ONERESEARCHBENCH: BENCHMARKING RESEARCH AGENTS FOR RESEARCH SYNTHESIS FROM REAL-WORLD SOURCE MATERIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Progress toward general-purpose AI agents requires capabilities beyond isolated question answering. Agents must operate reliably across diverse domains, handle complex user-provided materials, and complete long-horizon knowledge tasks that involve gathering evidence, integrating information, and producing trustworthy outputs. Existing research-agent benchmarks, however, largely start from research questions or open information needs, and give limited attention to research that begins from real-world source materials. We define *source-grounded research synthesis*, a task in which an agent must first understand provided materials, then conduct external research, integrate the resulting evidence, and produce a grounded long-form report. To evaluate this capability, we introduce **OneResearchBench**, containing 101 human-constructed tasks across six domains and 3,671 expert-written, task-specific rubric rules. The rubrics evaluate source understanding, external research, evidence grounding, technical accuracy, citation traceability, and structure quality. Our criterion-level evaluator achieves a Pearson correlation of 0.909 with expert scores on the human calibration subset. We evaluate eight model–harness configurations. Under direct report generation, no system scores above 44.75/100, with external research and citation traceability emerging as the largest capability gaps. A structured skill-driven research workflow improves Claude Opus 4.8 with Claude Code from 44.75 to 53.07 (+8.32), and GPT-5.5 with Codex from 43.03 to 47.90 (+4.87). These results show that source-grounded research synthesis remains challenging for current research agents. OneResearchBench provides a benchmark for measuring progress on this task.

1 INTRODUCTION

Progress toward general-purpose AI agents requires capabilities beyond isolated question answering. Recent agent benchmarks have placed models in environments that require tool use, web interaction, code modification, and realistic task execution: GAIA evaluates reasoning, multimodality, web browsing, and tool use, while SWE-bench tests whether models can resolve real-world GitHub issues (Mialon et al., 2024; Jimenez et al., 2024). These efforts show that agent evaluation is shifting from static QA toward more realistic and long-horizon task environments. In real-world knowledge work, such as research, business analysis, policy consulting, medical review, and professional auditing, users often begin with source materials accumulated in practice, such as papers, reports, and other research-oriented documents. Agents must understand these materials, identify missing information, verify and supplement evidence from external sources, synthesize evidence across sources, distinguish source facts from external context and their own inferences, and avoid unsupported conclusions when evidence is insufficient. We refer to this task as *source-grounded research synthesis*: producing a long-form research report by jointly reasoning over provided source materials and external evidence.

Existing benchmarks have advanced the evaluation of agents and research report generation from several directions. Deep Research benchmarks, such as ResearchRubrics and DeepResearch Bench, evaluate open-ended long-form reports in terms of quality, citation ability, and rubric compliance (Sharma et al., 2025; Du et al., 2025). However, these benchmarks mainly focus on prompt-to-report or web-search-to-report settings, with limited emphasis on research that begins from user-provided source materials. Other agent benchmarks, including GAIA, WebArena, and SWE-bench, evaluate

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.

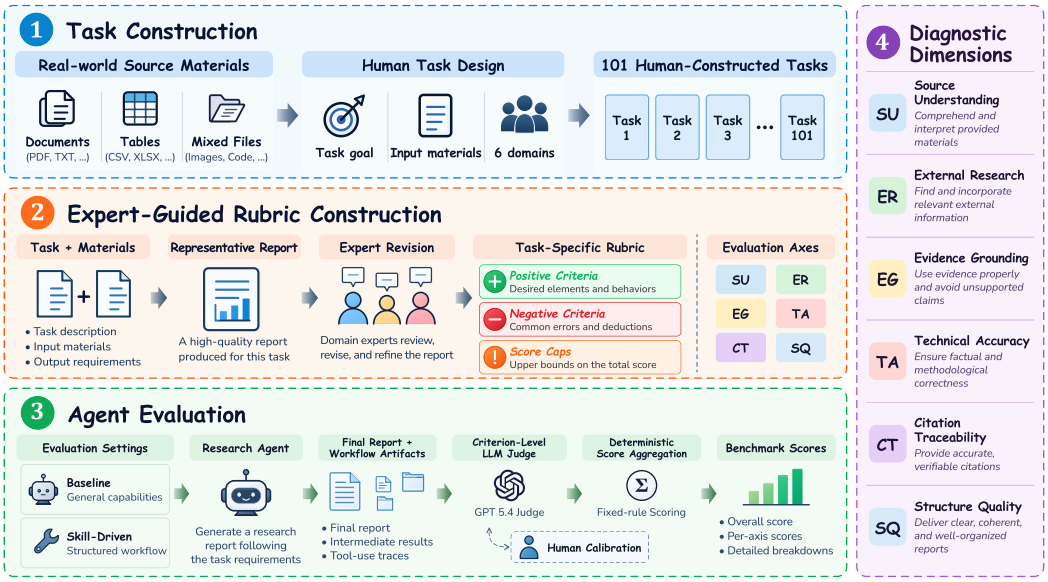

Figure 1: **Overview of OneResearchBench.** The benchmark is constructed from real-world source materials through human-designed tasks and expert-guided rubric construction, and evaluates research agents with criterion-level judging and diagnostic metrics.

tool use, web task completion, and real-world code repair (Mialon et al., 2024; Zhou et al., 2024; Jimenez et al., 2024). Yet they primarily target final answers, task success, or test-passing patches, leaving the combination of material understanding, external research, evidence integration, and report refinement insufficiently evaluated. This raises a key question: can current agents reliably perform source-grounded research synthesis?

To evaluate this task, we introduce OneResearchBench, a benchmark for source-grounded research synthesis. We construct 101 human-constructed tasks across six broad domains: STEM, humanities and social sciences, law and politics, economics, finance and business studies, medicine and life sciences, and arts and design. Each task begins with provided source materials and a research goal, and is paired with expert-written, task-specific rubrics. In total, OneResearchBench contains 3,671 rubric rules covering source understanding, external research, evidence grounding, technical accuracy, citation traceability, and structure quality. We use criterion-level LLM judging for scalable evaluation and validate the evaluator against expert scores, obtaining a Pearson correlation of 0.909 on the human calibration subset.

Using OneResearchBench, we evaluate eight representative model–harness configurations. Under direct report generation, no system exceeds 44.75/100, with external research and citation traceability as the largest capability gaps. On two representative configurations, our structured skill-driven workflow improves Claude Opus 4.8 with Claude Code from 44.75 to 53.07 (+8.32) and GPT-5.5 with Codex from 43.03 to 47.90 (+4.87).

Our contributions are as follows:

- We define **source-grounded research synthesis**, a research task in which agents must understand user-provided source materials, conduct external research, integrate evidence, and produce grounded long-form reports.
- We construct **OneResearchBench**, consisting of **101 human-constructed tasks** across six broad domains and **3,671 expert-written, task-specific rubric rules** covering source understanding, external research, evidence grounding, technical accuracy, citation traceability, and structure quality.
- We propose a structured **skill-driven research workflow** that organizes material understanding, external research, evidence synthesis, and report refinement. The workflow improves performance on both evaluated model–harness configurations.

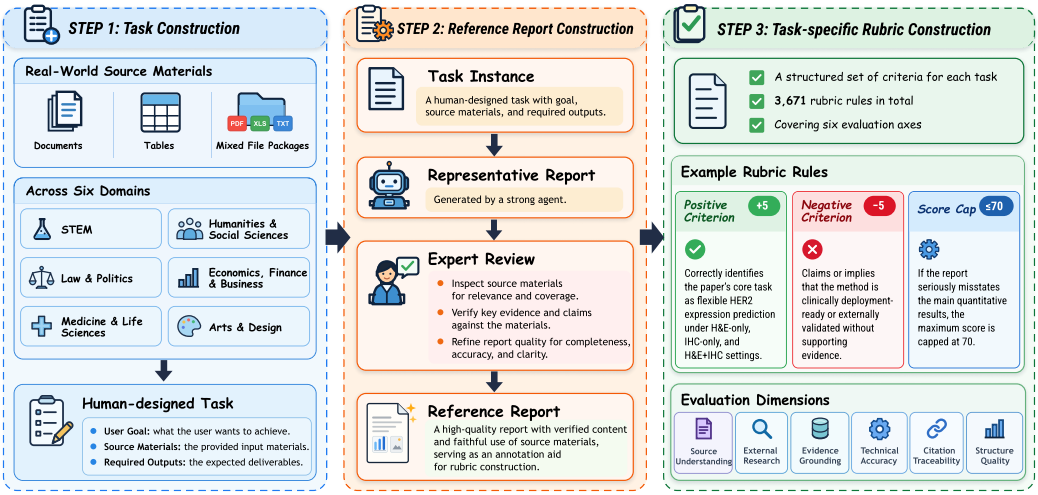

Figure 2: **Task construction and expert-guided rubric creation in OneResearchBench.** Real-world source materials are organized into human-designed research tasks. Representative reports are then generated and refined by experts, from which task-specific rubrics with positive criteria, negative criteria, score caps, and six evaluation dimensions are constructed.

2 OneResearchBench

OneResearchBench contains 101 human-constructed tasks across six broad domains for evaluating research agents on source-grounded research synthesis. Each task pairs provided source materials with a research goal and expert-written, task-specific rubrics, yielding 3,671 rubric rules in total.

2.1 Task Definition

Each task in OneResearchBench instantiates a *source-grounded research synthesis* problem. Unlike prompt-only research, the agent begins with user-provided source materials that serve as the primary grounding context. Given these materials and a research goal, the agent is expected to understand the provided evidence, identify information that requires further investigation, incorporate relevant external evidence, and produce a grounded long-form research report.

We represent each task as

$$\tau = (I, g, C), \tag{1}$$

where I denotes the provided source materials, g the research goal, and C the execution constraints, such as output language and output format. Each task is additionally paired with an expert-written, task-specific rubric R, which is used for evaluation.

We consider two execution settings. In the *direct-report* setting, the agent receives the task and produces the final report using its native planning and research behavior:

$$\pi(\tau) = \hat{y}. \tag{2}$$

We further study a *skill-driven* setting, in which the same task is executed with an explicit structured research skill $\mathcal{S}$:

$$\pi(\tau; \mathcal{S}) = \hat{y}. \tag{3}$$

The skill organizes the research process without changing the task definition or the task-specific evaluation rubric. Its detailed workflow is described in Section 2.4.

2.2 Task Collection and Domain Coverage

As shown in Figure 3(a), OneResearchBench contains 101 human-constructed tasks spanning six broad domains: STEM, Humanities and Social Sciences, Law and Politics, Economics, Finance and

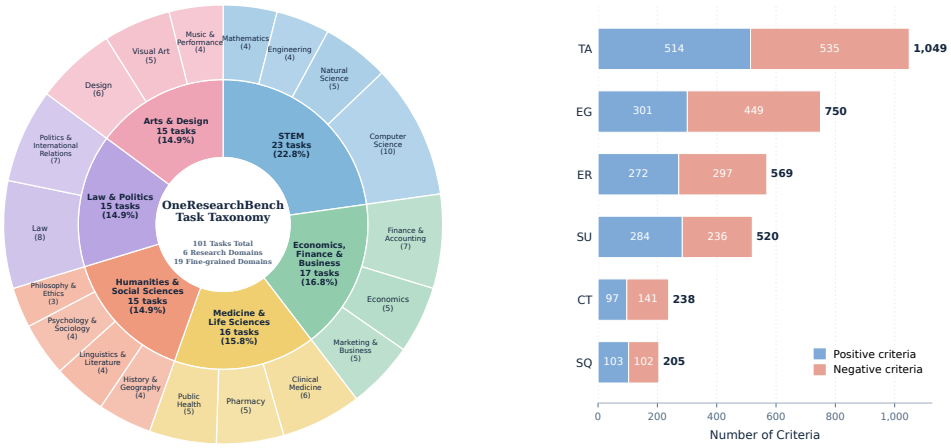

Figure 3: **Overview of benchmark diversity and rubric composition in OneResearchBench.** (a) The benchmark covers 101 human-constructed tasks across six broad research domains with fine-grained domain distributions. (b) Task-specific rubrics are organized across six evaluation dimensions, with 340 additional score caps for severe failures.

Table 1: Six evaluation dimensions in OneResearchBench. The maximum scores sum to 100 points.

Axis	Max.	Evaluation focus
Source Understanding (SU)	20	Accurate and sufficient understanding and use of the provided materials.
External Research (ER)	20	In-depth external research, retrieval, and integration of relevant evidence.
Evidence Grounding (EG)	20	Support for claims, effective use of evidence, and appropriate boundaries.
Technical Accuracy (TA)	15	Factual, numerical, methodological, and domain-specific correctness.
Citation Traceability (CT)	15	Whether citations are verifiable, correctly attributed, and support the associated claims.
Structure Quality (SQ)	10	Clarity, organization, coherence, and overall deliverability of the report.

Business Studies, Medicine and Life Sciences, and Arts and Design. These domains capture different research and evidence requirements. STEM and medical tasks often emphasize technical details, experimental findings, and numerical evidence; legal, political, and financial tasks require careful interpretation of rules, risks, and supporting evidence; while humanities, social sciences, and arts tasks place greater emphasis on conceptual analysis and structured interpretation.

The tasks are grounded primarily in scholarly papers and other research-oriented documents collected for concrete research topics. Rather than treating these materials as reference answers, each task requires agents to identify and interpret relevant source evidence, recognize missing information, and extend the provided materials through external research. This design evaluates research synthesis that begins from substantive source materials rather than from a standalone research prompt.

2.3 Expert-Guided Rubric Construction

Open-ended research reports rarely have a single canonical answer and may vary substantially in structure, argumentation, and presentation. We therefore construct a task-specific rubric for each task to capture its key information requirements, evidence expectations, common errors, and severe failure conditions. As illustrated in Figure 2, we first obtain a representative report using Claude Opus 4.7 with Cursor. Experts then review and refine the report by consulting the original source materials and, when necessary, additional external evidence. The resulting reference report serves as an annotation aid rather than a unique target answer. Based on the source materials, the refined report, and the identified failure modes, experts write the task-specific rubric rules.

Each rubric consists of three types of rules: positive criteria, negative criteria, and hard score caps. Positive criteria specify the information, evidence, and reasoning expected in a high-quality report, while negative criteria capture task-specific errors such as numerical misinterpretation, concept confusion, literature misuse, unsupported extrapolation, and missing limitations. Across the 101

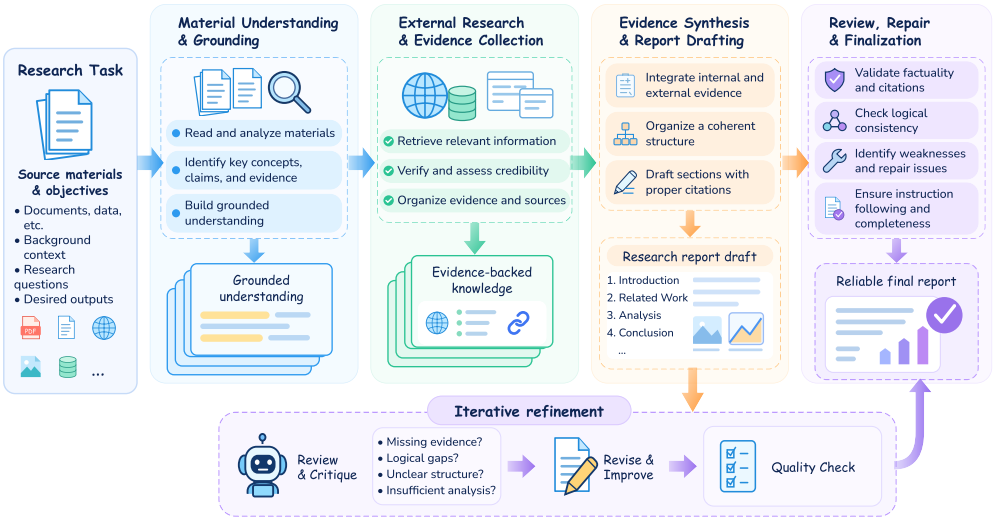

Figure 4: **Skill-driven research workflow for source-grounded research synthesis.** Given a research task with source materials and objectives, the workflow decomposes the synthesis process into four stages: material understanding and grounding, external research and evidence collection, evidence synthesis and report drafting, and review, repair, and finalization. An iterative refinement loop enables agents to improve report quality through structured review and revision.

tasks, experts write 1,571 positive criteria, 1,760 negative criteria, and 340 hard score caps, yielding 3,671 rubric rules in total. Positive and negative criteria are organized into six evaluation dimensions, summarized in Table 1.

Positive criteria use weights from $\{5, 4, 3\}$ for core, important, and general requirements, while negative criteria use $\{-5, -4, -3\}$ for severe, clear, and moderate errors. Positive criteria are accompanied by evidence anchors where applicable, linking evaluation requirements to relevant source materials, external evidence, or workflow artifacts. Hard score caps address severe failures that cannot be adequately captured by local penalties, such as failure to meaningfully use the provided materials, severe off-topic generation, or fabricated core citations. Together, these components provide weighted and traceable task-specific evaluation while allowing different report structures and valid analytical paths. Representative rubric annotations are provided in Appendix B.2.

2.4 Skill-Driven Research Workflow

For source-grounded research synthesis, we design a structured skill-driven workflow that organizes the research process into four stages as illustrated in Figure 4.

Material understanding and grounding. The agent analyzes the materials and constructs grounded notes capturing key facts, document structure, numerical evidence, and potential evidence gaps.

External research and evidence collection. Based on the user goal and grounded notes, the agent retrieves relevant external sources to complement and verify the provided materials. It assesses source credibility and organizes traceable evidence covering background, related work, comparative findings, methodological context, and limitations.

Evidence synthesis and report drafting. The agent integrates the provided materials and external evidence into a structured draft, while distinguishing source facts, external context, and analytical judgments.

Review, repair, and finalization. The agent reviews the draft for factual accuracy, citation traceability, logical consistency, evidence grounding, and completeness. It then identifies weaknesses such

as missing evidence, unsupported claims, or structural gaps and iteratively revises the report into the final output.

The workflow produces intermediate artifacts such as grounding notes, search records, evidence collections, drafts, and review history, enabling process-level analysis of how agents conduct research and synthesize evidence.

2.5 Evaluation Protocol and Metrics

OneResearchBench evaluates agent outputs through three steps: submission validation, criterion-level LLM judging, and deterministic score aggregation.

Submission validation. We first verify that the required outputs are present and valid. The direct-report setting requires only the final report, while the skill-driven setting additionally validates the intermediate artifacts produced by the workflow. An artifact is valid if it exists, is non-empty, and is parseable when structured. We define the artifact contract score S_{contract} as the percentage of valid required artifacts on a 0–100 scale. A missing or empty final report results in an invalid submission.

Criterion-level LLM judging. A fixed LLM judge independently evaluates each rubric rule. For positive criteria, it outputs a satisfaction score $s_i \in [0, 1]$; for negative criteria, an error-severity score $v_j \in [0, 1]$; and for hard score caps, whether the corresponding failure condition is triggered.

Deterministic score aggregation. For each evaluation axis a, the positive score is computed as

$$P_a = M_a \frac{\sum_{i \in Pos(a)} w_i s_i}{\sum_{i \in Pos(a)} w_i}, \tag{4}$$

where M_a is the maximum score of axis a, w_i is the criterion weight, and s_i is its satisfaction score. The negative penalty is

$$D_a = \sum_{j \in Neg(a)} |w_j| v_j. \tag{5}$$

The resulting axis score is $S_a = \max(0, P_a - D_a)$, and the total score before hard caps is $S = \sum_a S_a$. If any hard score cap is triggered, the content score is bounded by the strictest cap, $S_{\text{content}} = \min(S, C_{\min})$; otherwise, $S_{\text{content}} = S$.

For direct-report evaluation, the final score is S_{content}. For the skill-driven setting, failure to satisfy the artifact contract further bounds the final score:

$$S_{\text{final}} = \begin{cases} S_{\text{content}}, & \text{contract satisfied,} \\ \min(S_{\text{content}}, S_{\text{contract}}), & \text{otherwise.} \end{cases} \tag{6}$$

Metrics. We report S_{final} as the overall score together with the six evaluation dimensions in Table 1. For cross-axis comparison, each dimension is normalized by its maximum score and reported as a percentage. We further compare the LLM-based evaluator with expert scores in Section 3.4.

3 Experiments

In this section, we evaluate LLM agents on all 101 tasks in OneResearchBench. Section 3.1 describes the experimental setup, Section 3.2 presents the main results, Section 3.3 examines the skill-driven workflow, and Section 3.4 compares LLM-based evaluation with expert scores.

3.1 Experimental Setup

Evaluated systems and settings. We evaluate five contemporary LLMs: Claude Opus 4.8 (Anthropic, 2026), Qwen3.7 Max (Qwen Team, 2026), GPT-5.5 (OpenAI, 2026), DeepSeek V4 Pro (DeepSeek-AI, 2026), and Kimi K2.6 (Moonshot AI, 2026). All five models are evaluated with Claude Code, while GPT-5.5, Qwen3.7 Max, and DeepSeek V4 Pro are additionally evaluated with Codex. We compare a direct-report baseline using minimal instructions with our skill-driven setting, in which the agent follows the structured research workflow described in Section 2.4.

Table 2: Main results on OneResearchBench. SU, ER, EG, TA, CT, and SQ denote Source Understanding, External Research, Evidence Grounding, Technical Accuracy, Citation Traceability, and Structure Quality, respectively, with maximum contributions of 20, 20, 20, 15, 15, and 10 points to the overall score. Scores for individual dimensions are normalized to percentages for comparison, while Overall reports the original score on a 0–100 scale.

Model	SU (%)	ER (%)	EG (%)	TA (%)	CT (%)	SQ (%)	Overall
Claude Code							
Claude Opus 4.8	**77.0**	20.9	41.1	**56.7**	15.9	60.7	**44.75**
Qwen3.7 Max	65.6	**26.1**	33.9	51.5	16.0	57.9	41.02
GPT-5.5	70.3	8.8	45.1	42.7	7.7	59.4	38.32
DeepSeek V4 Pro	44.5	14.4	19.6	31.1	9.8	36.7	25.47
Kimi K2.6	61.8	11.3	21.1	30.6	8.9	52.3	30.01
Codex							
GPT-5.5	73.0	11.9	**50.7**	49.6	13.7	**64.1**	43.03
DeepSeek V4 Pro	68.5	18.0	30.6	45.2	**16.3**	53.3	37.99
Qwen3.7 Max	61.6	19.4	25.8	38.0	10.6	53.3	34.00

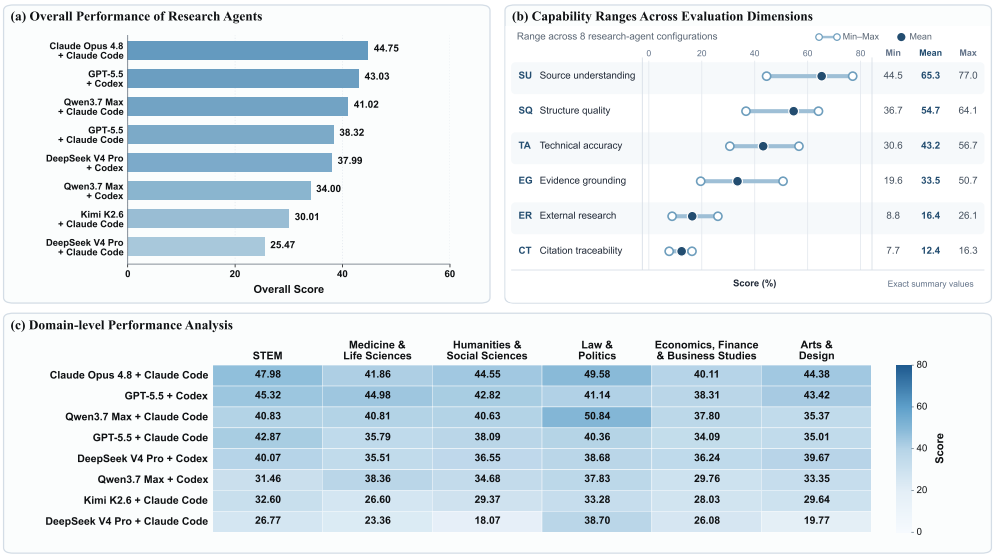

Figure 5: **Diagnostic analysis of research agents on OneResearchBench.** (a) Overall performance of representative model–harness configurations. (b) Capability breakdown across six evaluation dimensions. (c) Domain-level performance analysis across heterogeneous research scenarios.

LLM judge. We use GPT-5.4 as the fixed criterion-level judge for all benchmark evaluations and apply the deterministic score aggregation described in Section 2.5. We further validate the judge against expert scores in Section 3.4.

3.2 Main Results on OneResearchBench

Table 2 and Figure 5(a) summarize performance under the direct-report setting. Among the eight evaluated model–harness configurations, Claude Opus 4.8 with Claude Code achieves the highest overall score of 44.75, followed by GPT-5.5 with Codex at 43.03, while DeepSeek V4 Pro with Claude Code records the lowest score of 25.47. No evaluated configuration exceeds 45 out of 100, indicating that source-grounded research synthesis remains challenging for current agents.

Where do current agents fall short? Figure 5(b) reveals an uneven capability profile. Across the eight configurations, Source Understanding and Structure Quality average 65.3% and 54.7%, respectively, compared with 43.2% for Technical Accuracy and 33.5% for Evidence Grounding.

Table 3: Effect of the skill-driven workflow on two model–harness configurations. Δ denotes the gain from Direct to +Skill.

Model	**Claude Opus 4.8**	**GPT-5.5**
Harness	Claude Code	Codex
Direct	44.75	43.03
+Skill	**53.07**	**47.90**
Δ	+8.32	+4.87

Table 4: Pearson correlation between LLM-judge and expert scores on the human calibration subset. Computed on report-level scores.

Model–Harness Configuration	**Pearson r**
Claude Opus 4.8 + Claude Code	0.861
GPT-5.5 + Codex	0.853
DeepSeek V4 Pro + Claude Code	0.972
Kimi K2.6 + Claude Code	0.898
Qwen3.7 Max + Claude Code	0.841
All configurations	**0.909**

External Research and Citation Traceability are substantially weaker, averaging only 16.4% and 12.4%; even their highest scores reach just 26.1% and 16.3%. These results suggest a gap between understanding and organizing provided information and extending it into well-supported research synthesis. The weakness in Evidence Grounding further highlights difficulties in supporting claims and respecting evidence boundaries, beyond the limitations in external research and citation traceability. An organized report that covers the input materials is therefore not, by itself, sufficient evidence of research competence. By jointly evaluating source understanding, external research, and evidence use, OneResearchBench exposes limitations that assessments centered on material coverage and presentation alone would overlook.

How do agents perform across research domains? Figure 5(c) shows that aggregate rankings do not fully capture domain-specific strengths and weaknesses. Law & Politics is the highest-scoring domain for four configurations, while the remaining configurations perform best in STEM or Medicine & Life Sciences. The strongest overall configuration is not uniformly the strongest: GPT-5.5 with Codex outperforms Claude Opus 4.8 with Claude Code in Medicine & Life Sciences (44.98 vs. 41.86), despite ranking lower overall. These differences indicate that an aggregate score can obscure which research settings are relatively more challenging for a given configuration. They should be interpreted as configuration-dependent performance on the sampled tasks, rather than an intrinsic difficulty ordering of disciplines. By covering multiple domains with task-specific rubrics, OneResearchBench assesses the breadth of research capabilities and reveals weaknesses that could be masked by evaluation concentrated in a single domain.

3.3 Effect of the Skill-Driven Research Workflow

We examine whether explicit workflow guidance improves research synthesis beyond direct report generation. Table 3 compares the two settings for two representative configurations, keeping the model and agent harness fixed.

The workflow improves both evaluated configurations, with absolute gains of 8.32 points for Claude Opus 4.8 with Claude Code and 4.87 points for GPT-5.5 with Codex. These gains suggest that explicit organization of the research process can help agents make better use of their existing capabilities, without changing the underlying model or harness. Nevertheless, the resulting scores of 53.07 and 47.90 leave substantial room for improvement, even with structured workflow guidance. OneResearchBench thus provides a testbed not only for characterizing current capabilities, but also for evaluating interventions in how agents conduct research.

3.4 Human Calibration of LLM-Based Evaluation

We assess the reliability of the LLM-based evaluator by comparing its scores with expert scores on the same reports. We select a stratified subset of 24 tasks covering all six benchmark domains. For each task, we assess reports from five representative model–harness configurations, yielding 120 report-level samples in total.

Table 5: Comparison with representative research-agent benchmarks. **Provided Sources** indicates task-specific source materials given to the evaluated system; **Human-authored Rubrics** indicates expert-designed task-specific criteria; **Structured Workflow** indicates an explicit benchmark-provided research workflow used as an evaluation setting. Domain coverage reports broad/fine-grained domains when available.

Benchmark	Provided Sources	Human-authored Rubrics	Structured Workflow	Domain Coverage	Avg. # Rubric Rules / Task
ResearchRubrics (Sharma et al., 2025)	✗	✓	✗	9	26
DeepResearch Bench (Du et al., 2025)	✗	✗	✗	22	25
ExpertLongBench (Ruan et al., 2025)	✓	✓	✗	9	16.64
DR^3-Eval (Xie et al., 2026)	✓	✗	✗	3 / 13	–
LiveDRBench (Java et al., 2026)	✗	✗	✗	2 / 8	–
OneResearchBench (Ours)	✓	✓	✓	**6 / 19**	**36.35**

Expert evaluators score each rubric criterion using the same task-specific rubrics as the LLM judge. Human and LLM judgments are then aggregated using the same deterministic scoring procedure, and we compare the resulting report-level scores.

As shown in Table 4, the aggregated LLM-judge scores achieve a Pearson correlation of 0.909 with expert scores across all 120 reports. Correlations within individual configurations range from 0.841 to 0.972, indicating that the strong association also holds within configurations rather than only across the pooled samples. These results provide evidence that the evaluator captures expert-assessed differences in report quality on the sampled tasks.

4 Related Work

A growing body of work evaluates agents on open-ended research tasks. DeepResearch Bench contains 100 expert-designed research tasks and evaluates both report quality and citation behavior (Du et al., 2025). ResearchRubrics pairs realistic research prompts with fine-grained expert-written rubrics for evaluating factual grounding, reasoning, and clarity (Sharma et al., 2025). ExpertLongBench targets expert-level long-form generation across diverse professional domains with domain-specific rubrics and structured checklists (Ruan et al., 2025). More recent benchmarks explore richer research settings: DR^3-Eval evaluates multimodal, multi-file research grounded in user-provided materials (Xie et al., 2026), while LiveDRBench characterizes deep research through high-fan-out search and evaluates systems using an intermediate claim representation (Java et al., 2026). Table 5 summarizes the key differences between OneResearchBench and representative research-agent benchmarks. To our knowledge, existing benchmarks do not explicitly formulate *open-ended source-grounded research synthesis* as a unified task in which agents begin from real-world source materials, identify information gaps, conduct external research beyond a fixed evidence corpus, and integrate source and external evidence while preserving their evidential boundaries in a grounded long-form report. We introduce this task and evaluate it with expert-written, task-specific rubrics spanning six research capabilities, together with hard score caps for severe task-specific failures.

5 Conclusion

We define *source-grounded research synthesis* and introduce OneResearchBench to evaluate research agents on this task. The benchmark contains 101 human-constructed tasks across six broad domains, paired with expert-written task-specific rubrics comprising 3,671 rubric rules. Our evaluation shows that current research agents remain substantially challenged: while they perform relatively well in source understanding and report organization, external research, evidence grounding, and citation traceability remain major bottlenecks. We further show that a structured skill-driven research workflow improves performance on both evaluated model–harness configurations, demonstrating the potential of explicitly organizing material understanding, evidence collection, synthesis, and revision. OneResearchBench provides a testbed for measuring progress toward research agents that can reliably understand provided source materials, integrate external evidence, and produce grounded long-form reports.

AI USE STATEMENT

Generative AI models are explicit components of the benchmark construction, experimental, and evaluation pipelines in this work. During benchmark construction, Claude Opus 4.7 with Cursor is used to generate representative reports, which are subsequently reviewed and refined by human experts to support task-specific rubric construction. In our experiments, LLM agents generate the research reports being evaluated, and GPT-5.4 serves as the fixed criterion-level evaluator. The task-specific rubrics are written by human experts. For manuscript preparation, generative AI tools were used to assist with language polishing and improving clarity and presentation. All scientific claims, analyses, and reported results were reviewed and verified by the authors, who take full responsibility for the final manuscript.

ETHICS STATEMENT

OneResearchBench is constructed from publicly accessible scholarly papers and research documents. We record the provenance of benchmark source materials and respect applicable copyright and redistribution restrictions. Source documents are redistributed only when their licenses explicitly permit redistribution; otherwise, we release source identifiers and public links, together with task definitions, task-specific rubrics, evaluation code, and workflow implementations, so that users can obtain the original materials from their authorized sources. The benchmark includes tasks in medical, legal, and financial domains. Generated reports are intended solely as evaluation artifacts and should not be interpreted as professional medical, legal, financial, or policy advice.

REPRODUCIBILITY STATEMENT

We provide detailed documentation to support reproduction of OneResearchBench and the reported experiments. The benchmark construction, task definition, rubric design, evaluation protocol, and experimental setup are described in Sections 2.1–3.1. Appendix A provides additional benchmark and source-material details, Appendix B reports rubric statistics and representative examples, and Appendix C documents the direct and skill-driven execution settings, prompts, workflow organization, and intermediate artifacts. Appendix D further specifies the criterion-level LLM-judge setup and human calibration procedure. We will release the benchmark tasks, task-specific rubrics, evaluation code, and workflow implementation to support reproduction and further evaluation.

REFERENCES

Anthropic. Claude Opus 4.8 system card. System Card, May 2026. URL https://anthropic.com/claude-opus-4-8-system-card.

DeepSeek-AI. DeepSeek-V4-Pro GA release, August 2026. URL https://api-docs.deepseek.com/news/news260813/.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025. URL https://arxiv.org/abs/2506.11763.

Martin Ebers. Truly risk-based regulation of artificial intelligence: How to implement the EU's AI Act. *European Journal of Risk Regulation*, 16(2):684–703, 2025. doi: 10.1017/err.2024.78. URL https://doi.org/10.1017/err.2024.78.

Darius-Aurel Frank, Polymeros Chrysochou, Panagiotis Mitkidis, and Dan Ariely. Human decision-making biases in the moral dilemmas of autonomous vehicles. *Scientific Reports*, 9:13080, 2019. doi: 10.1038/s41598-019-49411-7. URL https://doi.org/10.1038/s41598-019-49411-7.

Ge Gao, Fan Xu, Jiangmin Xu, Guanghai Tang, and Zhenyu Liu. A survey of the influence of process parameters on mechanical properties of fused deposition modeling parts. *Micromachines*, 13(4): 553, 2022. doi: 10.3390/mi13040553. URL https://doi.org/10.3390/mi13040553.

Abhinav Java, Ashmit Khandelwal, Sukruta Midigeshi, Aaron Halfaker, Amit Jayant Deshpande, Navin Goyal, Ankur Gupta, Nagarajan Natarajan, and Amit Sharma. Characterizing deep research: A benchmark and formal definition. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/114e1dc345fe31b8b9b0c6f7b55a0644-Abstract-Conference.html.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=VTF8yNQM66.

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*, 2026. URL https://arxiv.org/abs/2601.08536.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=fibxvahvs3.

Kamiar Mohaddes, Ryan N. C. Ng, M. Hashem Pesaran, Mehdi Raissi, and Jui-Chung Yang. Climate change and economic activity: Evidence from U.S. states. *Oxford Open Economics*, 2:28–46, 2023. doi: 10.1093/ooec/odac010. URL https://doi.org/10.1093/ooec/odac010.

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

OpenAI. GPT-5.5 system card. System Card, April 2026. URL https://openai.com/index/gpt-5-5-system-card/.

Karl Pearson. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895. doi: 10.1098/rspl.1895.0041. URL https://doi.org/10.1098/rspl.1895.0041.

Chau Minh Pham, Zichao Wang, Puneet Mathur, Alexa Siu, Akriti Jain, Aparna Garimella, Ananya B. Sai, Nedim Lipka, Mohit Iyyer, and Varun Manjunatha. AnalystBench: Benchmarking professional long-form report generation with web-mined multimodal tasks. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23894–23926, 2026. doi: 10.18653/v1/2026.findings-acl.1197. URL https://aclanthology.org/2026.findings-acl.1197/.

Qwen Team. Qwen3.7: The agent frontier, May 2026. URL https://qwen.ai/blog?id=qwen3.7.

Kazem Rahmani, Rasoul Shavaleh, Mahtab Forouhi, Hamideh Feiz Disfani, Mostafa Kamandi, Rozita Khatamian Oskooi, Molood Foogerdi, Moslem Soltani, Maryam Rahchamani, Mohammad Mohaddespour, and Mostafa Dianatinasab. The effectiveness of COVID-19 vaccines in reducing the incidence, hospitalization, and mortality from COVID-19: A systematic review and meta-analysis. *Frontiers in Public Health*, 10:873596, 2022. doi: 10.3389/fpubh.2022.873596. URL https://doi.org/10.3389/fpubh.2022.873596.

Jie Ruan, Inderjeet Nair, Shuyang Cao, Amy Liu, Sheza Munir, Micah Pollens-Dempsey, Tiffany Chiang, Lucy Kates, Nicholas David, Sihan Chen, Ruxin Yang, Yuqian Yang, Jasmine Gump, Tessa Bialek, Vivek Sankaran, Margo Schlanger, and Lu Wang. ExpertLongBench: Benchmarking language models on expert-level long-form generation tasks with structured checklists. *arXiv preprint arXiv:2506.01241*, 2025. URL https://arxiv.org/abs/2506.01241.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d842425e4bf79ba039352da0f658a906-Abstract-Conference.html.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. ResearchRubrics: A benchmark of

prompts and rubrics for evaluating deep research agents. *arXiv preprint arXiv:2511.07685*, 2025. URL https://arxiv.org/abs/2511.07685.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. In *Advances in Neural Information Processing Systems*, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/77c33e6a367922d003ff102ffb92b658-Abstract-Conference.html.

Christian Spreafico and Daniele Landi. Using product design strategies to implement circular economy: Differences between students and professional designers. *Sustainability*, 14(3):1122, 2022. doi: 10.3390/su14031122. URL https://doi.org/10.3390/su14031122.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023. URL https://arxiv.org/abs/2305.16291.

Jiayu Wang, Yifei Ming, Riya Dulepet, Qinglin Chen, Austin Xu, Zixuan Ke, Frederic Sala, Aws Albarghouthi, Caiming Xiong, and Shafiq Joty. LiveResearchBench: A live benchmark for user-centric deep research in the wild. *arXiv preprint arXiv:2510.14240*, 2025. URL https://arxiv.org/abs/2510.14240.

Michael F. Wehner and Kevin A. Reed. Operational extreme weather event attribution can quantify climate change loss and damages. *PLOS Climate*, 1(2):e0000013, 2022. doi: 10.1371/journal.pclm.0000013. URL https://doi.org/10.1371/journal.pclm.0000013.

Qianqian Xie, Qingheng Xiong, He Zhu, Tiantian Xia, Xueming Han, Fanyu Meng, Jiakai Wang, Zhiqi Bai, Chengkang Jiang, Zhaohui Wang, Yubin Guo, Yuqing Wen, Jiayang Mao, Zijie Zhang, Shihao Li, Yanghai Wang, Yuxiang Ren, Junlan Feng, and Jiaheng Liu. DR^3-Eval: Towards realistic and reproducible deep research evaluation. *arXiv preprint arXiv:2604.14683*, 2026. URL https://arxiv.org/abs/2604.14683.

Tianze Xu, Pengrui Lu, Lyumanshan Ye, Xiangkun Hu, and Pengfei Liu. ResearcherBench: Evaluating deep AI research systems on the frontiers of scientific inquiry, 2025. URL https://arxiv.org/abs/2507.16280.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oKn9c6ytLx.

SUPPLEMENTARY MATERIAL

Contents

A Benchmark and Source-Material Details 14
A.1 Source Collection and Task Construction 14
A.2 Domain Coverage 14
A.3 Source-Material Scale 14
A.4 Representative Tasks and Source Materials 15
B Rubric Statistics and Examples 16
B.1 Rubric Composition and Statistics 16
B.2 Representative Rubric Examples 16
C Implementation and Skill Workflow Details 17
C.1 Direct and Skill-Driven Execution 17
C.2 Execution Prompts 18
C.3 Skill Pipeline and Intermediate Artifacts 18
D Evaluation and Human Calibration Details 19
D.1 LLM Judge Inputs and Prompt 20
D.2 Human–LLM Score Comparison 21
D.3 Robustness to Negative-Penalty Normalization 22
D.4 Hard-Cap Activation and Score Impact 23
E Additional Results and Case Studies 23
E.1 Dimension-wise Effect of the Skill Workflow 23
E.2 Qualitative Case Studies 24
F Extended Related Work 25

A BENCHMARK AND SOURCE-MATERIAL DETAILS

This section provides additional details on the construction and source materials of OneResearchBench. Section A.1 describes source collection and task construction, Section A.2 summarizes domain coverage, Section A.3 reports source-material scale, and Section A.4 presents representative tasks and sources.

A.1 SOURCE COLLECTION AND TASK CONSTRUCTION

The source materials in OneResearchBench were collected by annotators with relevant domain expertise. For each domain, annotators first identified representative research topics and searched public websites and publication platforms for materials that were closely aligned with the corresponding domain and topic. Scholarly papers form the primary source of benchmark materials, providing substantive technical, empirical, or conceptual content from which grounded research tasks can be constructed.

The collected materials serve as the starting point of each task rather than as reference answers. Based on the selected source material, annotators formulate a research goal that requires the agent to understand and synthesize the provided content and, when appropriate, extend the analysis through external research. Candidate materials are retained when they are relevant to the target domain and provide sufficient substantive content to support a meaningful research-synthesis task. Materials that are poorly aligned with the target topic or do not support a well-defined research task are excluded.

A.2 DOMAIN COVERAGE

OneResearchBench spans six broad research domains and 19 fine-grained domains, as summarized in Table 6. The taxonomy covers both technical and non-technical research settings, including scientific and engineering analysis, medical and financial research, legal and political studies, humanities and social sciences, and arts and design.

Table 6: Domain coverage of OneResearchBench. Numbers in parentheses denote the number of tasks.

Broad Domain	**Fine-Grained Domains**	**# Tasks**
STEM	Computer Science (10), Natural Science (5), Mathematics (4), Engineering (4)	23
Economics, Finance & Business	Finance & Accounting (7), Economics (5), Marketing & Business (5)	17
Medicine & Life Sciences	Clinical Medicine (6), Pharmacy (5), Public Health (5)	16
Humanities & Social Sciences	History & Geography (4), Linguistics & Literature (4), Psychology & Sociology (4), Philosophy & Ethics (3)	15
Law & Politics	Law (8), Politics & International Relations (7)	15
Arts & Design	Design (6), Visual Art (5), Music & Performance (4)	15
Total	**19 fine-grained domains**	**101**

STEM is the largest broad domain with 23 tasks, followed by Economics, Finance & Business with 17 tasks and Medicine & Life Sciences with 16. Humanities & Social Sciences, Law & Politics, and Arts & Design each contain 15 tasks. This distribution provides relatively balanced coverage across substantially different forms of research, while the 19 fine-grained domains further introduce variation in subject matter, evidence requirements, and analytical goals.

A.3 SOURCE-MATERIAL SCALE

The benchmark is primarily grounded in scholarly papers and other research-oriented documents. Across the PDF-based source materials, OneResearchBench contains 2,128 pages in total. An individual PDF contains 20.9 pages on average, with a median of 18 pages. The 25th and 75th

percentiles are 13 and 25 pages, respectively, and the 90th percentile is 33.9 pages. The longest source document contains 118 pages.

Table 7: Scale of PDF-based source materials in OneResearchBench.

Statistic	Value
Total PDF pages	2,128
Mean pages per PDF	20.9
Median pages per PDF	18
25th percentile	13
75th percentile	25
90th percentile	33.9
Maximum pages	118

These statistics highlight an important characteristic of OneResearchBench: agents generally begin with a non-trivial body of source material that must be understood before further research and report synthesis can be carried out. This differs from prompt-to-report settings in which the research process primarily begins from a question and open-web retrieval.

A.4 Representative Tasks and Source Materials

Table 8 presents representative examples from the six broad domains. The examples illustrate the diversity of research topics and source materials used in OneResearchBench. In each case, the source document provides the grounding context from which annotators construct a broader research-synthesis task.

Table 8: Representative tasks and source materials across the six broad domains.

Domain	Research Task	Representative Source Material	Pages
STEM	Process parameters and mechanical properties in FDM 3D printing	*A Survey of the Influence of Process Parameters on Mechanical Properties of Fused Deposition Modeling Parts* (Gao et al., 2022)	28
Economics, Finance & Business	Climate change and economic activity in U.S. states	*Climate Change and Economic Activity: Evidence from U.S. States* (Mohaddes et al., 2023)	11
Medicine & Life Sciences	Effectiveness of COVID-19 vaccines in reducing incidence, hospitalization, and mortality	*The Effectiveness of COVID-19 Vaccines in Reducing the Incidence, Hospitalization, and Mortality from COVID-19: A Systematic Review and Meta-Analysis* (Rahmani et al., 2022)	25
Humanities & Social Sciences	Human moral biases in autonomous-vehicle dilemmas	*Human Decision-Making Biases in the Moral Dilemmas of Autonomous Vehicles* (Frank et al., 2019)	19
Law & Politics	Risk-based regulation of artificial intelligence	*Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act* (Ebers, 2025)	20
Arts & Design	Product-design strategies for the circular economy	*Using Product Design Strategies to Implement Circular Economy: Differences between Students and Professional Designers* (Spreafico & Landi, 2022)	19

These examples illustrate that the benchmark is not restricted to a single style of academic research. The source materials range from empirical and technical studies to systematic reviews, policy and legal analysis, and research in the humanities and design. The resulting tasks require agents to first establish a grounded understanding of the provided material and then synthesize it into a broader research report, often together with additional external evidence.

B RUBRIC STATISTICS AND EXAMPLES

Section 2.3 describes the overall rubric-construction procedure. Here, we provide additional statistics and concrete examples of the resulting task-specific rubrics. Section B.1 summarizes the overall rubric statistics, while Section B.2 presents a representative task-specific rubric example.

B.1 RUBRIC COMPOSITION AND STATISTICS

Across the 101 tasks, OneResearchBench contains 1,571 positive criteria, 1,760 negative criteria, and 340 hard score caps, yielding 3,671 rubric rules in total. Table 9 summarizes their per-task distributions.

Table 9: Statistics of task-specific rubric rules in OneResearchBench.

Rule Type	Total	Mean / Task	Median	Min–Max
Positive Criteria	1,571	15.55	16	10–20
Negative Criteria	1,760	17.43	17	8–23
Hard Score Caps	340	3.37	3	3–7
All Rules	3,671	36.35	36	26–45

An average task contains approximately 33 positive and negative criteria, together with 3.4 hard score caps. Positive criteria specify task-specific information, reasoning, and research requirements, whereas negative criteria capture concrete failure modes, unsupported extrapolations, and other task-specific errors. Hard score caps address severe failures whose impact should not be fully compensated by success on unrelated criteria.

Explicit evidence anchors are attached to 1,515 of the 1,571 positive criteria (96.4%), linking them to relevant source materials, external evidence, or research artifacts. In the current rubric schema, negative criteria and hard score caps are instead specified directly through their task-specific failure conditions.

B.2 REPRESENTATIVE RUBRIC EXAMPLES

To illustrate the granularity and task specificity of OneResearchBench, we use an extreme-weather attribution task as a representative example. The task is grounded in *Operational extreme weather event attribution can quantify climate change loss and damages* (Wehner & Reed, 2022) and asks the agent to connect the provided source with external attribution research while respecting scientific uncertainty and evidence boundaries. Its complete rubric contains 15 positive criteria, 17 negative criteria, and four hard score caps, covering all six evaluation dimensions. Figure 6 shows selected excerpts from this rubric.

The complete rubric is included with the benchmark release. This example shows that correctly summarizing the provided source is not sufficient for a high score. The report must also conduct substantive external research, preserve appropriate evidence boundaries, and avoid scientifically unsupported conclusions. Conversely, a fluent and well-structured report may still receive substantial penalties when it misinterprets the source or overstates what the available evidence supports.

Cross-domain examples. Task-specific requirements vary substantially across research domains. In Medicine & Life Sciences, for example, one criterion penalizes concluding that a clinical prediction model is suitable for rule-out use merely because its discrimination is strong when sensitivity and likelihood ratios do not support that interpretation. In Economics, Finance & Business, an external research criterion requires a structured literature matrix covering the research question, methodology, key finding, limitation, and relationship of each external source to the source argument. In Humanities & Social Sciences, citation criteria distinguish questionnaire-based correlations from evidence of real platform-level propagation or causal effects. In Law & Politics, claim-specific citations must distinguish statements supported by the provided article from external authorities concerning fair use, text-and-data-mining exceptions, AI regulation, and related litigation. In Arts & Design, task-specific

Positive Criteria
pc03 — Source Understanding, +5. Correctly explains how event attribution results are expressed in the article: as changes in the occurrence frequency or probability of an event of observed magnitude, and/or as changes in event magnitude at a fixed estimate of observed probability. The report must explain that these are probabilistic or counterfactual statements, not deterministic claims that climate change solely caused an event.

pc08 — External Research, +5. Integrates external attribution-science literature in a method-specific way, including probabilistic event attribution, fraction of attributable risk, risk ratio/probability ratio, change-in-intensity framing, extreme value analysis, detection-and-attribution distinction, and uncertainty assessment. The report must use these sources to clarify the source article's argument, not merely repeat that climate change increases extremes.

Negative Criteria
nc03 — Technical Accuracy, -5. Treats attribution as deterministic causation. This includes saying that climate change solely caused an individual event, that attribution proves a single cause, or that an event would not have happened without climate change without presenting the probability/magnitude framing required by the source.

nc08 — External Research, -5. Provides no substantive independent external research beyond the source article. Repeating the article's own references to IPCC, WWA, Hurricane Harvey, heat mortality, Sandy, or WMO does not count unless the report uses external sources to evaluate, qualify, or extend the article's argument.

Hard Score Cap
h4 — Maximum score: 75. The report treats individual extreme weather events as deterministically caused by climate change or presents attribution results as absolute proof rather than probabilistic shifts, contradicting the source article's probabilistic framing.

Figure 6: **Representative task-specific rubric excerpts from OneResearchBench.** The example illustrates how positive criteria, negative criteria, and hard score caps jointly capture source understanding, external research, technical accuracy, and severe task-specific failures.

criteria may instead require the report to correctly distinguish experimental conditions, such as the four combinations of high/low aesthetics and high/low usability in a UI/UX study.

These examples demonstrate that OneResearchBench rubrics are not generic report-quality checklists. They encode task-specific factual requirements, methodological boundaries, external-research expectations, citation scope, and severe domain-specific failure conditions. This allows the benchmark to distinguish reports that are merely fluent or comprehensive from those that are substantively grounded in both the provided source materials and relevant external evidence.

C Implementation and Skill Workflow Details

This section provides additional implementation details for the two research settings evaluated in OneResearchBench. Section C.1 clarifies the execution differences between direct-report generation and the skill-driven setting, while Section C.2 presents the corresponding task-level prompts. Section C.3 further describes the organization of the structured research workflow and the intermediate artifacts produced during execution.

C.1 Direct and Skill-Driven Execution

For each benchmark task, both settings receive the same task instance, including the provided source materials, task-specific research goal, required outputs, and execution constraints. They differ only in how explicitly the research procedure is specified and executed.

In the **direct-report** setting, the agent receives a minimal instruction to inspect the provided materials and produce a clear, well-structured report. Relevant background information or external context may be incorporated when useful and available, but no specific research procedure is prescribed.

Table 10: Comparison between the direct-report and skill-driven execution settings.

Component	Direct-Report	Skill-Driven
Source materials	Benchmark-provided materials	Benchmark-provided materials
Research instruction	Minimal report-generation instruction	Explicit invocation of the `one-report` skill
Research procedure	Not explicitly prescribed	Structured multi-stage workflow
External research	Incorporated when useful and available	Explicitly integrated into the workflow
Review and refinement	Left to native agent behavior	Explicit iterative review and repair
Final output	English Markdown research report	English Markdown research report

Table 11: Principal intermediate artifacts produced by the skill-driven workflow.

Stage	Artifact	Role
Grounding	`grounded.md`	Records structured understanding of the provided source material and identifies information that should guide downstream research.
External Research	`lit.md`	Organizes retrieved external evidence and literature relevant to the grounded research topic.
Report Drafting	`summary.md`	Produces the main evidence-rich draft by integrating the provided source material with external research.
Review & Repair	`review_report.md`, `review_state.json`	Records review judgments, identified weaknesses, repair actions, and the state of iterative refinement.
Finalization	`research_report.md`	Stores the final reviewed research report used as the deliverable for evaluation and export.

In the **skill-driven** setting, the agent is explicitly instructed to load and follow the `one-report` skill. The skill structures the research process around material grounding, external research, evidence synthesis, and iterative review, with source inspection and verification integrated into the workflow. Thus, the skill-driven setting provides a more explicitly specified research procedure than direct report generation. Table 10 summarizes the main differences between the two settings.

C.2 Execution Prompts

For reproducibility, we provide the prompt templates used in the two evaluation settings. For each task, the task-specific research goal, required outputs, and execution constraints are instantiated identically in both settings. The direct-report prompt leaves the research procedure unspecified, whereas the skill-driven prompt additionally invokes the one-report skill and its runtime configuration. Figure 7 shows the core prompt templates; task-specific research goals, required outputs, and execution constraints are omitted for brevity.

C.3 Skill Pipeline and Intermediate Artifacts

The `one-report` skill serves as a top-level orchestration layer that coordinates grounding, external research, report drafting, review, and finalization. Rather than generating the final report in a single pass, the workflow maintains explicit intermediate artifacts so that each stage can be inspected and verified. For a standard single-source task, the main execution path is:

`Input` → `Grounding` → `External Research` → `Report Drafting` → `Review & Repair` → `Final Report`.

A shared `ground_id` is propagated across the workflow so that all artifacts generated from the same task remain associated with the same pipeline run. Table 11 summarizes the principal artifacts produced at each stage.

The grounding stage first converts the provided material into a structured grounded representation. Based on this artifact, the workflow generates research queries and invokes the external-research

Direct-Report Prompt

Reference Materials
```
The reference materials for this task are available in:
/workspace/input/
Inspect and use all relevant files in this directory.
```

Task
```
Produce a clear, well-structured long-form research report in English based on the
provided materials.
Add relevant background information or external context when useful and available.
Organize the report in a way that is easy for a reader to follow, and keep it focused
on the provided materials.
```

Completing Your Work
```
Save the complete final report to:
/workspace/result/research_report.md
```

Skill-Driven Prompt

```
Please use the existing one-report skill available in the current environment to
generate a full report from the provided source materials.
First load and follow the instructions of the one-report skill available in the current
environment.
```

Input
```
input_path: /workspace/input/
output_formats: md
research_mode: complex
```

Search settings
```
search_backend: native
```

Figure 7: **Prompt templates used in the direct-report and skill-driven settings.** The direct setting leaves the research procedure largely unspecified, whereas the skill-driven setting invokes the structured `one-report` workflow with its experimental runtime configuration.

stage. Under the native search configuration used in our experiments, retrieved sources must be opened and inspected before they are treated as research evidence.

The drafting stage then combines the grounded source representation with the external literature result to produce the main report draft. The review stage acts as a quality-control and refinement step: it evaluates the current draft, identifies issues requiring repair, applies revisions, and may repeat the review–repair process before producing the final report. Review states and histories are retained as explicit artifacts rather than being discarded after revision.

A workflow stage is considered complete only when its expected artifact has been successfully written. This artifact-based execution rule makes the research process inspectable and provides process-level signals in addition to the final report. Across all 202 skill-driven runs, the reported overall score is identical to the corresponding content score, indicating that the artifact contract does not alter any score reported in the skill-driven comparison. The complete skill specifications and lower-level skill implementations are included with the released code.

D Evaluation and Human Calibration Details

This section provides additional details on the criterion-level LLM judge and evaluation robustness. Section D.1 describes the judge inputs and grading prompt, Section D.2 reports human–LLM score comparisons, Section D.3 examines robustness to negative-penalty normalization, and Section D.4 analyzes hard-cap activation and its effect on final scores.

D.1 LLM JUDGE INPUTS AND PROMPT

All benchmark evaluations use GPT-5.4 as the fixed criterion-level judge with temperature 0. Each rubric criterion is evaluated independently rather than assigning a holistic score to the report.

For each judgment, the evaluator provides the judge with the task definition, an evaluator-side context pack, the final report, the complete rubric criterion in JSON format, and its criterion type (`positive`, `negative`, or `score_cap`). The final report is treated as the primary object of evaluation, while the context pack is used only to verify whether claims in the report are supported by the available source and literature evidence. Information appearing only in the context pack does not receive credit if it is absent from the final report.

Figure 8 shows the core instructions of the criterion-level judge. The full prompt, including detailed calibration rules and score ceilings, is included with the released evaluator.

```
You are an evaluator for a grounded research report benchmark.
You will evaluate exactly one rubric criterion at a time.

You will receive:
1. Task definition
2. Context pack from source grounding and literature notes
3. Final report
4. One rubric criterion in JSON format
5. Criterion type: positive, negative, or score_cap

Core grading rules:
- Evaluate only the given criterion.
- Use the final report as the main object being evaluated.
- Use the context pack only to verify whether claims in the final report are supported.
- Do not give credit for information that appears only in the context pack but is
missing from the final report.
- Do not reward verbosity, fluent writing, confident tone, or report length unless
they directly satisfy the criterion.
- Be strict about citation misuse, unsupported extrapolation, evidence-boundary errors,
and high-stakes claims.

Positive criteria:
Assign a fractional satisfaction score from 0.0 to 1.0.  Decompose multi-component
criteria into concrete sub-requirements and evaluate each separately.

Negative criteria:
Assign an error severity from 0.0 to 1.0.  A nonzero severity indicates that the
corresponding failure mode is triggered.

Hard score caps:
Determine whether the severe failure condition is triggered. A cap may be triggered
by a central failure even when other parts of the report are correct.

Return only valid JSON.
```

Figure 8: **Core instructions used by the criterion-level LLM judge.** The released evaluator contains the complete prompt with detailed calibration, evidence, coverage, depth, and grounding rules.

For positive criteria, the judge returns a satisfaction score, confidence, verdict, rationale, supporting quotes, missing elements, and an explicit decomposition of sub-requirements. For negative criteria, it returns an error-severity score together with a trigger decision, rationale, and report evidence. Hard score caps are evaluated through a separate trigger decision with an associated severity and rationale. For citation-traceability criteria, the same evidence-verification procedure is used to assess whether cited sources are verifiable, correctly attributed, and support the associated claims. These criterion-level outputs are then aggregated deterministically using the scoring procedure described in Section 2.5.

Table 12: Task-level comparison between GPT-5.4 judge and expert scores on the 24-task human-calibration subset. Each cell reports **LLM / Human**.

Task	O48-CC	G55-CX	DSV4-CC	K26-CC	Q37-CC
STEM					
Assembly Subroutine	57.3/59.4	61.1/67.9	55.0/41.7	44.7/40.1	47.4/47.3
MPC Power Plant	63.6/41.0	46.4/32.2	0.0/0.0	33.0/21.3	52.2/52.9
Kurepa Left Factorial	62.2/65.6	53.2/50.0	57.4/45.6	49.0/40.5	46.9/45.3
Extreme Weather Attribution	50.0/40.1	57.4/44.2	48.1/38.7	38.9/33.7	36.7/27.7
Humanities & Social Sciences					
Stylometry Network	42.9/27.2	51.9/48.0	8.6/0.1	28.4/21.1	50.0/36.5
Trustworthy AI Guidelines	33.4/21.1	42.7/25.3	9.0/0.8	17.9/18.9	33.4/28.1
Emotional Contagion	56.1/45.8	44.6/44.7	0.0/0.0	29.8/20.7	39.9/26.8
History & Culture	33.6/14.0	28.2/21.5	30.8/23.3	19.1/9.9	25.1/20.0
Law & Politics					
AI Training Fair Use	55.9/41.9	41.8/36.4	41.4/35.4	37.4/32.3	54.6/43.4
Data Protection Framework	33.1/18.8	45.8/37.4	39.9/41.4	10.2/12.1	49.4/49.2
BRI Adaptive Governance	55.0/43.5	47.3/30.8	50.3/45.0	26.6/21.4	56.8/61.4
IO Responsibility Attribution	50.2/39.9	33.6/18.2	37.9/36.0	37.5/34.5	53.7/41.6
Economics, Finance & Business Studies					
Trade & Inequality	47.4/30.9	36.0/20.8	34.1/25.8	17.4/9.2	45.1/28.2
Financial Statement Fraud	36.1/36.0	56.6/51.9	35.1/35.9	30.9/26.9	42.1/39.2
Green-Product Advertising	39.6/27.7	38.8/39.6	34.4/34.0	24.4/16.8	22.4/15.1
Digital Supply-Chain Resilience	31.8/16.9	30.6/26.3	34.1/23.7	23.4/13.0	37.7/26.0
Medicine & Life Sciences					
TB Activity in HIV	43.1/32.1	49.0/47.9	46.5/44.5	26.1/20.4	48.1/44.5
Neurocysticercosis POC Diagnosis	66.2/53.5	51.4/46.1	48.3/43.7	33.3/33.7	52.7/48.1
Pharmacovigilance ADE in EHR	46.7/31.9	59.1/47.4	1.5/0.0	32.6/20.4	43.0/29.2
Social-Media Health Misinformation	34.3/21.7	49.8/39.3	49.6/42.9	23.9/25.3	43.3/27.8
Arts & Design					
Impressionism	44.9/37.1	46.1/39.3	47.0/36.3	37.3/35.1	49.3/41.4
Composition & Eye Movement	33.6/16.3	34.9/23.4	0.0/0.0	16.2/9.0	52.7/32.3
E-commerce UI/UX	42.6/44.2	46.0/44.5	0.0/0.0	30.0/20.4	46.1/41.6
Conducting Gesture	58.2/38.4	48.2/33.4	36.6/22.6	43.1/29.6	27.9/22.4

Abbreviations: O48-CC = Claude Opus 4.8 + Claude Code; G55-CX = GPT-5.5 + Codex; DSV4-CC = DeepSeek V4 Pro + Claude Code; K26-CC = Kimi K2.6 + Claude Code; Q37-CC = Qwen3.7 Max + Claude Code.

D.2 Human–LLM Score Comparison

We further examine the relationship between automated and expert evaluation at the individual-task level. The calibration subset contains 24 tasks, with four tasks sampled from each of the six broad benchmark domains. For every task, reports from five representative model–harness configurations are evaluated, yielding 120 report-level comparisons.

We quantify the linear association between LLM-judge and expert scores using Pearson's correlation coefficient (Pearson, 1895). Given paired report-level scores $\{(x_i, y_i)\}_{i=1}^{N}$, we compute

$$r = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \tag{7}$$

where x_i and y_i denote the LLM-judge and expert scores for the same report, respectively.

Experts score the same reports using the same task-specific rubrics and deterministic aggregation procedure as the LLM evaluator. Using the unrounded report-level scores, Pearson correlations are 0.861 for Claude Opus 4.8 with Claude Code, 0.853 for GPT-5.5 with Codex, 0.972 for DeepSeek V4 Pro with Claude Code, 0.898 for Kimi K2.6 with Claude Code, and 0.841 for Qwen3.7 Max with Claude Code. Pooling all 120 reports yields a Pearson correlation of $r = 0.909$.

We additionally report rank correlation and uncertainty estimates. Across all 120 reports, the Spearman rank correlation between LLM-judge and expert scores is $\rho = 0.896$. Using 10,000 stratified task-cluster bootstrap resamples, where tasks are resampled within each domain and all five model–harness reports associated with a sampled task are retained together, we obtain a 95% confidence interval of $[0.894, 0.924]$ for Pearson's r and $[0.875, 0.916]$ for Spearman's ρ.

Table 12 reports the individual task scores. Each cell is shown as **LLM / Human**. For readability, task names are shortened in the table; the released calibration data retain the complete task identifiers.

Across all 120 report-level comparisons, the GPT-5.4 judge assigns an average score of 39.43, compared with 31.84 from expert evaluation, with a mean absolute difference of 8.04 points. Together with the Pearson correlation of 0.909, these results indicate that the automated evaluator tracks relative differences in report quality well, although its absolute scores are not perfectly calibrated to expert scores. We therefore use human calibration to characterize the evaluator rather than treating it as an exact replacement for expert scoring.

D.3 Robustness to Negative-Penalty Normalization

Negative criteria represent distinct task-specific failure modes, and our primary scoring rule applies their penalties additively. Since the number and total weight of negative criteria vary across tasks, we additionally test an alternative normalized formulation:

$$D_{t,a}^{\text{norm}} = M_a \frac{\sum_{j \in Neg(t,a)} |w_j| v_j}{\sum_{j \in Neg(t,a)} |w_j|}, \tag{8}$$

where M_a is the maximum score of axis a. We recompute pre-cap scores on the five Claude Code direct-report configurations, covering all 101 tasks, to isolate the effect of negative-penalty aggregation.

Table 13 shows that all five system rankings remain unchanged. The system-level Spearman correlation is $\rho = 1.000$ and Pearson correlation is $r = 0.995$. Across all 505 report-level comparisons, the corresponding correlations are $\rho = 0.953$ and $r = 0.967$.

Table 13: System-level sensitivity to negative-penalty normalization on the five Claude Code direct-report configurations. Δ is computed from unrounded scores.

Model	Original	Normalized	Δ
Claude Opus 4.8	44.75	41.95	-2.81
Qwen3.7 Max	41.02	37.60	-3.42
GPT-5.5	38.32	36.99	-1.34
Kimi K2.6	30.01	27.54	-2.47
DeepSeek V4 Pro	25.47	22.63	-2.84

Table 14 further shows that negative-rule density does not explain the weakest dimensions. Technical Accuracy and Evidence Grounding contain the most negative criteria, whereas External Research and Citation Traceability contain substantially fewer. Nevertheless, External Research and Citation Traceability remain the two lowest-scoring dimensions under both scoring formulations.

Table 14: Dimension-level sensitivity and negative-rule density. Scores and zero-clipping rates are percentages.

Axis	# Neg.	Neg./Task	$\overline{\lvert W^- \rvert}$	Original	Normalized	Zero Orig.	Zero Norm.
SU	236	2.34	10.70	63.84	59.30	6.73	8.91
ER	297	2.94	14.02	16.28	13.07	40.59	54.46
EG	449	4.45	21.46	32.13	31.74	17.03	19.60
TA	535	5.30	25.22	42.51	48.14	19.80	13.47
CT	141	1.40	6.56	11.67	4.42	42.57	77.62
SQ	102	1.01	3.50	53.39	46.33	9.11	17.82

The normalized formulation changes score magnitudes and the relative ordering of some intermediate dimensions, but leaves the system ranking unchanged and preserves the main diagnostic finding that External Research and Citation Traceability are the weakest dimensions. We therefore retain additive negative penalties as the primary scoring rule and report normalization as a robustness analysis.

D.4 Hard-Cap Activation and Score Impact

Hard score caps are designed to bound reports with severe task-specific failures. We analyze their empirical effect on the five Claude Code direct-report configurations, covering 505 report-level evaluations.

Across these runs, 64 reports (12.67%) trigger at least one hard cap. However, most of these reports already receive pre-cap scores below the corresponding cap thresholds because the same severe failures also lead to substantial losses under the ordinary rubric criteria. As a result, hard-cap activation rarely changes the final score. Only 1 of the 505 reports (0.20%) is actually score-limited by a cap, changing from 50.50 to 50.00.

Table 15: Hard-cap activation and score impact on the five Claude Code direct-report configurations.

Model	Cap-Triggered	Trigger Rate	Cap-Limited
Claude Opus 4.8	4	3.96%	0
Qwen3.7 Max	7	6.93%	0
GPT-5.5	3	2.97%	1
Kimi K2.6	7	6.93%	0
DeepSeek V4 Pro	43	42.57%	0
Total	**64 / 505**	**12.67%**	**1 / 505**

Thus, hard caps mainly serve as an upper-bound safeguard for severe failures rather than a major source of score variation in the reported results. Hard-cap activation is binary; when multiple caps are triggered, the strictest applicable threshold is used.

E Additional Results and Case Studies

This section extends the overall workflow comparison in Section 3.3 with additional analyses. Section E.1 examines the effect of the skill-driven workflow across the six evaluation dimensions, while Section E.2 presents qualitative case studies illustrating both the improvements and remaining limitations.

E.1 Dimension-wise Effect of the Skill Workflow

Table 16 compares the direct-report and skill-driven settings across the six evaluation dimensions. Both settings cover the same 101 tasks for each evaluated model–harness configuration. Dimension scores are normalized by their respective maxima, following Table 1.

Table 16: Dimension-wise comparison of Direct and +Skill on OneResearchBench. Dimension scores are percentages; Overall is on the original 0–100 scale. Bold indicates the higher score within each configuration.

Setting	SU (%)	ER (%)	EG (%)	TA (%)	CT (%)	SQ (%)	Overall
Claude Opus 4.8 + Claude Code							
Direct	**77.0**	20.9	41.1	**56.7**	15.9	60.7	44.75
+Skill	64.5	**42.7**	**58.5**	34.6	**50.4**	**72.0**	**53.07**
GPT-5.5 + Codex							
Direct	**73.0**	11.9	50.7	**49.6**	13.7	64.1	43.03
+Skill	64.2	**44.7**	**53.5**	23.9	**33.3**	**68.8**	**47.90**

Overall gains accompany stronger external research and citation support. The overall improvements of 8.32 and 4.87 points are accompanied by substantial increases in External Research and Citation Traceability. For Claude Opus 4.8 with Claude Code, these dimensions increase from 20.9% to 42.7% and from 15.9% to 50.4%, respectively. For GPT-5.5 with Codex, the corresponding changes are 11.9% to 44.7% and 13.7% to 33.3%. Evidence Grounding and Structure Quality also improve in both configurations, although the Evidence Grounding gain is much larger for Claude Opus 4.8. These results are consistent with the workflow's emphasis on literature collection, evidence integration, and report refinement.

Improvements are not uniform across capabilities or tasks. Source Understanding and Technical Accuracy decline in both configurations. In particular, Technical Accuracy drops from 56.7% to 34.6% for Claude Opus 4.8 and from 49.6% to 23.9% for GPT-5.5. Thus, better external-research and citation scores do not necessarily coincide with more accurate treatment of the provided materials and their technical content. Task-level gains are also heterogeneous: Claude Opus 4.8 improves on 73 tasks and declines on 28, while GPT-5.5 improves on 71 and declines on 30. These comparisons describe the complete workflow setting rather than isolate the effect of any individual stage. They also illustrate why OneResearchBench reports separate capability dimensions: an overall gain can otherwise obscure substantial regressions in particular capabilities.

E.2 Qualitative Case Studies

We examine two illustrative tasks evaluated with Claude Opus 4.8 and Claude Code. The first shows a large improvement in research synthesis; the second shows that technical limitations can persist despite a higher overall score. These are selected examples, not estimates of typical workflow gains. Figure 9 presents report excerpts and the corresponding evaluation evidence.

From source summary to literature-based synthesis. In the business-model innovation task, the direct report identifies the source's review scope and process framework, but its external references mostly support a reconstruction of the original article. The judge identifies limited study-by-study analysis and insufficient extension beyond the source's review period. The +Skill report instead organizes external studies by research question, method, findings, limitations, and relevance to the source framework, followed by a cross-study thematic assessment. Its overall score increases from 32.71 to 72.32, with a gain of 39.60 points computed before rounding. External Research rises from 0.0% to 72.5%, Evidence Grounding from 19.6% to 69.0%, and Structure Quality from 21.8% to 78.0%. The zero Direct ER score is the net result after penalties, not evidence that the report contains no external references. Technical Accuracy nevertheless decreases from 71.3% to 66.8%, so the large overall gain should not be read as improvement on every dimension.

Better evidence organization does not ensure technical fidelity. In the climate-and-economic-activity task, the overall score increases from 33.30 to 48.50. Source Understanding remains at 93.0% in both settings, while Evidence Grounding improves from 0.0% to 50.3% and Citation Traceability

(a) Business model innovation: extending beyond source summary

Overall: $32.71 \rightarrow 72.32$. **ER:** $0.0\% \rightarrow 72.5\%$. **EG:** $19.6\% \rightarrow 69.0\%$.

Direct report — excerpt:
"All substantive claims, statistics, and quotations are drawn from the article itself."

+Skill report — excerpt:
"The recursive demonstrations are method-bound: Wang's is a simulation with no field validation of parameters, and Ciccola's is a single interventionist case."

Evaluation evidence:
The +Skill report includes ten structured external-paper analyses and a thematic synthesis. The judge no longer triggers the criterion concerning missing paper-by-paper external analysis (`nc08`).

(b) Climate and economic activity: unresolved technical detail

Overall: $33.30 \rightarrow 48.50$. **CT:** $29.3\% \rightarrow 51.7\%$. **TA:** $0.0\% \rightarrow 1.7\%$.

+Skill report — excerpt:
"The authors use the half-panel Jackknife Fixed Effects (HPJ-FE) estimator of Chudik et al. (2018) to correct the bias and size distortion that afflict standard fixed-effects estimators in dynamic panels with weakly exogenous regressors."

Remaining issues identified by the judge:
The report names the estimator but does not adequately explain its lag structure, long-run coefficient construction, or bias-correction mechanism (`nc04`). Data provenance, deflation, and industrial-classification reconciliation are also insufficiently reconstructed (`nc02`).

Figure 9: **Illustrative cases from the Claude Opus 4.8 + Claude Code comparison.** Scores are shown as Direct → +Skill. Quoted passages are reproduced from the evaluated reports; diagnostic summaries paraphrase criterion-level judge records. The excerpts illustrate the reports' treatment of evidence, rather than independently verifying the external studies they cite.

from 29.3% to 51.7%. Nevertheless, the +Skill report receives only 1.7% in Technical Accuracy. The criterion-level judgments identify incomplete explanations of the underlying data construction and dynamic-panel estimation procedure. The low net score reflects these penalties and does not imply that every technical statement is incorrect. This case distinguishes recognizing a study's topic and broad argument from faithfully reconstructing the methodological details required by its rubric.

Together, the cases illustrate the distinction between improving the breadth and organization of research evidence and preserving precise, source-grounded technical analysis. OneResearchBench makes both aspects visible rather than treating a higher aggregate score as uniform progress in research competence.

F Extended Related Work

Additional research-agent benchmarks. Beyond the representative benchmarks compared in Table 5, several recent works study adjacent research settings. LiveResearchBench evaluates 100 expert-curated, user-centric tasks spanning daily life, enterprise, and academia, with an emphasis on dynamic live-web research and citation-grounded report evaluation (Wang et al., 2025). Researcher-Bench focuses on frontier scientific inquiry, comprising 65 research questions across 35 AI subjects and combining expert-designed rubrics with factual assessment (Xu et al., 2025). AnalystBench studies professional long-form report generation over large multimodal document collections using expert-validated quality checklists and groundedness evaluation (Pham et al., 2026). DeepResearch Bench II evaluates 132 open-web research tasks across 22 domains using 9,430 fine-grained binary rubrics derived from expert-written investigative reports through an LLM–human construction and refinement pipeline (Li et al., 2026).

These benchmarks substantially expand the evaluation of web research, scientific inquiry, document-grounded report generation, and fine-grained research quality. However, they do not formulate source-grounded research synthesis as the joint process studied in this work: beginning from user-

provided real-world materials, identifying what information remains missing, acquiring external evidence to address those gaps, and integrating source and external evidence into a grounded final report. OneResearchBench explicitly defines and evaluates this task, with task-specific criteria spanning source understanding, external research, evidence grounding, technical accuracy, citation traceability, and structure quality.

Agentic workflows for complex tasks. LLM-based agents increasingly extend beyond single-turn generation by decomposing complex tasks, interacting with external tools, maintaining intermediate states, and iteratively refining their outputs. ReAct interleaves reasoning traces with external actions, allowing agents to update their decisions based on observations from the environment (Yao et al., 2023). Toolformer studies how language models can learn when and how to invoke external APIs, demonstrating the potential of tool use to augment model capabilities beyond parametric knowledge (Schick et al., 2023). HuggingGPT further explores tool orchestration by using an LLM as a controller to coordinate specialized models for different subtasks (Shen et al., 2023).

A complementary line of work focuses on longer-horizon agent behavior and the reuse of experience across interactions. Voyager maintains reusable skills and iteratively expands its capabilities through exploration in an open-ended environment (Wang et al., 2023). Collectively, these systems illustrate several recurring design patterns in agentic workflows, including task decomposition, explicit tool interaction, intermediate-state maintenance, and reusable procedural knowledge. Such mechanisms provide an alternative to treating complex tasks as a single model invocation and enable agents to organize multi-stage execution over extended trajectories.

These ideas are particularly relevant to research-oriented tasks, which often require multiple interdependent activities such as understanding input materials, identifying missing information, retrieving external evidence, integrating findings, and revising an initial draft. In this setting, the quality of the final output depends not only on the underlying language model, but also on how the research process is organized and how information is carried across stages. Structured agentic workflows therefore provide a natural mechanism for studying whether explicit process organization can improve long-form, evidence-grounded research synthesis.