RESEARCHRUBRICS: A BENCHMARK OF PROMPTS AND RUBRICS FOR EVALUATING DEEP RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Research (DR) is an emerging agent application that leverages large language models (LLMs) to address open-ended queries. It requires the integration of several capabilities, including multi-step reasoning, cross-document synthesis, and the generation of evidence-backed, long-form answers. Evaluating DR remains challenging because responses are lengthy and diverse, admit many valid solutions, and often depend on dynamic information sources. We introduce RESEARCHRUBRICS, a standardized benchmark for DR that pairs realistic, domain-diverse prompts with expert-written, fine-grained rubrics to assess factual grounding, reasoning soundness, and clarity. We also propose a new complexity framework for categorizing DR tasks along three axes: conceptual breadth, logical nesting, and exploration. In addition, we develop human and model-based evaluation protocols that measure rubric adherence for DR agents. We evaluate several state-of-the-art DR systems and find that even leading agents like Gemini's DR and OpenAI's DR achieve under 59% average compliance with our rubrics, primarily due to missed implicit context and inadequate reasoning about retrieved information. Our results highlight the need for robust, scalable assessment of deep research capabilities, to which end we release RESEARCHRUBRICS (including all prompts, rubrics, and evaluation tools) to facilitate progress toward well-justified research assistants.

1 INTRODUCTION

An exciting development in the growing capabilities of large language models (LLMs) is the emergence of Deep Research agents: autonomous LLM-based systems that conduct multi-step web exploration, targeted retrieval, and synthesis to answer open-ended queries. Industry leaders have begun deploying such systems (e.g., OpenAI's "Deep Research" (OpenAI, 2025a) and Google's "Gemini Deep Research" (Google, 2025)), which have demonstrated strong performance on certain benchmarks (for instance, scoring 26.6% on the expert-level HLE benchmark Phan et al. (2025)). However, evaluating deep research agents poses significant challenges. Deep Research (DR) tasks are inherently open-ended: they require reasoning across multiple documents often with no single "correct" answer, and their outputs can be long and varied. Consequently, existing evaluation methods fall short: typical QA benchmarks, both general (Yang et al., 2018; Mialon et al., 2023; Phan et al., 2025; Krishna et al., 2025) and deep research specific (Java et al., 2025), focus on short, easily-verifiable factual answers and do not capture the long-form, multi-source synthesis required by DR (for example, "Which material has band gap=$0.9eV$, dislocation density=$4 \times 10^8 cm^{-2}$?" with the unique answer "GaN").

To better characterize these challenges, we introduce a **task complexity framework** for deep research. Each query can be described along three independent axes: (1) its **conceptual breadth** (the number and diversity of distinct topics or domains involved), (2) its **logical nesting depth** (the number of reasoning or decision steps required, including sub-questions and conditionals), and (3) its **exploration level** (the degree of open-

endedness or underspecification of goals). This tri-axial view highlights how DR queries differ from simpler QA tasks and motivates the need for a benchmark with fine-grained assessment criteria. Several recent efforts have sought to benchmark deep research agents, but each exhibits important limitations: for example, some benchmarks introduce LLM-generated rubrics and evaluation metrics reliant upon LLM-generated reference reports (Du et al., 2025) (thus raising concerns about circularity and limited oversight (Dorner et al., 2025)), while others are far more narrow in their scope, assessing only one specific angle of research in a technical domain (e.g., generating a "Related Works" section) (Patel et al., 2025; Li et al., 2025; Wan et al., 2025). In practice, however, users direct deep research systems toward a broad array of everyday topics, ranging from product comparisons to legal, financial, and health-related queries—underscoring the need for benchmarks that combine domain diversity with expert-authored, fine-grained rubrics capable of capturing the full spectrum of research performance.

To address these gaps, we introduce RESEARCHRUBRICS, which pairs realistic, diverse prompts with expert-authored, fine-grained rubrics for deep research. We curate queries from eight broad domains (including STEM, health, finance, legal, and common consumer questions) to reflect real-world use cases. Each prompt comes with a detailed rubric: in total, we provide 1,868 rubric criteria that check factual grounding, coherence of reasoning, completeness, relevance, and clarity of the answer. The benchmarks also include negative rubrics that specifically aim to penalize extraneous or incorrect content. Crucially, all rubrics are written and reviewed by human experts (not auto-generated), ensuring they capture nuanced, domain-specific requirements. We also develop evaluation protocols for both human and automated scoring. Following the LLM-as-judge paradigm, we use powerful LLMs to assess rubric compliance, and we systematically experiment with improving this process comparing binary vs. ternary grading for each criterion and the level of detail in the rubrics. Finally, we apply our framework to leading DR systems (OpenAI's DeepResearch (OpenAI, 2025a), Google Gemini's Deep Research (Google, 2025), and Perplexity's Deep Research (AI, 2025)). The results show that even the strongest agents fall below 59% average rubric compliance, revealing substantial room for improvement in multi-document synthesis and rigorous justification.

Our contributions are as follows:

- **A human-crafted benchmark for deep research.** We present RESEARCHRUBRICS, a suite of open-ended research tasks across diverse domains, each with an expert-written rubric (1,800 + total criteria). This is, to our knowledge, the first benchmark combining such domain breadth with fine-grained human evaluation for DR agents.
- **A task complexity framework.** We formalize deep research queries along three axes – **breadth**, **depth**, and **ambiguity** – to distinguish them from conventional QA tasks and to guide the construction of balanced benchmarks.
- **Rubric-based, open-ended evaluation.** We demonstrate that fine-grained rubrics provide rigorous, multi-dimensional evaluation of long-form, multi-source research answers that closely align with expert judgments, while also enabling diagnoses of model strengths and weaknesses.
- **Scalable LLM judging with ablations.** We introduce enhanced prompt design and scoring recommendations for LLM-as-a-judge evaluation that improve agreement with human evaluators and are validated through ablation studies.

By releasing RESEARCHRUBRICS and its tools, we aim to catalyze progress toward trustworthy, well-justified DR assistants for complex, open-ended research tasks in a multitude of domains.

2 RELATED WORK

Early benchmarks have largely taken two approaches: deriving or constructing tasks from static corpora or relying on expert-curated questions.

Derived Benchmarks AcademicBrowse (Zhou et al., 2025) and BrowseComp (Wei et al., 2025) assess retrieval from academic papers or the web, while ResearchBench (Liu et al., 2025) builds complex queries from static data. More recent work goes further and derives tasks from dynamic, real-world scenarios. DeepScholar-Bench (Patel et al., 2025) evaluates systems on related work writing using live queries from arXiv papers, though it is specialized to academic synthesis and uses automated metrics. ReportBench (Li et al., 2025) leverages published surveys as ground truth, measuring overlap with expert-written reviews but prioritizing replication. DeepResearch Arena (Wan et al., 2025) automatically curates 10,000 open-ended tasks from academic seminars, pairing them with adaptively generated rubrics, though automatic rubric generation can miss domain nuances. However, static datasets risk data leakage, cannot adapt to new information, and automatic rubric generation can miss domain nuances.

Expert Curated Benchmarks Expert-authored benchmarks include Humanity's Last Exam (HLE) (Phan et al., 2025), which provides 2,500 expert-written short-answer questions across advanced domains, but does not target more ambiguous / open-ended analysis directly, and DeepResearch Bench (Du et al., 2025), which introduced 100 PhD-level problems requiring long-form reports. DeepResearch Bench confirmed the difficulty of research tasks (no model exceeded 30%) but had a number of critical weaknesses, including using LLM-generated rubrics for specialized domans, evaluation metrics reliant upon LLM-generated reference reports and simplistic reference overlap metrics. ExpertLongBench (Ruan et al., 2025) similarly targets expert-level, long-form tasks across 9 domains with domain-specific rubrics, using the CLEAR framework for fine-grained assessment, though it depends on high-quality references.

In contrast to benchmarks that rely on static answer keys or coarse metrics, RESEARCHRUBRICS offers a middle ground: realistic research queries (academic and everyday domains) paired with expert-written rubrics assessing grounding, synthesis, reasoning, clarity, and citation usage. By using human-written rubrics with LLM judges, we avoid simplistic overlap measures while maintaining scalability. RESEARCHRUBRICS complements efforts like ExpertLongBench and DeepResearch Arena, emphasizing domain diversity and rubric quality.

3 OVERVIEW OF RESEARCHRUBRICS

RESEARCHRUBRICS consists of 75 single-turn prompts, each paired with a set of 20–60 prompt-specific rubric criteria. Every prompt and criterion in RESEARCHRUBRICS was written and iteratively refined by human experts to ensure clarity and relevance (no criteria were seeded or generated by LLMs). The prompts cover a wide range of topics and inquiry types to emulate real user questions that deep research agents receive. In total, the benchmark contains 1,868 unique rubric items, enabling a fine-grained assessment of open-ended, realistic research queries. Figs. 2 and 3 provides an overview of our benchmark design and evaluation process.[1]

3.1 DATA COLLECTION AND TASK DOMAINS

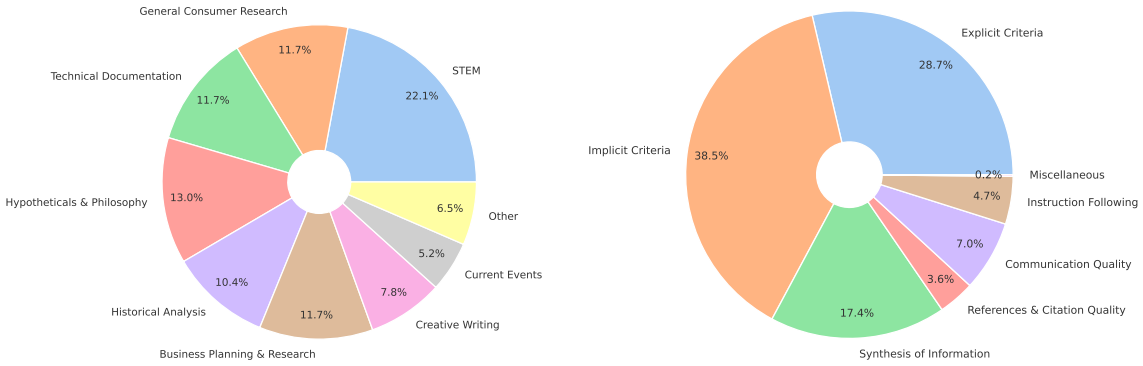
Our data collection pipeline consists of three expert participants, as shown in Fig. 3. In this context, we define an "expert" as an individual with a strong STEM background who is skilled in task design and evaluation, rather than a domain-specific specialist for each prompt. All participants in our data collection only chose and worked on domains they were familiar with.

Expert 1 initially proposes a prompt and a set of rubric criteria. The initial proposal is reviewed by Expert 2, who provides feedback and an updated prompt and rubrics. Experts 1 and 2 continue iterating until Expert 2 approves the pair of prompt and rubric. Finally, Expert 3 reviews the proposed prompt and rubrics and

[1] A selection of the benchmark tasks have been uploaded as supplementary material. A public release of the dataset will shortly follow. The code to auto-evaluate on this benchmark will also be publicly released soon.

makes final adjustments to the data. This three-participant setup ensures high quality in the final data, thanks to the numerous reviews of each component of each sample.

We curated prompts from **eight broad categories** (Fig. 1a) to maximize diversity, namely **General Consumer Research**, **STEM**, **Technical Documentation**, **Creative Writing**, **Creative Writing**, **Hypotheticals and Philosophy**, **Current Events**, and **Business Planning and Research**. For more details, see Table 8 in Section A.8.



(a) Distribution of task domains in our collected data. The dataset has a fairly even spread across the task domains.

(b) Distribution of rubric criteria. The majority of rubric criteria is either an Explicit Criterion (28.7%) or an Implicit Criterion (38.5%).

Figure 1: Overview of the dataset composition, showing the distribution of (a) task domains and (b) rubric evaluation criteria.

To compile realistic questions, we drew inspiration from user forums, Q&A sites, and brainstorming sessions, then had domain experts refine each prompt for clarity and appropriate difficulty. Our goal was to cover both **breadth** (many different domains) and **depth** (challenging multi-step problems) in the dataset. Each prompt was then assigned to one or more expert annotators to create the rubric: a list of criteria specifying what an ideal answer should include and common errors to avoid. The rubric creation process involved multiple passes to ensure completeness and remove ambiguity. We weighted each criterion based on its importance (see Section 3.3) and included negative criteria targeting likely pitfalls (e.g., factually incorrect statements, off-topic tangents, or disallowed content).

STEM and general consumer queries constitute the largest portions, reflecting both specialized academic topics and everyday research questions. Other categories provide targeted challenges (e.g., historical sources, creative synthesis, or real-time news retrieval). This diversity ensures that a DR agent must draw on a wide range of knowledge sources and adapt to different task structures.

3.2 Prompt Complexity Dimensions

Not all research prompts are equal—some involve a broader knowledge base, others require deeper reasoning, and others are underspecified and exploratory. We categorize each ResearchRubrics task along three orthogonal complexity dimensions: **Conceptual Breadth**, **Logical Nesting Depth**, and **Exploration** (Table 1). This framework helps ensure our benchmark covers a balanced mix of task types and allows analysis of where agents struggle most.

Every task in ResearchRubrics is annotated with a triplet of (Breadth, Depth, Ambiguity) labels. In our evaluations, we analyze model performance across these dimensions to see, for example, if a model strug-

Complexity Axis	Level	Examples
Conceptual Breadth	Simple	A math word problem or a factual lookup from one source.
	Moderate	A prompt combining two fields (physics concept applied in a medical device context).
	High	"Analyze the environmental, economic, and political factors affecting renewable energy adoption in Asia."
Logical Nesting	Shallow	"What is the capital of X country?"
	Intermediate	"Find the sales of Company A and Company B last year and determine who grew faster; then identify one reason for that difference."
	Deep	"Develop an evidence-backed investment strategy given current economic indicators, then stress-test it against at least two historical scenarios and suggest contingency plans."
Exploration	Low	"Summarize the methodology of the referenced paper." The task is clear-cut.
	Medium	"Discuss the benefits and risks of AI in healthcare."
	High	"I want to switch to a career with strong future growth, what should I consider?"

Table 1: Prompt complexity categories used to annotate each task in RESEARCHRUBRICS.

gles more with breadth (integrating many sources) or with depth (long reasoning chains). This also helps researchers filter the benchmark for specific experiment focuses (e.g., testing only high-depth reasoning tasks).

3.3 RUBRIC DESIGN AND EVALUATION SCHEME

RESEARCHRUBRICS is fundamentally a rubric-based benchmark: each prompt is judged against a tailored set of criteria that define the requirements of a good answer. Table 2 provides an overview of the types of rubric criteria we include. The criteria are grouped into six broad **evaluation axes** to cover different aspects of quality. The weights for the criteria are in the range -5 to 5, with each weight corresponding to a clear

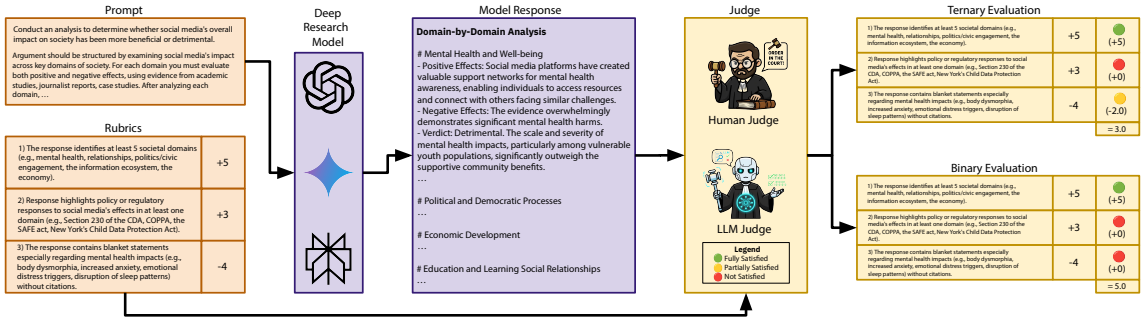


Figure 2: Overview of RESEARCHRUBRICS and its evaluation pipeline.

human preference (Critically Detrimental, Detrimental, Slightly Detrimental, Slightly Important, Important, Critically Important) as described in Table 7, to encourage better agreement of the model-based grader. Criteria with weights in the ranges $[-4, -5]$ and $[4, 5]$ are mandatory criteria, while criteria with weights in the range $[-3, 3]$ are optional. Criteria with mandatory weights consist of guidelines for a minimum viable response, while criteria with optional weights include "nice-to-have" behaviors. Negative criteria are carefully chosen by experts to target common mistakes a response may make and prevent the model from reward hacking for unnecessary length and detail in the responses.

Criterion	Description
Explicit Requirements	Checks whether the answer addressed all points explicitly asked in the prompt, and did so correctly. For example, if the prompt says "compare X and Y and recommend one," the rubric will have items to verify that the answer compared X vs. Y on relevant traits and made a clear recommendation.
Implicit Requirements	These criteria cover relevant points that a well-informed person might expect, even if not directly asked. For instance, a question about a medical treatment might implicitly require mentioning side effects or costs. We include such criteria to reward comprehensive answers that demonstrate a deep understanding of the context.
Synthesis of Information	These criteria evaluate the model's ability to connect and synthesize information across multiple sources or sub-parts of the query. Rather than just listing facts, does the answer draw novel insights or conclusions? For example, after gathering evidence from several studies, a synthesis criterion might check if the answer provided an integrated comparison or identified an overarching trend.
Use of References	For tasks that expect external citations or evidence, these criteria check whether the answer included specific appropriate references (e.g., particularly relevant URLs or academic citations by name) and whether those references actually support the claims.
Communication Quality	We include criteria on clarity, organization, and style, as a factually correct answer may still fail on these. These check whether the answer is well-structured (with logical flow, headings or bullet points if appropriate), whether it is concise yet complete, and whether it uses a tone and terminology suitable for the audience (e.g., not too much jargon if a casual medium, e.g., a blog is requested).
Instruction Following	If the user prompt contains specific instructions or constraints (formatting requirements, a request for a certain perspective, exclusion of some info, etc.), we include criteria to verify adherence. For example, if the prompt says "do not discuss Topic Z," a negative criterion will trigger if the answer mentions Topic Z. Following user instructions is critical for useful assistance.

Table 2: Rubric criteria used to evaluate responses.

Evaluation Methodology Each model response is evaluated against all the rubric criteria using a model as a grader, in an LLM-as-a-judge setup. The model-based grader outputs ternary judgment verdicts for each rubric, which are {`Satisfied`, `Partially Satisfied`, `Not Satisfied`}. This scoring process is the same for negative criteria, which are phrased so that the negative weights are applied to the sum if the negative criteria are met. The final task score is the weighted sum of all positive and negative weights, normalized by sum of the absolute weights (to prevent heavy penalization for any negative rubrics).

$$S_k = \frac{\sum_{r_i \in C} w_{r_i} m_{r_i}}{\sum_{r_i \in C} \text{abs}(w_{r_i})}, \qquad m_{r_i} = \text{Judge}(P_k, \text{Res}, r_i) \tag{1}$$

where S_k is the final task score for the task k with prompt P_k, C is the set of all criteria, w_{r_i} is the weight of criterion r_i (positive or negative), and m_{r_i} is the ternary indicator returned from the model-based judge Judge as a score dependent on the task prompt, model response and i-th rubric item with the following output space: $m_{r_i} = 1$ if criterion r_i is satisfied, $m_{r_i} = 0.5$ if criterion r_i is partially satisfied, and $m_{r_i} = 0$ if criterion r_i is not satisfied.

Human Consistency Analysis. Similarly to HealthBench (Arora et al., 2025), we utilize the Macro $F1$ score to validate the effectiveness of using a model-based grader as a proxy for human judgment. In our setup, we compare the ground truth predictions of experts and model-based graders for each task, and compute the $F1$ scores for each of the classes {`Satisfied`, `Partially Satisfied`, `Not Satisfied`}.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}, \text{ where precision} = \frac{TP}{TP + FP} \text{ and recall} = \frac{TP}{TP + FN} \tag{2}$$

We also run ablation studies to isolate the most significant factors in the level of alignment between the model-based grader and human judgments. For more details see Appendix A.7.

4 EXPERIMENTAL RESULTS

We evaluated several state-of-the-art deep research agents and baseline LLMs on RESEARCHRUBRICS.

Evaluated Agents & Models. We focus primarily on three commercial, closed-source state-of-the-art Deep Research Agents: OpenAI Deep Research OpenAI (2025a), Gemini Deep Research Google (2025), and Perplexity Deep Research AI (2025) for which we also curate gold standard, human-judged evaluations. These systems represent the most widely deployed frontier-level agents for retrieval-augmented reasoning and multi-step synthesis. For a comparative baseline, we also tested baseline LLMs with integrated search tools, using the Open Deep Search framework Alzubi et al. (2025), though these outputs are excluded from the human agreement study due to tradeoffs regarding resource constraints and their especially poor performance. Full results and per-category failure breakdowns are presented in Table 3, Table 4, and Table 5.

Implementation Details. For evaluation, we investigate the effectiveness of four LLMs-as-judges: GPT-4.o (OpenAI, 2025b), GPT-5 (OpenAI, 2025c), Claude-Opus-4.1 (Anthropic, 2025) and Gemini-2.5-Pro DeepMind (2025). Alignment with human annotations was measured using Macro F1 under both ternary {`Satisfied`, `Partially Satisfied`, `Not Satisfied`} and binary {`Satisfied`, `Not Satisfied`} grading regimes.

4.1 MAIN RESULTS: AGENT PERFORMANCE

Table 3 presents the overall rubric compliance scores for each system, calculated using the formula described in 1. Several clear trends emerge:

Table 3: Evaluation results for commercial DR agents and LLMs with search tools. The 'Final Score' column shows the final score averaged across all tasks, while the remaining columns report failure rates (in %) per category (ratios of how many rubrics of a specific category failed compared to all the failed rubrics).

			Final Score	Failure Rates (%)					
	Grader Verdicts	**Model**	**Overall**	**Comm. Quality**	**Explicit**	**Implicit**	**Instruct. Follow.**	**References**	**Synthesis of Info.**
		Deep Research Agent							
Human-evaluated	Ternary	Perplexity DR	0.490	14.2%	27.1%	47.9%	18.8%	17.3%	26.3%
		Gemini DR	**0.586**	15.6%	31.0%	45.1%	22.4%	16.6%	30.4%
		OpenAI DR	0.569	15.8%	28.0%	47.9%	22.5%	16.1%	28.3%
	Binary	Perplexity DR	0.425	10.9%	24.5%	43.2%	12.6%	13.3%	26.9%
		Gemini DR	**0.535**	13.0%	26.7%	42.2%	17.1%	13.4%	27.3%
		OpenAI DR	0.512	12.8%	25.2%	44.0%	15.6%	13.3%	26.9%
		LLM with Search Tools							
Model-evaluated	Binary	Claude-4.1-Opus	0.026	8.71%	28.85%	39.96%	8.56%	7.36%	21.45%
		Gemini-2.5-Pro	-0.008	8.34%	28.76%	39.42%	8.53%	7.12%	21.04%
		GPT-5	**0.163**	9.41%	26.92%	41.11%	10.85%	8.19%	24.97%

- **Agent Performance Variability:** Based on Table 3, Google's Gemini DR agent achieves the highest overall compliance under both the ternary ($\sim$0.59) and binary ($\sim$0.54) grading schemes, with OpenAI's DR agent a close second overall ($\sim$0.57 for ternary and $\sim$0.51 for binary), while Perplexity's agent lags with $\sim$0.49 overall for ternary (and $\sim$0.43 for binary). Implicit criteria account for the lion's share of rubric failures across the 3 agents, but when compared to rubric axes distribution in the benchmark1, most of the categories are overrepresented in the failure rates of the agents: references ($\sim$ 4% vs. 17%), instruction following ($\sim$ 5% vs. 21%), implicit ($\sim$ 38% vs. 47%), synthesis of Information ($\sim$ 17% vs. 28%) and communication quality ($\sim$ 7% vs. 14%); only explicit criteria seems to be represented consistently ($\sim$ 27% vs. 28%). This suggests that the agents are generally effective at relaying accurate facts (explicit criteria) but face challenges in integrating knowledge, reasoning implicitly and conveying it effectively.
- **Binary vs. Ternary Scoring:** Introducing ternary judgments doesn't seem to offer more granularity compared to binary verdicts, as we see only slight increases in performance through assigning partial credit. Switching to binary judgments reduces absolute scores as expected, but the relative ranking of agents remains unchanged. Therefore, rubric benchmarks can comprehensively be graded via binary judgments.
- **Baseline LLM+Tools:** Table 3 also reports results for LLMs with search tools under the binary rubric. These models are not directly comparable to the specialized Deep Research Agents, as they are prompted from scratch for each query and receive no fine-tuning on research tasks, but probe the importance of search in such an agent via the benchmark. Among them, GPT-5 achieved the highest final score (0.163), while Claude-4.1 and Gemini-2.5-Pro performed far lower overall (0.026 and -0.008, respectively, with similar patterns of relative strengths and weaknesses. A pattern we noticed was that LLMs with search tools often struggled with the hypothetical questions in the benchmark, as they often fell outside the purview of retrieved content. These results indicate that while LLMs with search tools can handle straightforward tasks and maintain good performance in communication and referencing, they remain far behind specialized agents in integrating knowledge and satisfying nuanced rubric criteria.

Overall, the results confirm that **no current system is close to "passing" our benchmark**. The best agents are around 60% compliance. Qualitatively, we found that agents often do well in gathering factual information (especially OpenAI and Gemini), but they falter in higher-order synthesis.

4.2 Human Consistency of LLM Judging

We next assess how closely the LLM-based rubric evaluations align with human judgments. We had 9 expert annotators manually grade the outputs of the three commercial agents on the entire benchmark (total of 225 responses). Table 4 summarizes the consistency results.

Table 4: Human consistency evaluation of Deep Research Agents under binary and ternary grading schemes. Values represent the Macro F1 scores between the human and model judgments.

	Agent	**Human Consistency Evaluation**			
		GPT-4o	GPT-5	Claude-4.1 (Opus)	Gemini-2.5-Pro
Binary	Perplexity DR	0.687	0.725	0.687	**0.742**
	Gemini DR	0.682	**0.742**	0.719	0.737
	OpenAI DR	0.679	0.725	0.707	**0.727**
Ternary	Perplexity DR	0.478	0.547	0.524	**0.554**
	Gemini DR	0.513	0.538	0.511	**0.549**
	OpenAI DR	0.488	**0.547**	0.545	**0.547**

Using the ternary rubric, human consistency with the top-performing agents ranged from approximately 0.478 to 0.554 (Macro F1), indicating moderate agreement among graders. Agreement was highest for Gemini-2.5-Pro, followed closely by GPT-5, reflecting more consistent judgments on clear successes and failures, while borderline cases still showed some variability. When collapsed to a binary setup, the Macro F1 values increased, with the highest consistency reached by 0.742 for Gemini-2.5-Pro (and occasionally matched by GPT-5) across different agents, demonstrating substantial agreement once partial credit was treated as a failure. Overall, these results suggest that human graders are generally reliable on these tasks, particularly for high-performing agents, and the level of agreement is comparable to prior studies of LLM evaluations on open-ended research tasks (e.g., (Arora et al., 2025; Du et al., 2025; Patel et al., 2025)).

4.3 Ablation: Examples and Rubric Augmentation

Table 5 compares agent performance under variations in rubric detail and LLM augmentation. Providing detailed examples slightly improved agreement with human judgments, with scores increasing by roughly 2–3 percentage points across agents, while overall model rankings remained stable. In contrast, LLM-augmented rubrics, where criteria were rephrased by an LLM to add more detail and examples, significantly lowered performance, suggesting that the original expert-written criteria were already clear and that automatic rephrasing may introduce ambiguity. These results reinforce practical recommendations to provide brief illustrative examples for potentially ambiguous rubric items to enhance consistency and involve expert guidance in writing rubrics as much as possible.

Table 5: Evaluation of the grader model (GPT-4o) alignment (Macro F1 scores) under different conditions of example detail and LLM augmentation. Columns 'Low' and 'High' report scores for prompts with minimal versus detailed examples, while 'Absent' and 'Present' indicate whether LLM augmentation was used.

Model	**Example Detail**		**LLM-Augmentation**	
	Low	High	Absent	Present
Perplexity DR	0.665	**0.687**	**0.687**	0.497
Gemini DR	0.680	**0.682**	**0.682**	0.565
OpenAI DR	0.652	**0.679**	**0.679**	0.518
Perplexity DR	0.418	**0.478**	**0.478**	0.332
Gemini DR	0.457	**0.513**	**0.513**	0.361
OpenAI DR	0.433	**0.488**	**0.488**	0.340

5 Conclusion and Future Work

We introduced ResearchRubrics, a new benchmark and evaluation framework for deep research agents that emphasizes fine-grained, human-aligned assessment. Through 75 diverse research challenges and over 1,868 expert-crafted rubric criteria, our benchmark provides a multi-dimensional lens on an agent's performance—checking not just factual recall, but the completeness, reasoning soundness, source usage, and clarity of its responses. Our experiments with state-of-the-art systems reveal that today's best agents achieve only around 60% compliance with these rigorous rubrics, leaving significant room for improvement. In particular, agents often fall short in integrating information across documents and in providing well-justified answers with proper citations. These findings echo the broader observation that current LLMs, despite their fluency, can struggle with the deeper validation and synthesis needed for trustworthy research assistance.

REFERENCES

Perplexity AI. Introducing perplexity deep research, 2025. URL https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research. Accessed: 2025-09-18.

Salaheddin Alzubi, Creston Brooks, Purva Chiniya, Edoardo Contente, Chiara von Gerlach, Lucas Irwin, Yihan Jiang, Arda Kaz, Windsor Nguyen, Sewoong Oh, Himanshu Tyagi, and Pramod Viswanath. Open deep search: Democratizing search with open-source reasoning agents, 2025. URL https://doi.org/10.48550/arXiv.2503.20201.

Anthropic. Claude opus 4.1, 2025. URL https://www.anthropic.com/news/claude-opus-4-1.

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025. URL https://doi.org/10.48550/arXiv.2505.08775.

Google DeepMind. Gemini 2.5 pro: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf, 2025. Accessed: 2025-09-25.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: Llm as judge won't beat twice the data. In The 13th International Conference on Learning Representations (ICLR) 2025, January 2025. URL https://arxiv.org/abs/2410.13341.

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.

Google. Gemini deep research — your personal research assistant, 2025. URL https://gemini.google/overview/deep-research/. Accessed: 2025-09-18.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, et al. Towards an ai co-scientist. arXiv preprint arXiv:2502.18864, 2025.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Meng Fang, Linyi Yang, Xuehai Li, Lijuan Shang, Shimin Xu, Jun Hao, Kun Shao, and Jiaxin Wang. Deep research agents: A systematic examination and roadmap. arXiv preprint arXiv:2506.18096, 2025.

Abhinav Java, Ashmit Khandelwal, Sukruta Midigeshi, Aaron Halfaker, Amit Deshpande, Navin Goyal, Ankur Gupta, Nagarajan Natarajan, and Amit Sharma. Characterizing deep research: A benchmark and formal definition. arXiv preprint, arXiv:2508.04183, August 2025. preprint.

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, 2025. URL https://doi.org/10.48550/arXiv.2409.12941.

Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. Reportbench: Evaluating deep research agents via academic survey tasks. arXiv preprint arXiv:2508.15804, 2025.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. arXiv preprint arXiv:2503.21248, 2025.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: A benchmark for general ai assistants, 2023. URL https://doi.org/10.48550/arXiv.2311.12983.

OpenAI. Introducing deep research, 2025a. URL https://openai.com/index/introducing-deep-research/. Accessed: 2025-09-18.

OpenAI. Introducing gpt-4.1, 2025b. URL https://openai.com/index/gpt-4-1/.

OpenAI. Introducing gpt-5, 2025c. URL https://openai.com/gpt-5/.

OpenAI. Introducing openai o3 and o4-mini, 2025d. URL https://openai.com/index/introducing-o3-and-o4-mini/.

Liana Patel, Negar Arabzadeh, Harshit Gupta, Ankita Sundar, Ion Stoica, Matei Zaharia, and Carlos Guestrin. Deepscholar-bench: A live benchmark and automated evaluation for generative research synthesis. arXiv preprint arXiv:2508.20033, 2025.

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

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.

Guijin Son, Jiwoo Hong, Honglu Fan, Heejeong Nam, Hyunwoo Ko, Seungwon Lim, Jinyeop Song, Jinha Choi, Gonçalo Paulo, Youngjae Yu, et al. When ai co-scientists fail: Spot-a benchmark for automated verification of scientific research. arXiv preprint arXiv:2505.11855, 2025.

Haiyuan Wan, Chen Yang, Junchi Yu, Meiqi Tu, Jiaxuan Lu, Di Yu, Jianbao Cao, Ben Gao, Jiaqing Xie, Aoran Wang, Wenlong Zhang, Philip Torr, and Dongzhan Zhou. Deepresearch arena: The first exam of llms' research abilities via seminar-grounded tasks. arXiv preprint arXiv:2509.01396, 2025.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Passos, William Fedus, and Amelia Glaese. Browsecomp: A simple yet challenging benchmark for browsing agents. arXiv preprint arXiv:2504.12516, 2025.

Ruixuan Xu and Jian Peng. A comprehensive survey of deep research: Systems, methodologies, and applications. arXiv preprint arXiv:2506.12594, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Junting Zhou, Wang Li, Yiyan Liao, Nengyuan Zhang, Tingjia Miao, Zhihui Qi, Yifan Wu, and Tongshuang Yang. Academicbrowse: Benchmarking academic browse ability of llms. arXiv preprint arXiv:2506.13784, 2025.

A APPENDIX

A.1 AI USE DISCLOSURE

LLMs were used as writing assistants, both to rephrase human-written text and generate initial drafts that were refactored by people.

A.2 EXTENDED RELATED WORK

The rapid emergence of deep research agents has been accompanied by several efforts to characterize and evaluate their capabilities. Recent surveys and roadmap papers highlight the promise and challenges of autonomous LLM-based research assistants. For example, Huang et al. (2025) provide a systematic examination of Deep Research agents, analyzing their tool integration and planning strategies, while Xu & Peng (2025) offer a comprehensive survey of deep research systems and applications. These works underscore the need for robust evaluation frameworks aligned with the complex, open-ended nature of research tasks.

Early benchmarks for deep research agents have largely taken one of two approaches: constructing tasks from static corpora or relying on expert-curated questions. In the first category, benchmarks like **AcademicBrowse** (Zhou et al., 2025) and **BrowseComp** (Wei et al., 2025) assess an agent's ability to navigate and retrieve information from academic papers or the web. AcademicBrowse focuses on literature-based queries (e.g., browsing academic papers for answers), and BrowseComp comprises over 1,200 web questions that demand multi-hop searching across sites. While these benchmarks test long-horizon retrieval and factual accuracy, their questions tend to have a predetermined scope or "ground truth" answers, which simplifies evaluation to matching reference facts. This limits their ability to capture the open-ended synthesis and exploratory aspect of real research inquiries. Another example is **ResearchBench** (Liu et al., 2025), which builds complex search questions from static data; however, static benchmarks risk data leakage (i.e., answers appearing in training data) and cannot adapt to newly emerging information.

The second category of benchmarks uses expert-authored tasks to evaluate research reasoning. **Humanity's Last Exam** (HLE) (Phan et al., 2025) is an expansive evaluation of 2,500 expert-written questions covering advanced domains ranging from mathematics to medicine. HLE revealed significant gaps in state-of-the-art models' knowledge, but it primarily consists of challenging short-answer questions, rather than multi-document analytical tasks. Closer to our setting, **DeepResearch Bench** (Du et al., 2025) introduced 100 PhD-level research problems across 22 fields (e.g., scientific analysis, legal reasoning), each requiring a long-form report. Their evaluation combines reference-based metrics and adaptive criteria, including measuring the number and accuracy of citations. This benchmark confirmed the difficulty of deep research tasks, where no model exceeded roughly 30% on their overall metrics, yet its scoring approach leans heavily on overlap with reference solutions and simple citation counts. Similarly, **ExpertLongBench** (Ruan et al., 2025) targets expert-level, long-form tasks in 9 domains (law, finance, healthcare, etc.), providing 11 complex prompts each accompanied by a domain-specific checklist or rubric. ExpertLongBench introduced the CLEAR evaluation framework, which extracts a structured checklist from both the model's output and a gold reference, then compares them for alignment. This method enables fine-grained assessment of content requirements, but it depends on high-quality reference outputs for each task. In contrast, our work uses expert-written criteria without assuming an ideal reference answer, and evaluates responses directly via LLM-as-a-judge – avoiding potential biases from any single ground-truth essay.

More recent benchmarks have moved toward dynamic, real-world research scenarios. **DeepScholar-Bench** (Patel et al., 2025) focuses on generative research synthesis: it draws live queries from recent arXiv papers and evaluates systems on writing a related work section by retrieving and summarizing up-to-date literature. Its evaluation emphasizes three axes (knowledge synthesis, retrieval quality, and verifiability), rewarding comprehensive coverage of relevant work and correct citation of sources. However, DeepScholar-Bench is specialized to academic writing tasks, and uses automated metrics (including LLM-generated

scores) which may introduce evaluation circularity. **ReportBench** (Li et al., 2025) takes another automated approach by leveraging existing survey articles as ground truth for evaluation. It generates academic survey-style prompts and measures the overlap between the AI agent's citations and statements and those in a published survey on the same topic. This provides a concrete correctness signal (since an expert-written literature review is treated as the gold standard), but inherently prioritizes replication of the reference content over creative or divergent but valid answers. Meanwhile, **DeepResearch Arena** (Wan et al., 2025) addresses the authenticity of research prompts: it automatically curates over 10,000 open-ended tasks from transcripts of academic seminars across 12 disciplines. By capturing questions that arise organically in expert discussions, DeepResearch Arena aims to evaluate agents on more ill-defined, exploratory problems. Their evaluation combines factual grounding checks with adaptively generated rubrics (checklists) to handle the breadth of tasks. One limitation, however, is that fully automatic rubric generation can miss domain nuances or implicitly favor certain solution paths.

In parallel to benchmarking efforts, researchers have begun exploring AI "co-scientist" systems that autonomously propose hypotheses or experimental plans beyond just information retrieval. Notably, Gottweis et al. (2025) present an **AI Co-Scientist** built on a multi-agent Gemini 2.0 system, which iteratively generates and refines scientific hypotheses (demonstrated in drug discovery and biology domains). The advent of such systems raises the stakes for evaluation: beyond finding correct facts, we must assess whether an AI's reasoning and conclusions hold up to expert scrutiny. Initial work in this vein includes benchmarks like SPOT (Son et al., 2025), which checks AI-generated scientific papers for logical errors or inconsistencies. Overall, as deep research agents expand from answering questions to performing nuanced scientific investigations, the need for **fine-grained, human-aligned evaluation** becomes ever more critical.

Our work builds directly on these prior insights. In contrast to previous benchmarks that either rely on static answer keys or on coarse-grained metrics, RESEARCHRUBRICS offers a new middle ground: a broad collection of realistic research queries (spanning academic and everyday domains) paired with expertly crafted rubrics that detail the requirements of a good answer. This approach enables evaluation of multiple dimensions – factual grounding, cross-source synthesis, reasoning validity, clarity, and citation usage – within a single unified framework. By using human-written rubrics and having LLM judges apply them, we avoid reward hacking based on simplistic overlap measures, while still achieving scalable scoring. RESEARCHRUBRICS is complementary to contemporaneous efforts like ExpertLongBench and DeepResearch Arena: those benchmarks target either highly specialized expert tasks or massive automatically generated task suites, whereas we prioritize diversity of domains and manually quality-checked criteria. Together, these efforts push toward a more rigorous and comprehensive assessment of deep research capabilities.

A.3 EXTENDED DATA COLLECTION AND TASK DOMAINS

We curated prompts from **eight broad categories** (Fig. 1a) to maximize diversity. These include:

- **General Consumer Research:** Complex decision-making or personal advice queries, e.g., finding an apartment given constraints, product comparisons, travel or event planning, personal finance, and legal advice;
- **STEM:** Scientific and technical questions requiring synthesis from academic papers or textbooks;
- **Technical Documentation:** Explaining code, APIs, or engineering concepts using official docs or manuals;
- **Creative Writing:** Long-form creative tasks incorporating researched facts or themes;
- **Hypotheticals and Philosophy:** Open-ended thought experiments, predictions, or ethical dilemmas requiring multi-perspective analysis;
- **Current Events:** Queries on recent or ongoing news topics that require up-to-date information;

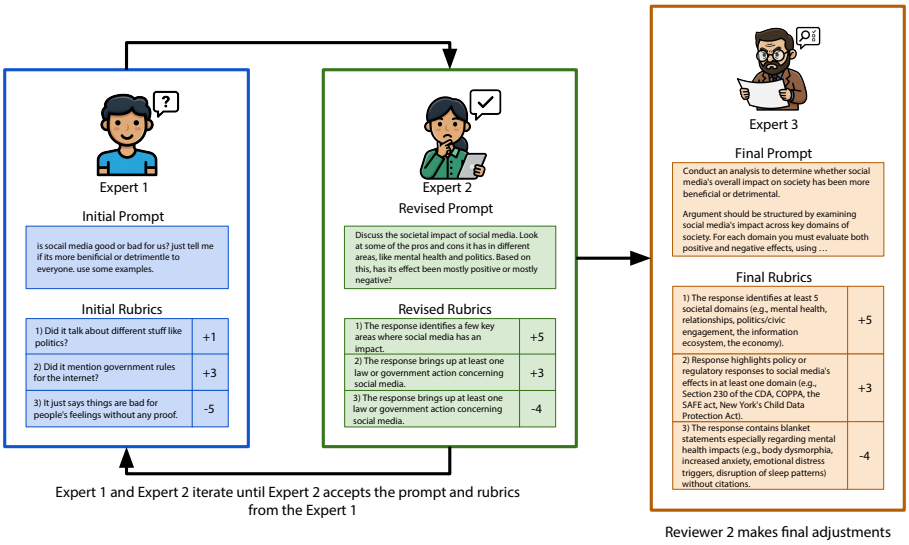


Figure 3: The three-stage pipeline for creating and refining prompts and rubrics. An initial draft by Expert 1 is iteratively improved with Expert 2 before a final review and adjustment by Expert 3.

- **Business Planning and Research:** Tasks related to business strategy, corporate finance, law, procurement, marketing, etc.;
- **Other:** A small number of prompts fell outside the groups defined above.

For more details, see Table 8 in Section A.8.

A.4 Extended Prompt Complexity Dimensions

Including examples

A.5 Extended Rubric Design and Evaluation Scheme

A.6 Extended Experimental Setup

Evaluated Agents & Models. We focus primarily on three commercial, closed-source state-of-the-art Deep Research Agents: OpenAI Deep Research OpenAI (2025a), Gemini Deep Research Google (2025), and Perplexity Deep Research AI (2025) for which we also curate gold standard, human-judged evaluations. These systems represent the most widely deployed frontier-level agents for retrieval-augmented reasoning and multi-step synthesis. Because their release cadence is not publicly documented, we fix the evaluation window to July 2025. Passing each benchmark instance through the agent produced a structured report as a PDF, which was then extracted as markdown and passed along as chunks to the LLM-as-a-judge framework to evaluate along the 6 axes described in the section above. For a comparative baseline, we also tested baseline LLMs with integrated search tools, using the Open Deep Search framework Alzubi et al. (2025), though these outputs are excluded from the human agreement study due to their especially poor. Additionally, we also investigate the tradeoffs in having binary vs. ternary criterion judgments, including 4 brief but

Complexity Axis	Level	Description & Examples
Conceptual Breadth	Simple	Single-domain focus or very narrow topic. Example: A math word problem or a factual lookup from one source.
	Moderate	Cross-domain with limited coupling. Example: A prompt combining two fields (physics concept applied in a medical device context). Requires some integration but scope is manageable.
	High	Spans disparate domains or many subtopics. Example: "Analyze the environmental, economic, and political factors affecting renewable energy adoption in Asia." Involves scientific data, market trends, and policy analysis across multiple countries.
Logical Nesting	Shallow	Direct Q&A or one-step inference. Example: "What is the capital of X country?" or a single tool query.
	Intermediate	Multi-step reasoning with a few (2–3) dependencies or conditionals. Example: "Find the sales of Company A and Company B last year and determine who grew faster; then identify one reason for that difference."
	Deep	Complex, recursive reasoning or planning. Example: "Develop an evidence-backed investment strategy given current economic indicators, then stress-test it against at least two historical scenarios and suggest contingency plans." Requires orchestrating many pieces of information and analysis.
Exploration	Low	Well-specified query, deterministic interpretation. Example: "Summarize the methodology of the referenced paper." The task is clear-cut.
	Medium	Some ambiguity or breadth, but manageable. Example: "Discuss the benefits and risks of AI in healthcare." Open-ended but with known key points to cover (privacy, accuracy, etc.).
	High	Very open-ended or vague prompt needing refinement. Example: "I want to change careers to something with strong future growth—what should I consider?" The agent must clarify criteria, explore multiple fields, and cannot rely on a fixed answer.

Table 6: Prompt complexity categories used to annotate each task in RESEARCHRUBRICS.

Score Range	Description
$[+4, +5]$	**Critically important** – A criterion without which the response is fundamentally flawed or incorrect. Required for a minimally viable response.
$[-5, -4]$	**Critically detrimental** – A criterion identifying an error so severe that it makes the response actively harmful, deeply unethical, or completely invalidates its reasoning.
$[+2 + 3]$	**Important** – A key feature of a strong response, but not absolutely essential.
$+1$	**Slightly Important** – A "nice-to-have" detail that improves a good response but does not significantly change overall quality.
-1	**Slightly Detrimental** – A minor issue, tangent, or stylistic weakness that does not impact core reasoning or validity.
$[-3, -2]$	**Detrimental** – A significant error that detracts from the response quality, introduces faulty logic, or offers poor advice, but does not make it fundamentally harmful.

Table 7: Rubric scoring scale with mandatory and optional criteria.

representative examples per rubric criterion, and LLM-augmentation applied to the rubrics. Full results and per-category breakdowns are presented in Table 3, Table 4, and Table 5.

Implementation Details. For evaluation, we investigate the effectiveness of four LLMs-as-judges: o3 (OpenAI, 2025d), GPT-5 (OpenAI, 2025c), GPT-4.1 (OpenAI, 2025b), and Claude-Opus-4.1 (Anthropic, 2025).

Agent responses were collected as PDFs and uploaded directly to the evaluator API calls as files, without normalization or reformatting. Maximum output length was capped at 10000 tokens to ensure comparability across models. Report-level assessments were generated by weighted averages of the scores given by the LLM-as-a-judge model scaled by the weighted of each rubric criterion. Alignment with human annotations was measured using Macro F1 under both ternary {`Satisfied`, `Partially Satisfied`, `Not Satisfied`} and binary {`Satisfied`, `Not Satisfied`} regimes.

A.7 Ablation Studies

- The impact of a reduced model output space, going from ternary ({`Satisfied`, `Partially Satisfied`, `Not Satisfied`}) to binary ({`Satisfied`, `Not Satisfied`}) verdicts by turning the `Partially Satisfied` verdicts to `Not Satisfied`
- The impact of the level of details in the rubrics by removing any representative examples in the rubrics that may provide additional levels of clarity to the model-based grader
- The impact of using LLMs to augment the expert-written rubrics

A.7.1 Prompts Used

Example Removal

> You are tasked with removing examples from rubric text while keeping everything else EXACTLY the same.
>
> Your job is to:
> 1. Identify portions of text that contain examples, typically in the form ”(e.g. example1, example2, etc.)” or similar
> 2. Remove ONLY these example portions
> 3. Keep all other text, formatting, punctuation, and structure EXACTLY as it was
> 4. Do not rephrase, reword, or change anything else
> 5. Do not add any new content
> 6. Simply return the text with the example portions removed
>
> Examples of what to remove:
> - ”(e.g., a diagnosis code block, a free-text note snippet without PHI, tabular data contexting text and numerical data)”
> - ”(eg. programmatic text extractions or more rigorous NLP and machine learning techniques)”
> - ”(e.g. (1) National Library of Medicine, (2) CDC Wonder, (3) publications from well-known universities)”
>
> Be very careful to maintain the exact same structure and wording for everything else.

Figure 4: Example removal prompt used in the experiments.

LLM Augmentation

A.8 Distribution

A.9 Rubrics Criteria Overview

You are an expert at improving evaluation rubrics to make them more detailed and concrete while keeping them concise.

CRITICAL FORMATTING REQUIREMENTS:
- Return exactly ONE cohesive sentence (NO newlines, NO line breaks)
- The rubric should be ONE SINGLE SENTENCE but can contain multiple phrases, subparts, clauses, and run-on components
- Do NOT create multiline, paragraph-style, or bullet-point rubrics

IMPORTANT: You will receive exactly ONE rubric to improve, and you must return exactly ONE enhanced version of that same rubric. Do not create multiple rubrics or variations.

Your job is to:
1. Keep ALL original information from the rubric EXACTLY as it was - do not delete or remove any core information or intent
2. Make the rubric more detailed and concrete by adding specific examples inline (e.g., specific patterns, formats, indicators)
3. Clarify vague terms with more precise descriptions within the same sentence flow
4. Add concrete criteria and benchmarks inline where applicable
5. Make the rubric as actionable and unambiguous as possible while staying concise

Focus on adding inline:
- Concrete examples in parentheses (e.g., specific technical details, data formats)
- Specific indicators to look for
- Clear boundary conditions
- Representative examples of what qualifies

Do not:
- Remove any original content
- Change the fundamental meaning or intent
- Add entirely new rubric categories
- Create multiple versions or variations
- Generate more than one rubric output
- Break the rubric into multiple sentences

Return only the single improved rubric as one cohesive sentence.

Figure 5: Example removal prompt used in the experiments.

Category (Approx %)	Description of Prompts
STEM (22.1%)	Science, technology, engineering, and math queries that require synthesizing information from textbooks, research papers, or technical reports (e.g., explaining a physics concept and related formula, summarizing the latest research on mRNA cancer vaccines and biotech startups.
General Consumer Research (11.7%)	Everyday research with complex constraints (e.g., finding an apartment under budget, multi-factor product comparisons, travel itineraries, personal finance or legal advice, health-related questions requiring reputable sources).
Technical Documentation (11.7%)	Prompts involving explanation of complex technical concepts, code, or APIs using official documentation or repositories (e.g., troubleshooting a programming error with library docs, comparing software architecture patterns).
Hypotheticals & Philosophy (13.0%)	Open-ended prompts asking for speculation, hypotheticals, or philosophical analysis, often requiring synthesis of diverse viewpoints (e.g., "How might society change if X. . . ?", ethical dilemmas, future predictions in technology).
Historical Analysis (10.4%)	Questions about historical events, figures, or periods that require pulling from archives, historical texts, and scholarly interpretations (e.g., analyzing causes of a historical conflict with primary source references).
Business Planning & Research (11.7%)	Prompts related to business or entrepreneurship (e.g., developing a go-to-market strategy, analyzing a company's financial health, legal considerations for a startup, HR or marketing plan), often requiring use of industry reports or case studies.
Creative Writing (7.8%)	Long-form creative tasks that incorporate factual elements or research (e.g., writing a historical fiction scene with accurate period details, or a sci-fi story grounded in real science).
Current Events (5.2%)	Prompts focused on recent or ongoing events, necessitating retrieval of up-to-date news or data (e.g., analysis of a recent policy change, comparison of current market trends).
Other (6.5%)	Miscellaneous prompts that do not neatly fit in the above categories, including cross-domain questions or niche topics.

Table 8: Distribution of prompt categories in RESEARCHRUBRICS.

Criterion	Definition
All Rubric Criteria	
Explicit Criteria	Whether all of the points that were explicitly asked for in the prompt were addressed correctly
Implicit Criteria	Context awareness of information that is relevant context for an answer to the prompt, but is not explicitly asked for in the prompt
Synthesis of Information	Criteria that evaluate the model's ability to reason and draw new insights/conclusions across different sources (not just facts relevant to a single source)
References	About any core / concrete citations/references that must be included for the answer to be complete
Communication Quality	Whether the response is well-structured and concise, and whether it uses a level of technical depth and vocabulary that is well-matched to the user.
Instruction Following	Many tasks involve specific user instructions, and criteria in this category check whether the model adheres to instructions.

Table 9: Overview of rubric criteria.