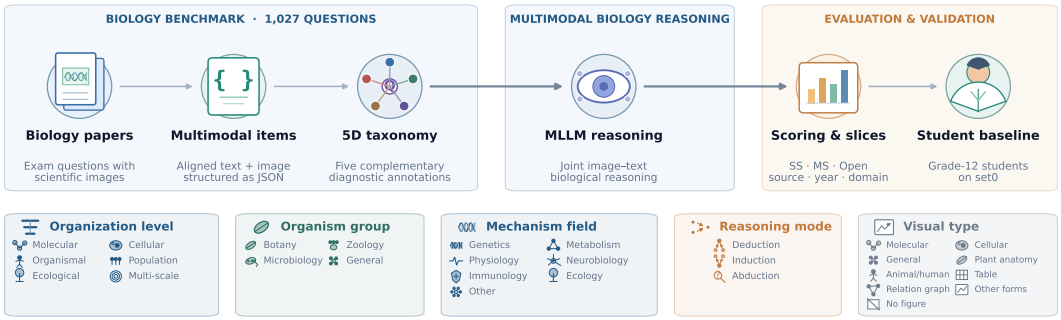


Figure 1: BioReasonBench construction and evaluation pipeline. Papers are converted into aligned text–image items, annotated along five dimensions, and evaluated on single-select, multiple-select, and multi-blank open-ended questions.

resources demonstrate that multimodal evaluation should go beyond plain image captioning and include diagrams, plots, and school-style questions. At the same time, most of them emphasize single-answer choice, short open responses, or broad cross-domain coverage. BioReasonBench uses biology-specific images, genuine multiple-select questions, and multi-blank open problems, so that local recognition can be separated from complete answers grounded in structured visual evidence (Table 1).

3 BioReasonBench

3.1 Overview and Construction

BioReasonBench is constructed from Chinese biology papers, including official and mock/practice sources from recent years. Source papers are parsed into structured JSON with exported images and tables. Stems retain inline image placeholders so multimodal prompts preserve image–text alignment. Figure 1 summarizes the construction and evaluation pipeline. The released benchmark comprises **1,027** questions with **2,238** open answer blanks across set0–set6: six official real-exam blocks (`set0`–`set5`, 527 items, exam years 2026–2022) plus one pooled mock subset (`set6`, 500 items, 2026 practice papers).

The benchmark contains 535 single-select, 194 multiple-select, and 298 open-ended items (2,238 blanks). Multiple-select questions require selecting the complete correct option set: 108 have two correct options, 85 have three, and only one has a single correct option. The open-ended questions alone contribute 2,238 answer blanks (3–14 per question, mean 7.5), and 291/298 comprise multiple subparts. Source split, question type, and biological taxonomy can therefore be analyzed independently rather than collapsed into a single aggregate score.

3.2 Source Splits by Type and Year

We partition items along two complementary axes: *source type* and *exam year*.

- **Official real exams** (`set0`–`set5`): official biology papers from 2026, 2025, 2024, 2023, and 2022 (527 items). `set0` contains 25 items. `set1`–`set5` cover additional official papers (502 items).
- **Mock pool** (`set6`): 2026 mock/practice papers reported as one subset (500 items).

3.3 Question Types and Visual Grounding

Single-select (SS) items require selecting one option from $\{A, B, C, D\}$. **Multiple-select (MS)** items require a set of correct options. Partial credit is awarded only for non-empty proper subsets with no wrong selections. **Open-ended** items contain multiple answer blanks (definitions, mechanisms, experimental conclusions), totaling 2,238 blanks across 298 questions in the full bank. Gold answers may include acceptable variants per blank.

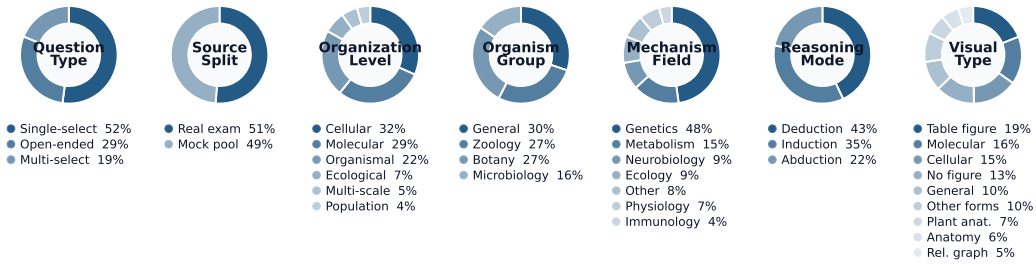


Figure 2: BioReasonBench composition. Seven donut charts show question type, source split, and the five diagnostic dimensions: organization level, organism group, mechanism field, reasoning mode, and visual type.

Visual evidence is central rather than decorative: 894/1,027 items contain images or tables spanning pedigrees, pathways, structures, setups, and quantitative displays. Appendix A presents six representative real questions across all three question types, including answer options for choice items, original Chinese text, English translations, gold answers, and five-dimensional annotations.

3.4 Taxonomy and Diagnostic Axes

A core design goal of BioReasonBench is that *subset structure is first-class*. The source papers are only a collection vehicle. The released taxonomy is defined in scientific rather than curricular terms. Every item is annotated on four complementary axes plus appearance-level visual type (Figure 2): organization level (Urry et al., 2017), organism group, mechanism field, and Peircean reasoning mode (Peirce, 1878). Figure 2 summarizes seven composition views including question type, source split, and the five taxonomy axes on the full set0–set6 bank.

Organization level. We coarsen the hierarchy of biological organization (Urry et al., 2017) into *molecular*, *cellular*, *organismal*, *population*, *ecological*, and *multi-scale*. This axis captures the dominant scale at which an item must be solved.

Organism group. We label the primary biological object of study as *botany*, *zoology*, *microbiology*, or *general* when no taxon is focal. Sparse virology and mycology cues are merged into microbiology, while paleontology-like items are folded into general.

Mechanism field. Orthogonal to taxon focus, we label the dominant mechanistic system as *genetics*, *immunology*, *neurobiology*, *physiology*, *metabolism*, *ecology*, or residual *other*.

Reasoning mode. We use Peirce's three modes of reasoning (Peirce, 1878): *deduction*, *induction*, and *abduction*. This axis captures whether an item is dominated by rule application, pattern generalization, or best-explanation reasoning.

Visual typology. For cleaner appearance-level analysis, we use a single-layer nine-way image partition with at least 50 evaluation items per class: *table images*, *molecular schematics*, *cellular schematics*, *general schematics*, *other image forms*, *plant anatomy schematics*, *animal/human anatomy schematics*, *relation graphs*, and *none*.

3.5 Evaluation Protocol

We use a unified automatic protocol for model outputs:

- **Single-select:** exact letter match. Report accuracy.
- **Multiple-select:** credit 1.0 if the predicted set equals gold, 0.5 if the prediction is a non-empty proper subset of gold with no extras, and 0 otherwise. We report both *exact-set* accuracy and *partial* (mean credit).
- **Open-ended:** official scoring uses an *LLM-as-judge* (Kimi-K2.5) that grades each blank as correct/incorrect given the gold answer(s) and model response. We report both judge

Table 2: Overall results on BioReasonBench set0–set6. Choice aggregates single-select and multiple-select items. MS part. is mean multiple-select credit. Open blk is blank-level accuracy. Open metrics use a Kimi-K2.5 LLM-as-judge.

Model	Choice	SS	MS exact	MS part.	Open Q	Open blk
Kimi-K2.5	**0.745**	0.800	**0.593**	**0.662**	**0.295**	**0.760**
Kimi-K2.6	0.737	**0.802**	0.557	0.647	0.282	0.716
Qwen3.7-Plus	0.686	0.766	0.464	0.585	0.232	0.681
Qwen3.5-Plus	0.586	0.688	0.304	0.461	0.191	0.651
GPT-5.4	0.530	0.697	0.067	0.098	0.185	0.719
Grok-4-20-NR	0.528	0.598	0.335	0.461	0.111	0.622
Mistral-Medium-3.5	0.497	0.579	0.268	0.420	0.034	0.527
GPT-5-mini	0.502	0.579	0.289	0.428	0.067	0.517
GLM-4.6V-Flash	0.451	0.542	0.201	0.420	0.024	0.504
GPT-4.1-mini	0.417	0.523	0.127	0.230	0.010	0.432
GLM-4.1V-Thinking-Flash	0.387	0.454	0.201	0.327	0.060	0.521
Grok-4-1-fast-NR	0.384	0.464	0.165	0.263	0.013	0.395
GLM-4V-Flash	0.366	0.460	0.108	0.325	0.003	0.395

question-level accuracy (all blanks in an item judged correct) and blank-level accuracy (Appendix E).

4 Experiments

4.1 Setup

We evaluate **13** MLLMs in a zero-shot multimodal setting: Chinese stem text, options (when present), and attached images/tables. The pool covers Kimi-K2.5/K2.6 (Kimi Team, 2026a;b), Qwen3.5-Plus / Qwen3.7-Plus (Qwen Team, 2026a;b), GPT-5.4 / GPT-5-mini / GPT-4.1-mini (OpenAI, 2026; 2025b;a), Grok-4-20-NR / Grok-4-1-fast-NR (xAI, 2025a;b), GLM-4.6V-Flash / GLM-4.1V-Thinking-Flash / GLM-4V-Flash (V Team, 2025; Zhipu AI, 2025), and Mistral-Medium-3.5 (Mistral AI, 2026). Every model receives the same instruction to write a step-by-step reasoning trace and then a final answer. The reasoning trace is the input to the error analysis below. Choice items use the exact/partial protocol above, and open items are graded by a Kimi-K2.5 LLM-as-judge.

4.2 Implementation Details

We query each family through its strongest publicly available multimodal API using a unified zero-shot prompt with Chinese stems, options, and attached images. Provider details, prompts, and code are released anonymously at `https://anonymous.4open.science/r/BioReasonBench-4F54`, with the LLM-judge and error-analysis prompts reproduced in Appendix B and Appendix C.

4.3 Overall Performance and Capability Separation

Table 2 establishes three results that organize the analyses below. First, no evaluated MLLM is reliable across all question types. Kimi-K2.5 leads choice accuracy (74.5%), multi-select exact accuracy (59.3%), open question accuracy (29.5%), and open blank accuracy (76.0%), yet still completes fewer than one third of open questions. Second, rankings are metric-dependent: GPT-5.4 reaches 71.9% open blank accuracy but only 6.7% exact multi-select accuracy, whereas other systems recover larger option sets but weaker open answers. Third, local correctness is much more common than complete correctness. The remaining sections test where this completion gap arises, whether it changes with source exposure, which biological and visual structures amplify it, and how far all models remain from students.

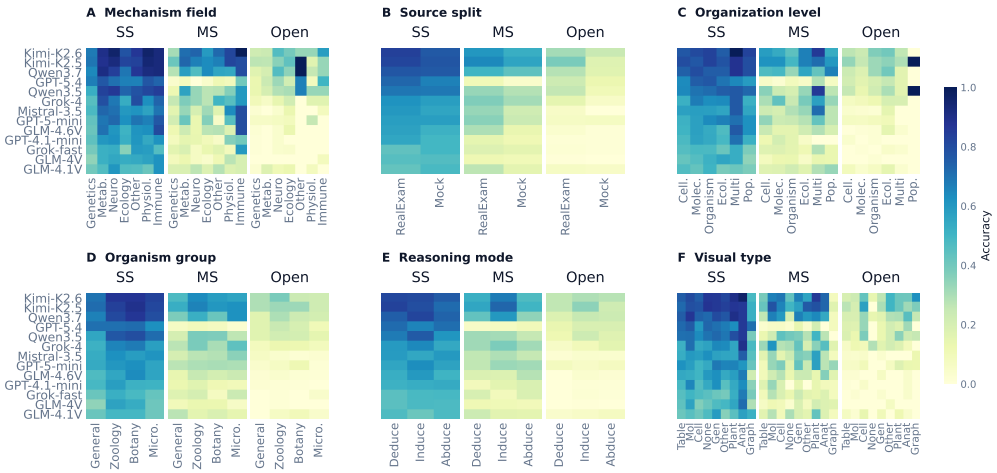


Figure 3: Heatmaps over SS, MS exact, and open accuracy by model rank.

4.4 Finding 1: Answer Completion Is the Primary Bottleneck

Across 13 models, mean accuracy falls from 61.2% on single-select to 28.3% on exact multi-select. Partial multi-select credit rises to 41.0%. This 12.7-point gap means that models frequently identify some correct statements without determining the complete answer set. Appendix G plots the same gradient. The gap is especially large for GLM-4.6V-Flash and GLM-4V-Flash (21.9 and 21.7 points), consistent with a systematic inclusion/stopping calibration problem rather than isolated knowledge errors. Open generation shows the same local-to-global failure in a different form: mean blank accuracy is 57.3%, but only 11.6% of questions have every blank correct. GPT-5.4 makes the separation particularly clear, combining 71.9% blank accuracy with 18.5% question accuracy and 6.7% multi-select exact accuracy. The principal bottleneck is the assembly of all required evidence into a complete answer. Such completeness indicates that the figure has been used throughout the reasoning, rather than merely recognized.

4.5 Finding 2: Difficulty Concentrates in Genetics and Relation-Structured Items

Figure 3 compares all 13 models across six diagnostic slices: mechanism field, source split, organization level, organism group, reasoning mode, and visual type. Read in order, each panel contains a harder subcategory. Across models, the most persistent failures align with genetics and relation-structured evidence. All panels share the same model order and $[0, 1]$ scale.

A. Mechanism field. Across models, genetics has the lowest mean choice accuracy (42.6%), well below immunology (72.2%), physiology (63.8%), and neurobiology (61.7%). This is consequential because genetics is also the largest field (490 items), so failures cannot be attributed to a small tail category. The contrast persists among the strongest systems: Kimi-K2.5 scores 61.7% on genetics choice items versus 90.3% on immunology, while Kimi-K2.6 scores 58.6% versus 93.6%. Genetics questions often require symbol-to-individual binding, generation and sex tracking, and multi-step inheritance constraint propagation. A model may know the Mendelian rule yet still fail when it must be grounded in a pedigree or combined with genotype bookkeeping.

B. Source split. Averaged across models, RealExam exceeds Mock on single-select (62.2% vs. 60.1%), multi-select exact (32.7% vs. 25.3%), and open question accuracy (14.4% vs. 7.8%). This likely reflects shifts in item construction and surface framing between official and practice sources.

C–E. Organization, organism group, and reasoning mode. These three axes do not separate difficulty. Mean single-select accuracy spans 57.9%–71.9% across organization levels and 54.3%–66.8% across organism groups, with no stable order from molecular to ecological scales or across taxa. Deduction, induction, and abduction likewise cluster (choice accuracy 51.3%–54.4%). The axes that do separate items are mechanism field and visual type, developed next.

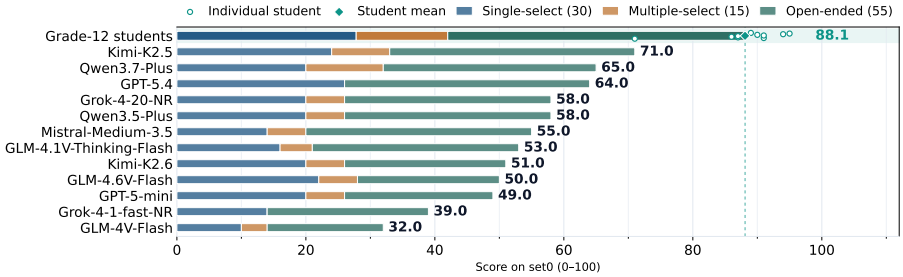


Figure 4: Final-year high-school students versus MLLMs on `set0`. Bars split the 100-point paper into single-select (30), multiple-select (15), and open-ended (55).

F. Visual type. The cross-model visual-type average is lowest for relation graphs (35.2% choice accuracy) and table images (46.2%), compared with 64.0% for animal/human anatomy schematics and 59.2% for molecular or plant-anatomy schematics. Text-only items are not uniformly easiest (52.5%), so the result is not a simple visual penalty. The hardest displays are those that require exact correspondence across nodes, edges, rows, columns, legends, and textual qualifiers. Relation graphs amplify a single localization error through downstream reasoning, while tables require row/column alignment across several conditions. More iconic anatomy images often provide stronger local semantic cues and impose less symbolic state tracking. Taken together, Figure 3 localizes the main bottleneck to genetics and relation-structured evidence rather than to any single organism group, organization level, or reasoning label.

4.6 Finding 3: Grade-12 Students Remain Substantially Ahead

As shown in Figure 4, ten final-year high-school students, randomly selected from a top-tier high school, completed `set0` (25 questions, 100 points: 15 single-select items worth 30, 5 multiple-select items worth 15, and open items worth 55), scoring 71–95 with a mean of 88.1 and median of 89.5. Their mean section scores are 27.8/30 single-select, 14.2/15 multiple-select, and 46.1/55 open. Model totals use the same caps. A correct single-select item scores 2 points. A multiple-select item scores 3 when the predicted set equals gold and 1 when it is a non-empty proper subset. Open points follow the students' per-blank rubric on Kimi-K2.5 judgments: questions 21–25 are worth 8, 10, 12, 11, and 14. Under this rubric, Kimi-K2.5 scores 24.0/30, 9.0/15, and 38.0/55, for 71.0/100, 17.1 points below the student mean. That total matches the lowest student score (71). The other eleven models with `set0` results score 32.0–65.0.

4.7 Finding 4: Errors Localize Beyond Factual Recall

Output-layer signatures (L1). Figure 5A quantifies observable failure signatures across incorrect model–item responses. Of 1,798 multi-select errors, 35.5% are under-selections, 28.3% are over-selections, and 35.9% both omit correct options and include incorrect ones. Among 3,420 failed open responses, 80.1% have multiple incorrect blanks, 16.0% are one blank short of a complete answer, and only 3.9% have no correct blanks. Thus most open failures contain some correct biological content, but the dominant problem is not a single missing token. Several linked components remain unresolved. Figure 5B–C places the same signatures in a cross-model view over all 13 MLLMs. Failure load is strongly type-dependent: open-ended items concentrate at the high-fail end (134 items missed by every model), multiple-select is intermediate, and single-select still retains a substantial easy tail. Items missed by at least 10 models (n=428) are enriched for genetics-heavy mechanism fields, indicating shared structured-reasoning bottlenecks rather than idiosyncratic model noise.

Mechanism-level breaks (L2). Observable L1 signatures do not name the causal break. We therefore attribute a primary L2 mechanism for 4,208 incorrect responses from eight models (Kimi-K2.5/K2.6, Qwen3.5/3.7-Plus, GPT-5.4, Grok-4-20-NR, Mistral-Medium-3.5, GLM-4.6V-Flash), using a locked taxonomy and a Kimi-K2.5 judge that contrasts model reasoning against the official explanation (Appendix C). Figure 7 reports the result. Appendix D gives four item-level walkthroughs with model outputs and attributed mechanisms. *Visual-symbol binding* (29.5%) and

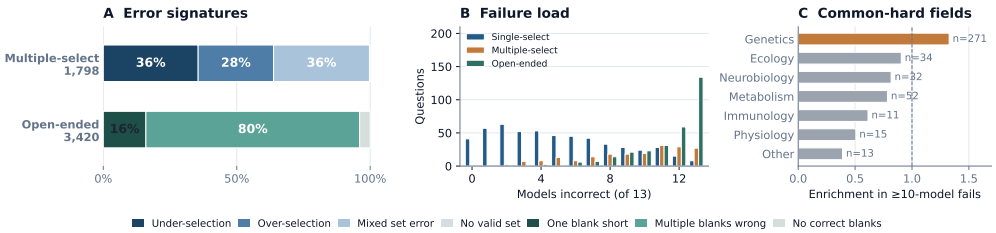


Figure 5: Output-level errors. (A) Signatures on incorrect responses. (B) How many of 13 models miss each item. (C) Mechanism fields enriched among items missed by at least 10 models. Blue, amber, and teal are single-select, multiple-select, and open-ended.

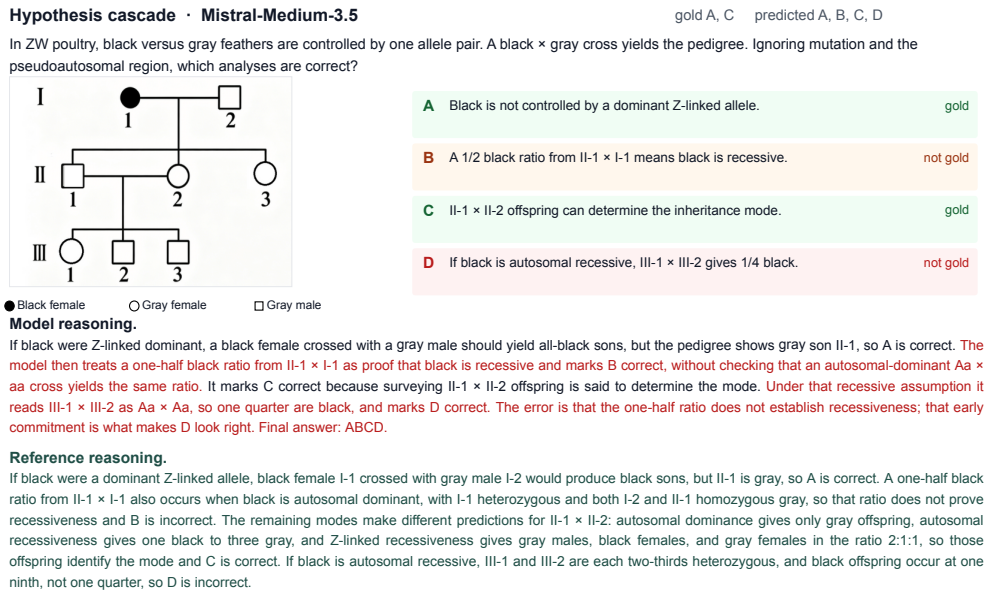


Figure 6: ZW pedigree reasoning from Mistral-Medium-3.5 (gold A, C). Red text marks the step that treats a 1/2 ratio as proof of recessiveness. The last paragraph is the official reference reasoning. Appendix D has the full bilingual card.

hypothesis cascade (27.3%) alone account for 56.8% of attributed failures, while *knowledge gap* is only 4.4%. The mix is strongly format-dependent: single-select errors concentrate in binding (36.5% within type); multiple-select errors are led by option-set calibration (26.6%) together with binding and cascade; open-ended errors shift toward cascade (31.0%), binding (28.1%), and slot misbinding (21.9%), with causal-chain gaps appearing almost only on open items. Genetics and relation-structured displays amplify cascade and binding rather than factual misses. Figure 7C shows that model profiles are largely shared: binding and cascade dominate across systems. GPT-5.4 is the clearest outlier, with a larger option-set calibration share, consistent with its large exact-to-partial gap; weaker models (Grok/Mistral/GLM) show higher cascade and knowledge-gap mass. These attributions locate the principal weakness beyond biological recall: systems struggle to bind symbols to roles, preserve intermediate state after an early commitment, assemble complete option sets, and map recovered entities into the correct answer slots. Figure 6 shows one such case beside the official reference reasoning: after a correct first option, the model locks a Z-linked recessive hypothesis and over-selects the pedigree. The trace already employs the appropriate inheritance terminology. The error is an inference that the pedigree does not warrant; once retained, it renders the remaining options apparently correct. The example is representative of the benchmark: locally appropriate biological language can persist after the supporting evidence has been abandoned. Further walkthroughs are in Appendix D.

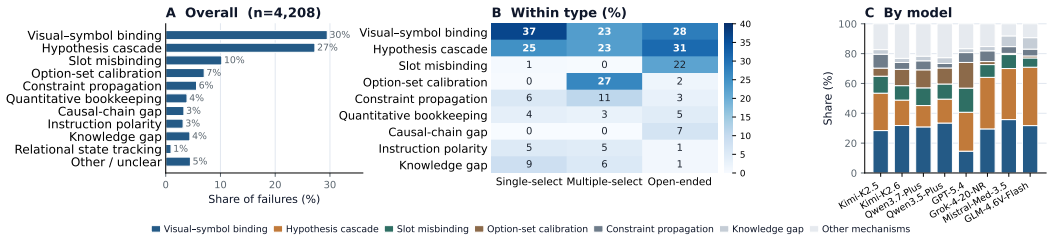


Figure 7: L2 mechanisms on 4,208 errors. (A) Overall share. (B) Within-type share for single-select (SS), multiple-select (MS), and open-ended items. (C) Composition by model. Binding is blue and cascade is amber in (C).

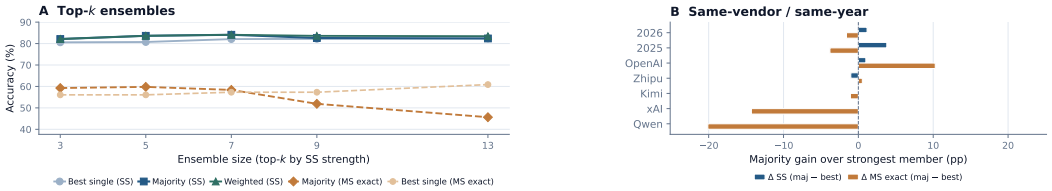


Figure 8: Ensemble gains over the strongest member.

4.8 Finding 5: Targeted Ensembles Provide a Practical but Limited Gain

On the choice items answered by all 13 models, an oracle reaches 98.5% single-select and 87.5% multi-select exact, so complementary correct answers exist (Figure 8). Unrestricted majority keeps the best single-select score but drops multi-select exact by 15.2 points. A Top-5 committee selected by single-select strength gains only 2.9 and 3.7 points on those two metrics. Correct answers are distributed across models, but the dominant errors are shared. Majority voting therefore tends to reproduce the same symbol-binding error or the same premature hypothesis. Aggregation does not compensate for a failure to preserve the evidence. Full committee breakdowns are in Appendix F.

4.9 Limitations

BioReasonBench is drawn from high-school biology exams rather than college or research materials. Nevertheless, this scope is already sufficient to establish the claim: on the same items, Grade-12 students complete the paper, whereas current MLLMs do not.

5 Conclusion

In this work, we introduced BioReasonBench, a diagnostic benchmark comprising 1,027 biology questions designed to evaluate whether MLLMs can perform structured reasoning over complex visual evidence. Through a fine-grained evaluation framework annotated across five diagnostic axes, our work distinguishes local visual perception from complete end-to-end reasoning and systematically pinpoints mechanism-level failure modes. Evaluation across 13 state-of-the-art MLLMs reveals a severe performance gap between visual description and complete scientific reasoning: while models achieve a mean accuracy of 61.2% on single-select items, their accuracy sharply drops to 28.3% on exact multiple-select questions and to a mere 11.6% on multi-blank open-ended tasks. This decline is particularly pronounced in genetics and relation-structured figures. Notably, even the top-performing model (71.0%) lags substantially behind Grade-12 students (88.1%) on the set0 subset. Error analysis demonstrates that these failures are driven by shared reasoning flaws, such as improper symbol-role assignment and stubborn adherence to early invalid hypotheses. Overall, BioReasonBench highlights that current MLLMs merely recover locally relevant information rather than reasoning through complex visual premises. We contend that future progress in multimodal scientific AI must be measured not by the fluency of visual description, but by the rigor and completeness of figure-based scientific reasoning.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools to implement the evaluation method and to interpret results: Kimi-K2.5 is the LLM-as-judge for open-ended blanks and the closed-set classifier that attributes L2 error mechanisms, as described in Section 3.5 and Appendices B and C. We have not used generative AI tools to generate synthetic datasets, develop theoretical models, formulate mathematical claims, or write proofs, and those tasks are not applicable to this benchmark study. The thirteen MLLMs in the experiments are systems under evaluation rather than authoring tools. Judge labels were checked against expert-graded student responses (Appendix E). We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

BioReasonBench is compiled from openly circulated Chinese gaokao biology papers and regional mock/practice papers. In the Chinese context, these materials are generally treated as public-domain educational resources for teaching and scientific analysis. We use them for non-commercial research only and exclude private information. The human comparison uses existing graded exam responses from ten Grade-12 students; no new private information is collected or released.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

Construction, taxonomy, and the exact / partial / LLM-judge protocol are described in Section 3 and Section 3.5. Model list, zero-shot setting, and the anonymous code release are in Section 4.1. Judge and error-analysis prompts are reproduced in Appendices B and C. Human agreement of the open-ended judge is reported in Appendix E, and full numeric tables are in Appendix F. Provider details, prompts, and code are released anonymously at `https://anonymous.4open.science/r/BioReasonBench-4F54`.

REFERENCES

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=d7KBjmI3GmQ`.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Jiayi Lei, Yao Fu, Maosong Sun, and Junxian He. C-Eval: A multi-level multi-discipline chinese evaluation suite for foundation models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 62991–63010, 2023. URL `https://proceedings.neurips.cc/paper_files/paper/2023/hash/c6ec1844bec96d6d32ae95ae694e23d8-Abstract-Datasets_and_Benchmarks.html`.

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

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2567–2577, 2019. doi: 10.18653/v1/D19-1259. URL `https://aclanthology.org/D19-1259/`.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European Conference on Computer Vision (ECCV)*, pp. 235–251, 2016. doi: 10.1007/978-3-319-46493-0_15.

Kimi Team. Kimi K2.5: Visual agentic intelligence, 2026a. URL https://arxiv.org/abs/2602.02276.

Kimi Team. Kimi K2.6: Advancing open-source coding, 2026b. URL https://www.kimi.com/blog/kimi-k2-6.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13299–13308, 2024. doi: 10.1109/CVPR52733.2024.01263. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_SEED-Bench_Benchmarking_Multimodal_Large_Language_Models_CVPR_2024_paper.html.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Pan Lu, Swaroop Mishra, Tony 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. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/11332b6b6cf4485b84afadb1352d3a9a-Abstract-Conference.html.

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. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=KUNzEQMWU7.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Mistral AI. Remote agents in vibe. powered by Mistral Medium 3.5, may 2026. URL https://mistral.ai/news/vibe-remote-agents-mistral-medium-3-5/.

OpenAI. GPT-4V(ision) system card. Technical report, OpenAI, 2023. URL https://openai.com/index/gpt-4v-system-card/.

OpenAI. Introducing GPT-4.1 in the API, apr 2025a. URL https://openai.com/index/gpt-4-1/.

OpenAI. Introducing GPT-5 for developers, aug 2025b. URL https://openai.com/index/introducing-gpt-5-for-developers/.

OpenAI. Introducing GPT-5.4, mar 2026. URL https://openai.com/index/introducing-gpt-5-4/.

Charles Sanders Peirce. Deduction, induction, and hypothesis. *Popular Science Monthly*, 13:470–482, 1878.

Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, feb 2026a. URL https://qwen.ai/blog?id=qwen3.5.

Qwen Team. Qwen3.7-Plus: Multimodal agent intelligence, may 2026b. URL https://qwen.ai/blog?id=qwen3.7-plus.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, Yannis Almirantis, John Pavlopoulos, Nicolas Baskiotis, Patrick Gallinari, Thierry Artiéres, Axel-Cyrille Ngonga Ngomo, Norman Heino, Eric Gaussier, Liliana Barrio-Alvers, Michael Schroeder, Ion Androutsopoulos, and Georgios Paliouras. An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015. doi: 10.1186/s12859-015-0564-6. URL https://doi.org/10.1186/s12859-015-0564-6.

Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky, and Jane B. Reece. *Campbell Biology*. Pearson, 11 edition, 2017.

V Team. GLM-4.5V and GLM-4.1V-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006. Also introduces the GLM-4.6V series including GLM-4.6V-Flash.

xAI. Grok 4, jul 2025a. URL https://x.ai/news/grok-4.

xAI. Grok 4 fast, sep 2025b. URL https://x.ai/news/grok-4-fast.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. 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 (CVPR)*, pp. 9556–9567, 2024. doi: 10.1109/CVPR52733.2024.00913. URL https://openaccess.thecvf.com/content/CVPR2024/html/Yue_MMMU_A_Massive_Multi-discipline_Multimodal_Understanding_and_Reasoning_Benchmark_for_CVPR_2024_paper.html.

Zhipu AI. GLM-4V-flash, 2025. URL https://docs.bigmodel.cn/cn/guide/models/free/glm-4v-flash.