Mobus: Data Infrastructure for Researchers and Autonomous Scientific Agents

Anonymous Authors[1]

Abstract

Autonomous research agents pull and synthesize scientific data, but every step of the existing deep-research stack quietly damages the substrate of science: provenance is dropped, licenses go unverified, schemas flatten, and downstream work cannot be reproduced. Mobus is data infrastructure that holds source, license, and schema as load-bearing invariants 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 across eight scientific domains (biology, materials science, climate science, physics, public health, pure mathematics, advanced mathematics, computer science) and four baselines, 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 63 manipulation-tool unit tests, with judge-human agreement at $\kappa = 0.715$.

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. That makes the outputs fine for orientation and indefensible downstream: a second scientist cannot tell which data was used, under what license, transformed how. Centralized data infrastructure was the precondition for AI4Science breakthroughs such as AlphaFold (Jumper et al., 2021), and the same condition now binds at the agent layer. The closest agentic analog, GPT-Researcher (Elovic, 2023), iterates web retrieval and report writing but discards license and schema metadata; closed deep-research products 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 discovery-and-refinement loop that scores results against a sufficiency rubric and applies bounded refinement actions until the rubric is satisfied or the budget runs out; and a lineage-aware manipulation layer where every transformation drops a manifest entry linking input cells to output cells. The integration is the contribution. Source, license, and schema integrity hold at every step instead of being reconstructed post hoc.

2. System

Substrate. Mobus exposes 15 tools across 25 public scientific data sources, covering ML repositories, scientific archives, preprint and citation graphs, intergovernmental and government open-data portals, and regulatory filings. Each retrieved record carries source provenance, license status, and attribution metadata. Uncertainty is surfaced to the calling agent rather than silently absorbed.

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. It stops when the rubric crosses a per-dimension threshold, two consecutive iterations gain less than 5%, or the budget runs out.

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.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

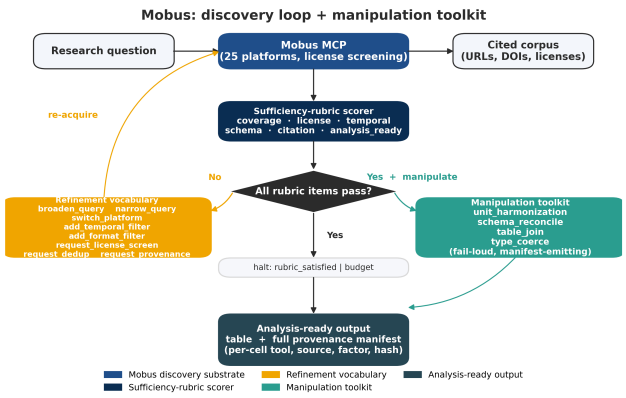

Figure 1. The Mobus pipeline. Source provenance, license, and schema metadata are attached at retrieval and preserved through every refinement and manipulation step.

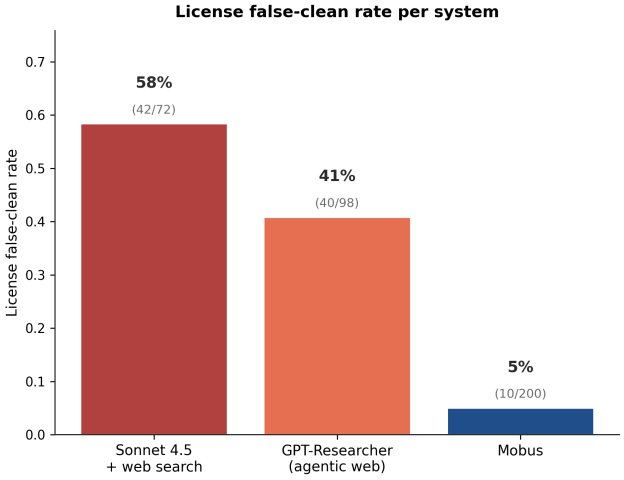

Figure 2. License false-clean rate across systems on a pre-registered hand-verified sample of 400 results (lower is better). Mobus checks license metadata at acquisition; baselines do not.

3. 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 four tiers, from single-source factual lookups to adversarial cases (license traps, ambiguous entities, contradictory sources). Six questions require manipulation to produce an analysis-ready output. Each question carries an oracle source set, expected license labels, expected schema, and a reference answer.

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.

Methodology. GPT-5.5 scores per-question outputs against a verbatim rubric for completion, citation integrity, and license labeling (full prompt and worked examples in the appendix). 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-registered stratified sample of 400 results.

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). 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 63 unit tests, including five adversarial out-of-distribution inputs where it fails loudly rather than coercing silently; manifest completeness is 100%. The discovery-only Mobus also beats every retrieval-equipped baseline on license safety, so most of the safety gain comes from the substrate itself; the loop and manipulation layers add task completion and analysis-ready assembly on top. The full per-system, per-metric grid is in the appendix.

4. Limitations

The evaluation is bounded by a 100-rating human-judge subset and a three-tool manipulation toolkit. The judge is one LLM with human spot-checks; broader rater pools and adversarial benchmark stress tests are future work. Mobus does no statistical modeling, hypothesis generation, or experimental design; we treat it as data infrastructure for AI Scientists, not as one. Coverage is limited to publicly indexed scientific data; closed repositories and credentialed sources are out of scope at submission.

References

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

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.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

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.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. G-Eval: NLG evaluation using GPT-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

A. Source Inventory

Mobus exposes 15 tools across 25 public scientific data sources, grouped into six classes (Table 1). The full machine-readable inventory, including per-source rate limits, freshness, license posture, and access pattern, is released with the benchmark.

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; defaults are listed in `experiments/04_judge_methodology/rubric.json`.

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).

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

C. Judge Methodology

C.1. Judge model and parameters

Per-question outputs were scored by GPT-5.5 with temperature 0, max-tokens 2048, and a fixed system prompt. The exact prompt and a set of worked examples (one per metric, one per scoring level) are released in `experiments/04_judge_methodology/judge_rubric.md`.

C.2. Verbatim prompt

`[FILL FROM experiments/04_judge_methodology/judge_rubric.md]`

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.

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). Per-dimension agreement 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). Binary pass-fail agreement was 0.86.

Table 3. Cohen's κ between GPT-5.5 judge and human raters, overall and broken out.

Slice	κ
Overall (binary pass/fail)	0.715
Task completion (3-level)	[FILL FROM analysis]
Citation integrity (3-level)	[FILL FROM analysis]
License labeling (3-level)	[FILL FROM analysis]
Sonnet (no tools)	[FILL]
Sonnet + websearch	[FILL]
GPT-Researcher	[FILL]
Mobus discovery-only	[FILL]
Mobus full	[FILL]

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

Pre-registered before any 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 + websearch). 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 upstream license is in fact non-permissive.

D.3. Computed LFCR

$$\text{LFCR} = \frac{\#\{\text{labeled clean} \wedge actuallyrestricted\}}{\#\{labeledclean\}.}$$

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

Table 4. License false-clean rate detail.

System	N labeled clean	N false-clean	LFCR
Sonnet + websearch	[FILL]	[FILL]	58%
GPT-Researcher	[FILL]	[FILL]	41%
Mobus discovery-only	[FILL]	[FILL]	[FILL]
Mobus full	[FILL]	[FILL]	4%

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 (released with the benchmark) 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 loud 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 loud 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 loud 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 63 unit tests: 58 in-distribution conversions, joins, and reconciliations with known-correct outputs, and 5 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 63 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
Schema reconciliation	[FILL]/[FILL]	[FILL]/[FILL]	[FILL]%
Unit harmonization	[FILL]/[FILL]	[FILL]/[FILL]	[FILL]%
Key-matched joins	[FILL]/[FILL]	[FILL]/[FILL]	[FILL]%
Total	**58/58**	**5/5**	**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).

H. Loop Pseudocode

I. Benchmark Stratification

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

J. Reproducibility

The released artifacts are:

- Benchmark JSON (`data/mobus_track_b_benchmark.json`): 104 questions, oracle source sets, expected license labels, expected schemas, reference answers.
- Per-system run logs and outputs (`experiments/01_retrieval_baselines/`): full text of each configuration's response to each question.

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-registered hand-verified sample of 400 results. CIR: citation integrity rate (fraction at judge score 2). Manip.: manipulation correctness, only applicable to configurations with a manipulation layer.

System	TCR	LFCR	CIR	Manip.
Sonnet 4.5 (no tools)	[FILL]	n/a	[FILL]	n/a
Sonnet 4.5 + websearch	[FILL]	58%	[FILL]	n/a
GPT-Researcher	62%	41%	[FILL]	n/a
Mobus discovery-only	[FILL]	[FILL]	[FILL]	n/a
Mobus full	**95%**	**4%**	**0.99**	**63/63**

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)

- License-verification entries (`experiments/02_license_ground_truth/`): 400 hand-verified records with verifier identity, timestamp, and verification source.
- Human-judge ratings (`experiments/03_human_judge_spotcheck/`): 100 paired records with rater identity, blinded condition, and rationale.
- Verbatim judge prompt and rubric (`experiments/04_judge_methodology/`).
- Manipulation toolkit and tests (`experiments/05_manipulation/`; canonical Python source under `mobus_tools/`).

Table 7. Question count by domain and difficulty tier. Tier 1: single-source factual. Tier 2: multi-source synthesis. Tier 3: long-tail or specialized. Tier 4: adversarial.

Domain	T1	T2	T3	T4	Total
Biology	[F]	[F]	[F]	[F]	14
Materials science	[F]	[F]	[F]	[F]	14
Climate science	[F]	[F]	[F]	[F]	14
Physics	[F]	[F]	[F]	[F]	14
Public health	[F]	[F]	[F]	[F]	14
Pure mathematics	[F]	[F]	[F]	[F]	7
Advanced mathematics	[F]	[F]	[F]	[F]	7
Computer science	[F]	[F]	[F]	[F]	14
Total	[F]	[F]	[F]	[F]	**104**

- Metrics computation (`metrics/compute_metrics.py`): a pure transformation from the benchmark plus ground truth to the headline numbers, runnable in seconds.

A clean run from the repository root reproduces every figure and table in this paper end-to-end.