DREAM: Deep Research Evaluation with Agentic Metrics

Elad Ben Avraham[1*†] **Changhao Li**[2*‡] **Ron Dorfman**[1*] **Roy Ganz**[1]
Oren Nuriel[1] **Amir Dudai**[1] **Aviad Aberdam**[1] **Noah Flynn**[1]
Elman Mansimov[1] **Adi Kalyanpur**[1] **Ron Litman**[1]
[1]AWS Agentic AI [2]Georgia Institute of Technology

Abstract

Deep Research Agents generate analyst-grade reports, yet evaluating them remains challenging due to the absence of a single ground truth and the multidimensional nature of research quality. Recent benchmarks propose distinct methodologies, yet they suffer from the *Mirage of Synthesis*, where strong surface-level fluency and citation alignment can obscure underlying factual and reasoning defects. We characterize this gap by introducing a taxonomy across four verticals that exposes a critical *capability mismatch*: static evaluators inherently lack the tool-use capabilities required to assess temporal validity and factual correctness. To address this, we propose `DREAM` (Deep Research Evaluation with Agentic Metrics), a framework that instantiates the principle of *capability parity* by making evaluation itself agentic. `DREAM` structures assessment through an evaluation protocol combining query-agnostic metrics with adaptive metrics generated by a tool-calling agent, enabling temporally aware coverage, grounded verification, and systematic reasoning probes. Controlled evaluations demonstrate `DREAM` is significantly more sensitive to factual and temporal decay than existing benchmarks, offering a scalable, reference-free evaluation paradigm.

1 Introduction

Large Language Models (LLMs) increasingly support autonomous, tool-using agents that perform complex, open-ended tasks. Among these, *Deep Research Agents* (DRAs) have emerged as a dominant paradigm. Given a broad query, they retrieve information from external sources, synthesize evidence, and produce long-form research reports. Such reports are inherently open-ended; unlike traditional question answering, where correctness can often be verified against a singular ground truth, deep research admits multiple valid trajectories. For a given query, two experts will inevitably produce distinct reports of potentially equal quality. This variability makes the assessment of deep research fundamentally complex, requiring a transition from single-answer correctness toward a high-dimensional evaluation of report quality.

Recognizing this challenge, the research community has shown growing interest in Deep Research Evaluation (DRE), with several recent works proposing benchmarks and datasets to assess DRAs (Coelho et al., 2025; Du et al., 2025; Patel et al., 2025; Sharma et al., 2025; Wan et al., 2025; Wang et al., 2025). However, a critical examination of these frameworks reveals a systematic limitation. While current approaches effectively assess surface-level dimensions, such as writing fluency and citation alignment, they remain largely insensitive to failures in *factual correctness*, *temporal validity*, and *substantive reasoning*. Therefore, fluent, well-cited reports can receive high scores despite containing obsolete information or flawed logic. We term this the *Mirage of Synthesis* – an illusion of quality created by surface-level coherence despite underlying factual and reasoning flaws.

To characterize this observation and unify the evaluation landscape, we propose a taxonomy that organizes DRE into four fundamental verticals: *Presentation Quality*, *Task Compliance*, *Analytical Depth*, and *Source Quality*. By mapping the metrics of existing benchmarks into this taxonomy, we trace their limitations to the fundamental design choices. Current benchmarks typically rely on either human-curated rubrics, which are highly reliable but prohibitively expensive, dataset-specific, and prone to becoming outdated, or LLM-as-a-judge paradigms limited to the LLM's static internal knowledge. These approaches and their associated citation-verification workflows operate as *static observers*, i.e., they lack access to external tools, temporal awareness, and the ability to

*Equal contribution
†Correspondence to: eladba@amazon.com
‡Work done during an internship at AWS Agentic AI

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9879–9904

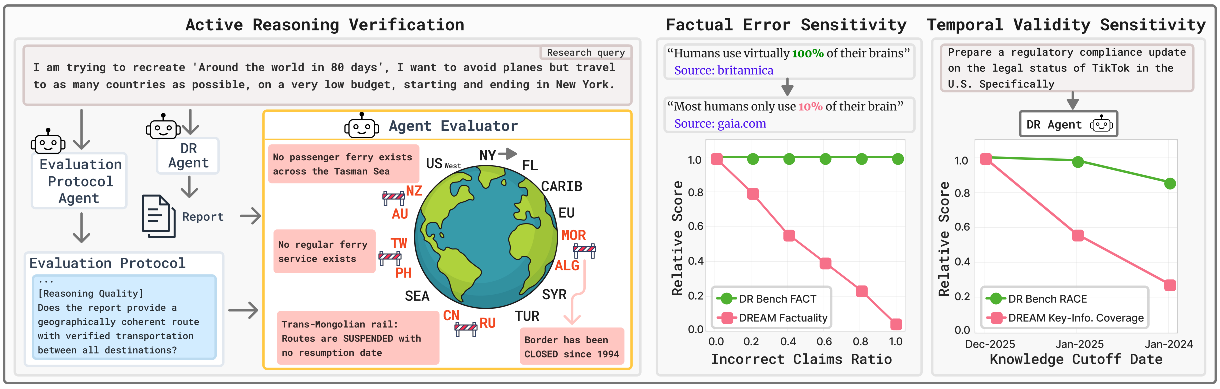

Figure 1: **Capturing Overlooked Dimensions of Research Quality.** DREAM actively verifies the reasoning of generated reports by probing external sources (left), detects factual errors injected in a controlled experiment (middle), and captures time-sensitive validity gaps by penalizing outdated reports (right).

independently gather and verify evidence. Our taxonomy thus reveals a unifying diagnosis: a capability mismatch, where evaluators lack the abilities required to assess the dimensions they purport to measure. This motivates the principle of **capability parity** – the evaluator should possess a similar set of capabilities as the researcher, including the ability to retrieve, verify, and reason over information.

We instantiate this principle in Deep Research Evaluation with Agentic Metrics (**DREAM**), a framework that makes evaluation itself agentic. DREAM structures assessment through an *evaluation protocol* that combines query-agnostic static metrics with query-adaptive ones constructed by a tool-calling agent. By independently researching the query and cross-referencing external evidence, DREAM provides temporally aware fact-checking and substantive depth assessment. This behavior is illustrated in Figure 1: the left panel depicts DREAM's active reasoning verification against external sources, while the middle and right panels demonstrate that DREAM's scores degrade appropriately as factual errors increase and report knowledge becomes outdated, whereas existing benchmarks remain largely insensitive to both.

In summary, our main contributions are:

- We examine existing DRE benchmarks to identify critical gaps and the "Mirage of Synthesis", a systemic failure where they lack the capacity to detect factual, temporal, and logical degradation.
- We propose a taxonomy to unify DRE into four key verticals and identify a capability mismatch in current benchmarks, where evaluators lack the active retrieval tools and reasoning required for independent, temporally-aware verification.
- We introduce DREAM, an agentic evaluation framework grounded in the principle of capability parity, replacing passive scoring with active verification via a two-phase Protocol Creation and Execution workflow.
- We empirically validate through controlled experiments a suite of agentic metrics—*Key-Information Coverage*, *Reasoning Quality*, and *Factuality*–showing that DREAM is substantially more sensitive to temporal degradation and extrinsic factual errors than existing benchmarks.

2 Deep Research Evaluation Landscape

Existing DRE benchmarks propose a variety of metrics to capture the multidimensional nature of research quality, yet they differ substantially in terminology, scope, and implementation. This fragmentation makes it difficult to reason about which aspects of deep research are well evaluated, which remain under-evaluated, and why certain failure modes persist across benchmarks. To address this, we introduce a unifying taxonomy that organizes existing criteria into a common structure and use it to analyze the current evaluation landscape.

2.1 A Unifying Taxonomy

To systematically derive the taxonomy, we employed an agentic pipeline that processed the evaluation metrics from the benchmarks summarized in Table 1. The pipeline extracted granular, leaf-level criteria from benchmark source code and documentation, embedded them semantically, and clustered them into coherent evaluation dimensions. This data-driven approach ensures the taxonomy is grounded in the actual implementation details of existing metrics, providing a structured, bottom-up synthesis of the landscape. Full details of this derivation process are provided in Appendix B.

Table 1: **Taxonomic breakdown of deep research evaluation benchmarks.** Metrics from existing benchmarks mapped to our four-dimensional taxonomy, alongside their creation and execution methodologies.

Benchmark	Vertical	Metric Name	Creation	Execution
DeepResearchGym (Coelho et al., 2025)	Presentation Quality	Clarity	N/A	LLM
	Task Compliance	Report Relevance	LLM + Human	LLM
	Analytical Depth	Insightfulness	N/A	LLM
	Source Quality	Retrieval Faithfulness	N/A	Workflow
DeepResearch Bench (Du et al., 2025)	Presentation Quality	RACE (Presentation Quality)	LLM	LLM
	Task Compliance	RACE (Inst., Comp.)	LLM	LLM
	Analytical Depth	RACE (Insight)	LLM	LLM
	Source Quality	FACT	N/A	Workflow
ResearchRubrics (Sharma et al., 2025)	Presentation Quality	Communication Quality	Human	LLM
	Task Compliance	Explicit/Implicit Req., Inst. Follow.	Human	LLM
	Analytical Depth	Synthesis of Information	Human	LLM
	Source Quality	Use of References	Human	LLM
DeepResearch Arena (Wan et al., 2025)	Presentation Quality	ACE	LLM	LLM
	Task Compliance	ACE	LLM	LLM
	Analytical Depth	ACE	LLM	LLM
	Source Quality	KAE	N/A	Workflow
LiveResearchBench (Wang et al., 2025)	Presentation Quality	Presentation & Organization	LLM + Human	LLM
	Task Compliance	Coverage & Comprehensiveness	LLM + Human	LLM
	Analytical Depth	Analysis Depth	N/A	LLM
	Source Quality	Factual Cons., Citation Assoc./Acc.	N/A	LLM / Workflow
DREAM (Ours)	Presentation Quality	Writing Quality	N/A	LLM
	Task Compliance	Key-Information Coverage	Agent	LLM
	Analytical Depth	Reasoning Quality	Agent	Agent
	Source Quality	Factuality, Citation Integrity	N/A	Workflow

Presentation Quality evaluates how effectively a report communicates its findings. This includes stylistic aspects such as clarity of expression and sentence fluency, as well as structural properties such as section organization and appropriate use of headings. Metrics in this vertical assess whether information is conveyed clearly and professionally, independent of content correctness or depth.

Task Compliance assesses whether a report fulfills the requirements of the research query. This includes explicit constraints (*e.g.*, format, scope, or requested comparisons) as well as implicit expectations of comprehensive coverage. Metrics in this vertical evaluate coverage breadth, recall of key information, and adherence to the query.

Analytical Depth captures the intellectual rigor of the report's synthesis. This includes the quality of logical reasoning, causal explanations, and critical evaluation of evidence. High analytical depth reflects coherent reasoning that goes beyond aggregating retrieved facts, demonstrating structured argumentation and meaningful synthesis.

Source Quality measures the reliability of the evidence supporting a report's claims. This vertical comprises two complementary aspects: *citation faithfulness*, which evaluates whether claims are accurately supported by their cited sources (intrinsic quality), and *factual correctness*, which evaluates whether claims are true with respect to external world knowledge regardless of citation support.

2.2 Diagnosing the Evaluation Landscape

In Table 1, we map existing DRE benchmarks onto our taxonomy and group them by evaluation paradigm—human-curated, closed-loop LLM-based, or citation-centered workflow-based—based on how their metrics are *created* and *executed*.

Human-Defined Evaluation Criteria. This paradigm includes benchmarks with human-defined evaluation criteria, specified either through manual rubrics or dataset-specific proxies. While reliable and interpretable, these approaches tightly couple evaluation to specific datasets, domains, or costly curation. ResearchRubrics (Sharma et al., 2025) evaluates all four verticals using manually constructed rubrics, requiring over 2,800 hours of expert annotation. DeepResearchGym (Coelho et al., 2025) encodes task compliance via dataset-specific behavioral proxies, such as whether reports cover documents historically clicked by users. LiveResearchBench (Wang et al., 2025) adopts a hybrid approach, combining LLM-generated checklists with human-in-the-loop verification.

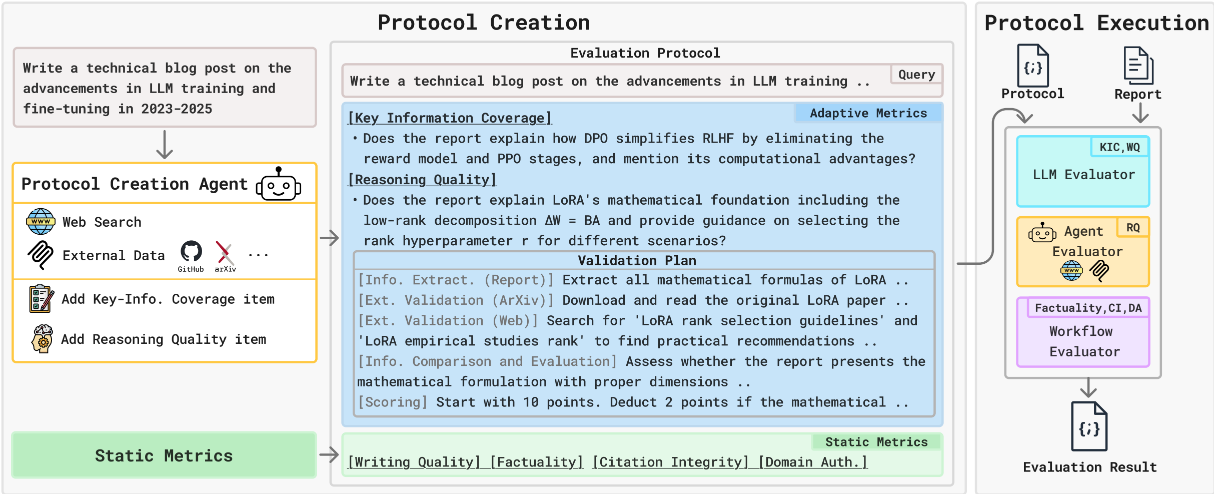

Figure 2: **DREAM Overview.** Our framework operates in two phases. **Left:** Protocol Creation, where query-independent Static Metrics are combined with Adaptive Metrics constructed by an agent equipped with web search tools and optional tools to access external data. **Right:** Protocol Execution, where each metric is routed to the appropriate evaluator, either an LLM, agent with tool access, or workflow.

Closed-Loop LLM-Based Evaluation. To improve scalability, several benchmarks replace human annotation with LLM-generated evaluation criteria. DeepResearch Bench (Du et al., 2025) uses RACE to generate weighted rubrics for presentation quality, instruction following, and depth, while DeepResearch Arena (Wan et al., 2025) removes references via ACE, prompting an LLM to generate a query-specific checklist. However, rubric and checklist construction is performed by static LLMs, which lack access to external tools, temporal context, or independent evidence gathering. This constrains their ability to evaluate grounded reasoning and evolving task requirements.

Citation-Alignment Workflows. Prevalent in the Source Quality vertical, this paradigm employs multi-step pipelines to verify alignment between claims and cited URLs. Benchmarks such as DeepResearch Bench (FACT), DeepResearchGym (Retrieval Faithfulness), and LiveResearchBench (Citation Accuracy) extract ⟨claim, URL⟩ pairs, retrieve the cited content, and utilize an LLM to judge alignment. While effective at detecting source misrepresentation, these workflows primarily measure *intrinsic* citation faithfulness rather than *extrinsic* factual correctness. Because verification is restricted to provided citations, they remain insensitive to claims supported by outdated or unreliable sources. This creates the *Citation-Alignment Fallacy*, where a report can contain entirely accurate citations while still being factually incorrect or obsolete with respect to the external world.

Systematic Imbalance and Evaluator Capability Mismatch. Applying the taxonomy reveals a consistent imbalance across evaluation paradigms. Presentation Quality and Task Compliance are extensively evaluated, while Source Quality is predominantly assessed through intrinsic citation alignment. In contrast, extrinsic factual correctness, temporal validity, and grounded reasoning receive little direct evaluation. This imbalance gives rise to the *Mirage of Synthesis*, where surface-level fluency and citation alignment are evaluated, while factual, temporal, and logical defects remain systematically unassessed. Crucially, this failure stems from a structural **capability mismatch**, where across all paradigms, evaluators lack critical capabilities available to the agents they assess, namely, to independently retrieve evidence, reason over competing sources, and incorporate temporal context.

3 DREAM: DRE with Agentic Metrics

To address the capability mismatch identified in Section 2, we introduce DREAM, a framework that enforces **capability parity** by making the evaluation process itself agentic. Given a research query, DREAM constructs a query-specific *evaluation protocol* (Section 3.1) and executes it using specialized evaluators (Section 3.2), as illustrated in Figure 2.

3.1 Phase 1: Protocol Creation

Because deep research questions admit multiple valid realizations, evaluation cannot rely on a single reference or a fixed rubric. DREAM addresses this by constructing a query-specific *evaluation*

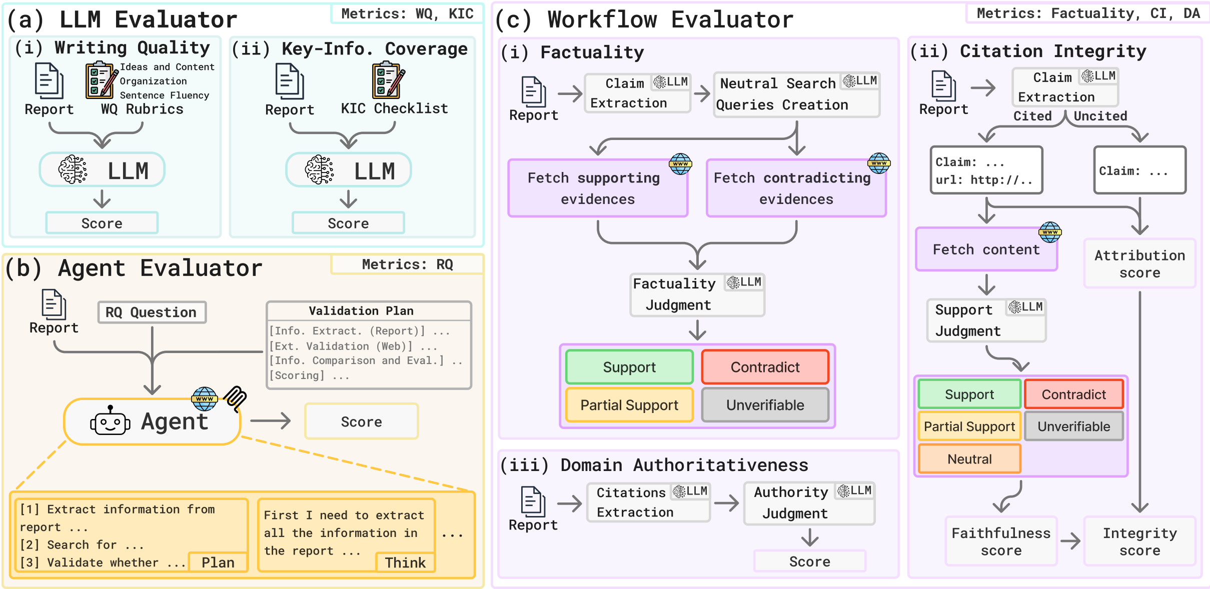

Figure 3: DREAM **Protocol Execution Evaluators.** (a) LLM Evaluator assesses writing quality (WQ) and key-information coverage (KIC); (b) Agent Evaluator evaluates reasoning quality (RQ) using external tools; (c) Workflow Evaluator performs factuality assessment via evidence retrieval, citation integrity (CI) verification through claim-source validation, and domain authoritativeness (DA) scoring via credibility assessment of extracted citations.

protocol that defines task-relevant evaluation criteria independently of any particular report. The protocol consists of two complementary classes of metrics: Static Metrics for universal quality standards, and Adaptive Metrics for task-specific depth. Full implementation details for the metric suite are provided in Appendix C.

Static Metrics. These query-agnostic criteria apply uniformly across tasks: (i) *Writing Quality (WQ)* assesses presentation quality along the fixed dimensions of Ideas and Content, Organization, and Sentence Fluency; (ii) *Factuality* validates claims against external knowledge independent of citation support; (iii) *Citation Integrity (CI)* verifies that claims are attributed to sources and supported by the cited content; and (iv) *Domain Authoritativeness (DA)* evaluates cited sources credibility to ensure reliance on reputable, high-quality domains.

Adaptive Metrics. To capture query-dependent expectations, we instantiate a *Protocol Creation Agent* as a CodeAgent (Wang et al., 2024b) equipped with retrieval tools including web search, ArXiv, and GitHub. Given a query, the agent performs lightweight tool selection to reduce noise, then conducts research to construct two metrics:

- **Key-Information Coverage (KIC):** The agent identifies essential facts by retrieving up-to-date sources and converting each key point into a verifiable *yes/no question*. This transforms coverage assessment into a grounded, temporally aware checklist that flags missing or outdated content.
- **Reasoning Quality (RQ):** The agent generates query-specific questions paired with *structured validation plans*. These plans specify the information to extract from both the report and independent sources, detailing how to cross-reference findings to ensure analytical depth is evaluated based on substantive reasoning. RQ is illustrated in the left panel of Figure 1 and in Figure 2.

3.2 Phase 2: Protocol Execution

Once a protocol is constructed, DREAM executes each metric using an evaluator whose capabilities match the metric's requirements. Evaluator selection follows the capability parity principle, i.e., each metric is routed to the simplest evaluator that possesses the capabilities required to close the diagnosed gap. As shown in Figure 3, DREAM employs three evaluator types.

LLM Evaluator. The LLM Evaluator is used for metrics that require judgment but not external tool use. It assesses Writing Quality using fixed rubrics to ensure calibration and evaluates Key-Information Coverage by verifying report content against the agent-generated yes/no checklist.

Agent Evaluator. The Agent Evaluator, instantiated as a CodeAgent, executes the Reasoning Quality metric. Unlike KIC, which focuses on coverage, RQ probes the coherence and validity of reasoning across the report. The agent autonomously follows the validation plan created in phase 1, retrieving external evidence as needed and assigning a final score based on evidentiary support.

Workflow Evaluator. The Workflow Evaluator implements three complementary verification pipelines. For Citation Integrity, the system computes the harmonic mean of *Claim Attribution* (the ratio of cited to total verifiable claims) and *Citation Faithfulness* (the alignment of cited claims with their source content). For Factuality, claims are extracted and evaluated independently of citations by generating **neutralized search queries** to retrieve external evidence. Finally, for Domain Authoritativeness, the extracted citations are assessed by scoring the reputation and reliability of referenced domains. Together, this suite distinguishes between internal citation adherence (CI) and extrinsic empirical truth (Factuality), ensuring that analytical depth is grounded in credible domains (DA).

4 Validation of DREAM

We evaluate DREAM through a set of complementary validation studies targeting distinct aspects of the evaluation framework. First, we assess the agent-constructed protocols via human judgment, establishing that the generated criteria are interpretable, relevant, and verifiable (Section 4.1). Next, we conduct targeted sensitivity analyses that isolate specific failure modes identified by our taxonomy, including temporal degradation, reasoning flaws, and extrinsic factual errors (Sections 4.2 to 4.4). Finally, we demonstrate that writing quality can be assessed in a reference-free manner while remaining aligned with human judgment (Section 4.5). Unless otherwise specified, we use queries from DeepResearch Bench (DRB) and Claude Sonnet 4.5 (Anthropic, 2025) as the base LLM.

4.1 Human Evaluation of Protocol Quality

A central claim of DREAM is that adaptive evaluation criteria can be constructed agentically, replacing manually authored rubrics. We therefore evaluate the quality of the agent-generated protocol metrics, KIC and RQ. We conducted a human study where expert and non-expert annotators rated generated items on relevance, clarity, and verifiability, as well as plan validity for RQ, using a 1–3 scale normalized to $[0, 1]$ (details are provided in Appendix E).

Table 2: **Human evaluation of KIC and RQ items.** Performance by metric and method. The agent with retrieval achieves the best performance across all axes.

Metric	Method	Rel.	Verif.	Clar.	Valid.	Avg.
KIC	LLM	0.84	0.73	0.80	–	0.79
	Agent	0.88	0.75	0.87	–	0.83
	Agent + Ret.	**0.94**	**0.91**	**0.92**	–	**0.92**
RQ	LLM	0.82	0.51	0.78	0.70	0.70
	Agent	0.89	0.67	0.88	0.92	0.84
	Agent + Ret.	**0.94**	**0.80**	**0.97**	**0.99**	**0.93**

We further performed a three-way ablation to isolate the impact of agentic structuring and retrieval.

As shown in Table 2, annotators rated the protocol criteria generated by the full agent highly (KIC: 0.92, RQ: 0.93), with RQ clarity (0.97) and plan validity (0.99) scoring particularly well. The ablation reveals that agentic structuring alone improves over a plain LLM (RQ average: 0.84 vs. 0.70). Furthermore, adding retrieval significantly boosts performance, notably increasing verifiability (KIC: 0.75 $\rightarrow$ 0.91, RQ: 0.67 $\rightarrow$ 0.80) and RQ plan validity (0.92 $\rightarrow$ 0.99). These results confirm the value of both structured multi-step reasoning and grounding in external evidence.

4.2 Temporal Awareness in KIC

We next evaluate the framework's sensitivity to *temporal obsolescence*, a core failure mode identified by our taxonomy. When temporal validity is central to a query, a competent evaluator should consistently penalize reports generated with outdated knowledge. We investigated this by selecting 20 temporally volatile queries (e.g., "TikTok US legal status") and generating three report variants for each: one using current information (Dec 2025), and two simulated with knowledge cutoffs of Jan 2025 and Jan 2024. We then evaluated all reports using both DREAM–KIC and DRB–RACE (Comprehensiveness and Insight), as reported in Table 4.

The results reveal a sharp contrast. DRB–RACE exhibits weak temporal sensitivity: Comprehensiveness fails to penalize moderately outdated reports (50.02 $\rightarrow$ 50.04 for Jan 2025), while Insight degrades only marginally (50.54 $\rightarrow$ 48.78). In contrast, DREAM–KIC degrades monotonically with information staleness, dropping from 79.35 (current) to 44.80 (Jan 2025) and 22.34 (Jan 2024). This performance gap is structural; DRB–RACE evaluates reports against static criteria (e.g., identifying a relevant law), which remain satisfiable even with obsolete facts. KIC instead encodes time-sensitive

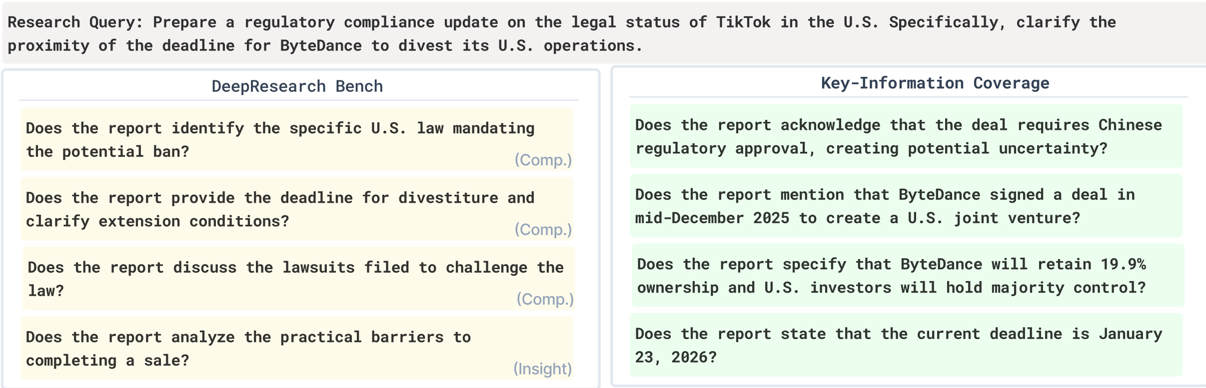

Figure 4: **Temporal Awareness in KIC Evaluation.** Comparison of evaluation criteria for a TikTok legal status query, showing DeepResearch Bench's static criteria (left) versus DREAM's KIC criteria (right) that incorporate time-sensitive facts (e.g., mid-December 2025 joint venture deal and January 23, 2026 deadline).

expectations derived from up-to-date evidence (Figure 4), causing outdated reports to fail explicit verification. Thus, effective temporal awareness requires agentic evaluation rather than static rubrics.

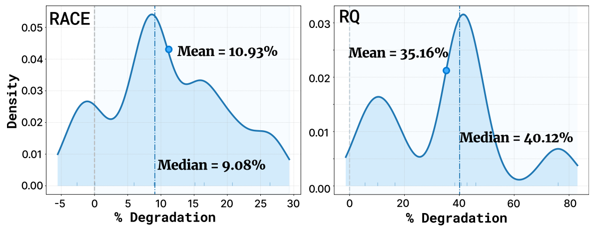

Figure 5: **Reasoning flaws detection.** Relative score degradations between well-reasoned and malformed reports. DREAM–RQ centers around 40.1% degradation, while RACE centers around 9.1%, with several malformed reports outscoring well-reasoned ones.

4.3 Detecting Reasoning Flaws

Analytical failures in deep research reports often occur within fluent, well-structured, and superficially coherent narratives. Detecting such flawed arguments requires evaluation methods that look beyond surface plausibility to explicitly probe reasoning validity. We examined whether existing metrics can distinguish sound reports from those containing subtle but substantive reasoning errors.

To study this, we designed a controlled experiment that isolates reasoning quality while holding surface fluency constant. We selected 10 complex queries spanning domains where analytical rigor is essential (e.g., policy analysis and technical comparisons), and generated two report variants for each: a standard version and a malformed variant. The malformed variant contained deliberately injected reasoning flaws, such as circular arguments or unsupported claims, while preserving a fluent structure (see full details in Appendix D.2).

Figure 5 reports the relative score degradation of malformed reports, revealing a significant gap. DRB–RACE exhibits weak and inconsistent sensitivity, with an average degradation of $\sim 9\%$, and frequently scores malformed reports higher than well-reasoned ones. By comparison, DREAM–RQ produces a consistent signal, centered at $\sim 40\%$ degradation, reliably penalizing reports with flawed reasoning despite their surface coherence.

4.4 Grounding Beyond Citation Alignment

Claims without citations are inherently invisible to citation-alignment workflows, motivating a reference-free factual verification metric such as DREAM–Factuality. More critically, cited claims can be faithful to their sources while remaining factually incorrect with respect to the external world.

We isolate this failure mode with a controlled corruption study. Starting from a base set of factual claims, we construct pairs consisting of a correct version with a supporting citation, and a plausible but incorrect variant with a matching citation (15 pairs; cf. Section D.3). We then sweep a corruption rate $r \in [0, 1]$, replacing an r fraction of correct claims with incorrect ones while preserving citation alignment by construction.

As shown in Figure 1 (middle), DRB–FACT scores remain invariant across all corruption levels, as the metric validates claims solely against the provided sources. In contrast, DREAM–Factuality degrades monotonically with increasing r, closely tracking the true error rate. This demonstrates that citation alignment is insufficient for factual evaluation, as detecting plausible, well-cited falsehoods requires access to external world knowledge.

4.5 Reference-Free Presentation Evaluation

Finally, we evaluate whether Presentation Quality can be reliably assessed without a reference report. Reference-based evaluation may introduce bias when references are suboptimal or outdated, often penalizing valid alternative formulations. To validate the reliability of DREAM's reference-free WQ metric, we measure its alignment with readability rankings established by DRB. Specifically, we utilize the DRB-RACE scores provided for 300 reports generated by six agents across 50 queries, which have been previously validated as a high-fidelity proxy for human preference.

We compute the rank correlation using Kendall's τ between the rankings induced by DREAM's WQ score and the DRB-RACE rankings. This analysis yields an average $\tau = 0.6$. Given that human inter-annotator agreement for subjective readability typically falls in the 0.5–0.7 range, this result confirms that DREAM provides a reliable signal comparable to established benchmarks.

5 Benchmarking Leading DRAs

Having validated DREAM, we now apply it to benchmark leading DRAs. Unlike current evaluations that rely on dataset-specific annotations, our framework is dataset-agnostic, enabling a unified comparison across diverse tasks. We employ three diverse datasets. DEEPRESEARCH BENCH (Du et al., 2025) is a bilingual benchmark comprising 50 English and 50 Chinese PhD-level questions across 22 research fields; we consider the English subset. LIVERESEARCHBENCH (Wang et al., 2025) focuses on timely information synthesis requiring access to recent sources, and we utilize the 80 publicly available queries.[1] Finally, ResearchRubrics (Sharma et al., 2025) provides 101 queries accompanied by expert-crafted rubrics.

We compare three open-source systems: LangChain Open Deep Research (LangChain AI, 2025) with GPT-5 as backbone model, Smolagents Open Deep Research (Roucher et al., 2025) with Claude Opus 4.6 as its backbone, and Tongyi Deep Research (Team et al., 2025). For brevity, full results are deferred to Appendix F (Table 10).

Results reveal several notable patterns across open-source DRAs. All three agents exhibit critically low Citation Integrity scores, exposing a systemic weakness in citation grounding, though the failure modes differ. Smolagents Open DR and Tongyi Deep Research achieve near-zero CI (4.78 and 1.03 in aggregate, respectively) primarily because they rarely ground claims to specific sources, reflecting negligible claim attribution rates. LangChain Open DR attains higher CI (15.92) thanks to substantially better attribution practices, yet is undermined by low citation faithfulness: attributed sources frequently fail to support the claims they are attached to (see Figures 8 and 9 for a detailed breakdown). Beyond citation behavior, Smolagents Open DR dominates in both source quality and synthesis, leading aggregate scores for Factuality (58.15), Writing Quality (63.97), Key-Information Coverage (75.95), and Reasoning Quality (69.16), achieving strong content quality despite its near-absent citation discipline. Tongyi Deep Research ranks second on Factuality (55.09) but trails on adaptive metrics (aggregate RQ of 45.48), while LangChain Open DR records the lowest Factuality (44.64) yet outperforms Tongyi on reasoning (RQ of 57.28). Overall, these results suggest that open-source DRAs can produce informative and well-written reports, but remain fundamentally limited by poor citation grounding, a critical gap for trustworthy research agents.

Robustness across Backbone Models. To evaluate the stability of our agentic metrics, we conducted a sensitivity analysis using different LLMs, specifically DeepSeek-V3.2 (**?**) and Kimi-K2.5 (**?**), as the backbone for protocol execution on DEEPRESEARCH BENCH. As shown in Table 12, our findings indicate that while absolute scores exhibit minor fluctuations depending on the judge's internal calibration—an expected variance across different LLM families—the relative performance rankings of the evaluated deep research agents remain highly robust. Notably, we observe perfect alignment across all metrics, with the slight exception of the Writing Quality (WQ) metric, for which Claude Sonnet 4.5 tended to provide more uniform scores, showing a less pronounced preference for specific agent compared to the other backbone models.

6 Conclusion

This work introduces a unifying taxonomy for Deep Research Evaluation that identifies a fundamental failure in existing methodologies: the *Mirage of Synthesis*. We demonstrate that current

[1] https://huggingface.co/datasets/Salesforce/LiveResearchBench

LLM-as-a-judge and reference-based benchmarks are often blinded by surface-level fluency and citation alignment, failing to detect deep-seated defects in factual correctness, temporal validity, and logical reasoning. These failures stem from a structural *evaluator capability mismatch*, where static evaluators lack the agency required to verify the very claims they are tasked to judge. To resolve this, we propose the principle of **capability parity** and instantiate it in **DREAM**, a framework that makes evaluation itself agentic. By transitioning from frozen rubrics to dynamic, tool-equipped evaluation protocols, DREAM provides temporally aware fact-checking and substantive reasoning validation. Our controlled evaluations confirm that these capabilities are essential for reliable assessment: DREAM detects temporal degradation and extrinsic factual errors that static benchmarks miss entirely, while reliably surfacing reasoning flaws masked by stylistic coherence. Ultimately, DREAM demonstrates that as AI agents gain the ability to research and reason over the open web, the frameworks used to judge them must evolve in kind. By grounding assessment in evidence and agency rather than surface form, DREAM provides a scalable and reference-free blueprint for evaluating the next generation of frontier deep research agents.

Limitations

While DREAM addresses the *evaluator capability mismatch*, several limitations remain. First, reliance on external tools introduces dependencies on third-party service availability and potential retrieval bias—a trade-off inherent to prioritizing temporal validity over closed-world consistency. Second, agentic evaluation is computationally intensive: multi-step verification and tool-interaction loops increase latency and cost compared to static judges. We view this as a necessary trade-off for scientific accuracy, though future work could explore optimization via caching or selective evaluation strategies. Finally, DREAM is a post-hoc evaluator of research outputs and does not directly assess intermediate research processes, such as search trajectory efficiency or source discovery dynamics. Extending evaluation to include process-level telemetry remains a promising future direction.

References

Anthropic. 2025. Claude sonnet 4.5. Accessed: 2025-10-15.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. Longwriter: Unleashing 10,000+ word generation from long context llms. *arXiv preprint arXiv:2408.07055*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

I Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, Pengfei Liu, and 1 others. 2023. FacTool: Factuality Detection in Generative AI–A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios. *arXiv preprint arXiv:2307.13528*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *Preprint*, arXiv:2110.14168.

João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, and 1 others. 2025. DeepResearchGym: A free, transparent, and reproducible evaluation sandbox for deep research. *arXiv preprint arXiv:2505.19253*.

Cohere. 2023. Introducing embed v3. Accessed: 2025-10-01.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv preprint*.

Assaf Elovic. 2025. GPT-Researcher: Autonomous agent designed for comprehensive online research. `https://github.com/assafelovic/gpt-researcher`. GitHub repository.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, and 1 others. 2025. Mind2Web 2: Evaluating Agentic Search with Agent-as-a-Judge. *arXiv preprint arXiv:2506.21506*.

Rujun Han, Yanfei Chen, Zoey CuiZhu, Lesly Miculicich, Guan Sun, Yuanjun Bi, Weiming Wen, Hui Wan, Chunfeng Wen, Solène Maître, and 1 others. 2025.

Deep researcher with test-time diffusion. *arXiv preprint arXiv:2507.16075*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *Preprint*, arXiv:2103.03874.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, and 1 others. 2025. Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, and Monica Lam. 2024. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9917–9955.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. 2025. Big-bench extra hard. *Preprint*, arXiv:2502.19187.

LangChain AI. 2025. LangChain Open Deep Research. https://github.com/langchain-ai/open_deep_research.

Zijian Li, Xin Guan, Bo Zhang, Shen Huang, Houquan Zhou, Shaopeng Lai, Ming Yan, Yong Jiang, Pengjun Xie, Fei Huang, and 1 others. 2025. WebWeaver: Structuring web-scale evidence with dynamic outlines for open-ended deep research. *arXiv preprint arXiv:2509.13312*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. *Preprint*, arXiv:2311.12983.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Liana Patel, Negar Arabzadeh, Harshit Gupta, Ankita Sundar, Ion Stoica, Matei Zaharia, and Carlos Guestrin. 2025. DeepScholar-Bench: A Live Benchmark and Automated Evaluation for Generative Research Synthesis. *arXiv preprint arXiv:2508.20033*.

Haoran Que, Feiyu Duan, Liqun He, Yutao Mou, Wangchunshu Zhou, Jiaheng Liu, Wenge Rong, Zekun Moore Wang, Jian Yang, Ge Zhang, Junran Peng, Zhaoxiang Zhang, Songyang Zhang, and Kai Chen. 2024. Hellobench: Evaluating long text generation capabilities of large language models. *Preprint*, arXiv:2409.16191.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, and Clémentine Fourrier. 2025. Open-source DeepResearch – Freeing our search agents. https://huggingface.co/blog/open-deep-research.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, and 1 others. 2025. ResearchRubrics: A Benchmark of Prompts and Rubrics For Evaluating Deep Research Agents. *arXiv preprint arXiv:2511.07685*.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Haiyuan Wan, Chen Yang, Junchi Yu, Meiqi Tu, Jiaxuan Lu, Di Yu, Jianbao Cao, Ben Gao, Jiaqing Xie, Aoran Wang, and 1 others. 2025. DeepResearch Arena: The First Exam of LLMs' Research Abilities via Seminar-Grounded Tasks. *arXiv preprint arXiv:2509.01396*.

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

Xiaoxuan Wang, Ziniu Hu, Pan Lu, Yanqiao Zhu, Jieyu Zhang, Satyen Subramaniam, Arjun R. Loomba, Shichang Zhang, Yizhou Sun, and Wei Wang. 2024a. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. *Preprint*, arXiv:2307.10635.

Xingyao Wang, Yangyi Chen, Lifan Yuan, Yizhe Zhang, Yunzhu Li, Hao Peng, and Heng Ji. 2024b. Executable code actions elicit better llm agents. In *Forty-first International Conference on Machine Learning*.

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

Yuhao Wu, Ming Shan Hee, Zhiqiang Hu, and Roy Ka-Wei Lee. 2025a. LongGenBench: Benchmarking Long-Form Generation in Long Context LLMs. In *The Thirteenth International Conference on Learning Representations*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Zijia Wang, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025b. Writingbench: A comprehensive benchmark for generative writing. *Preprint*, arXiv:2503.05244.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

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. 2024. Webarena: A realistic web environment for building autonomous agents. *Preprint*, arXiv:2307.13854.

Mingchen Zhuge, Changsheng Zhao, Dylan Ashley, Wenyi Wang, Dmitrii Khizbullin, Yunyang Xiong, Zechun Liu, Ernie Chang, Raghuraman Krishnamoorthi, Yuandong Tian, and 1 others. 2024. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*.

A Extended Related Work

The rapid development of DRAs has outpaced the capabilities of existing evaluation frameworks. Our work addresses this gap by proposing a holistic evaluation protocol tailored to these complex systems.

Deep Research Systems. The domain of autonomous research has evolved from simple retrieval agents to complex systems capable of synthesizing long-form reports. In addition to the proprietary deep research systems (like those considered in Section 5), open frameworks have pioneered transparent architectures. STORM (Shao et al., 2024) introduced the paradigm of "outline-driven retrieval", using multi-perspective question asking to simulate expert dialogue. This was further extended by CO-STORM (Jiang et al., 2024) to include human-agent collaboration. In the open-source domain, GPT-RESEARCHER (Elovic, 2025) established a baseline for parallelized plan-and-solve execution, where a static planner decomposes tasks for concurrent retrieval agents. WEBWEAVER (Li et al., 2025) improved upon this with dynamic outline optimization, while TTD-DR (Han et al., 2025) proposed a test-time diffusion framework for iterative report refinement. Despite this proliferation, a recent survey (Huang et al., 2025) highlights critical limitations in current evaluation benchmarks, notably their "restricted access to external knowledge".

From Static Judges to Agentic Evaluation To scale evaluation beyond human annotation, recent work has coalesced around the *LLM-as-a-Judge* paradigm (Liu et al., 2023; Zheng et al., 2023). However, these static judges typically focus on final outcomes, failing to capture the complex, multi-step decision-making inherent to autonomous agents. This limitation has spurred the evolution toward *Agent-as-a-Judge* frameworks, designed to verify the execution process itself rather than just the result. Zhuge et al. (2024) pioneered this approach for code generation, proposing recursive agentic evaluation to assess intermediate steps without excessive manual labor. More recently, MIND2WEB-2 (Gou et al., 2025) extended this to the web domain, arguing that the complexity of long-horizon tasks exceeds the capacity of simple LLM calls and necessitates an agentic evaluator to trace navigation trajectories. However, these frameworks remain confined to specific domains like code generation or web navigation and have not yet been applied to the open-ended information synthesis required in deep research.

While our primary focus is on the holistic evaluation of DRAs, an extensive body of work targets the isolated sub-skills that underpin this domain, ranging from atomic factuality to long-form text generation.

Factuality and Atomic Verification. To address hallucination, recent works have focused on granular verification. FACTSCORE (Min et al., 2023) introduced the paradigm of decomposing long-form text into atomic facts for independent verification. Similarly, FACTOOL (Chern et al., 2023) pioneered the use of tool-augmented frameworks for detecting factual errors. However, these methods focus strictly on atomic truthfulness and do not evaluate the synthesis, argumentation, or structural coherence of a complete research report.

Retrieval-Augmented Generation (RAG) and Citation. While standard RAG benchmarks like RGB (Chen et al., 2024) evaluate retrieval accuracy, more specialized frameworks focus specifically on attribution. ALCE (Gao et al., 2023) and RAGAS (Es et al., 2024) established the standards for measuring citation recall and faithfulness. However, to maintain reproducibility, these benchmarks typically rely on *static snapshots* of the web or closed corpora where ground-truth documents are pre-defined. This constraint allows for recall calculation but fails to capture the complexity of open-ended deep research, which operates on the live, dynamic web. Unlike RAG tasks where the goal is often to retrieve a specific "gold" document, deep research admits multiple valid search trajectories. Consequently, evaluation cannot rely on retrieval recall against a fixed index, but must instead assess the extrinsic validity of autonomously discovered evidence and its synthesis into a coherent narrative. Additionally, it does not take into account verticals such as Presentation Quality or Analytical Depth.

Web Interaction and Reasoning Benchmarks. For web interaction, WEBARENA (Zhou et al., 2024) and BROWSECOMP (Wei et al., 2025) evaluate browsing skills and goal-directed navigation in controlled environments. Beyond these, benchmarks such as GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), GAIA (Mialon et al., 2023), SCIBENCH (Wang et al., 2024a), and BIG-BENCH HARD (Kazemi et al., 2025) focus on quantitative reasoning and domain-specific problem solving.

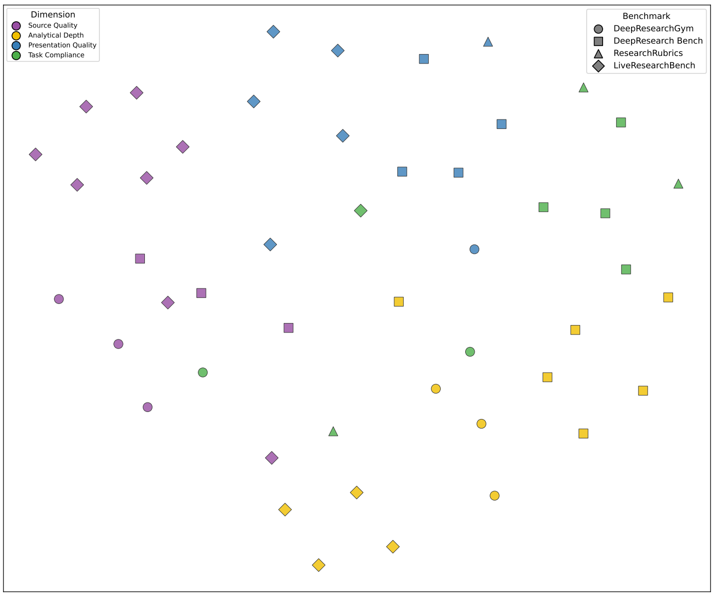

Figure 6: **Agentic Taxonomy Sub-Dimension Clustering Visualization.** Each point represents an evaluation criterion from one of four deep research benchmarks (shape), colored by its LLM-assigned taxonomy dimension. Criteria assigned to the same dimension tend to form localized regions in embedding space, suggesting coherent semantic groupings across the four verticals.

Long-form Text Generation. Finally, benchmarks such as LONGWRITER (Bai et al., 2024), HELLOBENCH (Que et al., 2024), WRITINGBENCH (Wu et al., 2025b), and LONGGENBENCH (Wu et al., 2025a) evaluate models specifically on coherence, length compliance, and stylistic quality. These frameworks primarily assess *Presentation Quality*, often relying on static references or intrinsic LLM-based judging without external verification of the content's truthfulness.

B Agentic Taxonomy Pipeline

As deep research evaluation gains momentum, multiple benchmarks have been proposed, each targeting overlapping yet distinct evaluation dimensions. While these efforts share common goals, differences in terminology, granularity, and evaluation focus make direct comparison difficult. To better understand the structure of existing evaluation practices, we develop an automated pipeline that organizes evaluation criteria from prior benchmarks into a unified, benchmark-agnostic taxonomy.

The pipeline consists of two stages. In the first stage, multiple LLM-based agents analyze each benchmark repository independently—DeepResearchGym (Coelho et al., 2025), DeepResearch Bench (Du et al., 2025), ResearchRubrics (Sharma et al., 2025), and LiveResearchBench (Wang et al., 2025). Agents examine benchmark documentation, evaluation prompts, rubrics, and source code to extract leaf-level evaluation criteria alongside short natural language descriptions. Across the four benchmarks, this process yields 48 distinct criteria. All LLM-based stages in the pipeline use `Claude Opus 4.5` with fixed decoding settings (temperature=0) to ensure consistency across extraction and aggregation.

In the second stage, a single LLM clusters the extracted criteria into higher-level evaluation dimensions based on semantic similarity rather than benchmark provenance. To discourage benchmark-specific groupings, we impose a constraint that each resulting dimension must contain criteria originating from multiple benchmarks. We refer to individual extracted items as *criteria*, which are grouped into *dimensions* and summarized as a small set of high-level *verticals*. This process reveals four recurring evaluation verticals observed across benchmarks:

- **Presentation Quality**: Writing clarity, structural organization, formatting consistency, and readability.
- **Task Compliance**: Adherence to instructions, coverage of requested topics, and fulfillment of explicit and implicit task requirements.
- **Analytical Depth**: Reasoning rigor, critical evaluation, originality of insights, and synthesis across multiple perspectives.

- **Source Quality**: Proper citation usage, source verification, and the extent to which claims are supported by credible, traceable evidence.

To assess the semantic coherence of the induced taxonomy, we embed all extracted criteria using Cohere's `embed-english-v3.0` (Cohere, 2023) model and visualize the resulting representations using UMAP (Figure 6). Criteria assigned to the same taxonomy dimension by the LLM tend to form localized regions in embedding space, providing qualitative evidence that the identified pillars correspond to coherent semantic groupings rather than artifacts of benchmark-specific phrasing.

C Metrics Details

In this section, we outline the implementation details and computation procedures for the suite of metrics employed in DREAM, categorized into Static Metrics (Writing Quality, Factuality, Citation Integrity, and Domain Authoritativeness) and Adaptive Metrics (Key-Information Coverage and Reasoning Quality).

C.1 Writing Quality

Writing Quality (WQ) is a static, reference-free metric that evaluates the stylistic and structural presentation of the research report. It is assessed by an LLM Evaluator using a fixed rubric that scores the report along three weighted dimensions.

- **Dimensions and Weights:** The score is computed as the average of three primary dimensions, each composed of weighted sub-dimensions, detailed in Table 3:
  - *Ideas and Content* (33%): Evaluates clarity of the main idea, relevance of details, information density, and conceptual synthesis.
  - *Organization* (33%): Assesses heading structure, logical grouping of bullet points, and structural coherence.
  - *Sentence Fluency* (33%): Measures rhythm/variety, transition smoothness, and readability/flow.
- **Scoring Procedure:** For a given task t, the LLM Evaluator assigns a score $S_{t,d} \in [0, 100]$ for each dimension $d \in \{\text{Ideas and Content}, \text{Organization}, \text{Sentence Fluency}\}$ based on the sub-dimension weights. The final Writing Quality score for task t is:

$$\text{WQ}_t = \frac{1}{3} \sum_{d} S_{d,t}$$

C.2 Factuality

Factuality evaluates whether the content of reports generated by the DRA is factually accurate with respect to external ground truth. The implementation utilizes a multi-stage verification pipeline that explicitly seeks both supporting and contradicting evidence for extracted claims.

1. **Key Factual Claim Extraction:** An LLM extracts the $N = 30$ most salient factual claims from the report with respect to the research question. The extraction is context-aware, incorporating the current date to resolve temporal references (e.g., "current status").
2. **Neutralized Query Construction and Search:** For each extracted claim, the system uses an LLM to generate multiple **neutralized** search queries. This step is critical to avoid *confirmation bias*: rather than searching for the specific values or predicates in the claim (e.g., searching for "inflation dropped to 2%"), the system generates open-ended queries (e.g., "current inflation rate") to maximize the retrieval of both supporting and contradictory evidence. These queries are executed via a Web Search Tool, and results are deduplicated to form a pool of candidate source content.
3. **Dual-Stream Evidence Extraction:** Unlike standard RAG verification, we perform two distinct extraction passes on the retrieved content,
   - *Supporting Stream:* The LLM extracts passages that explicitly confirm the claim.
   - *Opposing Stream:* The LLM actively scans for passages that contradict or refute the claim.

Table 3: **Writing Quality Rubric.** The exact descriptions used as system prompts for the LLM Evaluator to score each sub-dimension.

Dimension	Sub-dimension (weight)	Evaluation Prompt (Exact Description)
Ideas and Content	Main Idea Clarity (0.25)	This dimension assesses the clarity and specificity of the main idea expressed in the section summary. A high-quality section will present a **focused and well-articulated central idea** that is tightly aligned with the report question. Summaries lacking precision, or that simply list general topics without insight or framing, should be penalized. Do not give high scores if the main idea is vague, overgeneralized, or merely implied.
	Detail Relevance (0.25)	This dimension focuses on how well the supporting details in the summary reinforce the main idea. Bullet points should be specific, relevant, and purposefully selected. Low-quality summaries may include off-topic, overly generic, or redundant details that do not support the section's main message. Do not reward high scores based on the amount of content alone—focus on alignment and purpose.
	Information Density (0.25)	This dimension measures the **information richness** of the section summary. High-density summaries use each bullet to convey important, non-obvious, and topic-specific content. Shallow summaries repeat known facts, use vague language, or include fluff. Length alone should not be rewarded—focus on content value per line.
	Conceptual Synthesis (0.25)	This dimension evaluates the **structural and conceptual integration** in the summary. Look for signs of synthesis such as: grouping related points, identifying contrasts, cause-effect relationships, or thematic framing. Poor summaries are unordered lists with no visible logic. Do not reward correctness alone—this dimension rewards insight, not just content.
Organization	Heading Structure (0.3)	This dimension assesses the use and clarity of headings in the section. High-quality summaries include headings that meaningfully segment the content, reflect topic hierarchy, and help orient the reader. Avoid rewarding default, generic, or misaligned headings. Headings should reflect actual conceptual boundaries.
	Bullet Grouping Logic (0.4)	This dimension evaluates the internal logic of bullet groupings. High-quality summaries group related points together according to thematic, temporal, causal, or hierarchical logic. Low-quality groupings mix unrelated ideas, interrupt flow, or reflect no discernible principle.
	Structural Coherence (0.3)	This dimension assesses whether the section's structure contributes to a logical, easy-to-follow reading experience. A coherent structure will show consistent flow from one part to the next, maintain logical transitions between bullet blocks, and avoid jarring shifts. Low-scoring sections often feel fragmented, with unclear order, repetition, or misplaced content.
Sentence Fluency	Rhythm & Variety (0.3)	This dimension evaluates how naturally and dynamically the sentences flow. Strong writing features **variation in sentence length and structure**, avoiding repetitive patterns. Rhythm refers to the pacing and cadence of the prose—whether it reads with natural emphasis or becomes monotonous. High-scoring writing feels expressive and crafted, not just correct.
	Transition Smoothness (0.3)	This dimension focuses on **how smoothly the sentences connect** to each other. High-quality prose includes **natural linking phrases, varied connectors, and logical sequencing**. Low-scoring writing jumps between ideas, or has jarring, abrupt shifts between sentences. Do not reward correctness alone—this dimension targets flow between thoughts.
	Readability & Flow (0.4)	This dimension evaluates the overall **readability and flow** of the paragraph. High-scoring writing reads smoothly aloud and requires little effort to follow. Low-scoring writing may include awkward phrasing, overcomplex or confusing sentence structures, or poor pacing. This metric captures the **global fluency** felt by readers, especially in multi-sentence passages.

4. **Factuality Judgment:** The Judge LLM reviews the claim alongside the aggregated *Supporting* and *Opposing* evidence. It assigns one of four labels:
   - `Supported`: Evidence explicitly confirms the claim.
   - `Partially Supported`: Evidence supports some aspects of the claim but differs on details or is mixed.
   - `Contradicted`: Evidence clearly refutes the claim.
   - `Unverifiable`: Evidence is insufficient, indirect, or too weak to make a reliable determination.

Factuality Scoring: The Factuality score for task t, denoted as F_t, is calculated by weighting verified claims based on their support labels (1.0 point for `Support` and 0.5 point for `Partial Support`):

$$\mathrm{F}_t = \frac{N_{\text{supp},t} + 0.5 N_{\text{part},t}}{N_{\text{supp},t} + N_{\text{part},t} + N_{\text{con},t}} \, .$$

The overall Factuality score F is computed as the average across all tasks in the dataset:

$$\mathrm{F} = \frac{1}{|T|} \sum_{t \in T} \mathrm{F}_t \, .$$

C.3 Citation Integrity

Citation Integrity (CI) evaluates the trustworthiness of a report by determining whether its claims are both explicitly attributed to citations and faithfully supported by the content of those sources. Relying on either metric in isolation is insufficient, as a report may achieve perfect citation coverage by hallucinating sources, or conversely, be factually accurate but lack transparency. We therefore decompose Citation Integrity into two constituent metrics—Claim Attribution (CA) and Citation Faithfulness (CF)—before unifying them into a single score.

The evaluation begins with a Verifiable Claim Extraction phase, where we identify all factual and argumentative assertions within the report. Crucially, an LLM filters out non-verifiable content—such as procedural meta-talk (e.g., “The following section discusses...”), subjective commentary, and common knowledge—to establish a precise set of verifiable claims that forms the basis for evaluation.

Claim Attribution (CA). This metric quantifies the extent to which the agent grounds its reasoning in external sources. It is defined as the ratio of distinct *cited* claims to the total number of verifiable claims. A score of 1.0 indicates that every verifiable claim is explicitly linked to a source URL, maximizing the report's auditability.

Citation Faithfulness (CF). This metric evaluates the **veracity** of the provided evidence. For the subset of claims that are cited, CF determines whether the content of the source genuinely supports the claim. The workflow proceeds by retrieving the content of the associated URL for every cited claim and employing an LLM Judge to compare the text against the claim. The Judge assigns one of five labels: `Supported`, `Partially Supported`, `Neutral`, `Contradicted`, or `Unverifiable`. The citation faithfulness score for task t, denoted CF_t, is defined as:

$$\mathrm{CF}_t = \frac{N_{\text{supp},t} + 0.5 N_{\text{part},t}}{N_{\text{supp},t} + N_{\text{part},t} + N_{\text{neu},t} + N_{\text{con},t}} \, .$$

The overall CF score is then computed as the average across tasks:

$$\mathrm{CF} = \frac{1}{|T|} \sum_{t \in T} \mathrm{CF}_t .$$

Unified Citation Integrity Score. To penalize trade-offs between attribution and faithfulness, we compute the final CI score as the harmonic mean of the two components:

$$\mathrm{CI} = \frac{2 \cdot \mathrm{CA} \cdot \mathrm{CF}}{\mathrm{CA} + \mathrm{CF}} \, ,$$

where CA is the average claim attribution score across all tasks in the dataset. This formulation ensures that a high CI score is only achievable when an agent consistently supports its claims with valid, corroborating evidence; a failure in either (e.g., citing nothing, or citing everything incorrectly) will reduce the score.

C.4 Domain Authoritativeness

Domain Authoritativeness (DA) evaluates the credibility and trustworthiness of the external sources cited by the agent. Unlike Citation Faithfulness, which checks if a specific URL supports a specific claim, DA assesses the reputation of the source domain itself, penalizing reliance on low-quality or unverifiable outlets (e.g., social media, clickbait farms) even if the content matches the claim.

The evaluation process proceeds as follows:

1. **Domain Extraction and Deduplication:** First, we aggregate all unique URLs cited across the report (derived from the CI pipeline). We extract the root domain from each URL (e.g., `https://www.nature.com/articles/xyz` $\rightarrow$ `nature.com`) and deduplicate them to ensure each source is evaluated only once per report.
2. **LLM-Based Authority Assessment:** An LLM Judge evaluates each unique domain using a rubric that assesses its overall reputation, credibility, and trustworthiness. The judge considers factors such as institutional backing, historical reliability, and editorial standards to classify the domain into a category (e.g., Government, Academic, News, Commercial) and assign an integer score $S_d \in [1, 10]$ based on the following scale:
   - **Definitive Authority (9–10):** Gold-standard sources with institutional credibility (e.g., government agencies, top-tier academic institutions).
   - **High Authority (7–8):** Trustworthy and credible sources (e.g., established news organizations).
   - **Moderate Authority (4–6):** Acceptable but not ideal sources (e.g., general commercial sites).
   - **Low Authority (1–3):** Unreliable sources with questionable credibility (e.g., social media platforms, unverified blogs).
3. **Scoring Formulation:** The scores are normalized to a $[0, 1]$ interval. The Domain Authoritativeness score for a task t, denoted as DA_t, is the average normalized score of all domains D_t cited in that task:

$$\text{DA}_t = \frac{1}{|D_t|} \sum_{d \in D_t} \frac{S_d}{10}$$

   The final metric is the average across all tasks in the dataset.

C.5 Key-Information Coverage

Key-Information Coverage (KIC) is an adaptive metric that measures whether the report addresses essential, query-specific facts. Unlike static comprehensiveness metrics, KIC uses an agent to retrieve up-to-date external knowledge to generate the evaluation criteria.

- **Protocol Creation (Adaptive):** For each query, the *Protocol Creation Agent* (equipped with web search tools) identifies K essential facts required for a complete answer. Each fact is converted into a yes/no question q_k (e.g., "Does the report mention the Jan 2026 deadline?") grounded in retrieved evidence.
- **Evaluation:** The LLM Evaluator checks the generated report against each question q_k. Let $v_{k,t} \in \{0, 1\}$ be the verification result for question k on task t.
- **KIC Scoring:** The score is the recall rate of these key facts:

$$\text{KIC}_t = \frac{1}{K} \sum_{k=1}^{K} v_{k,t}$$

C.6 Reasoning Quality

Reasoning Quality (RQ) evaluates the logical rigor and analytical depth of the report. It focuses on the validity of arguments, the synthesis of disparate sources, and the avoidance of logical fallacies.

- **Validation Plan Generation:** The Protocol Creation Agent generates a set of open-ended, query-specific *Challenging Questions*. For each question, it generates a structured *Validation Plan* consisting of: (1) Extraction of reasoning chains from the report, (2) External verification using tools (e.g., web search, ArXiv, GitHub), and (3) Comparison criteria.

- **Agentic Execution:** The *Agent Evaluator* executes the validation plan. It actively retrieves external evidence to verify the report's reasoning steps. Unlike KIC which checks for the presence of facts, RQ checks the soundness of the logic connecting them.
- **RQ Scoring:** The Agent Evaluator assigns a score $R_t \in [0, 10]$ based on a deductive rubric (starting at 10, deducting points for logical gaps, unsupported inferences, or ignored counter-evidence). The final normalized score is:

$$\mathrm{RQ}_t = \frac{R_t}{10}$$

D Controlled Experiments Details

D.1 Temporal Awareness in KIC

To generate the reference reports for RACE, we utilized Smolagents' open Deep Research agent (Roucher et al., 2025), an open-source CodeAgent-based framework. We performed three independent generation runs per query to create distinct report versions: one reflecting current information, and two simulating outdated knowledge bases with cutoffs set to January 1st, 2025, and January 1st, 2024, respectively. These simulated cutoffs were enforced by appending specific temporal constraints to the system prompt and strictly filtering results from the web search tool to exclude any content published after the target dates.

In Table 4, we present the complete evaluation results for DRB–RACE Comprehensiveness, Insight, and DREAM–KIC across all 20 queries. The data indicates that DRB–RACE lacks consistent temporal sensitivity. While Comprehensiveness scores do drop on average for the Jan-2024 cutoff (e.g., Query 2 "DOGE agency authority"), the metric is effectively blind to the Jan-2025 cutoff, with average scores remaining virtually identical to the baseline. Furthermore, RACE metrics frequently exhibit paradoxical behavior, where scores increase for outdated reports (e.g., Query 1 "Current papal leadership"). In contrast, DREAM–KIC demonstrates a far stronger alignment with the information lag. Although not strictly monotonic (scores fluctuated in two instances – Queries 11 and 15) DREAM–KIC typically imposes drastic penalties as the knowledge horizon recedes, effectively capturing the absence of specific, recent developments that the baseline metrics miss.

To complement the results in Figure 4, the complete set of DRB–RACE Comprehensiveness and Insight criteria as well as KIC checklist items is provided in Table 5.

Table 4: **Temporal sensitivity analysis.** Comparison of DRB-RACE and DREAM-KIC scores across current (Upd.) and simulated outdated knowledge cutoffs (Jan 2025, Jan 2024) for the DRA. Unlike baselines, DREAM-KIC scores degrade on reports generated with older knowledge cutoffs, effectively capturing information lag.

ID	Topic	DRB-RACE Comp.			DRB-RACE Insight			DREAM-KIC		
		Upd.	Jan-25	Jan-24	Upd.	Jan-25	Jan-24	Upd.	Jan-25	Jan-24
1	Current papal leadership & policies	48.29	53.80	55.75	48.55	52.12	55.23	64.00	0.00	0.00
2	DOGE agency authority & funding	54.17	40.46	23.73	56.44	33.56	17.62	73.33	13.33	0.00
3	US-China LLM competition (DeepSeek)	50.62	47.10	44.75	50.14	45.90	39.76	84.38	56.13	31.25
4	US-Venezuela military tensions	46.96	47.07	49.15	44.38	44.32	47.28	46.88	6.12	0.00
5	Sudan civil war status	48.32	48.72	0.00	46.21	48.46	0.00	81.67	52.00	0.00
6	UK Prime Minister's challenges	55.57	50.59	48.58	55.64	48.17	42.11	64.00	43.80	36.20
7	Boeing CEO & Starliner mission	50.03	51.77	49.86	49.87	52.62	44.32	94.44	78.00	5.33
8	Nvidia market cap ranking	52.30	50.26	49.29	48.41	47.49	42.67	78.00	38.89	22.44
9	Latest Best Picture winner	51.44	48.22	51.18	48.60	46.31	51.08	84.87	6.13	6.13
10	Ukrainian Kursk occupation	50.28	46.63	40.69	48.43	44.93	40.43	80.00	43.33	16.67
11	Protein structure prediction SOTA	48.82	49.73	45.86	44.78	49.05	41.73	65.75	71.88	18.88
12	US Strategic Bitcoin Reserve	48.34	79.21	82.53	52.59	73.29	60.24	81.33	33.00	11.44
13	Digital Euro implementation	51.44	42.92	41.17	56.50	42.30	38.28	97.00	25.00	21.75
14	Tesla Cybercab deployment	46.77	44.88	40.64	53.36	45.73	39.58	75.00	60.71	39.14
15	GLP-1 drug supply status	51.09	47.05	42.01	50.07	49.30	41.09	73.70	82.78	56.13
16	TikTok divestiture deadline	48.61	51.67	43.71	53.70	60.86	44.90	95.83	50.00	16.58
17	SpaceX Starship IFT results	47.47	47.86	47.46	51.22	48.01	45.85	81.25	49.62	34.25
18	Taylor Swift Eras Tour impact	50.62	47.21	43.18	53.06	47.71	45.78	92.20	60.20	56.00
19	Paramount-Skydance merger status	50.87	51.61	47.71	48.40	40.44	51.33	73.33	36.67	10.00
20	Disney CEO succession	48.32	53.94	48.89	50.51	55.10	44.31	100.00	88.35	64.53
	AVG	**50.02**	**50.04**	**44.81**	**50.54**	**48.78**	**41.68**	**79.35**	**44.80**	**22.34**

Table 5: **Task Alignment Criteria Comparison.** DeepResearch Bench RACE rubric criteria alongside DREAM KIC Checklist items for the query: "Prepare a regulatory compliance update on the legal status of TikTok in the U.S. Specifically, clarify the proximity of the deadline for ByteDance to divest its U.S. operations to avoid a nationwide app store ban."

#	RACE Criteria	DREAM KIC Checklist Items
	Comprehensiveness	
1	**Identification and Explanation of Core Legislation**: Assesses if the article clearly identifies the specific U.S. law mandating the potential ban (e.g., the 'Protecting Americans from Foreign Adversary Controlled Applications Act') and explains its key provisions. This is the foundational context for the entire update.	Does the report mention the Protecting Americans from Foreign Adversary Controlled Applications Act (PAFACA) as the specific law requiring ByteDance to divest TikTok?
2	**Detailed Breakdown of the Divestiture Deadline and Extensions**: Evaluates if the article provides the exact initial deadline for divestiture and, crucially, clarifies the conditions and duration of any potential presidential extensions. This directly addresses the task's specific requirement to clarify the deadline's 'proximity.'	Does the report state that the current divestiture deadline is January 23, 2026?
3	**Coverage of Ongoing Legal Challenges**: Checks if the article discusses the lawsuits filed by TikTok/ByteDance or other parties to challenge the law. This is essential for a comprehensive understanding of the current 'legal status,' as these proceedings could impact the deadline and enforcement.	Does the report accurately calculate and convey that approximately 31 days or one month remains until the deadline from late December 2025?
4	**Analysis of Key Stakeholder Positions**: Assesses whether the report covers the stances and actions of key stakeholders, including the U.S. Executive and Legislative branches, ByteDance, and the Chinese government (regarding its approval of tech exports). This provides a complete picture of the geopolitical and corporate landscape.	Does the report mention that ByteDance signed a deal in mid-December 2025 to create a U.S. joint venture, with closing scheduled for January 22, 2026?
5	**Explanation of the Ban's Enforcement Mechanism**: Evaluates if the article explains how the ban would be implemented—specifically, the nature of the nationwide app store ban and its implications for app distribution and hosting services. This clarifies the practical consequences of non-compliance.	Does the report specify the ownership structure of the joint venture deal, including that ByteDance will retain 19.9% and U.S. investors will hold majority control?
6	**Inclusion of Relevant Historical Context**: Checks for a brief summary of previous U.S. government actions or attempts to regulate TikTok (e.g., under the Trump administration). This provides depth and demonstrates a thorough understanding of the long-standing nature of the issue.	Does the report mention that the Supreme Court upheld the TikTok divestiture law in January 2025?
	Insight	
7	**Analysis of Legal Challenges and Timeline Impact**: Evaluates if the article moves beyond merely stating that lawsuits exist to deeply analyze the core legal arguments (e.g., First Amendment rights), assess their potential to secure an injunction or stay, and logically connect these legal outcomes to the potential alteration of the divestiture deadline. This is the primary factor defining the deadline's stability.	Does the report mention that the original deadline was January 19, 2025, and that TikTok briefly went dark on that date?
8	**Synthesis of Factors into Forward-Looking Scenarios**: Assesses the ability to integrate the legal, political, and business analyses into a coherent forecast of potential outcomes (e.g., successful sale, ban proceeds after court losses, extended legal stalemate). This demonstrates a holistic understanding and provides the most valuable strategic insight.	Does the report document that President Trump granted multiple extensions to the TikTok deadline throughout 2025?
9	**Depth of Divestiture Feasibility Assessment**: Evaluates the analysis of the practical barriers to a sale, such as the technical complexity of separating TikTok's recommendation algorithm, the financial valuation challenges, and the need for approval from the Chinese government. This insight is crucial for determining if the deadline is realistically achievable.	Does the report explain the specific consequences of missing the deadline, including prevention of app store updates and distribution?
10	**Contextualization of the Deadline's Political Significance**: Assesses if the article effectively frames the deadline within the broader U.S.-China geopolitical tensions and U.S. domestic political motivations. This provides insight into the 'why' behind the law and the political will to enforce it, which influences any potential for extensions or negotiated solutions.	Does the report mention the bipartisan congressional support for the TikTok divestiture law?
11	**Identification of Key Signposts for Future Monitoring**: Evaluates whether the analysis identifies specific, critical future events or indicators (e.g., key court dates, statements from potential buyers, Chinese regulatory responses) that readers should watch to anticipate how the situation will evolve. This transforms the analysis into an actionable intelligence tool.	Does the report explain the national security concerns about data sharing with China that motivated the law?
12		Does the report mention that TikTok has approximately 170 million U.S. users?
13		Does the report mention the valuation of TikTok's U.S. operations in the context of the deal?
14		Does the report acknowledge that the deal requires Chinese regulatory approval, creating potential uncertainty?

D.2 Detecting Reasoning Flaws

We curated ten research queries spanning domains where analytical reasoning is critical, including policy analysis, technical system comparisons, and causal explanations. Each query was designed to require multi-step reasoning and evidence synthesis, making them suitable for evaluating reasoning quality

Table 6: Research queries for the "Reasoning Flaws Detection" experiment in Section 4.3.

ID	Query
1	How will a 10% price cut affect unit sales next month for an e-commerce product line?
2	What assumptions about inflation, interest rates, and currency stability most influence predictions of stock market volatility in emerging economies over the next five years?
3	How will the adoption of renewable energy technologies impact global oil demand over the next decade?
4	What assumptions about hospital staffing, patient flow, and technology adoption most influence predictions of emergency department wait times over the next five years?
5	How will AI-powered recommendation engines affect consumer purchase diversity in online retail markets over the next two years?
6	What assumptions about monetary policy, housing supply, and demographic trends most influence predictions of urban housing prices over the next decade?
7	How will climate change affect agricultural crop yields in major exporting countries over the next 20 years?
8	What assumptions about vaccine distribution, mutation rates, and healthcare access most influence predictions of pandemic recovery times?
9	How will quantum computing adoption influence financial risk modeling practices over the next 15 years?
10	How would a heatwave next month affect ice cream demand?

detection. The complete query set is provided in Table 6. To generate the reports, we used Smolagents' open Deep Research agent (Roucher et al., 2025). For each query, we generated two report variants using identical source materials. High-quality reports were generated following standard analytical practices with sound logical structure, properly supported claims, and coherent argumentation. Malformed reports were systematically injected with reasoning flaws while preserving surface-level fluency and plausibility; injected flaws included unsupported causal claims, circular reasoning, false equivalences, and cherry-picked evidence. This controlled design yielded 20 total reports (10 high-quality, 10 malformed) forming 10 matched pairs, enabling direct comparison while controlling for query-specific factors. Each report was evaluated using both DREAM–RQ and DRB–RACE frameworks.

D.3 Grounding Beyond Citation Faithfulness

To rigorously test the hypothesis that citation-alignment metrics are blind to well-cited falsehoods, we constructed a controlled dataset focusing on *extrinsic* factual errors, that is, claims that are false in reality but are accompanied by a supporting citation. We manually curated 15 adversarial claim pairs. Each pair consists of:

- A **Ground Truth** variant (c_{true}) accompanied by a valid source URL.
- A **Plausible Hallucination** variant (c_{false}) accompanied by a matching, misleading URL.

Crucially, the matching URLs for the false variants were selected to satisfy standard citation faithfulness checks, serving as specious support for the incorrect claims. These sources span various modes of invalidity, such as outdated information—for instance, citing a 2019 study claiming 46% of Bitcoin transactions are illicit, which directly contradicts current data showing less than 1%. Other examples leverage persistent myths, such as the mechanic's rule of thumb to change oil every 3,000 miles despite modern engineering standards, or fringe narratives, like articles framing 15-minute cities as government-imposed lockdown zones. This setup ensures that the falsity stems from the *content* of the claim relative to objective reality, rather than a mismatch between the claim and its provided citation. The complete set of pairs, contrasting ground truth with aligned misinformation, is listed in Table 7.

To isolate the sensitivity of the metrics, we bypassed the claim-source extraction phase and directly constructed synthetic evaluation payloads. For a given corruption level $r \in [0, 1]$, we assembled a test batch of size $N = 15$ by selecting $r \times N$ false variants and $(1 - r) \times N$ true variants. These batches were then fed directly into the evaluation pipelines of both DRB–FACT and DREAM–Factuality.

Table 7: **Adversarial Factuality Dataset.** The 15 adversarial claim pairs used to measure factual sensitivity. Ground Truth (T) variants are backed by reliable evidence, while False (F) variants are paired with specious citations that satisfy text-alignment criteria but contradict objective reality.

ID	Topic	Variant	Claim	URL Domain
1	15-Min Cities	T	15-minute cities are an urban planning concept to place essential services within walking distance; they do not restrict movement.	fullfact
		F	The '15-minute city' is a lockdown plan to restrict residents' movement and fine them for leaving their zones.	spiked-online
2	GMO Safety	T	Commercially available GMO foods are safe for consumption and do not pose health risks.	natl-academies
		F	There is no scientific consensus on GMO safety; they pose potential long-term risks to health and biodiversity.	ensser
3	Climate Change	T	Rapid warming since the mid-20th century is primarily driven by human activities, specifically fossil fuel emissions.	nasa
		F	Climate forecasting has fundamental uncertainties regarding natural variability; warming may not be as catastrophic as models predict.	judithcurry
4	Homeopathy	T	Homeopathy is a pseudoscience with no medicinal effect beyond placebo.	nhs
		F	Homeopathy is a clinically effective medical treatment for various conditions beyond the placebo effect.	hri-research
5	Dino–Human	T	Non-avian dinosaurs went extinct approximately 65 million years before modern humans evolved.	usgs
		F	Humans and dinosaurs lived on Earth at the same time, approximately 6,000 years ago.	answersingenesis
6	Moon Landing	T	On July 20, 1969, Apollo 11 landed on the Moon; Neil Armstrong and Buzz Aldrin became the first humans to walk on its surface.	britannica
		F	The 1969 Apollo moon landings were staged events, as evidenced by photographic anomalies.	time
7	5G Health	T	5G frequencies are non-ionizing and safe for humans within international guidelines.	who
		F	5G technology will massively increase exposure to radiofrequency radiation, which is proven harmful to humans.	ehtrust
8	Bermuda Triangle	T	There is no evidence that disappearances occur more frequently in the Bermuda Triangle than in other well-traveled ocean areas.	noaa
		F	The Bermuda Triangle is a deadly zone where ships and planes disappear at a frequency far exceeding statistical probability.	smu
9	Knuckle Cracking	T	Knuckle cracking creates a popping sound due to cavitation; there are no known detrimental effects or links to arthritis.	houstonmethodist
		F	Cracking knuckles consistently can wear away cartilage and increase risk of developing arthritis.	nih
10	10% Brain	T	Humans use virtually 100% of their brains throughout the day, even during sleep.	britannica
		F	Most humans only use 10% of their brain capacity, leaving vast potential untapped.	gaia
11	Google Energy	T	A Google search uses about 0.0003 kWh, orders of magnitude less energy than boiling a kettle.	google
		F	Two Google searches generate the same amount of CO2 as boiling a kettle for tea.	hindustantimes
12	EV Emissions	T	Even accounting for battery manufacturing, total lifetime GHG emissions of an EV are lower than a comparable gasoline car.	epa
		F	Widespread EV adoption may increase emissions due to energy-intensive mining and refining of battery materials.	manhattan-inst
13	Crypto Illicit	T	Illicit activity accounts for less than 1% of total cryptocurrency transaction volume.	chainalysis
		F	About 46% of bitcoin transactions are involved in illegal activity, transforming black markets.	repec
14	Earth Shape	T	The Earth is an oblate spheroid, confirmed by satellite imagery, gravity, and centuries of astronomical observation.	wikipedia
		F	The Earth is a stationary plane; the 'South Pole' is an impenetrable ice wall surrounding the world.	gutenberg
15	Oil Change	T	Modern vehicles using synthetic oils can go 7,500–10,000 miles between changes, far surpassing the 3,000-mile guideline.	kbb
		F	Engine oil must be changed every 3,000 miles because additives break down, causing sludge and reducing lubrication.	artmorse

The divergence in performance, clearly visualized in Figure 1 (middle), highlights a fundamental difference in evaluation scope. DRB–FACT proved insensitive to the corruption, with citation accuracy scores remaining effectively constant ($\approx 100\%$) across the entire sweep. Restricted to verifying consistency with the provided URL, the metric correctly identified that the false claims matched their sources, but failed to detect the underlying misinformation. In sharp contrast, DREAM–Factuality exhibited a near-linear degradation inversely proportional to the corruption rate r. By independently retrieving fresh evidence from the web, the evaluator successfully identified the external contradictions for the false variants, correctly rejecting the misleading citations provided in the input.

E Additional Details of Human Evaluation

E.1 Human Evaluation Effort

The evaluation process entails a rigorous workflow to ensure reliable consistency. For each task within DeepResearch Bench (Du et al., 2025), the core data points expected in the report were first identified, followed by a comprehensive review of the three corresponding protocols. This substantial effort ensures that the resulting feedback is sufficiently detailed to serve as a robust ground truth for calculating precision and recall against human judgments.

E.2 Protocol Creation Methods

To assess the effectiveness of our approach, we compare three strategies for creating evaluation protocols:

- **Direct LLM generation:** The LLM is directly prompted with the task description to produce evaluation criteria in a single step, relying solely on its internal knowledge with no access to external resources.

- **Agent without external knowledge:** An agentic system generates evaluation criteria through multi-step reasoning. Although the agent does not have access to external knowledge, the multi-step process enables deeper analysis compared to direct LLM generation.

- **Agent with external knowledge (DREAM):** Our final setting equips the agent with both multi-step reasoning and access to external knowledge capabilities. The agent can query external sources (e.g., web search, and ArXiv/GitHub for domain-specific knowledge) to supplement its analysis.

E.3 Human Evaluation Protocol

Human evaluators are instructed to assess each evaluation point in the generated protocols based on relevance, clarity, and verifiability, with validation soundness additionally included for RQ. The detailed definitions of these criteria are provided in Table 8. To ensure a fair comparison, protocols produced under the three different strategies are evaluated using the same criteria. Each evaluation point is rated on a 1–3 scale (where higher scores indicate better quality), which is then normalized to a 0–1 scale. We further analyze inter-annotator agreement across evaluators in Sec. E.4 to confirm human assessment reliability.

Table 8: **Human evaluation rubric**. Criteria used to guide expert assessment of created protocols.

Dimension	Scoring Guidelines
Relevance	Directly addresses the specific requirements and key aspects of the research query.
Verifiability	Can be objectively confirmed using accessible external evidence or well-defined logic.
Clarity	Is formulated precisely and unambiguously to ensure consistent interpretation.
Validation	Is methodologically correct, free of circular reasoning, and capable of rigorously confirming the answer.

E.4 Inter-Annotator Agreement Analysis

We compute Kendall's coefficient of concordance (Kendall's W) to measure annotator agreement on protocol quality. Kendall's W quantifies the level of concordance among raters when assessments are ranked; here, it reflects the degree to which annotators share similar preferences across the different creation strategies. As reported in Table 9, annotators exhibit moderate agreement on Relevance and Clarity ($W \approx 0.58$), and substantial agreement on Verifiability and Validation Soundness ($W > 0.75$). The overall agreement is statistically significant across all indicators.

We attribute the variance in agreement primarily to the similarity between the *Direct LLM* and *Agent without retrieval* baselines. Since both rely solely on the same backbone model's internal knowledge, their outputs are often qualitatively similar, making it difficult for evaluators to consistently rank one over the other. However, the *Agent with retrieval* is consistently distinguished from these baselines, confirming that the human preference for retrieval-augmented protocols is reliable.

Table 9: Agreement analysis among human annotators, measured using Kendall's W. All results are statistically significant ($p < .05$).

Indicator	Kendall's W	p-value
Relevance	0.58	.030
Verifiability	0.78	.009
Clarity	0.58	.030
Validation Soundness	0.86	.006
Average	**0.69**	**.016**

F Detailed DREAM Benchmarking Results

In this section, we present complete evaluation results for three Deep Research Agents (DRAs) across our three benchmarks, complementing the results in Section 5. A key advantage of the DREAM framework is its dataset-agnostic design, which enables us to normalize scores across heterogeneous tasks and compute a unified aggregate performance profile.

Performance Overview. Table 10 details the performance of each system on static metrics (WQ, Factuality, CI, DA) and adaptive metrics (KIC, RQ), alongside a composite score averaging results across all metrics. As discussed in Section 5, distinct system profiles emerge: Smolagents Open DR leads in most metrics (except CI), though behavior may depend on the dataset, e.g., on DeepResearchBench, Tongyi Deep Research has higher WQ and Factuality.

Table 10: **DREAM evaluation.** Static and adaptive metric scores for leading DRAs across three datasets.

		Static Metrics				**Adaptive Metrics**	
Dataset	**DRA**	**WQ**	**Fact.**	**CI**	**DA**	**KIC**	**RQ**
DeepResearchBench	LangChain Open DR	63.69	46.02	**14.43**	**86.32**	64.99	53.34
	Smolagents Open DR	63.30	57.64	2.45	9.44	**74.15**	**63.68**
	Tongyi Deep Research	**63.95**	**58.58**	0.00	0.00	58.53	40.56
LiveResearchBench	LangChain Open DR	60.75	38.11	**11.90**	**84.00**	63.66	59.04
	Smolagents Open DR	**61.71**	**56.90**	9.67	24.41	**73.55**	**68.82**
	Tongyi Deep Research	60.03	51.69	2.92	5.71	55.17	46.95
ResearchRubrics	LangChain Open DR	62.53	49.13	**19.81**	**90.00**	68.25	57.84
	Smolagents Open DR	**66.60**	**59.38**	2.03	0.00	**78.74**	**72.15**
	Tongyi Deep Research	61.92	56.06	0.00	0.00	59.87	46.75
Aggregate	LangChain Open DR	62.17	44.64	**15.92**	**85.08**	65.96	57.28
	Smolagents Open DR	**63.97**	**58.15**	4.78	17.96	**75.95**	**69.16**
	Tongyi Deep Research	61.72	55.09	1.03	3.38	57.95	45.48

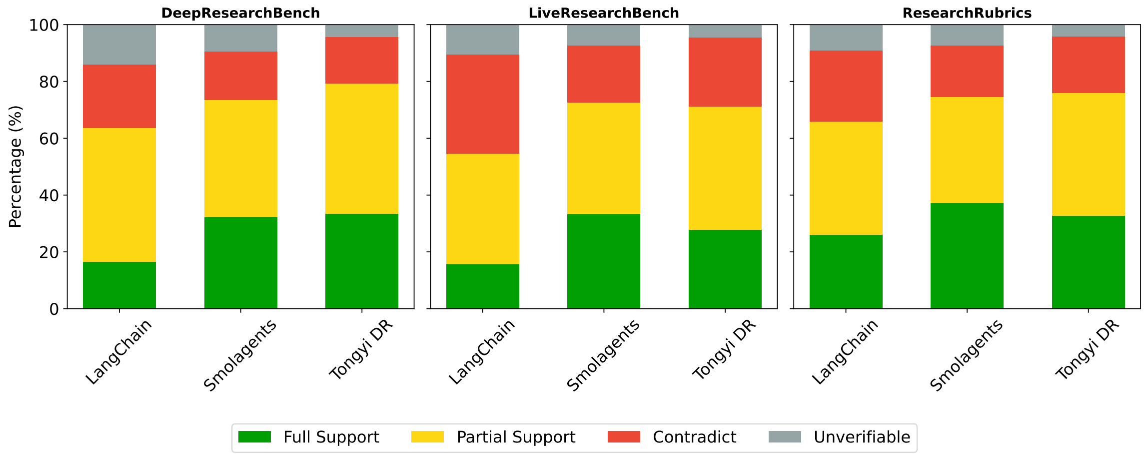

Figure 7: **Factuality Label Distribution.** Distribution of factuality judgments (Full Support, Partial Support, Contradict, Unverifiable) for each model across three datasets.

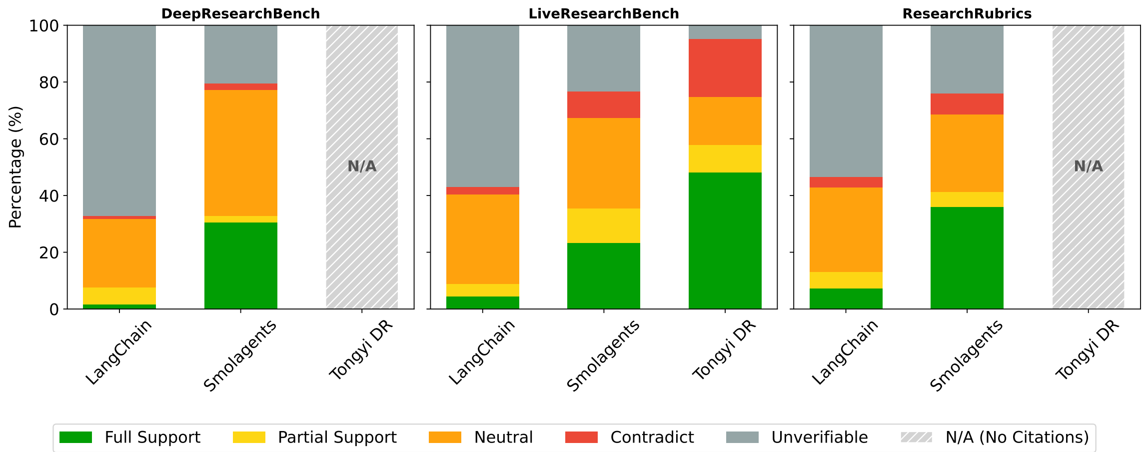

Figure 8: **Citation Faithfulness Label Distribution.** Distribution of citation faithfulness labels assessing the alignment between claims and cited source text across three datasets.

Source Quality Analysis. The detailed breakdown of our Source Quality-related metrics reveals clear distinctions in verification capabilities. Figure 7 presents the distribution of Factuality labels across all three benchmarks. Smolagents typically achieves higher proportion of "Full Support" judgments (green bars), indicating a superior ability to generate claims that are independently corroborated by retrieved external evidence. In contrast, LangChain exhibits higher rates of "Unverifiable" or 'Contradict' claims. A similar pattern emerges in Figure 8, which details the Citation Faithfulness (CF) labels. Tongyi Deep Research produces no citations at all on DEEPRESEARCHBENCH and RESEARCHRUBRICS (marked N/A), while LangChain and Smolagents show high proportions of "Unverifiable" labels, indicating that cited sources frequently cannot corroborate the associated claims. Figure 9 visualizes the components of Citation Integrity by plotting CF against Citation Attribution (CA). This scatter plot reveals two distinct failure modes: LangChain attributes claims to sources frequently (high CA, $\approx$ 75–80) but with poor faithfulness (low CF, $\approx$ 10–20), meaning it cites often but inaccurately. Conversely, Smolagents and Tongyi Deep Research rarely attribute claims at all (low CA, $\approx$ 5–15), though when they do, faithfulness is moderately higher (CF $\approx$ 35–55). Neither strategy yields reliable citation behavior, underscoring that open-source DRAs have yet to bridge the gap between source retrieval and faithful evidence grounding.

Report Length Analysis. Figure 10 and Table 11 present the word count statistics (Mean $\pm$ Std, Min, Max) and distributions, respectively, for each agent. Smolagents Open DR consistently generates the longest reports across all datasets, averaging $\approx$ 3,000–3,700 words, with high variance and heavy-tailed distributions (e.g., $\pm$ 2,066 words on RESEARCHRUBRICS, with outliers exceeding 14,500 words). LangChain Open DR and Tongyi Deep Research produce more concise output in the 1,400–1,800 word

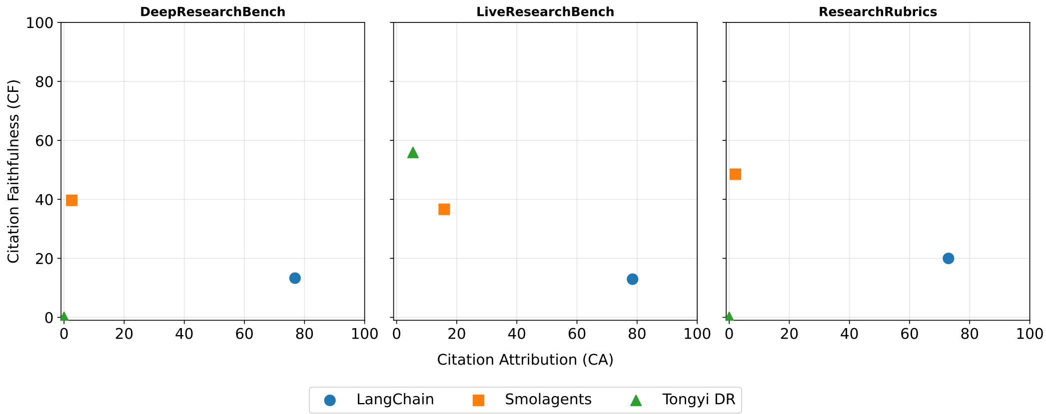

Figure 9: **Citation Integrity Components.** Citation Faithfulness and Citation Attribution visualized for each model across three benchmarks.

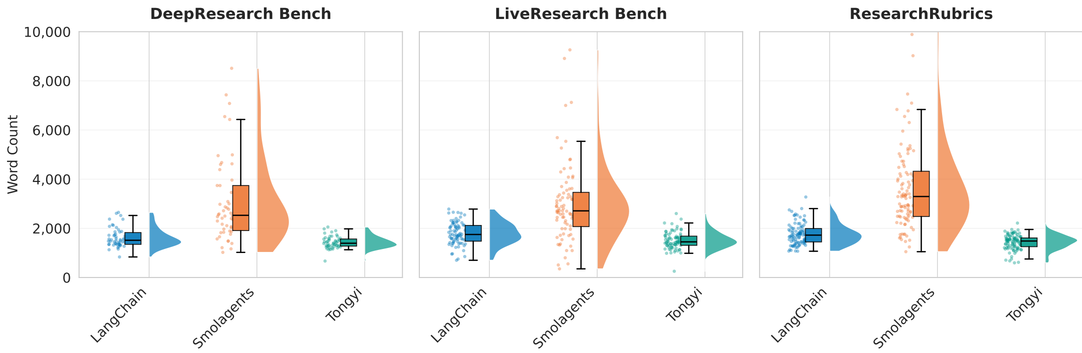

Figure 10: **Report length distributions across DRAs and datasets.** Word count distributions across DRAs and datasets. Smolagents Open Deep Research consistently exhibits the highest variance and mean length, while LangChain Open Deep Research and Tongyi Deep Research are more concise and consistent.

range, with Tongyi being the most consistent (Std $\approx$ 250–350 words).

Robustness Across Backbone Models. We examined the impact of the underlying LLM backbone used to execute DREAM's evaluation protocol on DEEPRESEARCH BENCH. We compared the default model, Claude Sonnet 4.5, against DeepSeek-V3.2 and Kimi-K2.5 across our suite of static and adaptive metrics. The findings detailed in Table 12 highlight DREAM's robust evaluation signal. Even though absolute scores may change due to varying internal grading thresholds, the relative performance rankings of the evaluated deep research agents remain highly consistent. A minor exception is the Writing Quality (WQ) metric, for which Claude Sonnet 4.5 did not provide strong differentiation among agents, whereas

Table 11: **Report length statistics across DRAs and datasets.** Word counts are reported as Mean $\pm$ Standard Deviation. Smolagents consistently produces the longest reports, while Tongyi Deep Research is the most concise.

Dataset	**DRA**	**Mean $\pm$ Std**	**Min**	**Max**
DeepResearchBench	LangChain Open DR	1628.3 $\pm$ 398.3	836	2650
	Smolagents Open DR	3092.0 $\pm$ 1728.0	1022	8511
	Tongyi Deep Research	1431.0 $\pm$ 248.7	670	2055
LiveResearchBench	LangChain Open DR	1792.0 $\pm$ 459.9	701	2784
	Smolagents Open DR	3009.3 $\pm$ 1627.8	352	9261
	Tongyi Deep Research	1505.2 $\pm$ 350.9	254	2602
ResearchRubrics	LangChain Open DR	1763.2 $\pm$ 430.7	1069	3275
	Smolagents Open DR	3735.2 $\pm$ 2065.6	1050	14554
	Tongyi Deep Research	1436.1 $\pm$ 301.4	597	2214

DeepSeek-V3.2 and Kimi-K2.5 yield identical rankings.

Table 12: **Backbone Model Comparison.** Metric scores on DEEPRESEARCHBENCH across three judge backbones. Agent rankings remain largely consistent regardless of the judge model used.

		Static Metrics				Adaptive Metrics	
Judge Model	**DRA**	**WQ**	**Fact.**	**CI**	**DA**	**KIC**	**RQ**
Claude Sonnet 4.5	LangChain Open DR	63.69	46.02	**14.43**	**86.32**	64.99	53.34
	Smolagents Open DR	63.30	57.64	2.45	9.44	**74.15**	**63.68**
	Tongyi Deep Research	**63.95**	**58.58**	0.00	0.00	58.53	40.56
DeepSeek V3.2	LangChain Open DR	66.19	47.35	**17.24**	**86.14**	63.86	67.68
	Smolagents Open DR	**67.10**	54.45	6.02	60.72	**73.57**	**71.89**
	Tongyi Deep Research	65.27	**57.46**	0.00	6.22	57.00	56.20
Kimi K2.5	LangChain Open DR	61.90	48.96	**17.70**	**84.63**	60.53	57.50
	Smolagents Open DR	**64.43**	58.76	3.25	13.30	**69.74**	**63.00**
	Tongyi Deep Research	58.82	**61.55**	0.00	0.00	51.01	43.27