Mobus: Data Infrastructure for Researchers and Autonomous Scientific Agents

Harry Ilanyan [1] **Hrant Virabyan** [1]

Abstract

Autonomous research agents increasingly retrieve and synthesize scientific data, but current systems often drop the metadata needed for auditable downstream use: provenance, license status, schema structure, and transformation history. Mobus is data infrastructure that carries source, license, and schema as first-class fields across iterative discovery, bounded refinement, and lineage-aware manipulation, producing analysis-ready datasets where every cell has traceable, license-checked lineage. On a 104-question benchmark spanning eight scientific domains and four comparison systems, Mobus reaches 95% task completion against 62% for the strongest agentic baseline (GPT-Researcher), drives the license false-clean rate down to 4% from the 41 to 58% of retrieval-equipped baselines, hits 0.99 citation integrity, and passes all 60 manipulation-tool unit tests.

1. Introduction

Today's research agents synthesize answers without keeping the substrate of the science itself. Licenses are rarely surfaced, citations are often hallucinated (Liu et al., 2023), schemas disappear, and transformations stay opaque, so outputs are fine for orientation and indefensible downstream: a second scientist cannot tell which data was used, under what license, transformed how. The closest agentic analog, GPT-Researcher (Elovic, 2023), iterates web retrieval and report writing but discards license and schema metadata; closed deep-research products (Google, 2024; OpenAI, 2025) do the same.

Mobus is data infrastructure for end-to-end scientific data acquisition, refinement, and analysis-ready assembly. It has three pieces: a multi-source substrate accessed through the Model Context Protocol (Anthropic, 2024), a rubric-scored discovery-and-refinement loop with bounded refinement actions, and a lineage-aware manipulation layer where every transformation writes a manifest linking input to output cells. None of the individual constraints is new (Section 2); the contribution is the enforcement point. Mobus makes them machine-checked, load-bearing invariants inside an autonomous acquisition loop, where failing checks are refused loudly rather than coerced silently, so source, license, and schema integrity hold at every step rather than being reconstructed post hoc.

Mobus covers the data-acquisition stage of the AI Scientist workflow, with human oversight at license uncertainty, sufficiency-threshold failures, and out-of-distribution transformations. We position it as the data substrate for AI Scientists, not a replacement for modeling, experimentation, or writing.

2. Related Work

Datasheets (Gebru et al., 2021), data statements (Bender & Friedman, 2018), and the FAIR principles (Wilkinson et al., 2016) define the metadata that should travel with data; audits show it rarely does, with license omission above 70% and license error rates above 50% on popular hosting platforms (Longpre et al., 2023). Those are documentation standards and post-hoc audits; Mobus performs the same checks at acquisition time. On the agent side, retrieval-augmented generation (Lewis et al., 2020), GPT-Researcher (Elovic, 2023), and closed deep-research products (Google, 2024; OpenAI, 2025) retrieve and synthesize without reporting per-source license status, while agentic-science systems (Lu et al., 2024; Skarlinski et al., 2024) target hypothesis generation and literature synthesis, not license- and lineage-preserving data acquisition. Benchmarks such as GAIA (Mialon et al., 2024) and BrowseComp (Wei et al., 2025) score answer-finding, not license, provenance, or schema fidelity; we know of none that does, so we build one (Section 4).

[1]University of California, Berkeley, CA, USA. Correspondence to: Harry Ilanyan <harry_ila@berkeley.edu>, Hrant Virabyan <hrant@berkeley.edu>.

Accepted at ICML 2026 AI for Science Workshop. Copyright 2026 by the author(s).

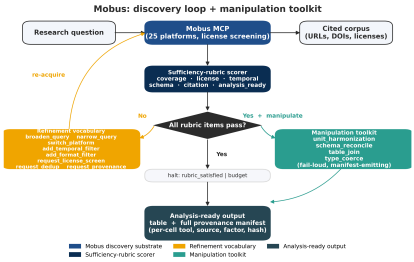

Figure 1. The Mobus pipeline. Source provenance, license, and schema metadata are attached at retrieval and preserved through every refinement and manipulation step.

3. System

Substrate. Mobus exposes 15 tools across 25 public scientific data sources, covering ML repositories, scientific archives, preprint and citation graphs, open-data portals, and regulatory filings. Each retrieved record carries source provenance, license status, and attribution metadata.

Loop. Given a research question, Mobus drafts a structured data hypothesis, retrieves candidates in parallel, and scores them against a multi-dimensional sufficiency rubric: coverage of the question, license cleanliness, source diversity, schema match, sample adequacy, freshness, and lineage completeness. When the rubric falls short, Mobus picks from a bounded refinement vocabulary (*broaden*, *narrow*, *swap source class*, *re-rank*, *decompose*, *escalate*) and re-acquires; stopping rules are in Appendix B.

Manipulation. Three tools so far: schema reconciliation, unit harmonization, and key-matched joins. Each takes typed input, writes a transformation manifest (input cells, operation, parameters, source pointer), and refuses to silently coerce when the input is out of distribution. The output is a single analysis-ready table whose every cell traces back to a specific licensed source through the manifest.

4. Evaluation

Benchmark. 104 questions across eight domains: biology, materials science, climate science, physics, public-health epidemiology, pure mathematics, advanced mathematics, and computer science. Difficulty spans multi-source synthesis, long-tail or specialized retrieval, and adversarial cases (license traps, ambiguous entities, contradictory sources; stratification in Appendix I). Six questions require manipulation to produce an analysis-ready output. Each question carries an oracle source set, expected license labels, an expected schema, and a reference answer. To our knowledge, no agent benchmark scores license or provenance fidelity (Section 2), so the benchmark is self-built.

Systems. Five end-to-end configurations: Sonnet 4.5 with no tools (the no-retrieval floor), Sonnet 4.5 with native web search, GPT-Researcher (Elovic, 2023), a discovery-only Mobus (substrate without loop or manipulation, isolating their contribution), and the full Mobus. GPT-Researcher is fully instrumentable: open source, with per-source license claims we can sample into hand-verification. Closed deep-research products emit no per-source license labels, so LFCR cannot be computed for them.

Methodology. GPT-5.5 scores per-question outputs against a verbatim rubric for completion, citation integrity, and license labeling (verbatim prompt in Appendix C). To check the LLM-as-judge protocol (Liu et al., 2023), the authors hand-rated 100 paired (human, judge) outputs across all five configurations and got Cohen's $\kappa = 0.715$ (Cohen, 1960), in the substantial-agreement band of Landis & Koch (1977). License labels were hand-verified upstream for a pre-specified stratified sample targeting 400 results (counts in Appendix D). Citation integrity is judge-assessed source support, not independent verification of each source; the false-clean rate is computed separately against hand-verified upstream ground truth, not via the judge.

Headline results. Mobus reaches 95% task completion against 62% for GPT-Researcher, the strongest agentic baseline. The largest gap is on license safety. The false-clean rate, results the system labels permissive that are in fact restricted, is 4% for Mobus, 41% for GPT-Researcher, and 58% for Sonnet with web search (Figure 2, Appendix D). The reason is architectural: Mobus is the only evaluated system that pulls and validates license metadata at acquisition rather than guessing afterward. Citation integrity reaches 0.99. The manipulation toolkit passes all 60 unit tests, including 11 adversarial out-of-distribution inputs where it fails loudly rather than coercing silently; manifest completeness is 100%. Discovery-only Mobus already beats every retrieval-equipped baseline on license safety, so most of the safety gain comes from the substrate; the loop and manipulation layers add task completion and analysis-ready assembly. The full per-system, per-metric grid is in the appendix.

5. Limitations

The benchmark is self-constructed; oracle source sets and hand-verified license ground truth constrain the grading, but that bias remains. Sample sizes are modest (104 questions, 100 paired human–judge ratings, 370 verified license records; per-system agreement slices are n=20). Spot-checks were author-rated under blinded labels, not by independent domain experts, and the judge is a single LLM. The manipulation layer is narrow: three tools, six manipulation questions, and a 60-test suite that establishes fail-loud behavior on known cases, not robustness under distribution shift. Closed-product audits and a backbone-scale ablation are future work. Mobus is not a full AI Scientist: it is the acquisition, verification, and analysis-ready assembly layer for publicly indexed scientific data.

References

Anthropic. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol, 2024.

Bender, E. M. and Friedman, B. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Elovic, A. GPT-Researcher: An autonomous agent for comprehensive online and local research. https://github.com/assafelovic/gpt-researcher, 2023. Open-source repository.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723.

Google. Try Deep Research and our new experimental model in Gemini, your AI assistant. https://blog.google/products/gemini/google-gemini-deep-research/, 2024.

Landis, J. R. and Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153.

Longpre, S., Mahari, R., Chen, A., Obeng-Marnu, N., Sileo, D., Brannon, W., Muennighoff, N., Khazam, N., Kabbara, J., Perisetla, K., et al. The Data Provenance Initiative: A large scale audit of dataset licensing & attribution in AI. *arXiv preprint arXiv:2310.16787*, 2023.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., and Scialom, T. GAIA: A benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

OpenAI. Introducing deep research. https://openai.com/index/introducing-deep-research/, 2025.

Skarlinski, M. D., Cox, S., Laurent, J. M., Braza, J. D., Hinks, M. M., Hammerling, M. J., Ponnapati, M., Rodriques, S. G., and White, A. D. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Wei, J., Sun, Z., Papay, S., McKinney, S., Han, J., Fulford, I., Chung, H. W., Passos, A. T., Fedus, W., and Glaese, A. BrowseComp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*, 2025.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., et al. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018, 2016. doi: 10.1038/sdata.2016.18.

A. Source Inventory

Mobus exposes 15 tools across 25 public scientific data sources, grouped into six classes (Table 1).

Table 1. Mobus source inventory by class. Counts are sources, not tools; several sources are accessed through more than one tool.

Class	Sources
ML repositories	Hugging Face (datasets and models), Kaggle, OpenML, UCI, OpenReview
Scientific archives	Zenodo, figshare, Harvard Dataverse, DataCite
Preprint and citation graphs	arXiv, OpenAlex, Semantic Scholar, Europe PMC, Crossref
Open-data portals	data.gov, U.S. Census, Eurostat, World Bank, WHO, NASA, Socrata
Regulatory filings	SEC EDGAR
Code and cloud catalogs	GitHub, AWS Open Data, Google Dataset Search

B. Sufficiency Rubric and Refinement Vocabulary

B.1. Sufficiency rubric

Each retrieved candidate set is scored on seven dimensions, each on a 0–2 scale:

- **Coverage of question.** Does the set address every sub-question? 0: gaps. 1: partial. 2: complete.
- **License cleanliness.** Fraction of results with a known, permissive license. 0: under 50%. 1: 50–80%. 2: 80%+.
- **Source diversity.** Entropy over source distribution. 0: single-source. 1: two or three. 2: four or more distinct sources.
- **Schema match.** Do retrieved schemas match the schema required by the hypothesis? 0: mismatch. 1: partial. 2: full.
- **Sample adequacy.** Where applicable, does the result set have enough samples for the intended use? 0: clearly under. 1: borderline. 2: adequate.
- **Freshness.** Where applicable, are results recent enough? 0: stale. 1: borderline. 2: current.
- **Lineage completeness.** Per-result source provenance, license, attribution. 0: missing. 1: partial. 2: full.

The loop terminates when every dimension scores at least 1, the average crosses 1.6, two consecutive iterations gain less than 5% on the average score, or the budget runs out. Per-dimension thresholds and weights are configurable.

Table 2. Lowest-scoring dimension to preferred refinement action.

Lowest-scoring dimension	Preferred actions
Coverage	decompose, broaden
License	swap source class, re-rank
Source diversity	swap source class, broaden
Schema match	narrow, re-rank
Sample adequacy	broaden, swap source class
Freshness	re-rank, narrow
Lineage	swap source class, escalate

B.2. Refinement vocabulary

When the rubric falls short, the loop selects from six bounded actions:

- *broaden*: relax the most restrictive constraint in the data hypothesis.
- *narrow*: add a constraint to filter low-precision results.
- *swap source class*: re-target acquisition at a different class of sources (for example, archives instead of repositories).
- *re-rank*: re-rank the existing candidate pool by an alternate criterion (recency, citation count, license cleanliness).
- *decompose*: split the question into sub-questions and run acquisition for each.
- *escalate*: surface the partial result set and the rubric gap to a human-in-the-loop checkpoint.

Action selection is rubric-driven: the dimension with the lowest score maps to a small set of preferred actions (Table 2).

C. Judge Methodology

C.1. Judge model and parameters

Per-question outputs were scored by GPT-5.5 with temperature 0, max-output-tokens 800, top-p 1.0, response format `json_object`, and a fixed system prompt. Three calls per (question, system) cell: one for completion, one for citation integrity, one for license labeling.

C.2. Verbatim system prompt

```
You are an evaluation judge for
a dataset-discovery benchmark.
You will be shown one (question,
system_response) pair at a time.
Your job is to assign a single
integer score from {0, 1, 2} for the
metric named in the user message,
with a one-sentence rationale.  You
must respond with a single JSON
```

```
object of shape {"score": <0|1|2>,
"rationale": "<one sentence>"}.

Hard rules: (1) Score the cell
on its own merits. Do not reason
about other systems' responses. Do
not assume the system has access
to tools you have not been told
about. Do not penalize legitimate
refusals. (2) Use the rubric
for the named metric exactly;
the boundaries between scores
are anchored; do not interpolate.
(3) If the response cites sources,
treat the cited URLs/DOIs as claims
the system makes about the world;
you do not need to verify them
yourself, but if the response itself
acknowledges that a citation is
unverified or speculative, that fact
is part of the response. (4) If
the response contains explicit
license claims, those are claims the
system has made; the license_quality
metric is about whether those
claims are internally consistent
and well-supported by the cited
platform's first-party metadata
convention, NOT about checking
upstream truth. (5) The rationale
must be one sentence and must name
the rubric anchor that determined
the score (e.g., "Anchor 1: oracle
recall < 30%").

Refuse to score, and return
{"score": -1, "rationale":
"<reason>"}, only if the response is
malformed (truncated, non-text, in
another language than the question).
Do not refuse for low quality.
```

User-prompt templates per metric carry full anchor language for the 0, 1, and 2 levels.

C.3. Scoring dimensions

The judge scores each output on three axes, each on a 0/1/2 scale:

- **Task completion (TCR).** 0: unusable. 1: partial. 2: usable as a starting point for further research. Binary pass is score 2.
- **Citation integrity (CIR).** 0: hallucinated or wrong. 1: real but loosely supports the claim. 2: real and accurately supports the claim. Reported as fraction at score 2.
- **License labeling.** 0: missing or wrong. 1: partial. 2: correct labels for every cited source. Used to compute false-clean rate against the hand-verified ground truth in Section D.

Table 3. Agreement between GPT-5.5 judge and human raters on the 100 paired ratings, overall and per system. Per-system rows have $n = 20$ and should be read as indicative; the overall binary $\kappa = 0.715$ is the load-bearing number.

Slice	κ
Overall (binary pass/fail)	0.715
Overall (3-level, unweighted)	0.753
Overall (3-level, quadratic-weighted)	0.837
Sonnet (no tools)	0.500
Sonnet + web search	0.600
GPT-Researcher	0.886
Mobus discovery-only	0.524
Mobus full	0.615

C.4. Human-judge agreement

To validate the judge protocol, the authors hand-rated 100 paired (human, judge) outputs across all five configurations. Cohen's $\kappa = 0.715$ overall, which falls in the substantial-agreement band of Landis & Koch (1977). Overall and per-system agreement are in Table 3. Disagreements were dominated by partial-credit cases (judge giving 1 where the human gave 2 or vice versa).

Exact 0/1/2 agreement was 84.0% and binary pass-fail agreement was 86.0% across the 100 paired ratings.

D. License-Verification Sampling Protocol

The license false-clean rate (LFCR) is the headline safety metric. It measures the rate at which a system labels a result as having a permissive license (CC-BY, CC-0, MIT, Apache, ODbL, public domain) when the upstream license is in fact restricted, unknown, or incompatible.

D.1. Sampling

The sampling plan was pre-specified before any license verification began. From each of 20 questions drawn at stratified random across the eight domains, we sampled up to five results per system that the system labeled permissive. The full target was 100 verifications per system, 400 total across the four license-labeling systems (Mobus full, Mobus discovery-only, GPT-Researcher, Sonnet + web search). Sonnet without tools is excluded from the LFCR table because it does not return citable sources.

D.2. Verification procedure

For each sampled result, the verifier resolved the URL or DOI, located the upstream license statement on the publishing platform, and recorded the actual license string and source. If the result did not resolve (404, link rot, paywall), it was logged as *unresolved* and not counted toward LFCR. Discrepancies between the system-claimed license and the upstream license were flagged *false-clean* when the

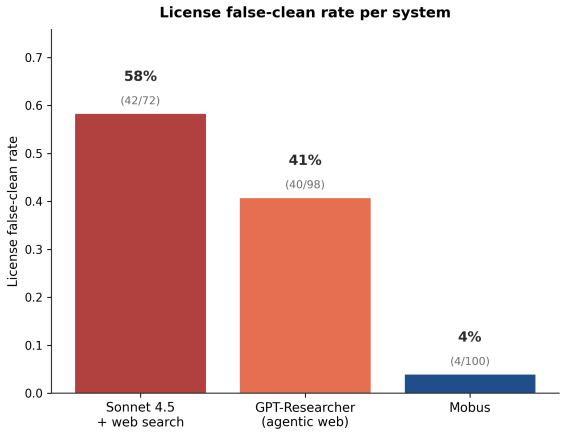

Figure 2. License false-clean rate across systems on a pre-specified hand-verified sample (per-system denominators in Table 4; lower is better). Mobus checks license metadata at acquisition; baselines do not.

Table 4. License false-clean rate detail. "Labeled clean" is the number of system-claimed permissive entries that fell into the hand-verification sample for that system. LFCR is the share of those entries whose upstream license is in fact non-permissive.

System	Labeled clean	False-clean	LFCR
Sonnet + web search	72	42	58%
GPT-Researcher	98	40	41%
Mobus discovery-only	100	6	6%
Mobus full	100	4	4%

upstream license is in fact non-permissive.

D.3. Computed LFCR

$$\mathrm{LFCR} = \frac{\#\{\text{labeled clean} \wedge \text{actually restricted}\}}{\#\{\text{labeled clean}\}}.$$

Per-system results are in Table 4.

E. Human Spot-Check Protocol

We hand-rated 100 paired (human, judge) outputs to validate the LLM-as-judge methodology. Twenty questions were drawn at stratified random across the eight domains, and for each, the outputs of all five configurations were rated against the same rubric used by the judge, yielding 100 paired ratings. Raters were the authors. Each rater used a written instruction sheet and was blind to which configuration produced each output. Discrepancies between human and judge were not adjudicated; they were used to compute Cohen's κ at the binary and 3-level granularities reported in Section C.

F. Manipulation Toolkit

F.1. Tools

Three tools at submission. Each takes typed inputs, writes a transformation manifest, and refuses to silently coerce on out-of-distribution input.

- **Schema reconciliation.** Aligns column names, types, and units across heterogeneous source schemas. Fails loudly on column collisions and type mismatches that the user has not explicitly resolved.
- **Unit harmonization.** Converts numerical values across unit systems with a known conversion (temperature, mass, length, pressure, time, energy, currency-with-date). Fails loudly on unrecognized units.
- **Key-matched joins.** Performs inner, left, right, and outer joins on declared keys with optional fuzzy matching (Levenshtein, soundex) on a per-column basis. Fails loudly when a fuzzy threshold is crossed without explicit user opt-in.

F.2. Manifest schema

Every operation appends an entry to a transformation manifest:

- `operation`: tool name plus operation identifier.
- `input_cells`: list of (source_id, row_id, column) tuples.
- `parameters`: the exact parameters passed to the tool.
- `output_cells`: list of (output_id, row_id, column) tuples.
- `provenance_pointer`: pointer back to the upstream license-verified source for each input cell.
- `timestamp` and `tool_version`.

The manifest is the artifact that lets a downstream reader reconstruct every cell of an analysis-ready output back to a specific licensed source record.

F.3. Test suite

The toolkit ships with 60 unit tests: 49 in-distribution conversions, joins, and reconciliations with known-correct outputs, and 11 adversarial out-of-distribution inputs (unrecognized units, ambiguous keys, type collisions). Adversarial cases are passing if and only if the tool fails loudly with an informative error rather than producing a silent coercion. All 60 tests pass at submission. Manifest completeness is 100% across the test set. Per-tool breakdown in Table 5.

Table 5. Manipulation-toolkit unit-test breakdown.

Tool	In-dist.	Adv. fail-loud	Manifest
Unit harmonization	31/31	5/5	100%
Schema reconciliation	11/11	3/3	100%
Key-matched joins	7/7	3/3	100%
Total	**49/49**	**11/11**	**100%**

G. Per-System Per-Metric Grid

Table 6 reports every metric across every configuration on the 104-question benchmark. Cells marked `n/a` are not applicable because the configuration lacks the relevant capability (manipulation correctness for systems without a manipulation layer).

Table 6. Full per-system, per-metric grid on the 104-question benchmark. TCR: task completion rate (binary pass at judge score 2; validated against human raters at $\kappa = 0.715$). LFCR: license false-clean rate from the pre-specified hand-verified sample (Appendix D). CIR: citation integrity rate (fraction at judge score 2). Manip. tests: manipulation-toolkit unit tests passed (Appendix F); only applicable to configurations with a manipulation layer.

System	TCR	LFCR	CIR	Manip. tests
Sonnet 4.5 (no tools)	11%	n/a	0.37	n/a
Sonnet 4.5 + web search	50%	58%	0.68	n/a
GPT-Researcher	62%	41%	0.74	n/a
Mobus discovery-only	77%	6%	0.99	n/a
Mobus full	**95%**	**4%**	**0.99**	**60/60**

H. Loop Pseudocode

Algorithm 1 gives the full discovery-and-refinement loop; halting conditions match Appendix B.

Algorithm 1 Mobus discovery-and-refinement loop.

Input: research question q, budget B (iterations, tokens, wall-clock).
Output: analysis-ready dataset D with transformation manifest M, or escalation.
$h \leftarrow$ HYPOTHESIZEDATA(q)
$R \leftarrow \emptyset$, $i \leftarrow 0$, $prev \leftarrow -\infty$
while $i < B$ **do**
  $C \leftarrow$ ACQUIRE(h) parallel across sources
  $R \leftarrow R \cup C$
  $s \leftarrow$ SCORERUBRIC(R, h)
  if s satisfies all dimensions **then**
    break
  else if $s - prev < 0.05$ **for two iterations then**
    break
  end if
  $a \leftarrow$ SELECTACTION(s)
  $h \leftarrow$ APPLYACTION(h, a)
  $prev \leftarrow s$, $i \leftarrow i + 1$
end while
if question requires manipulation **then**
  $(D, M) \leftarrow$ MANIPULATE(R, h)
else
  $(D, M) \leftarrow (R, \emptyset)$
end if
return (D, M)

I. Benchmark Stratification

Table 7 reports the number of questions per domain and per difficulty tier.

Table 7. Question count by domain and difficulty tier. Tier 2: multi-source synthesis. Tier 3: long-tail or specialized. Tier 4: adversarial cases including license traps. The six manipulation questions that require analysis-ready assembly are distributed across tiers according to source complexity, not concentrated in any single tier. (No tier 1 questions in the v2 benchmark; all questions require at least multi-source synthesis.)

Domain	T2	T3	T4	Total
Biology	10	5	1	16
Materials science	10	4	0	14
Climate science	10	4	2	16
Physics	4	10	0	14
Public health	13	2	1	16
Pure mathematics	0	7	0	7
Advanced mathematics	2	5	0	7
Computer science	9	5	0	14
Total	**58**	**42**	**4**	**104**