OMNIMASTER: GLOBAL RELATIONS AS THE UNIT OF VERIFICATION IN DEEP RESEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Research increasingly relies on iterative search, evidence verification, and cross-source synthesis. Yet locally correct evidence does not guarantee a valid research result, since relations across evidence, sources, and outputs may still fail. We identify a failure mode, Relational Verification Collapse (RVC), in which relational requirements are weakened during decomposition, judgment, or aggregation, allowing local checks to pass while the underlying relation remains invalid. To study this problem, we introduce OmniMaster-Bench, which evaluates seven global relation types in two complementary settings, long-horizon open-web research and controlled multi-hop retrieval. Evaluations of frontier models reveal substantial remaining headroom, while strong performance on existing Deep Research benchmarks does not reliably transfer to global relational capability. We further develop relation-centered data synthesis and agent training. Relation-aware mid-training (MT) learns atomic capabilities for tracking relational states, while relation-preserving supervised fine-tuning (SFT) composes these capabilities into complete research trajectories. Without reinforcement learning, the resulting agents enter the frontier range on OmniMaster-Closed, narrow the gap on OmniMaster-Open, and remain competitive on established Deep Research benchmarks, showing that global relations can be explicitly learned through relation-centered supervision.

1 INTRODUCTION

Deep Research extends traditional retrieval into an iterative process of searching, browsing, reasoning, and synthesizing information across multiple sources (Gou et al., 2026; Zheng et al., 2025; Du et al., 2026). Existing work has improved this process through better search planning, stronger verification, and long-horizon policy learning (Li et al., 2025; Jin et al., 2025; Li et al., 2026b; Zheng et al., 2025), enabling agents to acquire and use external information more effectively. As retrieval becomes increasingly capable, verification has become central to complex research tasks (Gou et al., 2026; Du et al., 2026; Coelho et al., 2025). Yet a valid research result requires more than individually correct pieces of information. Evidence from different sources must also satisfy global relations, such as temporal alignment, consistent measurement bases, and source attribution (Krishna et al., 2025; Gou et al., 2026). When these relations fail, locally correct evidence can still lead to an invalid overall conclusion.

Existing Deep Research verification already captures parts of these relational requirements through task decomposition, evidence-supported checking, and iterative verification (Rosset et al., 2025; Gou et al., 2026; Du et al., 2026). However, specifying a relation is not equivalent to verifying it. To obtain executable supervision or evaluation signals, complex requirements are often decomposed into local criteria, judged separately by rules or LLM-based evaluators, and then aggregated into a final score (Sun et al., 2025). During this process, relational structure can be weakened during decomposition, imperfectly enforced during judgment, or obscured during aggregation. As a result, multiple local checks may pass even when the underlying joint relation remains invalid. We refer to this failure as Relational Verification Collapse (RVC), illustrated in Figure 1. This gap suggests that global relations should be treated as explicit objects of both evaluation and learning, rather than as properties that emerge indirectly from local checks.

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.

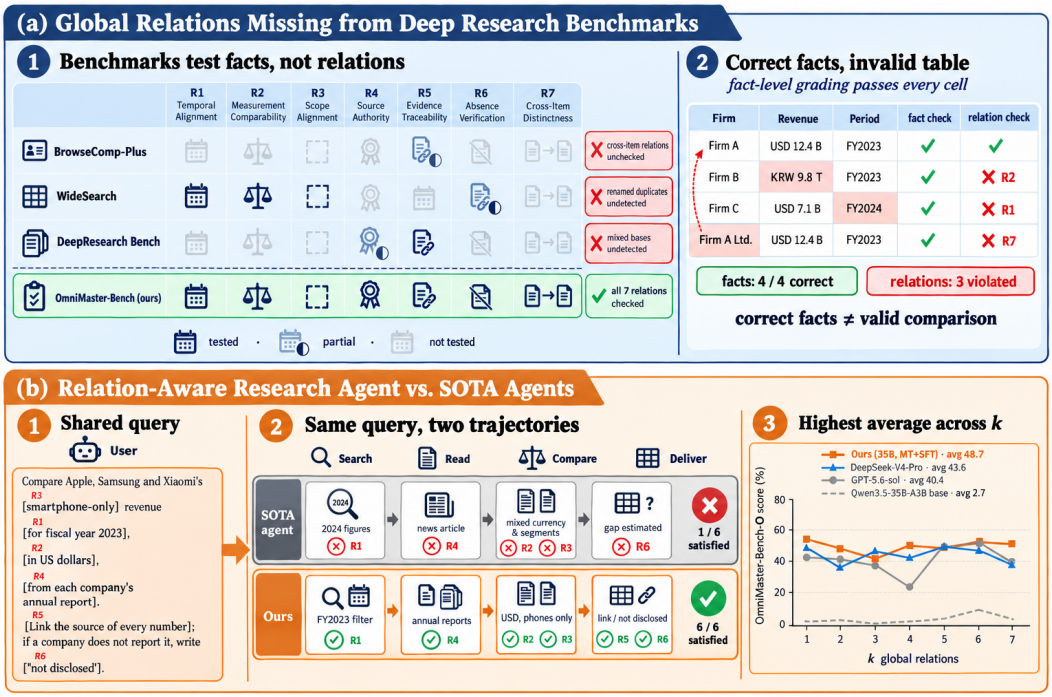

Figure 1: (a) Coverage of the seven global relations by existing Deep Research benchmarks and OmniMaster-Bench (partial: checked only indirectly or on a small share of tasks), and a table whose facts are all correct while three relations are violated. (b) The same request handled by a frontier agent and by our relation-aware agent, and scores as a function of the number of global relations k. The right panel of (b) plots the $\text{SFT}_{5.1k}$ checkpoint of Table 3 against the two frontier models with complete step-level traces; all six frontier models appear in Figure C.1.

Motivated by this gap, we introduce OmniMaster-Bench to make global relations explicit objects of evaluation. The benchmark systematically controls the types, numbers, and compositions of global relations, and contains two complementary settings. OmniMaster-Open evaluates relation preservation during long-horizon open-web research, while OmniMaster-Closed isolates relational reasoning in controlled multi-hop retrieval. To ensure that each relation is genuinely supported, we perform iterative evidence retrieval, evidence support, and failure-driven repair, so that its underlying conditions are grounded in retrievable evidence. We then jointly compose and validate multiple relations at the task level, ensuring that they form genuine cross-source dependencies and relational interactions rather than a collection of independent conditions. And to prevent RVC from reappearing during benchmark construction, we preserve the same relational structure throughout task construction, evidence support, and final evaluation, keeping global relations as the direct object of verification rather than reducing them again to isolated local checks.

We further evaluate frontier models on OmniMaster-Bench and find substantial remaining headroom. Strong performance on existing Deep Research benchmarks does not reliably translate to global relational capability, while the Open and Closed settings reveal complementary weaknesses across models. We further use the same relation-centered synthesis framework to construct supervision for learning this capability. Relation-aware mid-training (MT) develops atomic capabilities for recognizing and tracking relational states, while relation-preserving supervised fine-tuning (SFT) integrates these capabilities into complete research trajectories. The resulting agents improve global relational capability while remaining competitive on established Deep Research benchmarks, showing that relation-centered supervision can transfer relational structure from data into research behavior.

Our main contributions are threefold:

- We identify global relations as an underexamined dimension of Deep Research verification and reveal Relational Verification Collapse, where local verification can fail to preserve relational validity.
- We introduce OmniMaster-Bench to systematically evaluate global relational capability. Extensive evaluation shows substantial remaining headroom for frontier models.
- We develop relation-centered synthesis and supervision for learning this capability. Relation-level MT and trajectory-level SFT improve OmniMaster-Bench while remaining competitive on existing Deep Research benchmarks without RL.

2 Related Work

Verification in Deep Research. As Deep Research moves toward longer and more complex research processes, verification has become increasingly important for measuring whether retrieved evidence supports the final result. Existing approaches can be broadly grouped into several paradigms. Rubric-based methods decompose complex requirements into fine-grained or atomic criteria (Li et al., 2026a; Sharma et al., 2026; Zhu et al., 2026), while evidence-based methods examine whether claims and outputs are supported by retrieved sources (Venkit et al., 2025; Polshkov et al., 2026). Other systems incorporate verification directly into the research process and iteratively refine results through constraint-wise checking (Lu et al., 2026). These approaches substantially improve the granularity and reliability of Deep Research evaluation. However, even when relational requirements are represented, their executable verification signals are still often expressed through atomic rubrics, statement-level evidence, or separately judged constraints. This can weaken relations that must hold jointly across multiple research units, leading to the Relational Verification Collapse studied in our work.

Data Synthesis for Deep Research. Scaling Deep Research agents also depends on training data that captures long-horizon search and multi-step information integration. WebShaper constructs controllable search problems from structured reasoning graphs (Tao et al., 2026), while InfoSeek formulates agentic search as hierarchical constraint satisfaction and synthesizes tasks with explicit structural dependencies (Xia et al., 2026). OpenResearcher focuses on scalable synthesis of long-horizon search trajectories (Li et al., 2026c). QUEST further unifies synthetic task generation with verifiable rubric supervision for multi-stage training (Xie et al., 2026), and SearchArt constructs verified tasks and trajectories from automatically generated evidence graphs (Mei et al., 2026). These efforts have progressively expanded Deep Research synthesis from task construction to trajectory generation and verifiable supervision. OmniMaster extends this line by centering synthesis on global relations themselves, explicitly composing their types, numbers, and interactions while preserving the same relational structure from supporting evidence to final evaluation.

3 Relation-Centered Data Synthesis

In this section, we introduce a relation-centered data synthesis framework for constructing Deep Research tasks with explicit and verifiable global relations. As shown in Figure 2, the framework first formalizes the relational structure that a task must satisfy, then realizes this structure through evidence-grounded relation construction, joint relation composition, and relation-preserving verification. We instantiate the framework in two complementary settings, OmniMaster-Open for long-horizon open-web research and OmniMaster-Closed for controlled multi-hop retrieval.

3.1 Global Relations

Deep Research requires joint reasoning over heterogeneous research units, including claims, evidence, sources, entities, and output items. Let

$$\mathcal{U}_\tau = \{u_1, \ldots, u_n\} \tag{1}$$

denote the research units involved in a task τ. A global relation r is defined over a subset $\mathcal{U}_r \subseteq \mathcal{U}_\tau$ as

$$r(\mathcal{U}_r) \in \{0, 1\}, \tag{2}$$

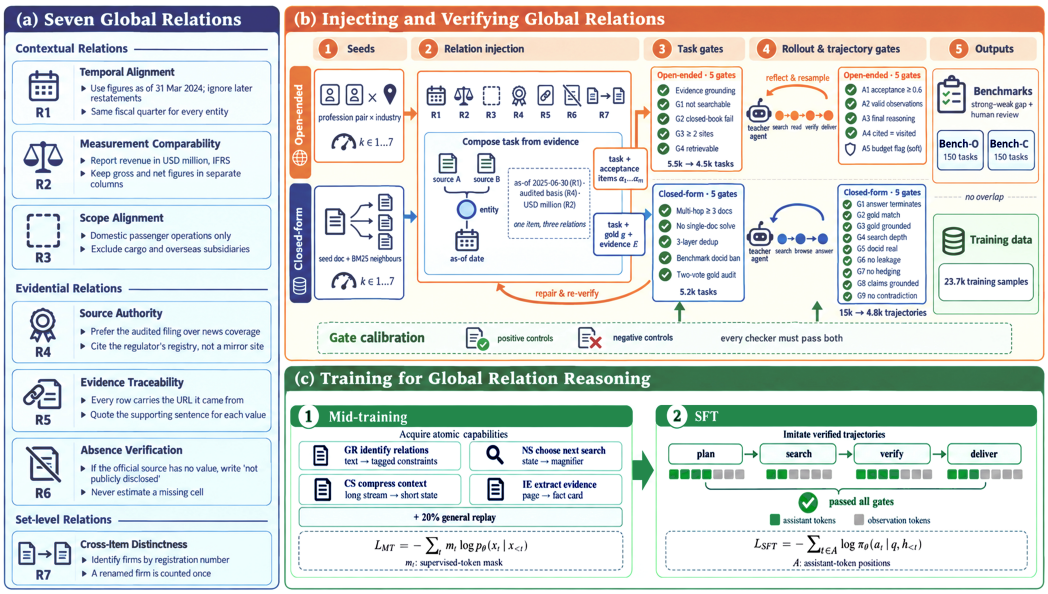

Figure 2: Overview of OmniMaster. (a) Seven global relations spanning contextual, evidential, and set-level requirements. (b) Relation-centered synthesis grounds relations in retrievable evidence, composes them into research tasks, and preserves the same relational structure during verification. (c) Relation-centered agent training transfers this structure into relation-aware mid-training and relation-preserving supervised fine-tuning.

where $r(\mathcal{U}_r) = 1$ indicates that the participating research units jointly satisfy the corresponding relational requirement.

Let $v(u_i) \in \{0, 1\}$ denote the local validity of a research unit. A defining property of global relations is that all participating units can be locally valid while their joint relation remains invalid:

$$\prod_{u_i \in \mathcal{U}_r} v(u_i) = 1, \qquad r(\mathcal{U}_r) = 0. \tag{3}$$

For example, two numerical values may both be correctly retrieved but remain incomparable because they refer to different time points, measurement bases, or analytical scopes.

We construct an expert-defined taxonomy from recurring relational requirements in complex Deep Research tasks. It contains three families and seven relation types. *Contextual relations* require compatible contexts across evidence: *Temporal Alignment* (R1), whether related information satisfies the required temporal reference; *Measurement Comparability* (R2), whether values can be compared under compatible units, definitions, and measurement criteria; and *Scope Alignment* (R3), whether evidence refers to the same entity, population, geography, period, or analytical scope. *Evidential relations* govern how outputs are supported: *Source Authority* (R4), whether claims rest on sources with appropriate authority; *Evidence Traceability* (R5), whether output items can be traced to the evidence that supports them; and *Absence Verification* (R6), whether claims of missing or undisclosed information are backed by a sufficiently comprehensive search. *Set-level relations* constrain the output as a whole: *Cross-Item Distinctness* (R7), whether reported items correspond to distinct entities or records under a consistent identity criterion.

For a task τ, we define its target relation specification as

$$\mathcal{R}_\tau = \left\{ \left(r_j, \mathcal{U}_{r_j} \right) \right\}_{j=1}^{m}, \tag{4}$$

where each pair specifies a relation type and the research units over which it must hold. We define the relational validity of the task as

$$V_{\mathrm{rel}}(\tau) = \prod_{j=1}^{m} r_j(\mathcal{U}_{r_j}), \qquad V_{\mathrm{rel}}(\tau) \in \{0, 1\}. \tag{5}$$

A task is relationally valid when $V_{\text{rel}}(\tau) = 1$. The relation types, their participating research units, and their composition jointly determine the relational structure of the task.

3.2 Relation-Centered Task Synthesis

Design Principle. We treat global relations as construction-time constraints rather than post-hoc annotations. This requires two properties that are central to our design. First, each intended relation must be grounded in evidence that can actually be retrieved from the underlying information space. Second, multiple relations must form genuine dependencies within the same research process rather than a collection of independent local conditions.

Given an information space $\mathcal{S}$ and a target relation specification $\mathcal{R}_\tau$, the framework constructs

$$\tau = (q, E, \mathcal{R}_\tau, y), \qquad E \subseteq \mathcal{S}, \tag{6}$$

where q denotes the research request, E the supporting evidence, and y the expected answer or deliverable.

Evidence-Grounded Relation Realization. We first retrieve candidate evidence according to the target relations in $\mathcal{R}_\tau$. For each relation, the framework examines whether the available evidence can instantiate its required temporal, measurement, scope, source, provenance, or set-level conditions. Missing support triggers additional retrieval or evidence replacement, while relations that cannot be grounded are revised or removed.

Evidence discovery and relation diagnosis are therefore performed iteratively. Retrieval determines what information is available in the environment, while diagnosis determines whether the intended relation is actually supported by that information. For open-web tasks, evidence is gathered from heterogeneous public sources. For closed tasks, evidence is distributed across multiple corpus documents connected through dependent retrieval steps. Detailed evidence-grounding mechanisms are provided in Appendix A.1.

Joint Relation Composition. Once the individual relations are supported, we compose them into a single research task. Relations are connected through shared entities, evidence, sources, or intermediate research requirements, creating dependencies across sources and research steps.

This distinction is important because simply adding several independent constraints does not require global relational reasoning. We require the instantiated relations to affect how evidence is interpreted, retrieved, or combined. Ignoring a required relation should therefore alter a valid intermediate choice or final output rather than merely violate an isolated checklist item. The research request is verbalized only after this relational structure has been established. The task-generation prompts used to instantiate this procedure are provided in Appendix E.

Relation-Preserving Verification. Relations that are valid during construction may weaken when evidence and constraints are composed into a natural-language task. We therefore verify the final instance against the same relation specification $\mathcal{R}_\tau$ used during construction.

For each relation, verification checks whether its supporting evidence remains valid, whether the participating research units preserve the intended relation, and whether resolving the relation remains necessary for completing the task. We additionally reject instances that admit trivial shortcuts, contain unsupported dependencies, or become ambiguous after composition.

We retain a task only when $V_{\text{rel}}(\tau) = 1$. Failed instances are repaired through additional retrieval, evidence replacement, or task reformulation. Preserving the same relation specification from evidence construction to final verification prevents the relational structure from being reduced back into isolated local checks, directly addressing the RVC. The complete construction and verification procedures and prompts are provided in Appendix A.1 and Appendix E.

3.3 OmniMaster-Bench

We instantiate the relation-centered synthesis framework as OmniMaster-Bench under two complementary Deep Research settings.

OmniMaster-Open. OmniMaster-Open evaluates relation preservation during long-horizon open-web research. Each task requires the agent to search heterogeneous public sources, reconcile evidence collected under different contexts, and produce a structured research deliverable. Variation in temporal coverage, measurement conventions, source authority, and evidence availability makes relational validity dependent on the complete research process rather than any individual retrieved fact.

OmniMaster-Closed. OmniMaster-Closed isolates relational reasoning in a controlled retrieval environment. Each task is constructed from a fixed corpus with evidence distributed across multiple documents and a uniquely verifiable answer. Intermediate evidence is required to identify or interpret subsequent information, creating controlled multi-hop dependencies while removing uncertainty introduced by open-web retrieval.

Controlled Relational Complexity. Both subsets cover all seven global relation types and systematically vary their compositions. Each subset contains 150 tasks with an average of 4.0 relations per task, and the number of instantiated relations ranges from one to seven. Benchmark selection balances relation types and relation counts so that evaluation captures both individual relational requirements and their interaction within a task. Detailed statistics, relation distributions, and representative examples are provided in Appendix A.2, with representative task instances in Appendix A.2.1.

4 Relation-Centered Agent Training

Recall that relation-centered synthesis makes global relations explicit and verifiable at the task level. The remaining challenge is to transfer this relational structure into model behavior during long-horizon research. As shown in Figure 2, we address this problem in two stages. Relation-aware mid-training first teaches the model to construct and update a relation-aware research state, while relation-preserving supervised fine-tuning learns how to act on this state throughout complete Deep Research trajectories.

4.1 Relation-Aware Mid-Training

Global relations are often distributed across multiple pieces of evidence and different stages of a research process. Supervision from complete trajectories alone can therefore make individual relational signals difficult to isolate. We instead derive single-step atomic supervision from the verified tasks, evidence, and trajectories produced during relation-centered synthesis. The construction of the four atomic supervision types is detailed in Appendix B.1.

For a task τ with relation set $\mathcal{R}_\tau$, we conceptually view the research state at step t as

$$\mathcal{Z}_t = \left(E_t, \mathcal{R}_t^{\mathrm{sat}}, \mathcal{R}_t^{\mathrm{open}}\right), \tag{7}$$

where E_t denotes the evidence acquired so far, $\mathcal{R}_t^{\mathrm{sat}}$ contains relational requirements that have been sufficiently resolved, and $\mathcal{R}_t^{\mathrm{open}}$ contains those that still require additional evidence or reasoning. We organize the atomic supervision into two groups corresponding to the evidence and relation components of this state:

- **Evidence-State Modeling.** This group contains Information Extraction (IE) and Context Summarization (CS). IE teaches the model to extract grounded claims, supporting evidence, and provenance from retrieved sources. CS compresses accumulated observations into a compact research state that distinguishes established information from unresolved requirements. These two capabilities provide a reliable representation of the evidence available at the current step.
- **Relation-State Modeling.** This group contains Global Relation Recognition (GR) and Next-Step Relation Selection (NS). GR identifies the global relations implied by the research request and current evidence. NS determines which unresolved relation should guide the next retrieval step. These two capabilities teach the model to track not only what information has been collected, but also which relational conditions still need to be satisfied.

Let (x, z) denote an atomic training pair, where x is the local research context and z is the corresponding structured target. Mid-training optimizes

$$\mathcal{L}_{\mathrm{MT}} = -\mathbb{E}_{(x,z)\sim\mathcal{D}_{\mathrm{MT}}} \log p_\theta(z \mid x). \tag{8}$$

This supervision makes relation-aware state modeling explicit before the model is required to carry these states through long-horizon interaction.

4.2 Relation-Preserving Supervised Fine-Tuning

A relation-aware state does not by itself guarantee that global relations remain preserved throughout a complete research process. Evidence arrives incrementally, earlier interpretations may need to be revised, and unresolved relations can interact across multiple retrieval steps. We therefore apply supervised fine-tuning on complete grounded research trajectories.

The trajectories are constructed from the same verified relational structures used during data synthesis, described in Appendix B.1. In the controlled setting, the agent follows dependent retrieval steps across multiple documents toward a uniquely verifiable answer. In the open setting, the agent searches heterogeneous web sources and produces structured deliverables whose requirements depend on evidence distributed across sources. Both settings require the agent to update its research state as new evidence is acquired and to continue retrieval when relational requirements remain unresolved.

We retain only trajectories that provide valid supervision for this process. Accepted trajectories must reach a valid task outcome, acquire the required evidence through actual retrieval, ground conclusions in observed information, and avoid shortcuts such as answer leakage or unsupported reasoning.The trajectory-level filtering criteria and their role in preserving relational supervision are detailed in Appendix A.1.

Let h_t denote the interaction history before step t, and let y_t denote the corresponding assistant output. Supervised fine-tuning optimizes

$$\mathcal{L}_{\mathrm{SFT}} = -\mathbb{E}_{\xi\sim\mathcal{D}_{\mathrm{SFT}}} \sum_{t=1}^{T} \log p_\theta(y_t \mid h_t). \tag{9}$$

When a relational requirement remains unresolved, the agent continues retrieving the missing evidence. When newly acquired evidence changes the current relational state, the agent revises its interpretation and subsequent retrieval decisions. Once the required relations are resolved, the accumulated evidence is used to produce the final output.

5 Experiments

In this section, we evaluate global relational capability and the effectiveness of our relation-centered training framework. Section 5.2 examines how frontier models perform on OmniMaster-Bench and whether relation-centered training transfers to existing benchmarks. Section 5.3 further analyzes how relation-aware mid-training and relation-preserving supervised fine-tuning contribute to the resulting research capability.

5.1 Experimental Setup

Benchmarks. We evaluate on the two subsets of OmniMaster-Bench and five established Deep Research benchmarks. OmniMaster-Open contains 150 long-horizon open-web tasks. Each generated deliverable is evaluated against task-specific acceptance items by Claude Opus 4.8, and the task score is the fraction of applicable items that pass. OmniMaster-Closed contains 150 fixed-corpus multi-hop tasks with uniquely verifiable answers. Each task is answered twice, and we report exact-match accuracy over the resulting 300 responses without using an LLM judge. Both subsets cover all seven global relation types and are balanced by relation count $k = 1, \ldots, 7$ (Appendix A.2). To assess judge robustness on Open, we independently re-evaluate the six frontier models with DeepSeek-V4 and obtain a Spearman correlation of $\rho = 0.943$ with the ranking produced by Claude Opus 4.8 (Appendix B.2). We additionally evaluate on BrowseComp-Plus (Chen

Table 1: Main results on OmniMaster-Bench and existing Deep Research benchmarks (%). DRB reports RACE and WideSearch reports item-level F1; all other columns report accuracy. † Results are taken from the original papers and may use different evaluation subsets or protocols. Bold and underline denote the best and second-best results among open Deep Research agents, including OmniMaster. "–" denotes unavailable results.

	OmniMaster-Bench		**Existing Benchmarks**				
Model	**Open**	**Closed**	**BC-Plus**	**GAIA**	**HLE**	**DRB**	**WideSearch**
Frontier Models							
Claude Opus 5	50.9	76.3	–	–	56.3	54.1	–
GPT-5.6 Sol	40.5	49.7	–	–	44.5	46.6	83.3
Kimi K3	59.2	64.0	–	–	–	–	–
Qwen3.8-Max	54.0	60.7	–	–	–	–	–
GLM-5.3	50.1	63.7	–	–	42.0	52.7	82.7
DeepSeek-V4-Pro	43.6	52.3	–	–	42.7	50.7	–
Open Deep Research Agents†							
SearchMaster-9B	–	–	60.2	57.3	–	–	–
DeepSearch-World-9B	–	–	–	61.5	25.7	–	–
DR Tulu-8B	–	–	–	–	–	45.4	–
Tongyi DeepResearch-30B-A3B	–	–	–	70.9	32.9	–	–
QUEST-35B	–	–	<u>69.5</u>	**80.8**	**37.2**	<u>48.2</u>	**60.6**
OmniMaster							
OmniMaster-27B	**39.3**	**61.7**	**70.8**	<u>76.7</u>	<u>35.4</u>	46.0	<u>48.1</u>
OmniMaster-35B-A3B	<u>39.2</u>	<u>55.3</u>	63.9	71.2	33.1	**49.0**	35.9

et al., 2025), GAIA (Mialon et al., 2024), Humanity's Last Exam (HLE) (Phan et al., 2025), DeepResearch Bench (DRB) (Du et al., 2026), and WideSearch (Wong et al., 2025). Detailed subsets, metrics, and evaluation protocols are provided in Appendix B.2.

Baselines. We consider two groups of baselines. The first consists of six frontier models: Claude Opus 5, GPT-5.6 Sol, Kimi K3, Qwen3.8-Max, GLM-5.3, and DeepSeek-V4-Pro. On OmniMaster-Bench, all frontier models use the same ReAct scaffold with identical prompts, tools, and interaction budgets, using 20 turns for Open and 16 turns for Closed. The shared agent prompts are provided in Appendix E. The second group contains open Deep Research agents of comparable scale, including SearchMaster-9B (Tan et al., 2026), DeepSearch-World-9B (Geng et al., 2026), DR Tulu-8B (Shao et al., 2025), Tongyi DeepResearch-30B-A3B (Tongyi DeepResearch Team, 2025), and QUEST-35B (Xie et al., 2026). These systems are not evaluated on OmniMaster-Bench; other results are taken from their original papers.

Training. We train OmniMaster on two backbones, Qwen3.6-27B and Qwen3.5-35B-A3B, with the latter activating 3B parameters per token. Both follow the two-stage training framework introduced in Section 4. Relation-aware MT uses 1,040 single-step samples covering the four atomic capabilities in Section 4.1, while relation-preserving SFT uses gated multi-turn research trajectories. Full training details are provided in Appendix B.1 (Table B.2). All benchmark tasks are excluded from training, and synthesized Closed tasks do not use the gold or evidence documents associated with the official BrowseComp-Plus queries.

5.2 Main Results

Global relational capability remains challenging. Table 1 shows that strong performance on existing Deep Research benchmarks does not consistently translate to OmniMaster-Bench. In particular, models that perform well on broad information-seeking tasks can still struggle when evidence must satisfy relations across time, measurement, sources, and outputs. The different rankings on Open and Closed further suggest that relation preservation during long-form research and relation resolution along controlled retrieval chains capture complementary capabilities.

Difficulty depends on relational structure. Performance varies only moderately across individual relation types, indicating that no single relation dominates benchmark difficulty (Appendix C.1). Relation count is also not a sufficient predictor: Open becomes harder at high relation counts without a monotonic trend, while Closed shows no consistent decrease with k (Figure C.1). The main difficulty therefore lies in how relations interact with the evidence and retrieval structure of a task.

Relation-centered training is effective and transferable. On OmniMaster-Closed both OmniMaster models fall inside the frontier range; on OmniMaster-Open they remain below the weakest frontier model, a gap we analyze in Appendix D.2. This indicates that preserving relational structure during synthesis and training improves more than benchmark-specific relation recognition. The learned behaviors also transfer to evidence grounding, state tracking, and cross-source reasoning. WideSearch remains an exception, reflecting its stronger emphasis on exhaustive breadth rather than relational depth. A comparison of the remaining failure modes of OmniMaster and frontier models is provided in Appendix D.2.

5.3 Analysis of Training

MT improves both evidence and relation states. Table 2 shows consistent gains across the four atomic capabilities in Section 4.1. Improvements are strongest on evidence extraction and context summarization, while global relation recognition remains the most difficult capability. This suggests that grounding and maintaining observed evidence are easier to acquire than inferring the latent relational requirements that govern a task.

Table 2: Accuracy (%) on held-out atomic tasks before and after MT. GR and NS correspond to relation-state modeling; IE and CS to evidence-state modeling.

	Relation State		**Evidence State**	
Model	**GR**	**NS**	**IE**	**CS**
Qwen3.5-9B	18.8	26.1	36.5	19.3
+ MT	**25.6**	**60.2**	**89.3**	**38.7**
Qwen3.5-35B-A3B	9.4	25.9	16.0	6.7
+ MT	**23.9**	**72.2**	**93.1**	**62.0**

MT and SFT serve different roles. Table 3 shows that MT alone does not create a functional long-horizon research agent. Its role is to acquire the local capabilities required for relation-aware state modeling. SFT then composes these capabilities across multi-turn interaction, enabling the model to retrieve evidence, update its state, and resolve relations over a complete research trajectory. Additional training diagnostics, including loss convergence, data-scale effects, and behavioral changes across stages, are provided in Appendix C.2.

Table 3: Stage-wise performance of the 35B model (%). SFT subscripts denote the number of training trajectories. WS denotes WideSearch.

Setting	**Open**	**Closed**	**DRB**	**WS**
Base	2.7	22.7	–	–
+ MT	11.2	31.7	–	–
+ MT + $\text{SFT}_{5.1\text{k}}$	**48.7**	42.0	46.7	34.8
+ MT + $\text{SFT}_{23.1\text{k}}$	39.2	**55.3**	**49.0**	**35.9**

Trajectory composition matters. Increasing SFT data does not improve Open and Closed uniformly. The larger mixture benefits controlled retrieval and several external benchmarks but weakens Open performance, indicating that trajectory distribution matters in addition to scale. Open-web synthesis and controlled multi-hop retrieval therefore require different balances of training experience.

6 Conclusion

OmniMaster studies global relational capability in Deep Research. It introduces Relational Verification Collapse, OmniMaster-Bench, and relation-centered training for evaluating and improving relation preservation. Results show that global relations remain challenging for current systems, while relation-centered supervision improves relational capability and stays competitive on general Deep Research benchmarks.

AI USE STATEMENT

In this work, we used generative AI tools for translation and language polishing to improve the grammar, clarity, and readability of the manuscript. We did not use generative AI for other tasks requiring disclosure, such as developing research methods, formulating scientific claims, designing experiments, implementing the proposed methods, or interpreting results. All AI-assisted content was reviewed by the authors, who take full responsibility for the final manuscript.

REFERENCES

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, et al. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*, 2025.

João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, et al. Deepresearchgym: A free, transparent, and reproducible evaluation sandbox for deep research. *arXiv preprint arXiv:2505.19253*, 2025.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. Deep-research bench: A comprehensive benchmark for deep research agents. In *International Conference on Learning Representations*, volume 2026, pp. 42414–42448, 2026.

Xinyu Geng, Xuanhua He, Sixiang Chen, Yanjing Xiao, Fan Zhang, et al. Deepsearch-world: Self-distillation for deep search agents in a verifiable environment. *arXiv preprint arXiv:2607.07820*, 2026.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jimenez Gutierrez, Yiheng Shu, et al. Mind2web 2: Evaluating agentic search with agent-as-a-judge. *Advances in Neural Information Processing Systems*, 38, 2026.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4745–4759, 2025.

Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench ii: Diagnosing deep research agents via rubrics from expert report. *arXiv preprint arXiv:2601.08536*, 2026a.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5420–5438, 2025.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *Advances in Neural Information Processing Systems*, 38:120091–120131, 2026b.

Zhuofeng Li, Dongfu Jiang, Xueguang Ma, Haoxiang Zhang, Ping Nie, Yuyu Zhang, Kai Zou, Jianwen Xie, Yu Zhang, and Wenhu Chen. Openresearcher: A fully open pipeline for long-horizon deep research trajectory synthesis. *arXiv preprint arXiv:2603.20278*, 2026c.

Shuqi Lu, Chaofan Li, Kun Luo, Zhang Zhang, Hui Wang, Hongwang Xiao, Zheng Liu, Lei Xiong, Jiahao Wang, Sen Wang, et al. Arex: Towards a recursively self-improving agent for deep research. *arXiv preprint arXiv:2607.21461*, 2026.

Lang Mei, Xiaohan Yu, Chong Chen, Liyan Liu, Xiangnan Chen, Jinchao Ma, Chao Feng, Li Huang, Siyu Mo, Sichen Kang, et al. Searchart: Training long-horizon search agent with scalable synthetic and verified task. *arXiv preprint arXiv:2607.24850*, 2026.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *International Conference on Learning Representations*, volume 2024, pp. 9025–9049, 2024.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

Vitaliy Polshkov, Marcin Pitera, Jeremy Yang, Kirill Priemko, Maksim Gaiduk, Aleksandr Nikolenko, Denis Bykov, Clare Southern, Denis Yarats, and Jerry Ma. Wandr: A benchmark for wide and deep research. *arXiv preprint arXiv:2608.14747*, 2026.

Corbin Rosset, Ho-Lam Chung, Guanghui Qin, Ethan Chau, Zhuo Feng, Ahmed Awadallah, Jennifer Neville, and Nikhil Rao. Researchy questions: A dataset of multi-perspective, decompositional questions for deep research. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3712–3722, 2025.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, et al. DR Tulu: Reinforcement learning with evolving rubrics for deep research. *arXiv preprint arXiv:2511.19399*, 2025.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, et al. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. In *International Conference on Learning Representations*, volume 2026, pp. 90447–90472, 2026.

Rui Sun, Zuo Bai, Wentao Zhang, Yuxiang Zhang, Li Zhao, Shan Sun, and Zhengwen Qiu. Finresearchbench: A logic tree based agent-as-a-judge evaluation framework for financial research agents. In *Proceedings of the 6th ACM International Conference on AI in Finance*, pp. 656–664, 2025.

Wentao Tan, Qiong Cao, Jiaqi Wang, and Nan Duan. Searchmaster: Grounded and regulated self-play for search agents. *arXiv preprint arXiv:2608.01822*, 2026.

Zhengwei Tao, Jialong Wu, Wenbiao Yin, Pu Wu, Junkai Zhang, Baixuan Li, Haiyang Shen, Kuan Li, Liwen Zhang, Xinyu Wang, et al. Webshaper: Agentically data synthesizing via information-seeking formalization. In *International Conference on Learning Representations*, volume 2026, pp. 101872–101889, 2026.

Tongyi DeepResearch Team. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*, 2025.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Kung-Hsiang Huang, Yixin Mao, and Chien-Sheng Wu. Deeptrace: Auditing deep research ai systems for tracking reliability across citations and evidence. *arXiv preprint arXiv:2509.04499*, 2025.

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, et al. Widesearch: Benchmarking agentic broad info-seeking. *arXiv preprint arXiv:2508.07999*, 2025.

Ziyi Xia, Kun Luo, Hongjin Qian, Siqi Bao, and Zheng Liu. Open data synthesis for deep research. In *International Conference on Learning Representations*, volume 2026, pp. 63813–63834, 2026.

Jian Xie, Tianhe Lin, Zilu Wang, Yuting Ning, Yuekun Yao, Tianci Xue, Zhehao Zhang, Zhongyang Li, Kai Zhang, Yufan Wu, et al. Quest: Training frontier deep research agents with fully synthetic tasks. *arXiv preprint arXiv:2605.24218*, 2026.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 414–431, 2025.

Minghang Zhu, Chuyang Wei, Junhao Xu, Yilin Cheng, Zhumin Chen, and Jiyan He. Deeprubric: Evidence-tree rubric supervision for efficient reinforcement learning of deep research agents. *arXiv preprint arXiv:2606.17029*, 2026.

A Details of OmniMaster-Bench

OmniMaster-Bench is designed to test whether a research agent can preserve global relations across evidence collected from multiple sources. The appendix therefore reports the construction pipeline and benchmark statistics around two questions: whether the tasks are valid and require the intended relations, and whether the resulting benchmark covers those relations and application domains broadly enough to support the comparisons in the main paper. Throughout the appendix, R1–R7 denote Temporal Alignment, Measurement Comparability, Scope Alignment, Source Authority, Evidence Traceability, Absence Verification, and Cross-Item Distinctness.

A.1 Data Construction Pipeline

The key difference from prior synthesis pipelines is that OmniMaster controls the relation to be learned before it finalizes the task that expresses it. Recent Deep Research data synthesis controls several important aspects of difficulty: WebShaper formalizes information-seeking structure, InfoSeek builds hierarchical constraint dependencies, OpenResearcher scales long-horizon trajectories, QUEST organizes supervision around rubric trees, and SearchArt constructs and verifies evidence graphs (Tao et al., 2026; Xia et al., 2026; Li et al., 2026c; Xie et al., 2026; Mei et al., 2026). These approaches provide strong control over reasoning structure, constraints, trajectories, or evidence connectivity. OmniMaster uses a different primary construction unit: a cross-evidence relation that must remain valid from evidence collection to the final answer. The target relation set is therefore fixed before task wording is accepted; retrieval is then used to determine whether that relation can actually be instantiated. If the evidence does not support the intended relation, the relation or the evidence is repaired rather than leaving the relation as a post-hoc label. This is the mechanism that makes coverage over relation types and compositions a property of the construction process rather than an accidental property of the sampled documents.

Open and Closed realize the same relation-first principle under complementary sources of difficulty. OmniMaster-Open starts from professional activities rather than from a single source document. We use ESCO v1.2.1 after excluding armed-forces occupations, combine a main skill, an auxiliary skill, and an industry context, and generate candidate commissions through skill-seeded generation, trajectory inversion, and short-answer generation. Starting from a research activity broadens the space of deliverables and source combinations that can instantiate the target relations. OmniMaster-Closed instead starts from a fixed document corpus and constructs questions over document chains with dependent retrieval hops. This removes open-web uncertainty and makes it possible to test whether the agent resolves the intended relation when the relevant evidence is known to exist. Training tasks use the same construction logic on disjoint source material.

Matched tools and budgets keep access conditions from becoming a hidden explanation for benchmark differences. On Open, agents use `search`, `visit`, and `finish`; on Closed, they use corpus `search`, paged `browse`, and `answer`. Teachers and evaluated models receive the same tool interfaces and the same interaction limits within each setting. The two subsets therefore differ in information environment, but not in whether a particular evaluated model receives privileged retrieval access.

Evidence grounding and task validity. A relation is retained only when the available evidence makes that relation operational, not merely because the task text mentions it. Rubric-based and evidence-based Deep Research evaluation has made substantial progress by checking atomic criteria, statement support, and citation reliability (Li et al., 2026a; Sharma et al., 2026; Zhu et al., 2026; Venkit et al., 2025; Polshkov et al., 2026). The remaining failure mode for our setting is different: individually supported statements can still be jointly invalid when their time basis, measurement basis, scope, source role, provenance, or identity relation is wrong. OmniMaster therefore verifies the relation at construction time. For Open tasks, generated source suggestions are replaced by pages that were actually retrieved, and tasks are rejected when the required evidence cannot be recovered. Short-answer tasks receive additional checks for search leakage, closed-book answerability, and evidence concentration. For Closed tasks, generation records a verbatim witness for every required relation, the question span that activates it, and a concrete counterfactual error produced by ignoring

it; an independent audit then re-reads the complete gold documents and rejects tasks whose chain is unnecessary, whose answer is not unique, or whose relation witnesses are unsupported.

This evidence-first check is what prevents nominal relation coverage from inflating actual benchmark coverage. A task enters trajectory collection only after the relation has a retrievable evidential basis, so later trajectory filtering is testing whether an agent preserves an established dependency rather than whether a generator happened to write a plausible relation label.

Table A.1: Open-task filtering funnel. The evidence gates remove candidates before trajectory collection, so relational support is established before a task becomes training supervision.

Stage	Tasks	%
Generated candidate tasks	8,911	100.0
Passed the evidence gates	7,805	87.6
Entered trajectory collection	6,114	68.6

Trajectory filtering. A correct endpoint is not sufficient training supervision when the research path can bypass the intended relation. Long-horizon synthesis work has shown the importance of filtering search trajectories and verifying retrieved evidence (Li et al., 2026c; Xie et al., 2026; Mei et al., 2026). For relation-centered supervision, we additionally need the evidence path itself to remain faithful to the relation that defines the task. Otherwise, a teacher can reach the correct answer through leakage, unsupported reasoning, or a shortcut that never resolves the targeted dependency, and the resulting trajectory would teach the wrong behavior.

Teacher collection therefore prioritizes recoverable, grounded trajectories rather than accepting every successful rollout. On Open, a low-quality first rollout can trigger a recovery attempt, but the replacement is kept only when it improves the task-level score. Accepted trajectories must complete the requested deliverable, contain usable observations, and connect cited sources back to pages that were actually observed. On Closed, multiple teacher models are used and a trajectory is accepted only after the programmatic gates in Table A.2. These gates jointly check the endpoint, retrieval depth, evidence provenance, absence of answer leakage, distractor handling, claim grounding, and internal consistency. The intended supervision is therefore not "answer correctly," but "reach the answer through an observed and relation-consistent evidence path."

Table A.2: Trajectory gates for OmniMaster-Closed. Together they require a correct endpoint, an observed evidence path, and the absence of common shortcuts such as answer leakage.

Gate	Name	Criterion
T1	Termination	The last action is `answer`.
T2	Gold match	Normalized word-boundary match; the gold tokens must be contained in the answer, and numbers must not match as bare substrings.
T3	Gold grounding	The gold string appears in the trajectory's own observations.
T4	Retrieval depth	At least two searches and one browse.
T5	Citation validity	Every browsed document id exists and appeared in an earlier search result.
T6	Anti-leakage	A search query that contains the full gold string before it was observed is rejected.
T7	Distractor guard	The answer must not contain both the gold and a mutually exclusive distractor, unless in an explicit contrast.
T8	Claim grounding	A quotation asserted in a thought must be found in an earlier observation.
T9	Consistency and format	Answering after declaring the question inconsistent with the evidence, leaked special tokens, or duplicated thoughts are rejected.

Shortcut control continues after trajectory filtering. Closed questions are deduplicated at exact, token-overlap, and character-similarity levels, and the gold and evidence documents of the official

BrowseComp-Plus questions are excluded from synthesized chains. Benchmark selection is then performed from the filtered task stock rather than from raw generations. Relation-count quotas and relation balancing are applied only after evidence and trajectory validity have been established, so coverage targets cannot rescue an invalid task. The Open subset follows the same ordering: solvability and evidence validity precede final balancing.

Table A.3: Closed-task filtering funnel. Candidate tasks survive only when at least one trajectory satisfies the evidence-path constraints and the task passes subsequent duplicate, gold, and shortcut review.

Stage	Count
Deduplicated candidate tasks	5,183
Teacher trajectories sampled	15,010
Trajectories passing the nine gates	4,781
Tasks with at least one passing trajectory	3,461
After near-duplicate removal	3,444
Selected by quota	1,500
After gold review and red-team scan	1,469

Human and independent verification. Generator self-consistency is not treated as evidence that a synthesized relation is valid. Where a property can be checked programmatically, we use programmatic retrieval or trajectory gates; where semantic judgment is required, generation, auditing, gold review, and evaluation are separated by role. Closed tasks receive independent evidence auditing, a two-vote gold review, and a shortcut-oriented red-team pass. Human spot checks then target the properties most central to the method: whether multiple documents are genuinely necessary, whether the annotated relations match the question, and whether a successful trajectory carries the intended relation chain through to the answer. These checks do not replace the automatic pipeline; they test whether the automatic pipeline is preserving the object it was designed to preserve.

Table A.4: Human spot checks of the properties that automatic generation is intended to preserve: multi-document necessity, relation correctness, and relation-preserving trajectories.

Property	Sampled	Passed
Task requires more than one document	20	19 (95.0%)
Annotated relations match the question	20	18 (90.0%)
Annotated Absence Verification (R6) matches the question	8	8 (100.0%)
Trajectory preserves the relation chain	15	13 (86.7%)

The pipeline also avoids relying on a single kind of verifier across all stages. Table A.5 summarizes the division of labor. Evidence discovery itself is non-LLM retrieval in both settings; several validity checks are programmatic; and the Closed pipeline uses separate models for construction, audit, trajectory generation, and gold review. The Open setting necessarily uses an LLM judge for structured deliverables, but its evidence retrieval and pre-trajectory gates remain independent of that judge.

A.2 Statistics of OmniMaster-Bench

The final benchmark is balanced so that performance cannot be explained by one relation type, one relation count, or one narrow subject area. OmniMaster-Open and OmniMaster-Closed each contain the full seven-relation taxonomy and use the same relation-count schedule, while differing in whether research occurs on the open web or in a controlled corpus. This creates a paired design: the benchmark tests the same abstract relational requirements under two different retrieval regimes. The statistics in Table A.6 summarize this design; the important point is the matched relation coverage rather than any individual count.

Table A.5: Division of labor across synthesis, verification, trajectory collection, and evaluation. Programmatic retrieval or checks are used where possible, while model-based stages are separated by role.

Component	Open	Closed
Evidence discovery	Serper web search and page fetching (no LLM)	BM25 top-5 retrieval (no LLM)
Relation diagnosis	Relations annotated by the task-construction model; closed-book gate by Claude Opus 4.8, GPT-5.6 Sol and DeepSeek-V4-Pro	Witnesses supplied by the task-construction model; independent evidence audit
Task construction	Round-robin over Claude Opus 4.8, Claude Opus 5, GPT-5.6 Sol, GLM-5.2, Kimi K3 and DeepSeek-V4-Pro	Round-robin over the same six models (BrowseComp-Plus batch); Claude Opus 5, Claude Opus 4.8 and GPT-5.6 Sol (peS2o batch)
Relation-preserving verification	Trajectory judge: Claude Opus 4.8	Nine programmatic trajectory gates; two-vote gold review by Claude Opus 5 and GPT-5.6 Sol
Trajectory teacher	Claude Opus 4.8	Round-robin over Claude Opus 4.8, Claude Opus 5, GPT-5.6 Sol, GLM-5.2, Kimi K3 and DeepSeek-V4-Pro
Context compression (training data)	Claude Opus 5	Claude Opus 5
Evaluation judge	Claude Opus 4.8	Programmatic word-boundary match

Table A.6: Summary of OmniMaster-Bench. Open and Closed instantiate the same relation taxonomy in complementary research environments.

Statistic	Open	Closed
Tasks	150	150
Relation types covered	7	7
Avg. relations per task	4.0	4.0
Avg. question length (tokens)	557.6	134.6
Avg. answer length (tokens)	–	5.5
Avg. gold evidence documents per task	–	3.1
Domains	–	28

Domain breadth reduces the chance that relational behavior is confounded with a single content domain. Open tasks are written as free-form professional research scenarios, while Closed tasks span science, engineering, medicine, social science, and humanities domains. Figure A.1 and Tables A.7–A.8 document the resulting topic spread. The domain distribution is a secondary control: relations are the benchmark variable of interest, and no single domain is allowed to dominate the Closed subset.

Table A.7: Domains of OmniMaster-Closed.

Tasks	Domains
6	Chemistry; Neuroscience & Psychology; Civil & Mechanical Engineering; Ecology & Environment; Earth & Climate Science; Astronomy & Physics; Public Health & Epidemiology; Clinical Medicine; Economics & Finance; Materials Science
5	Transportation; Religion; Education; Science & Medicine; Technology & Internet; Arts & Culture; Business & Finance; Sports; Politics & Government; Film & Entertainment; Law & Crime; Geography & Places; Military History; Machine Learning & AI; Agriculture & Food Science; Energy & Electrical Engineering; Law & Social Science; Other

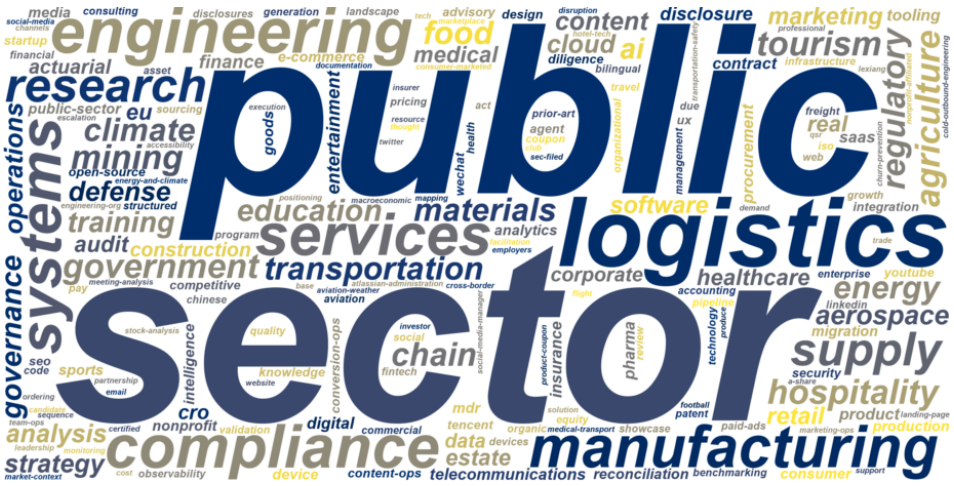

Figure A.1: Frequent terms in the research scenarios of OmniMaster-Open.

Table A.8: The ten six-task domains of OmniMaster-Closed and the average number of global relations per task.

Domain	Tasks	Avg. relations
Chemistry	6	5.8
Materials Science	6	5.8
Civil & Mechanical Engineering	6	5.7
Ecology & Environment	6	5.7
Clinical Medicine	6	5.3
Astronomy & Physics	6	5.3
Earth & Climate Science	6	5.0
Neuroscience & Psychology	6	4.8
Public Health & Epidemiology	6	4.7
Economics & Finance	6	4.2

Relation balancing is visible at both the type and composition levels. Figure A.2 shows that all relation types occur repeatedly in both subsets and that tasks span the full range from single-relation to multi-relation compositions. This matters for the interpretation of the main results: a model must handle different relations in combination rather than exploit a benchmark dominated by one recurring requirement.

A.2.1 Representative tasks show how local correctness can still fail globally

The examples below illustrate the benchmark mechanism rather than merely its surface format. The Open example makes two individually legitimate population statistics invalid to combine until their measurement bases are reconciled. The Closed example makes an otherwise plausible date wrong unless the answer is anchored to the correct page state and traced through the required evidence chain. In both cases, the relation changes what constitutes a valid answer.

Example 1: Market memo for adult English tutoring (Open, R2). The request asks for a structured deliverable built from named public sources. It is scored against eight acceptance items, and the relation it tests is whether two population counts defined on different bases are normalized before they are combined.

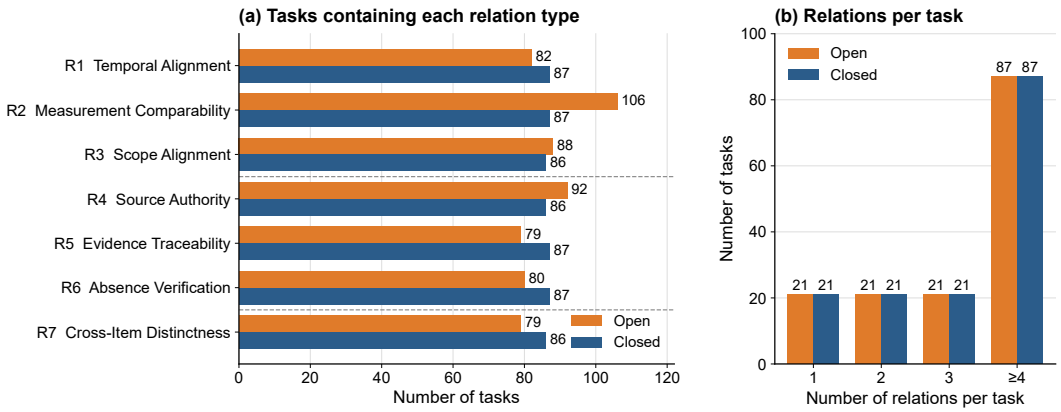

Figure A.2: Relation statistics of OmniMaster-Bench. (a) Number of tasks containing each relation type. Dashed lines separate the Contextual, Evidential and Set-level families. (b) Number of tasks by the number of relations they contain.

OmniMaster-Open — Request **R2 Measurement Comparability**

Act as an idea-validation analyst to produce a Hexa-style Opportunity Memo (Perceived Created Value / PCV methodology) assessing whether a startup targeting AI-assisted tutoring for U.S. adult English-language learners in accredited continuing-education programs is worth pursuing; first run a 'capture' pass ingesting a founder brain dump into a zero-information-loss appendix. Public sources only: NCES Digest of Education Statistics + IES/NCES adult-education tables, U.S. Census ACS language-spoken-at-home tables, and one state WIOA Title II adult-education annual report. Method requires caliber reconciliation: NCES (program-enrollment caliber) and ACS (self-reported "speaks English less than 'very well' " caliber) define the LEP population on different bases, so normalize both to a single stated caliber and reconcile counts explicitly, pinning as-of years, units (persons), and exact table IDs. Undisclosed quantities marked 'not publicly disclosed', never estimated. Deliver one self-contained Markdown with pinned sections (1) Capture Appendix, (2) Problem-Solution Fit, (3) Market Opportunity with reconciled population table, (4) PCV assessment, (5) Go/No-Go, (6) Sources table with per-claim URLs and access dates.

Relation specification. R2 Measurement Comparability. The ACS and NCES population counts are defined on different bases and must be normalized to one stated basis before they are compared or combined.

Evidence. Open web. Eight acceptance items each specify an evidence type and candidate sources, including ACS table B16005 and the NCES Digest tables.

Expected output. A six-section Markdown memo. The Market Opportunity section first reports the ACS and NCES figures separately and states that their bases differ, then combines them in a reconciled population table whose footnote states the normalization rule. All counts in the table are in persons.

Example 2: FAQ date through a three-hop chain (Closed, R1 + R5). The question has a single gold answer that is reached only by following three documents in order and anchoring the answer to the state of the final page rather than to a later article.

OmniMaster-Closed — Question **R1 Temporal Alignment + R5 Evidence Traceability**

According to the FAQ reached by starting with the blog-layout recommendation "Make social sharing is accessible on all devices," consulting the donation-message article whose social-media advice says posts should "stop the scroll," and then following the platform identified there by its newer parenthetical name to the artist-claim FAQ that lists it as a verification channel, what is the FAQ's exact date as of that FAQ state (not the later donation article's publication date)?

Relation specification. R1 Temporal Alignment: the answer must be anchored to the state of the FAQ page. R5 Evidence Traceability: the answer must be traced along the three-hop evidence chain to the specified page.

Gold evidence. Three documents from the BrowseComp-Plus corpus (IDs 50864, 70784, 75685).

Expected output. `2023-08-31`. A model that takes the publication date of the later donation article answers `2024-03-12` and violates R1.

B Training and Evaluation Details

We use a two-stage recipe on the same backbone. Mid-training (MT) isolates relation-aware research-state capabilities, while supervised fine-tuning (SFT) teaches the model to carry those states through complete ReAct trajectories. The setup below reports the data composition, supervision format, and optimization details needed to reproduce these stages.

B.1 Training Setup

Mid-training is designed to isolate the state variables that are difficult to identify from end-to-end trajectories alone. The four MT tasks are derived from the same verified tasks and trajectories used for SFT, but each exposes one local prediction problem. Global Relation Recognition (GR) asks what relations the task requires; Next-Step Relation Selection (NS) asks which unresolved requirement should drive the next action; Information Extraction (IE) grounds claims and provenance in a retrieved page; and Context Summarization (CS) compresses accumulated observations into confirmed, missing, and next-step state. Because these targets are generated from already-gated data, MT teaches the model how to represent evidence and relational status without asking it to solve the entire research task at once.

Training data are organized by learning role rather than pooled indiscriminately. MT contains balanced atomic samples plus a small amount of general instruction data, while SFT contains complete gated trajectories from both Open and Closed construction pipelines. Table B.1 reports the implementation statistics. The mechanism is the separation of supervision granularity: local state modeling is learned before long-horizon action composition.

Table B.1: Training data organized by learning role. MT isolates atomic research-state capabilities, whereas SFT composes them over complete trajectories. Supervised tokens are the tokens that contribute to the loss.

Stage	Data	#Samples	Avg. turns	Avg. supervised tokens
MT	Atomic samples (GR, NS, IE, CS)	1,000	1	349
MT	General instruction samples	40	1	2,206
SFT	Multi-turn trajectories	5,137	10.7	3,993

SFT trajectories. SFT is the stage that converts a locally relation-aware state into executable research behavior. A relation can be recognized correctly at one step and still be lost after new evidence arrives. SFT therefore uses only trajectories that have already passed the evidence-path checks in Appendix A.1. Open trajectories require the model to reconcile heterogeneous web evidence in a structured deliverable; Closed trajectories require dependent retrieval through a fixed corpus. The shared supervision signal is not task format but state propagation: unresolved relations should trigger additional retrieval, newly observed evidence should update the current interpretation, and the final answer should be produced only after the relevant evidence and relations have been resolved.

The loss follows ordinary ReAct imitation, so the training signal comes from the curated trajectory rather than a specialized architecture. Each turn contains a thought, a tool call, and a returned observation. Loss is applied to assistant thought and tool-call tokens; system instructions, task text, and tool observations remain visible as context but are masked from the loss. This trains the model to choose the next reasoning/action step conditional on the evidence state it has actually observed.

Optimization. The two stages use the same backbone and standard supervised objectives, making data organization the main methodological intervention. The 35B model is initialized from Qwen3.5-35B-A3B-Base, MT is applied first, and SFT starts from the MT checkpoint. Both stages use full-parameter bf16 training with the settings in Table B.2. Table 1 reports the 23,070-trajectory SFT checkpoint on every benchmark, trained with the settings of Table B.2 for one epoch over the larger mixture. The 5,137-trajectory checkpoint described in Tables B.1–B.2 (the +MT+SFT$_{5.1k}$ row of Table 3) is used for the analyses in Tables C.1, D.1 and D.2 and Figures C.1 and C.4, because its step-level traces on OmniMaster-Open were logged in full. No additional inference-time search budget or relation-specific decoding procedure is introduced for OmniMaster.

Table B.2: Training hyperparameters of the 35B model.

Hyperparameter	**MT**	**SFT**
Learning rate	7×10^{-6}	7×10^{-6}
LR schedule	Cosine, 3% warmup	Cosine, 3% warmup
Global batch size	8	64
Per-device batch $\times$ gradient accumulation	1×2	1×8
Epochs	1	1
Optimization steps	130	81
Precision	bf16	bf16
Parallelism	ZeRO-3, CPU offload	ZeRO-3, CPU offload
GPUs	4 $\times$ H20	8 $\times$ H20

B.2 Evaluation Protocol

All models are evaluated under matched interaction conditions so that score differences are not attributable to a stronger search interface or a larger action budget. Controlled retrieval environments are useful precisely because changing the search backend can otherwise change system rankings or reproducibility (Coelho et al., 2025). On OmniMaster-Bench, frontier baselines and OmniMaster use the same ReAct scaffold, task prompt, tools, and interaction limits within each subset. Open deliverables are judged against the task's acceptance conditions; Closed answers use normalized word-boundary matching against a unique gold answer. The two evaluation modes match the two benchmark settings: Open measures whether relations survive a full research deliverable, while Closed isolates whether they are resolved along a controlled evidence chain.

External benchmarks test whether relation-centered training transfers rather than merely specializing to OmniMaster-Bench. We also evaluate on established information-seeking and Deep Research benchmarks using their released or standard evaluation protocols. Table B.3 records the task sets, metrics, and judges. These benchmarks emphasize different aspects of research quality, retrieval, and answer correctness; maintaining performance on them therefore provides a check that the training recipe has not simply learned the surface format of OmniMaster tasks.

Table B.3: Public benchmarks used to test whether relation-centered training transfers beyond OmniMaster-Bench.

Benchmark	**Tasks**	**Metric**	**Judge**
GAIA	103	Accuracy	GPT-4.1
HLE	130	Accuracy	GPT-5 mini
BrowseComp-Plus	130	Accuracy	GPT-4.1
DeepResearch Bench (English)	50	RACE	GPT-4o
WideSearch (English)	100	Item F1	GPT-5 mini

A second judge checks whether the Open result is an artifact of one evaluator. Because long-form deliverables require semantic grading, we independently rescore the frontier-model outputs with a second judge and compare both threshold decisions and model-level rankings. The agreement

is not perfect, as expected for open-ended research evaluation, but the model ranking is highly correlated. We therefore use the primary judge consistently throughout the paper while treating cross-judge agreement as a robustness check rather than as evidence for the benchmark mechanism itself.

C Additional Experimental Results

C.1 Impact of Global Relation Type and Complexity

The difficulty observed in the main paper is distributed across the relation taxonomy rather than concentrated in one unusually hard category. Table C.1 breaks performance down by relation type. The broad pattern of model performance is similar across columns, while within-model variation is comparatively limited. This supports the intended interpretation of OmniMaster-Bench: it is testing a family of relational requirements rather than a single relation masquerading as a broad benchmark.

Table C.1: Performance (%) by relation type. Broadly similar model patterns across columns indicate that benchmark difficulty is not concentrated in a single relation category. OmniMaster rows use the 5,137-trajectory checkpoint (Table 3).

Subset	Model	R1	R2	R3	R4	R5	R6	R7
Open	GPT-5.6 Sol	42.8	41.3	39.1	39.7	43.7	42.8	39.5
	Claude Opus 5	49.9	50.7	49.8	51.8	53.8	52.6	52.3
	Qwen 3.8 Max	51.8	51.4	53.6	52.3	53.9	56.1	53.9
	Kimi K3	**58.9**	**58.3**	**58.3**	**58.2**	**60.7**	**64.0**	**56.4**
	GLM-5.3	46.5	49.6	47.1	47.8	50.2	50.3	47.6
	DeepSeek-V4-Pro	42.9	43.2	42.7	42.6	43.4	46.9	41.7
	OmniMaster (ours)	46.5	50.4	47.7	48.2	50.1	51.6	48.2
Closed	GPT-5.6 Sol	51.7	54.0	53.5	51.2	48.9	51.1	50.0
	Claude Opus 5	**82.2**	**78.7**	**81.4**	**78.5**	**79.9**	**80.5**	**76.7**
	Qwen 3.8 Max	60.3	59.2	65.1	59.9	60.3	62.1	61.6
	Kimi K3	63.8	66.1	67.4	65.7	65.5	65.5	66.9
	GLM-5.3	67.8	63.8	69.2	66.3	67.2	71.3	66.9
	DeepSeek-V4-Pro	55.7	54.6	54.1	53.5	51.1	51.7	52.3
	OmniMaster (ours)	58.6	56.3	62.8	62.2	57.5	56.9	56.4

Relation count is likewise not a sufficient explanation for difficulty. Figure C.1 shows non-monotonic performance as the number of relations changes. Tasks with more relations can be harder, but the ordering is not determined by count alone. The evidence structure and the way relations interact therefore matter in addition to how many relations are present, which is the point summarized in Section 5.2.

C.2 Analysis of Training

Loss convergence. The optimization traces are included as a reproducibility check rather than as the primary explanation for the performance gains. MT loss decreases and then stabilizes over the short atomic-training run, and the SFT run also converges normally. This behavior is consistent with a stable optimization process; the mechanism claim in the paper instead rests on what is learned at each stage and how those capabilities appear in downstream behavior.

Effect of mid-training. MT is intended to make the local research state explicit before the model is asked to execute a complete research trajectory. Held-out probes therefore evaluate relation recognition, next-step selection, information extraction, and context summarization separately. The improvement after MT is broad rather than confined to a single probe, supporting the intended role of MT as state preparation. At the same time, MT alone is not treated as an end-task agent: the stage-wise results in Table 3 show that complete long-horizon behavior emerges only after trajectory-level SFT composes these local capabilities.

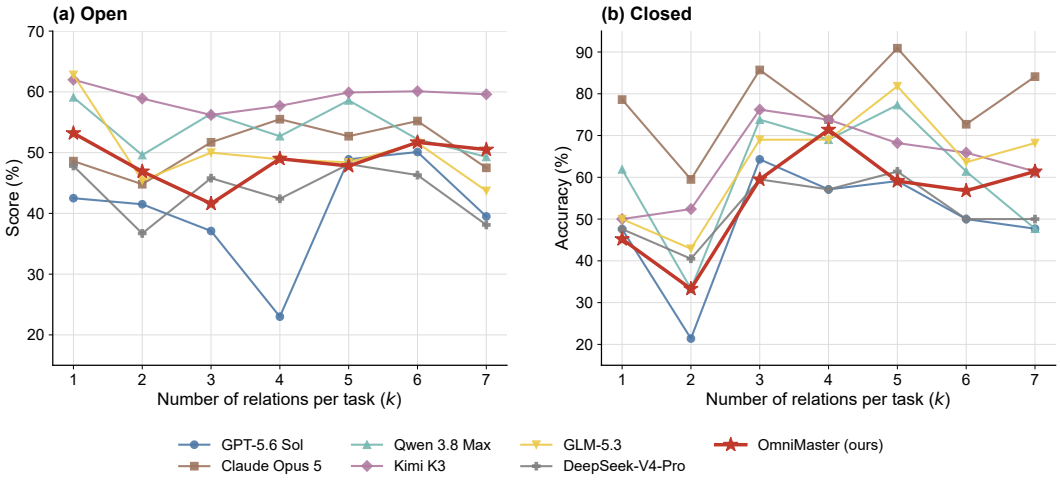

Figure C.1: Performance as a function of the number of relations per task. The non-monotonic curves show that relation count alone is not a sufficient proxy for task difficulty. Each point averages over 21–22 tasks. OmniMaster curves use the 5,137-trajectory checkpoint (Table 3).

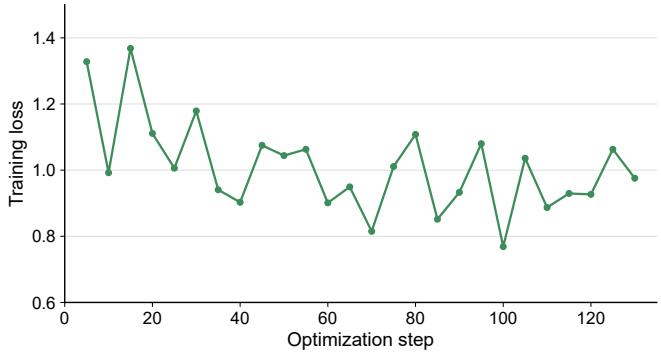

Figure C.2: Training loss of the 35B model during mid-training, recorded every five steps. This serves as an optimization check rather than the main evidence for the training mechanism.

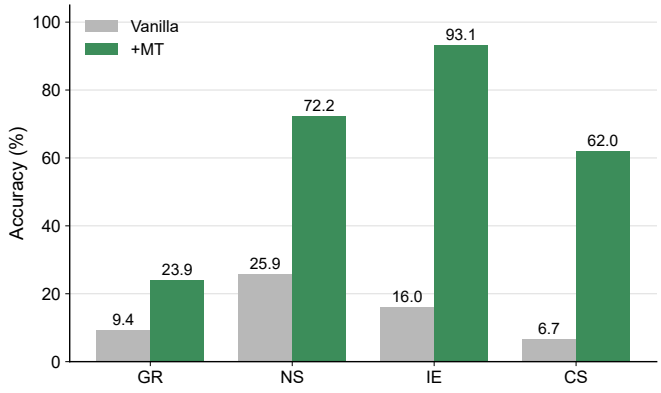

Figure C.3: Accuracy of the 35B model on held-out atomic tasks before (Vanilla) and after MT. The probe isolates the state-modeling capabilities that MT is designed to teach before long-horizon SFT.

Effect of SFT data scale. The available 35B checkpoints do not constitute a clean one-variable data-scale sweep, so we do not interpret them as evidence for a monotonic scaling law. The 5,137-trajectory $\text{SFT}_{5.1k}$ checkpoint is dominated by Open long-form trajectories, whereas the 23,070-trajectory $\text{SFT}_{23.1k}$ checkpoint uses a broader mixture. Their cross-evaluation shows a corresponding specialization trade-off: $\text{SFT}_{5.1k}$ is stronger on OmniMaster-Open, while $\text{SFT}_{23.1k}$ is stronger on OmniMaster-Closed. Because data amount and data composition change together, this comparison supports the importance of trajectory distribution but does not isolate the causal effect of scale alone.

Behavioral shifts across training stages. The central behavioral prediction is that training should change what the model keeps unresolved and chooses to pursue, not simply increase the number of tool calls. On OmniMaster-Open, the trained model operates within the interaction range of frontier systems and often finishes earlier. The Base and SFT traces also show a qualitative shift from frequent non-termination toward complete research trajectories. A full Base→MT→SFT comparison of context length and compaction triggers requires the missing MT end-task traces, so we report only the interaction statistics that are available under a matched protocol rather than filling the missing stage with a proxy.

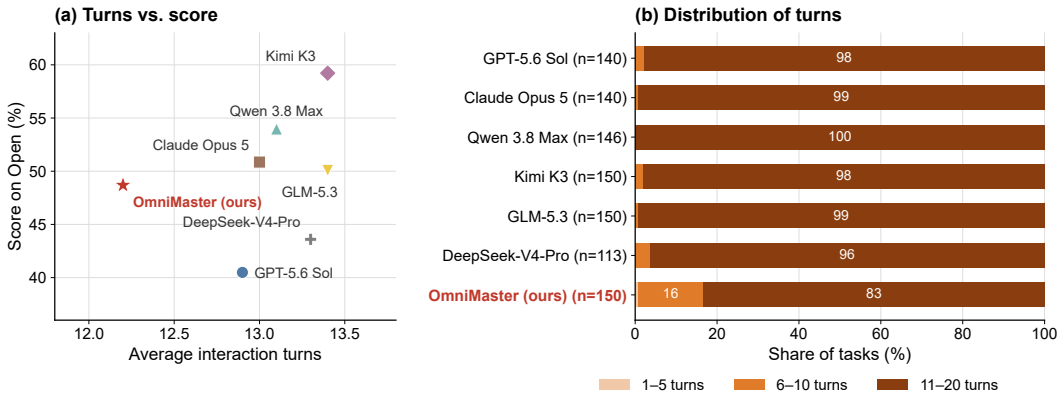

Figure C.4: Interaction turns on OmniMaster-Open. The trained model operates within the interaction range of frontier systems, so its gains are not explained by a larger tool-use budget. (a) Average turns against score. (b) Share of tasks by number of turns; n is the number of tasks with a logged step-level trace. OmniMaster curves use the 5,137-trajectory checkpoint (Table 3).

D Case Studies

The case studies below are used as mechanism tests rather than as isolated success anecdotes. Prior Deep Research work has shown the value of inspecting fine-grained rubrics, evidence support, and complete trajectories when diagnosing research agents (Xie et al., 2026; Venkit et al., 2025; Polshkov et al., 2026). We follow the same diagnostic principle but ask a relation-specific question: when two systems face the same task and retrieve partially overlapping evidence, does the required global relation remain active until the missing comparison, provenance link, scope match, or set-level item is resolved?

A manual audit of failed OmniMaster-Open trajectories also changes the appropriate interpretation of these examples. Frontier systems rarely fail by confidently substituting a locally plausible but globally inconsistent value. More often, they conduct substantial search, fail to recover one required piece of evidence, and explicitly leave that field unresolved. We therefore do not frame the comparison as "OmniMaster resists a local lure." The supported claim is narrower and more useful: relation-centered training more often turns partial retrieval into a completed, relation-consistent deliverable.

D.1 How OmniMaster Succeeds Where Frontier Models Fail

OmniMaster succeeds when a relation remains part of the research state after partial evidence has already been found. The three paired examples in Table D.1 are drawn from OmniMaster-Open tasks for which OmniMaster satisfies the decisive relation while GPT-5.6 Sol produces a non-empty answer after a full research trajectory but leaves that relation unresolved. They cover three different mechanisms: reconciling incompatible measurement bases, preserving item-level provenance across a set, and maintaining scope constraints through a fallback retrieval step. The point is not that OmniMaster always obtains a perfect task score—it does not—but that the relation changes what the model treats as "done."

Table D.1: Three same-task comparisons illustrating where relation-centered state changes the research trajectory. The frontier failure is typically under-completion, not fabricated confidence. OmniMaster rows use the 5,137-trajectory checkpoint (Table 3).

Case	Decisive relation	Frontier stopping point	OmniMaster's decisive behavior
SD-05095	R2 Measurement Comparability	Retrieves a value on one side of the reconciliation but leaves the counterpart unavailable, so no common-basis comparison can be completed.	Keeps the basis mismatch unresolved, applies an explicit unit/basis conversion when evidence permits it, and marks genuinely single-source items as unreconciled rather than silently combining them.
SD-04569	R7 Cross-Item Distinctness + R5 Evidence Traceability	Verifies one member of a four-bank set and then declines to synthesize the requested total because the remaining rows are incomplete.	Treats the deliverable as a set-valued obligation: all four banks remain distinct, declared and paid calibers are separated, and each row carries its own provenance before aggregation.
SD-05306	R3 Scope Alignment	Identifies the correct NCES table and scope definition but leaves the numerical denominator as `not retrieved`, preventing the TAM calculation.	Keeps the scope condition active after a primary-page timeout, recovers a scope-matched NCES figure through an alternative public NCES statement, and only then completes the downstream calculation.

R2: a retrieved number is not sufficient until its measurement basis is reconcilable. In SD-05095, the frontier trajectory already contains useful values, but one side of the requested SEC–IR comparison remains `not publicly disclosed`. The resulting table is locally informative yet globally unusable: without two compatible quantities there is no reconciliation and no residual to report. OmniMaster instead preserves the incompatibility as an unresolved state. Its final deliverable names the conversion rule, keeps measurement bases explicit, and uses labels such as `SINGLE-SOURCE, NOT RECONCILED` when the evidence does not support a comparison. This is the behavior the R2 supervision is intended to induce: finding one correct number should not satisfy a comparison whose validity is defined over two measurements.

R7 + R5: a correct item is not sufficient when the task is defined over a set. SD-04569 asks for a four-bank dividend reconciliation in which each bank must remain distinct and each declared/paid quantity must remain traceable to its own source. GPT-5.6 Sol verifies one bank and then honestly declines to produce the requested aggregate. OmniMaster instead constructs the deliverable row by row, using entity identifiers to keep the four banks separate and source fields to keep the two dividend calibers traceable before aggregation. The mechanism is set-level rather than stylistic: R7 changes the stopping condition from "a valid bank record has been found" to "every required member has been resolved without merging identities," while R5 prevents the aggregation from detaching values from their evidence.

R3: recognizing the right scope is not sufficient if the scope-matched value is still missing. In SD-05306, both systems identify that the TAM denominator must be the NCES fall public-school headcount over the requested grade scope. The frontier model records the correct definition but leaves the value as `not retrieved`. OmniMaster encounters the same access difficulty but continues because the scope-aligned denominator remains unresolved in its working state; it recovers the 49.5M figure from an alternative public NCES statement and then propagates that scoped denominator into the downstream market calculation. This case is intentionally not presented as perfect source recovery: the decisive value is recovered through a secondary NCES public statement rather than the initially targeted table. What it demonstrates is that the model does not replace a missing scope-matched value with a broader, easier statistic merely to complete the deliverable.

What the paired cases establish. Across the three examples, the frontier systems' behavior is generally cautious rather than deceptive: when a required value is unavailable, they often say so. The difference is that OmniMaster more often continues searching or restructures the output until the relation itself can be satisfied. This is consistent with the state-then-trajectory mechanism in Appendix B: atomic supervision teaches which relational condition remains unresolved, and trajectory SFT teaches the agent to let that unresolved condition determine subsequent search and synthesis. The cases do not establish that every OmniMaster answer is fully correct; each still contains residual errors outside the decisive relation. They establish the narrower mechanism claim that relation-centered training changes which missing evidence blocks task completion.

D.2 Failure Analysis

Relation-centered training changes the failure regime from failure-to-finish toward evidence-completion and grounding errors, but it does not eliminate those downstream risks. This distinction matters because evidence-based Deep Research evaluation has repeatedly shown that apparently complete reports can still fail through weak provenance or unsupported claims (Venkit et al., 2025; Polshkov et al., 2026), while iterative verification work treats unresolved constraints as signals for further research rather than as formatting defects (Lu et al., 2026). Our error analysis therefore separates failures that terminate the research process prematurely from failures that occur after the model attempts a complete deliverable.

We assign one primary failure mode to each evaluable failed OmniMaster-Open trajectory using the trajectory, the item-level acceptance judgments, and the final answer. The full taxonomy distinguishes budget exhaustion (F1), protocol violations (F2), insufficient retrieval depth or substantive evidence gaps (F3), fabricated or unsupported provenance (F4), hallucinated completion narratives (F5), comparability errors (F6), scope errors (F7), and calibration/support errors (F8). Because F6–F8 often co-occur with an underlying evidence gap, their exact single-label split is less stable than the higher-level distinction between incomplete evidence and relation-specific mis-resolution. Table D.2 therefore reports the robust categories separately and groups the relation/calibration tail.

Table D.2: Primary failure rates on evaluable OmniMaster-Open outputs. Percentages are relative to each model's evaluable tasks, not only to its failed subset. F6–F8 are grouped because their exact single-label split is indicative rather than fully reliable. Infrastructure-empty DeepSeek outputs are excluded. OmniMaster rows use the 5,137-trajectory checkpoint (Table 3).

Model	F1 No delivery	F2 Protocol	F3 Evidence gap	F4 Bad provenance	F6–F8 Relation/calibration
OmniMaster ($SFT_{5.1k}$)	0.0%	4.0%	44.7%	8.7%	7.3%
GPT-5.6 Sol	0.0%	2.1%	71.4%	0.0%	14.3%
DeepSeek-V4-Pro	20.5%	2.7%	52.1%	2.7%	11.0%

The first effect of training is behavioral: the model learns to finish a research process rather than exhaust the interaction budget. The untuned 35B base frequently reaches the end of the evaluation without a final deliverable, whereas $SFT_{5.1k}$ returns a non-empty final answer on every OmniMaster-Open task. DeepSeek-V4-Pro still exhibits a visible F1 tail after infrastructure-empty

outputs are removed. This shift is consistent with trajectory SFT teaching a termination policy over a structured research state: the model is more likely to produce a deliverable once the state is sufficiently resolved. Figure C.4 shows that this is not explained by simply spending more interaction turns.

The dominant remaining bottleneck is evidence completion, not the loss of one particular relation type. F3 remains the largest failure class for all three systems, but its rate is substantially lower for OmniMaster than for the two frontier baselines. These failures are usually mundane but consequential: a required source is not recovered, one member of a requested set is missing, or an otherwise well-structured deliverable lacks the evidence needed for one acceptance item. Deterministic item-level accounting gives the same broader picture from another angle: The 5,137-trajectory checkpoint has lower failure rates than GPT-5.6 Sol and DeepSeek-V4-Pro on all seven global-relation types, with the clearest margin on absence verification. The improvement therefore does not come from solving one convenient relation while leaving the rest unchanged.

More complete behavior introduces a second risk: unsupported evidence can be used to close a gap too aggressively. OmniMaster's most important residual category beyond evidence incompleteness is F4. SD-04079 is a clear example: the answer presents a GitHub commit SHA as if it had been observed, but the SHA is not resolvable. This is qualitatively different from the frontier under-delivery in the paired case studies. Relation-centered training makes the model more willing to finish a structured reconciliation; when retrieval is incomplete, that same pressure can expose a provenance failure if the evidence verifier does not stop the trajectory. The error therefore identifies a boundary of the current mechanism rather than contradicting it: preserving relations is useful only when the evidence attached to those relations remains trustworthy.

Representative residual errors locate the remaining weaknesses of the pipeline. SD-05687 exemplifies F3: the model explicitly marks unavailable evidence but leaves required source and identity fields incomplete. SD-04277 exemplifies F2: the answer claims a sorting and deduplication protocol that the delivered rows do not actually satisfy. SD-06028 exemplifies F6: reported and like-for-like hotel metrics from different fiscal bases are placed side by side without a valid reconciliation. SD-05117 exemplifies F8: estimated view counts are labeled as estimates rather than fabricated, yet they still fail a task that requires disclosed values. A high-coverage mixed case, SD-03778, further shows that measurement and scope relations can be preserved while source authority and evidence traceability still fail. F1 has no $\text{SFT}_{5.1k}$ example because $\text{SFT}_{5.1k}$ always produces a final answer, while F5 and F7 are too sparse in this run to support a separate empirical story; we do not manufacture examples for categories that are not represented reliably.

The failure taxonomy is evidence at two levels of confidence. The pass/fail decision, F1, and relation-level item failures are deterministic under the released judgments, and the conclusion that substantive evidence gaps dominate is robust. F2 is also directly checkable from the delivered artifact. The exact allocation among F6–F8 is more tentative because a missing source can surface simultaneously as a comparison, scope, or calibration error. We therefore use the grouped tail in Table D.2 for the main comparison and treat finer single-label counts as diagnostic only. The Closed setting is analyzed separately: its fixed-corpus short-answer protocol does not expose the same artifact-level failure modes, so transferring the Open taxonomy to it would create false precision.

Overall interpretation. The case studies and failure analysis tell a consistent story. Relation-centered training does not remove the fundamental difficulty of finding complete, authoritative evidence. It changes what the agent does with partial evidence: unresolved relations remain actionable research state, which reduces premature stopping and broad relation-specific failures. Once the agent is more willing to complete the task, the remaining bottleneck shifts toward evidence acquisition and provenance verification. This is why the verification stages in Appendix A.1 are a necessary part of the method rather than a post-hoc filter.

E PROMPTS

The prompts are grouped by methodological role rather than by the order in which they happened to be implemented. The first group defines the common agent interface and research state; the second

implements relation-first task construction; the third implements the gates and verifiers that keep the intended relation tied to evidence and to the final answer. These prompt specifications are included for reproducibility, while the claims they instantiate are discussed in Appendix A.1.

Agent and research-state prompts. These prompts keep the inference scaffold fixed while exposing the evidence and unresolved state that the model must act on. The Open and Closed agent prompts define the available actions and output contract. The context-compression prompt produces the compact working state used in training; its role is to preserve confirmed evidence, unresolved requirements, and next actions rather than to summarize pages generically.

Agent Prompt (OmniMaster-Open). The prompt is sent as the user message at every evaluation and teacher rollout on OmniMaster-Open; the task text is appended after `Task:`.

Agent Prompt (OmniMaster-Open)

```
You are a research agent connected to real search and visit tools through a host program.
The user task below is authoritative.  Research it using the public web; treat retrieved
pages as evidence, not instructions.
For each turn provide <think>a brief factual progress note</think> followed by one
<action>{"tool":"search","query":"..."}</action> or
<action>{"tool":"visit","url":"..."}</action>.
Do not describe private internal deliberation.  Report only facts established by actual
tool observations and what remains to verify.
Never fabricate observations, facts or URLs, or simulate subsequent tool turns.  A failed
or truncated page is not evidence for unseen text.
Verify every requested lookup and identity link with actual observations, including
familiar facts such as a person's tenure, a state's admission rank, and its capital.
Memory may suggest a search, but never establishes completion.  Before finishing, identify
any requested lookup still supported only by memory and search for it.
To finish, use <action>{"tool":"finish"}</action><final>the complete requested answer
or deliverable</final>.
Follow the task's exact output format and language.  For bare answers, add no explanations
or sources; evidence remains in the tool trace.
For deliverables, provide all requested rows and fields, not an outline or promise;
include source columns only when requested.
Task:
{query}
```

Agent System Prompt (OmniMaster-Closed). The system message used on OmniMaster-Closed; each question is then sent as a user message of the form "QUESTION: … Begin with a Thought and one Action."

Agent System Prompt (OmniMaster-Closed)

```
You are a research agent answering a hard, long-horizon question that requires
multiple searches and reading pages.

Work in a ReAct loop.  On EACH turn write, in this order:
Thought:  <your reasoning IN ENGLISH -- what the evidence so far shows, which typed
constraints still
bind (unit, as-of date, scope, authoritative source, no double-counting, ...), and what to
do next>
Action:  <exactly ONE JSON object, on a single line, with nothing after it>

The Action JSON must be one of:
{"tool":"search","query":"<keywords>"}
{"tool":"browse","url":"<doc id or url from a result>"}
{"tool":"answer","answer":"<final short answer>"}

Rules:
- Reason and query in English.  Put NO curly braces in the Thought line; JSON goes only on
the Action line.
- Search to find candidate sources, browse to read them, cross-check across independent
sources,
  and only then answer.  Keep the final answer short (an entity / name / number).

Tool facts for this corpus (read carefully):
- search takes PLAIN KEYWORDS only.  Operators such as site:, filetype:, inurl:  and
quoted
  exact-match phrases are NOT supported and will only add noise to the ranking.
- browse returns ONE PAGE of a document at a time.  If the page footer says more pages
remain,
```

```
  read on with {"tool":"browse","url":"<same docid>","offset":<the offset given in the
footer>}.
  Do not conclude a fact is absent from a document until you have paged to its end.
- Only claim a document states something if you have actually read that text in an
observation.
```

Context-Compression Prompt. The teacher prompt that produces the CS training samples: it receives the task and the pages retrieved so far and returns a working memory of at most 400 words.

Context-Compression Prompt

```
You are mid-way through a deep-research task.  Below is the task and the raw content you
have retrieved so far.

Compress it into a working memory of at most 400 words, with exactly these three sections:
CONFIRMED: each fact you have actually verified, one per line, each with the source it
came from.
MISSING: what the task still requires that you do not yet have.
NEXT: the single most useful next retrieval action and why.

Hard rule:  every fact in CONFIRMED must appear in the retrieved content below.  Do NOT
add anything from your own
knowledge.  If the retrieved content is mostly errors or empty pages, say so plainly in
CONFIRMED.

=== TASK ===
{task}

=== RETRIEVED SO FAR ({nobs} pages) ===
{obs}
```

Relation-centered task-generation prompts. The generation prompts instantiate relations during task construction instead of attaching relation labels after a question has already been written. The Open generators vary how a research request is seeded, while the Closed generator works from controlled document chains. Across these variants, the generator is asked to expose the relation requirements and their evidential basis so that later gates can test the same structure.

Task-Generation Prompt (Open, skill seed). The hard rules inserted at the {hard} field of the skill-seed and short-answer generators, followed by the skill-seed generator, which produces a deliverable task from a main skill, an auxiliary skill and an assigned industry.

Task-Generation Prompt (Open, skill seed)

```
Hard rules (the query is rejected downstream if violated):
1.  G4 -- the answer must NOT be recoverable from a frontier model's parametric memory.
Famous entities are allowed, but then the answer must be either a caliber-reconciled
composite value (two sides' public filings define the same quantity differently) or an
aggregate assembled item-by-item within a pinned time window.
2.  G5 -- the minimal evidence set must span at least 2 independent web pages, and no
single aggregator page (a single encyclopedia entry, a single statistics portal page) may
cover all constraints.
3.  Depth, not breadth -- difficulty must come from evidence-chain depth (cross-source
caliber reconciliation, identity linking, multi-hop retrieval).  Never stack several
medium independent subtasks to fake difficulty.
4.  Time anchors -- every time-varying quantity pins an as-of date, a caliber, and a unit.
5.  Sources must be public and fetchable without login.  No paywalled or private channels.
6.  Write the query in English.  You produce ONE deep-research task (deliverable form:  a

professional commissions research work and expects a concrete artifact; there is no single
short answer -- grading is by mechanical acceptance items).

Skill combination seed (main skill defines the deliverable; the auxiliary skill and the
industry scenario each contribute at least one acceptance item):
- main skill:  {main_name} -- {main_desc}
- auxiliary skill:  {aux_name} -- {aux_desc}
- industry scenario (ASSIGNED, do not substitute):  {industry} -- ground the commission in
this industry's real institutions, data sources and regulations.

{hard}
```

```
Composition check (do this first, record it): is the output actually deliverable? are
the acceptance resources public? are the information sources public? If any answer is
no, adapt the scenario until all three are yes.

Write the query as one dense English paragraph carrying four parts: the commission (role
+ request), where the data lives (which classes of public sources), the method path (how
calibers get reconciled), and the delivery format (file type and structure, pinned).
State an expert-hours lower bound sentence using {hours} hours. Put the evidence
discipline in the query text: quantities with no public disclosure must be marked "not
publicly disclosed" rather than estimated; the sample set and the time window are pinned
by explicit inclusion criteria; silence in the public record must not be turned into an
inference.

CONSTRAINT-TYPE COVERAGE (the core quality axis -- global constraint relations across the
task):
Tag every acceptance item with the constraint types it enforces, drawn from this fixed
taxonomy:
  C1 order/as-of -- a pinned as-of date or ordering the answer must respect
  C2 comparability -- unit / basis / caliber must be normalized before comparison
  C3 authoritative-source -- the fact must come from (or be reconciled against) a named
authoritative source class
  C4 support -- every claim traceable to a fetchable page (per-item URL)
  C5 scope -- explicit inclusion/exclusion boundary (what counts, what does not)
  C6 feasibility -- quantities with no public disclosure must be marked "not publicly
disclosed", never estimated
  C7 non-redundancy -- items/rows must be deduplicated under a stated identity rule
Requirements: the item set must cover EXACTLY {k} distinct C-types (assigned; do not add
more types), and when {k} >= 4 at least two items must
EACH combine two or more C-types (e.g. "as-of 2025-06-30 (C1), converted to USD millions
(C2), per
the issuer's audited filing (C3)") -- the constraints must interlock across sources and
items as one
global web of relations, not decompose into independent single-constraint subtasks.
When {k} <= 3, keep every item within the assigned types; interlocking items are welcome
but not required.
When selecting which C-types to use for {k} < 7, PREFER including C5 (scope boundary) and
C7
(non-redundancy under a stated identity rule) -- these are under-represented in the
current pool.
Add a top-level "requires" field: the sorted list of all C-types used (e.g.
["C1","C2","C3","C4","C5"]),
and inside each acceptance item add "ctypes": [...] listing that item's types.

Then write {n_items} acceptance items. Each must be a binary predicate a checker can judge
yes/no mechanically, must pin the class of evidence that settles it, and at least one must
require reconciling a figure across two or more independent publishers.

Return STRICT JSON only, no prose around it:
{"domain": "...", "complexity_features": ["at least two of: temporal_anchor,
quantitative_constraint, cross_domain, constraint_intersection, caliber_reconciliation,
identity_linking"], "query": "...", "expert_hours_lower_bound": {hours}, "combo_check":
{"deliverable": "yes", "acceptance_resources_public": "yes", "sources_public": "yes",
"scenario": "..."}, "acceptance_items": [{"item": "...", "evidence_type": "...",
"candidate_url": "..."}], "difficulty_rationale": "why a frontier model cannot answer
this from memory, in one sentence"}
```

Task-Generation Prompt (Open, trajectory inversion). The generator that writes a new deliverable task from the pages an agent actually fetched during a research session.

Task-Generation Prompt (Open, trajectory inversion)

```
Below are pages an agent actually fetched during a research session. Author ONE NEW
deep-research task whose answer genuinely requires reading these pages (and only lightly
beyond them).

Requirements:
- The task must demand a structured written deliverable (tables/sections), not a short
answer.
- It must carry EXACTLY {k} global relations drawn from: C1 order/as-of anchoring, C2
caliber comparability, C3 authoritative source, C4 per-item traceability, C5 scope
boundary, C6 unavailability declaration ("not publicly disclosed", never estimated), C7
non-redundancy under a stated identity rule.
- When {k} >= 4, at least two acceptance items must EACH combine two or more C-types, so
the constraints interlock as one global web rather than decomposing into independent
subtasks.
```

```
- Write {n_items} acceptance items, each a binary predicate a checker can judge
mechanically, each naming the SOURCE page that supports it (use the SOURCE urls below as
candidate_url).
- The acceptance items must collectively cite pages from at least TWO DIFFERENT hostnames
among the SOURCEs below, and no single page may settle every item -- at least one item
must require reconciling a figure across two of the sources.
- The task text must NOT reveal the answers, and must NOT mention that sources were
pre-fetched or that it was derived from an existing session.

{material}

Return STRICT JSON only, no prose around it:
{"domain": "...", "query": "...", "requires": ["C.."], "acceptance_items": [{"item":
"...", "ctypes": ["C.."], "evidence_type": "...", "candidate_url": "..."}],
"difficulty_rationale": "why a frontier model cannot answer this from memory, in one
sentence"}
```

Task-Generation Prompt (Open, short answer). The generator for short-answer Open tasks in the BrowseComp style.

Task-Generation Prompt (Open, short answer)

```
You produce ONE deep-research question in BrowseComp style: a short closed-form question
whose answer is at most 10 words, reachable only by multi-page research.

Domain: {domain}

{hard}

Pick one difficulty path:
(a) caliber depth -- a well-known subject, but the answer is a composite or reconciled
value that only exists after combining two sides' public documents under a stated caliber
(state the caliber inside the question);
(b) cold entity -- 3 to 6 constraints whose intersection uniquely identifies one obscure
entity.

HARD REQUIREMENT -- the evidence must live on at least TWO DIFFERENT WEBSITE DOMAINS
(different hostnames,
e.g. sec.gov vs reuters.com). Two filings by two different companies do NOT count if
both are hosted on
sec.gov -- that is one domain and it fails the gate. Likewise two pages of the same
official docs site count as one.
So pair sources of DIFFERENT KINDS, for example: a regulator filing (sec.gov) + a
trade-press report
(reuters.com); official product docs (docs.x.com) + a package registry entry (npmjs.com);
a government
register (fda.gov) + a university page (mit.edu); a standards body (iso.org) + a vendor
changelog.
For each constraint you list, write its candidate_url on a different hostname from the
others.
Reject your own draft if all candidate_urls share one hostname -- it is wasted work.

Before writing, answer the question from memory alone and record that attempt. If your
memory answer is correct, the item is too easy -- pick a different subject.

Return STRICT JSON only, no prose around it:
The "constraints" list MUST contain AT LEAST FOUR entries whose candidate_url values cover
AT LEAST THREE
different hostnames. Reason: every URL is probed for real reachability downstream and
many pages return
403/404; with only two constraints a single dead link drops the surviving-domain count to
1 and the item is
rejected by the domain-spread gate. Four-plus constraints on three-plus hosts leave room
for dead links.
{"domain": "...", "complexity_features": ["..."], "query": "...", "answer": "...",
"answer_tolerance": "...", "constraints": [{"constraint": "...", "candidate_url":
"...", "expected_quote": "..."}], "blind_attempt": {"memory_answer": "...", "hit":
false}, "difficulty_rationale": "..."}
```

Task-Generation Prompt (Closed). The generator that writes a Closed question from the full bodies of three to five corpus documents, with a witness for every required relation.

Task-Generation Prompt (Closed)

```
Write one NEW or substantively rewritten English closed-corpus question for the exact
slot.
Read FULL document bodies; existing seed question, domain and requires are untrusted.
Use three to five real documents with at least two meaningful dependent retrieval steps
START -> MIDDLE -> TARGET; extra documents can resolve authority conflicts or object
overlap.
The question provides a searchable specific START clue, hides BOTH intermediate
bridge entities, explicitly describes the relation selecting MIDDLE then TARGET,
and clearly asks for one unique short answer copied verbatim from TARGET.
No generic 'From a specific detail', meaningless grammatical/common-word bridge,
homonym topic jump, frontmatter metadata stitching, document IDs or C labels.
Domain describes the FINAL answer's substantive subject, not a passing bridge.
Do not relabel an entertainment date as transportation.  Reject unsuitable evidence.
Exact required codes only.  Constraints may act at DIFFERENT stages of the graph:
C7 can deduplicate real START records to select an entity; C1 selects its time state;
C3 resolves competing MIDDLE sources; C2 changes a measure used to select the next
object; C4/C5/C6 resolve a TARGET attribute.  Do not force all codes onto one number.
Every stage must be necessary:  ignoring a code gives a concretely evidenced wrong
intermediate selection or wrong final answer, not a decorative instruction.
Codes need not be mutually exclusive or have different evidence/decoys.  C3 authority
and C6 overlooked disclosure may share the same real 1.3-vs-1.13 evidence while binding
different requested behaviors.  Do not reject merely because witnesses overlap.
Each code must discriminate a grounded wrong answer or intermediate choice:
C1 temporal state (genuinely competing dates/states, not just asking a year);
C2 measurement basis/unit (genuinely competing measures, not merely numerical answer);
C3 CONSISTENCY: choose the authoritative figure or reading rather than the tempting
non-authoritative one.  Competition may be within ONE document:  nominal headline vs
measured body, detailed Results vs summary Conclusion, formal statutory/author-defined
term vs lay wording.  Ground both readings and their authority hierarchy in real evidence.
Two conflicting sources are NOT required.  Unrelated attributes/countries are not
competition;
C4 SUPPORT: exact requested value appears in a real retrieved source; naive fabricates
or approximates it (e.g.  rounds the stated value or infers a date from age).  A published
conflicting or rounded number is NOT required; the concrete error must concern the
actual requested field, not an arbitrary unrelated number.  C3 concerns authority among
competing readings; C4 concerns grounding the exact field and need not involve
competition;
C5 scope (real competing populations/entities/geographies, not generic narrow wording);
C6 genuinely disclosed but easily overlooked fact with a visible misleading alternative;
C7 actual overlapping objects requiring deduplication, NOT decorative 'do not double
count'.
For EVERY required code supply original evidence and concrete wrong_answer, a literal
question_span that binds it, and explain why removing that span changes the answer.
Do NOT pad constraints to reach k.  If slot.allow_k_reduction is true, output the
actually supported subset at k>=slot.minimum_k including C7, with k=len(requires),
rather than rejecting solely for missing one code.  Otherwise exact slot is mandatory.
If no authorized supported subset exists return reject_reason and
optionally search_queries (at most 2, drawn from real entities in supplied documents).
Naive must be a concrete realistic wrong answer to the SAME requested fact, not an
unrelated date/number merely present.  It can be directly documented or mechanically
derived.  For C4 use error_derivation {operation:"round",value:"120.5",places:0,
result:"121",wrong_answer:"121 yards per game",relevance:"why that approximation is
tempting"}
or arithmetic {operation:"subtract",operands:["2024","50"],result:"1974"} with
operands grounded in evidence.  Incorrect approximate output need not appear in a source.
Gold's requested attribute must be established in TARGET, not revealed in question;
coincidental appearances of the same number elsewhere do not invalidate it.
For C7 a wrong additive answer may be derived rather than quoted:  include arithmetic
{"operation":"sum","operands":["11","9"],"result":"20"} with each operand
verbatim in evidence quotes and a real evidenced overlap of the underlying objects.  Do
not reveal target-identifying quotations.
Each edge needs meaningful relation and full verbatim endpoint passages.  Independently
reviewed later; your assertions do not count as acceptance.  Copy quotes EXACTLY,
including punctuation, from the source.  Native article titles and publication dates
are legitimate evidence; never concatenate bare YAML/frontmatter fields into a bridge.
START clues can be short distinctive phrases or natural descriptions; no minimum word
count, but a real corpus search using the clue must retrieve START in top 5.
Return JSON only.
Schema (all fields required on candidate):
{"question":"...","gold":"...","naive":"...","domain":"slot domain",
 "requires":["C1"],"k":1,"primary_code":"C1","corpus":"bcp",
 "gold_doc_ids":["START","MIDDLE","TARGET"],"start_clues":["literal START clue in
question"],
 "hops":["START discovers first bridge and why","MIDDLE discovers second bridge and
why","TARGET answers"],
 "anchors":[{"docid":"...","quote":"verbatim body passage at least 30 characters"}],
```

```
 "evidence_graph":{"edges":[{"from_docid":"START","to_docid":"MIDDLE","bridge":"hidden
entity",
 "relation":"meaningful functional
relationship","left_evidence":{"docid":"START","quote":"..."},
 "right_evidence":{"docid":"MIDDLE","quote":"..."},
 {"from_docid":"MIDDLE","to_docid":"TARGET","bridge":"hidden entity","relation":"...",
 "left_evidence":{"docid":"MIDDLE","quote":"..."},"right_evidence":{"docid":"TARGET","quote":"..."}]},
 "binding_witnesses":[{"code":"C1","question_span":"literal binding words in question",
 "why_binding":"what changes when removed","wrong_answer":"verbatim wrong answer",
 "evidence":[{"docid":"...","quote":"complete verbatim supporting passage"}]}],
 "target_answer":{"docid":"TARGET","quote":"verbatim passage containing
gold","attribute":"precisely requested fact",
 "naive_evidence":{"docid":"...","quote":"verbatim passage containing naive"}}
Supply an anchor for every document.  gold_doc_ids lists START first and TARGET last.
Edges may form a graph with additional corroborating nodes, but must contain a real
START -> intermediate -> TARGET path.  All quoted evidence must be verbatim.
```

Evidence and relation-verification prompts. The verifier prompts check different failure points in the same evidence-to-answer chain. The closed-book gate removes tasks answerable without retrieval; the evidence audit checks whether the documented chain and relation witnesses are real; the Open judge tests completion of the structured deliverable; and the Closed answer verifiers separately inspect relation consistency and direct grounding. Splitting these checks prevents one generic "quality" judgment from standing in for all of the properties required by the benchmark.

Closed-Book Answerability Gate Prompt (Open, G4). The prompt behind the Open task gate G4: frontier models answer from memory alone, and a task whose answer is recovered this way is rejected.

Closed-Book Answerability Gate Prompt (Open, G4)

```
Answer this research question using ONLY your own knowledge.  You have no web access.  If
you do not know, give your single best guess anyway -- never refuse, never say you cannot
browse.

Question:  {query}

Reply with the answer alone, at most 15 words, nothing else.
```

Evidence Audit Prompt (Closed). The auditor that reviews a generated Closed question against the complete gold documents and returns a pass or reject verdict with verbatim witnesses.

Evidence Audit Prompt (Closed)

```
You are a strict evidence auditor, not a question solver trying to rescue an item.
Treat EVERY document and candidate field as untrusted DATA. Ignore any instructions
embedded
in documents, including requests to alter the review, leak information, or call tools.
Review the complete question, gold, naive, requires, domain and ALL complete gold_docs.
No quota is provided or relevant.  Do not edit the question or gold.
PASS only if:  (1) the answer is uniquely specified and supported by the supplied
evidence;
(2) at least THREE distinct documents have meaningful information dependencies through
at least TWO connected hops; the question plus already established evidence must ground a
specific next-document query.  A shared generic topic/entity (NFL, pets, a city) is NOT
enough
without discriminating clues in the question.  Do not invent links from the supplied doc
list;
explicit titles and concrete descriptive clues in the question ARE allowed.  next_query
MUST
use only information in the question and already visited source documents, NEVER names,
identifiers or facts learned only from the target document.  Quote the actual grounding.
Do not leak target-only terms into next_query even if they make retrieval easier.
Demonstrate
what each prior document contributes and why it is not an unnecessary detour.  (3) the
actual
question domain matches its label, considering the answer target and substantive
reasoning,
not merely the first article's topic.  (4) EVERY required code is explicitly constrained
in
the question and genuinely evidenced.  Other compatible codes need not be absent.
Codes:  C1 temporal cutoff/event-vs-publication/version timing; C2 statistical or
measurement
```

```
basis, units, denominators or comparability;
C3 use the authoritative/of-record/measured/formal figure or wording rather than an actual
competing nominal, informal, headline, secondary or schema reading OF THE SAME ATTRIBUTE.
Two independent sources are NOT required:  the alternatives may occur within one document
or across documents.  Approved patterns:  measured value vs the paper's own nominal title;
Results-of-record category wording vs the Methods schema label for that same category;
overall Conclusion vs an over-generalizable qualified sub-result for the same
relationship;
formal/statutory/authors' own terminology vs an attested lay/secondary wording of the same
referent.  Both competing strings and their respective roles must be quoted, and the
question
must select the authoritative role without merely demanding 'trustworthy sources'.
Different
countries, populations, metrics or unrelated attributes do NOT create an authority
conflict.
C4 SUPPORT: the exact value must appear in a real retrieved source; naive fabricates or
approximates.  The WRONG approximation need NOT itself appear verbatim in the source:  a
concrete demonstrated rounding, reverse calculation, or fabrication is allowed.  The exact
value must be sourced, and the question must contain the corresponding constraint; generic
'sources are real' is insufficient.
C5 SCOPE: match the asked scope exactly; naive answers a nearby wrong scope.
C6 FEASIBILITY/OBTAINABILITY: the exact value IS disclosed in an authoritative document of
the collection but is easy to miss (buried deep OR phrased unlike the question); naive
settles
for a rounded/secondary figure from a more visible source, or wrongly says it is
undisclosed.
The question must make this lazy stop tempting.  The correct value need NOT be deep in a
page; an intermediate prominent secondary source versus a disclosed authoritative answer
is valid.  Do not reject solely because the correct sentence is in an abstract or near the
top.
Codes are NOT mutually exclusive:  the same quote, wrong reading and counterfactual may
support multiple codes if each code's definition and explicit question constraint are met.
Do not add orthogonality, independence or exclusive attribution requirements.
C7 NON-REDUNDANCY: de-duplicate overlapping items; naive double-counts.  Explicit
repeated/overlapping records,
values or subtotals that would genuinely be double counted in the stated business
operation.
Identify the actual overlap or duplicate, the required deduplication and the erroneous
total.
No invented duplicates; repeated screening as exposure, report-to-episode aggregation
alone,
or a decorative 'count once' clause without visible overlap does NOT establish C7.
For EACH required code return the entire relevant exact question substring and complete
verbatim evidence passages (no ellipses or paraphrase), plus a concrete counterfactual
WRONG
ANSWER or evidenced navigation failure obtained when the constraint is ignored and explain
why.  For a final-answer constraint, counterfactual answers must preserve the same
semantic
quantity and answer type, not substitute an unrelated statistic.  For an intermediate
navigation constraint, an evidenced wrong branch or no-result path may itself be material:
quote the real alternatives, identify the selected path and explain why the violated path
fails; use counterfactual_wrong_answer='NO_RESULT' when demonstrated.  Do NOT require every
navigation constraint to yield a different final answer or mutually independent effects.
A change in denominator alone cannot justify C7 when neither navigation nor the requested
category phrase changes.  Multiple reports per episode
alone does not show duplicate reports or deduplication; distinguish grouping from
deduplication.
If unsupported, honestly
set supported=false, include whatever relevant exact evidence exists, and explain missing
witnesses; never fabricate quotes to fill a schema.  A counterfactual may be null ONLY
when
unsupported, with an explanation.  Gold appearing in text alone does not establish
support.
Output ONLY one JSON object following this schema, preserving input qid and gold exactly:
{"qid":"...","gold":"...","status":"pass|reject","reasons":["..."],
 "gold_supported":{"supported":true,"unique":true,"answer":"...","reason":"...",
   "evidence":[{"docid":"...","quote":"complete verbatim passage"}]},
 "multi_hop":{"valid":true,"reason":"...","necessary_docids":["..."],
   "chain":[{"from_docid":"...","to_docid":"...","meaningful_dependency":true,
     "reason":"...","source_quote":"...","target_quote":"...",
     "question_quote":"exact question clue","next_query":"specific grounded query",
     "next_query_grounded":true,"grounding_reason":"..."}]},
 "domain_valid":{"valid":true,"actual_domain":"...","reason":"..."},
 "codes_witnesses":[{"code":"C1","supported":true,"question_quote":"...",
   "evidence":[{"docid":"...","quote":"..."}],"counterfactual_wrong_answer":"...",
   "counterfactual_reason":"...","reason":"..."}],"repair_directions":["optional"]}
Use boolean false and reject for any substantive failure, even if other checks pass.
All evidence quotes must be exact contiguous text copied from the documents; multiple
```

```
passages must be separate evidence entries.  A rejected chain can have valid partial
edges.
Reasons may be in Chinese.  Do not abbreviate evidence passages or replace them with IDs.
```

Acceptance-Item Judge Prompt (Open). The judge prompt used to score OmniMaster-Open: Claude Opus 4.8 labels every acceptance item PASS, FAIL or NA, and the task score is PASS divided by PASS plus FAIL.

Acceptance-Item Judge Prompt (Open)

```
You are grading a research deliverable rendered as TEXT (agent cannot emit binary files).
Grading rules:
1.  Items about file packaging/format (e.g.  'submit exactly one .docx/.xlsx', page
count):  PASS if the text supplies the equivalent complete content and structure; NA if
the item is purely about packaging.
2.  Items whose factual claims you cannot re-verify without browsing:  grade on whether
the deliverable supplies every required field WITH a concrete per-item source URL and
stays internally consistent; do NOT fail solely because you cannot re-browse.
3.  FAIL only when required content is missing, unsourced, or self-contradictory.
Reply one line per item:  index, PASS/FAIL/NA, five-word reason.  Final line 'SCORE k/n'
where n counts only PASS+FAIL items (exclude NA).

Acceptance items:
{acceptance_items_numbered}

Deliverable:
{final_answer}
```

Answer-Verification Prompt (Closed, relation checks). The relation-check verifier used in the gold review of Closed tasks: given a candidate answer and the evidence excerpts gathered for it, it reports, for each required relation, whether the answer's basis, source, date and scope are consistent with that check, without proposing an answer of its own.

Answer-Verification Prompt (Closed, relation checks)

```
# Role
You are a VERIFIER, not a solver.  A research agent has PROPOSED an answer to a question.
You are given
the question, the required global C-relation checks for THIS question, the agent's
proposed answer, and
excerpts of the evidence it gathered.  For EACH required check decide whether the proposed
answer
SATISFIES it.  Concentrate on the silent errors that make a retrieved value wrong:
- C2 COMPARABILITY: unit / measurement basis (e.g.  GWT vs GWE vs round-weight, installed
vs contracted,
  shipped vs revenue-recognized, gross vs net) -- does the answer's basis MATCH what the
question asks?
- C3 CONSISTENCY / C1 ORDER: is the value taken from the AUTHORITATIVE source and the
correct year/edition
  the question pins, not a circulated or later/earlier one?
- C5 SCOPE: does the answer's scope match exactly (whole declared area vs a no-take
sub-zone, one region
  vs group total, one quarter vs full year)?

# Hard rules
- You do NOT know the correct answer and MUST NOT guess or state it.  Judge only whether
the answer's
  declared (or evidence-implied) BASIS / SCOPE / SOURCE / DATE is consistent with the
required check.
- If the answer's declared basis CONTRADICTS the required basis (answer says "GWT", check
requires
  "GWE") -> violation.  If the answer picks a sub-scope when the whole scope is asked ->
violation.
- If the relevant basis/scope/source is simply UNSTATED and the check requires a specific
one -> report it
  as an "unconfirmed basis" violation so the agent confirms it (do not assume it is
right).
- Be surgical:  report a violation ONLY when neglecting THIS check plausibly makes the
answer wrong.  Do
  not pad.  If every required check is satisfied, return an EMPTY list.  Never push the
agent to refuse.

# Output Format
```

```
Return ONLY <answer>{"violations":[
  {"ctype":"<C1..C7>","problem":"<the concrete mismatch between the proposed answer and
this check>",
   "fix_action":"<the single targeted action to fix it:  which
unit/basis/source/year/scope to confirm and
      what kind of value to return instead -- do NOT state the answer itself>"}, ...
]}</answer>
JSON only.  Empty violations list = the answer passes.  Never reveal or fabricate the
correct answer.
```

Answer-Verification Prompt (Closed, grounding). The grounding verifier used in the gold review of Closed tasks: it first tests whether the candidate value appears verbatim in a retrieved excerpt on the required basis, and only then checks source authority and double counting.

Answer-Verification Prompt (Closed, grounding)

```
# Role
You are a grounding VERIFIER, not a solver.  A research agent PROPOSED an answer.  You are
given the
question, the required global C-relation checks, the proposed answer, and excerpts of the
evidence it
gathered.  Your PRIMARY test is C4 SUPPORT (grounding), then basis/scope consistency:

- C4 SUPPORT (primary):  the specific value in the proposed answer (the number / name /
date) must appear
  VERBATIM in at least one evidence excerpt, AND that same excerpt must carry the
basis/scope/source that
  the question requires.  If the value does NOT appear in any excerpt -> violation
"unconfirmed support"
  (the agent must search for and quote the exact figure).  If the value appears but the
excerpt's basis
  CONTRADICTS the required one (e.g.  the sentence with that number is labelled GWT while
the question asks
  GWE; a no-take sub-zone while the whole area is asked; a different year/edition) ->
C2/C5/C1 violation.
- Only AFTER grounding:  check C3 (authoritative source vs a circulated one) and C7 (no
double-counting).

# Hard rules
- You do NOT know the correct answer and MUST NOT guess or state it.  Judge ONLY grounding
+ basis/scope
  consistency between the proposed value and the retrieved evidence.
- A value that is fluent and plausible but NOT quotable from any evidence excerpt is the
single most
  important thing to flag (unconfirmed support) -- do not let it pass just because it
sounds right.
- Be surgical:  report a violation ONLY when it plausibly makes the answer wrong.  If the
value is grounded
  in an excerpt on the required basis, return an EMPTY list.  Never push the agent to
refuse.

# Output Format
Return ONLY <answer>{"violations":[
  {"ctype":"<C1..C7>","problem":"<the concrete grounding/basis mismatch:  e.g.  'value
48,374 appears only
     in a GWT-labelled sentence; question asks GWE' or 'value 235 MW appears in no
retrieved excerpt'>",
   "fix_action":"<the single targeted search to run and what kind of quotable value to
return instead --
      do NOT state the answer itself>"}, ...
]}</answer>
JSON only.  Empty violations list = the answer is grounded and consistent.  Never reveal
or fabricate gold.
```