DR^3-Eval: Towards Realistic and Reproducible Deep Research Evaluation

Anonymous ACL submission

Abstract

Deep Research Agents (DRAs) aim to solve complex, long-horizon research tasks involving planning, retrieval, multimodal user-file understanding, and report generation, yet their evaluation remains challenging due to unstable live-web environments and ambiguous task definitions. We propose **DR^3-Eval**, a realistic and reproducible benchmark for evaluating deep research agents on multimodal, multi-file report generation. DR^3-Eval is constructed from authentic user-provided materials and paired with a per-task static research sandbox corpus that simulates open-web complexity while remaining repeatable and fully verifiable, containing supportive documents, distractors, and noise. Moreover, we introduce a multi-dimensional evaluation framework measuring Information Recall, Factual Accuracy, Citation Coverage, Instruction Following, and Depth Quality, and validate its alignment with human judgments. Experiments with our developed multi-agent system DR^3-Agent across multiple state-of-the-art language models demonstrate that DR^3-Eval is highly challenging and reveals persistent bottlenecks in evidence retrieval, citation coverage, and factual reliability.

1 Introduction

Recent advances in large language models have enabled the development of Deep Research Agents (DRAs), which aim to autonomously perform complex, long-horizon research tasks involving planning, iterative information retrieval and synthesis of structured reports (Li et al., 2025; Schmidgall et al., 2025; Xu and Peng, 2025; OpenAI, 2025; Google, 2025; Team, 2025; xAI Team, 2024; ByteDance, 2025). Unlike traditional question-answering systems, DRAs must operate under uncertainty, reason over heterogeneous and noisy information sources, and integrate evidence into coherent analytical outputs. As these capabilities rapidly improve, establishing realistic, reproducible, and deterministic evaluation protocols has become increasingly important.

Recently, in Figure 1, several benchmarks have been proposed to evaluate deep research agents (Wang et al., 2025; Han et al., 2026; Sharma et al., 2025; Patel et al., 2025; Li et al., 2026; Ruan et al., 2025) . For example, DeepResearch Bench (Du et al., 2025) emphasizes open-ended report generation and long-horizon reasoning, but typically relies on live web access, making results difficult to reproduce and prone to evaluation ambiguity, because search rankings, page availability, and retrieved content can change over time. DR-Bench (Abaskohi et al., 2025) improves structure by focusing on enterprise-style report generation from curated documents, yet largely omits explicit modeling of noisy or misleading information that is common in real-world research. More recently, DeepResearchGym (Coelho et al., 2025) introduces a sandbox-based framework that replaces live web access with fixed local corpora, but its tasks are purely text queries and lack grounding in authentic user research workflows. As a result, *there remains a gap between real-world research complexity—which involves multimodal user-provided materials, noisy and misleading information, and implicit research intent—and the environments used to evaluate DRA performance.*

To address these limitations, we introduce **DR^3-Eval**, a benchmark to reconcile realism, controllability, and reproducibility for deep research. DR^3-Eval targets report-generation tasks grounded in real-world needs, constructed from authentic multimodal user files that users have encountered in practice. Following the evaluation paradigm of DeepResearchGym and BrowseComp-Plus (Chen et al., 2025), we localize the retrieval corpus into a static, controlled sandbox rather than directly evaluating agents on the live web. Each task is paired with a per-case sandbox corpus that simulates the open web as a deep-research environment while re-

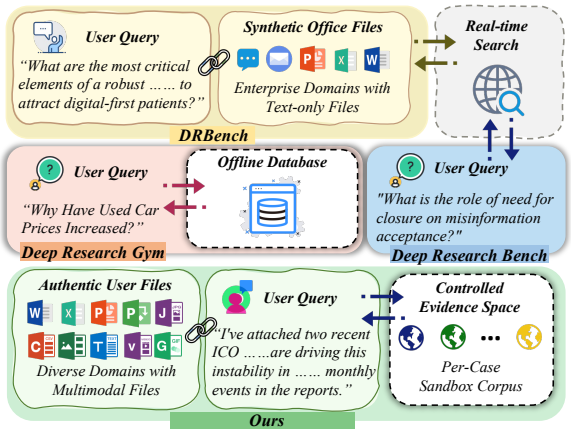


Figure 1: Comparison of different benchmarks. Given text queries, **Deep Research Bench** executes text queries via real-time search, and **DeepResearchGym** retrieves from a global offline database. **DRBench** incorporates user files (text modality) as input but relies on real-time search and focuses on the enterprise domain. In contrast, **DR^3-Eval** processes queries and files in a controlled sandbox corpus on diverse domains.

maining fully static and verifiable. Within this sandbox, documents are carefully curated to include evidential sources, confounding documents, and ambient noise. Besides, a key feature of DR^3-Eval is its reverse-construction methodology: instead of posing open-ended questions with uncertain answerability, we derive each query from verified documents, ensuring every task admits a single, well-defined solution path. To support fine-grained assessment, we propose an evaluation framework that measures Information Recall, Factual Accuracy, Citation Coverage, Instruction Following, and Depth Quality.

Moreover, to show the utility of DR^3-Eval, we have developed **DR^3-Agent**, a multi-agent research system adapted to the benchmark's closed-world setting, which can also take text user queries and corresponding user files (including text, image, video, audio, etc.) as input. Extensive experiments across state-of-the-art language models reveal that DR^3-Eval is highly challenging and exposes failure modes that are obscured by existing benchmarks. Overall, the contributions are as follows: (1) We propose DR^3-Eval, a realistic, reproducible, and multimodal user-file benchmark for evaluating deep research agents in the report generation setting. (2) We introduce a controlled sandbox-based task construction pipeline that balances real-world complexity with verifiable evaluation. (3) We build the DR^3-Agent and conduct comprehensive benchmarking of frontier and open-weight models.

Benchmark	Task	Domain	UF	SC	MM	RS	MF	RC
GAIA (Mialon et al., 2023)	QA	General	✓	✗	✓	✓	✗	✗
HLE (Phan et al., 2025)	QA	General	✓	✗	✓	✗	✓	✗
BrowseComp-Plus (Chen et al., 2025)	QA	General	✗	✓	✗	✓	✗	✓
DocBench (Zou et al., 2024)	QA	General	✓	✗	✓	✓	✗	✗
MMLongBench-Doc (Ma et al., 2024)	QA	General	✓	✗	✓	✗	✗	✗
DRBench (Abaskohi et al., 2025)	Report	Enterprise	✓	✓	✗	✓	✓	✓
Deep Research Bench (Du et al., 2025)	Report	General	✗	✗	✗	✓	✗	✗
Deep Scholar Bench (Patel et al., 2025)	Report	Academic	✗	✗	✗	✓	✗	✗
DEER (Han et al., 2026)	Report	General	✗	✗	✗	✗	✗	✗
LiveResearch Bench (Wang et al., 2025)	Report	General	✗	✗	✗	✓	✗	✗
DeepResearchGym (Coelho et al., 2025)	Report	General	✗	✓	✗	✓	✗	✗
DR^3-Eval (Ours)	Report	General	✓	✓	✓	✓	✓	✓

Table 1: Comparisons of different representative benchmarks. Column abbreviations: **UF**=User Files, **SC**=Sandbox Corpus, **MM**=Multi-modal, **RS**=Real Scenario, **MF**=Multi-File, and **RC**=Reverse Construction. Unlike prior work, DR^3-Eval combines user files and sandbox corpus, proposing a realistic, reproducible, and multimodal benchmark for evaluating deep research agents in report-generation settings.

2 Related Work

Deep Research Agent. The capabilities of deep Research Agents (DRAs) have evolved beyond the traditional question-answering paradigm (Chen et al., 2026) to autonomously planning long-horizon workflows, navigating heterogeneous web sources, and ultimately synthesizing information into structured reports (OpenAI, 2025; Google, 2025; Team, 2025; AI, 2024; xAI Team, 2024). Currently, mainstream deep research systems have demonstrated powerful capabilities in handling complex research tasks, but they are mostly available as closed-source commercial products; open-source efforts, in contrast, emphasize modularity and reproducibility (Li et al., 2025; Yang et al., 2025; Qiao et al., 2025; MiroMindAI, 2025).

Deep Research Benchmark. Early efforts primarily addressed general reasoning and tool-use in QA scenarios (Mialon et al., 2023; Phan et al., 2025), which have recently evolved into complex information-seeking tasks in open-web environments. However, striking an ideal balance between reproducibility and realism remains inherently challenging. Benchmarks relying on live web access (Wang et al., 2026; Han et al., 2025) provide high ecological validity but suffer from temporal volatility, where fluctuating search results make performance comparisons inconsistent over time. Conversely, existing sandbox-based benchmarks (Coelho et al., 2025) ensure stability but omit the multimodal complexity and the confounding noise inherent in authentic research. In Table 1, DR^3-Eval aims to address the above limitations by reconciling real-world research complexity with a reproducible evaluation protocol.

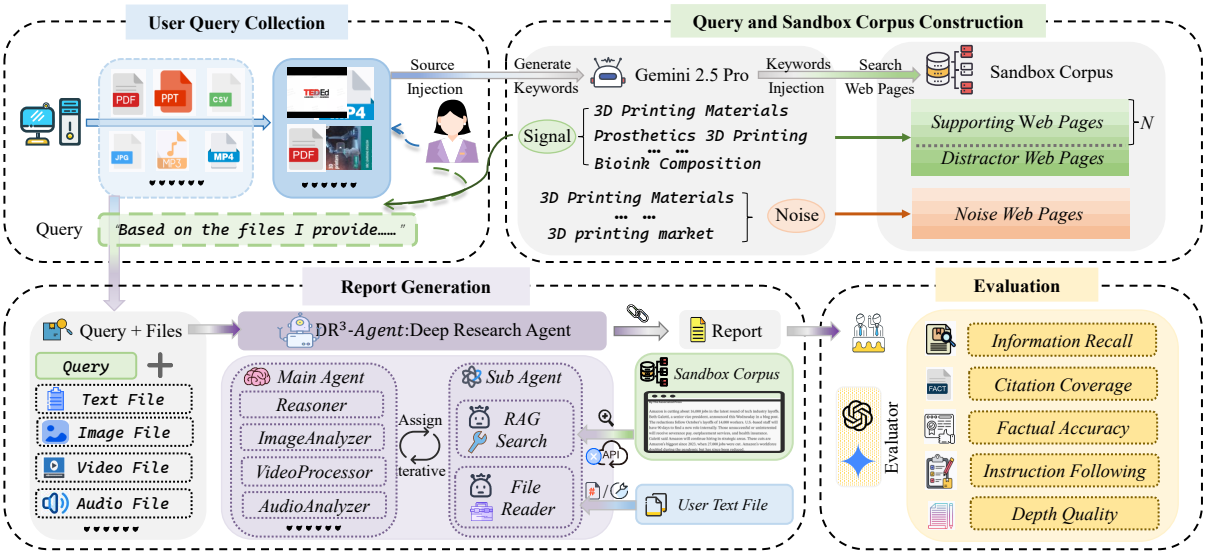


Figure 2: Overview of the DR^3-Eval framework. (1) **Data construction** synthesizes search paths from real-world multimodal files via a divergent-convergent mechanism, establishing a static sandbox with controlled signal-to-noise ratios and backward-derived queries. (2) Our **DR^3-Agent** adopts a hierarchical multi-agent architecture where a perception-enhanced Main Agent coordinates global reasoning while specialized sub-agents execute iterative sandbox retrieval and file parsing. (3) **Evaluation protocol** utilizes a multidimensional metric suite to comprehensively assess performance in both evidence acquisition and analytical report generation.

3 DR^3-Eval

3.1 Data Construction

The construction process involves five stages; detailed prompts are provided in Appendix A.

Stage 1: Grounding in Real-World Needs. We recruited human volunteers and tasked them with providing intrinsically relevant material sets, and all collected materials then underwent an automated sanitization process, where an automated script first identified and redacted personally identifiable information (PII), followed by a manual cross-validation by a separate group of annotators.

Stage 2: Distilling Search Paths. We design a two-stage "divergent-convergent" process (Design Council, 2005; Yao et al., 2023) to generate search keywords. We leverage Gemini-2.5-Pro to perform an open-ended analysis of the user files, generating an initial set of candidate keywords, simulating a wide-ranging "brainstorming" session in the divergent stage. Subsequently, in the convergent stage, the model divides them into two categories: (1) Signal Keywords, which collectively point toward the core solution path; and (2) Noise Keywords, which are thematically related but designed to lead to irrelevant or misleading information.

Stage 3: Building Research Sandbox. We utilize the keywords to retrieve up to 100 web results for each keyword via the Google Serper API. We categorize all processed documents into three types: (1) Supportive: high-relevance results from signal keywords, whose content is manually verified to provide necessary and sufficient evidence to answer the query; (2) Distractor: also from signal keywords, but their content is confirmed to be outdated, one-sided, or inaccurate; (3) Noise: results from noise keywords for evaluation environments with varying signal-to-noise ratios. The distribution is shown in Figure 3(d). To simulate the "long-tail effect" in real-world deep research, we set five different context lengths: 32k, 64k, 128k, 256k, and 512k tokens. All settings include the supportive web pages. Additionally, the number of distractor web pages increases proportionally with the total context length, and the remaining token quota is filled with noise web pages.

Stage 4: Constructing Query. Inspired by the backward design approach used in works like BrowseComp (Wei et al., 2025) for building QA benchmarks, we synthesize the query based on the pre-determined evidential documents, integrated with the signal keywords. In this way, we ensure each query requires joint reasoning over the user files and specific web evidence.

Stage 5: Quality Control. Starting from 280 candidate queries, we enforce a validation protocol to guarantee the rigorousness of candidate queries, yielding a final pass rate of 35.7%. (1) Queries must guide agents toward signal keywords without verbatim disclosure, excluding 30 queries. (2)

A "leave-one-out" verification is employed to ensure that the final answer is strictly contingent on combining the initial user files with specific web evidence, filtering out 105 queries. (3) Queries are disqualified if the core factual claims of the golden insight are directly retrievable from public search engines, eliminating 40 queries that can be answered without the user files. (4) Each query undergoes manual inspection to eliminate ambiguity and guarantee a singular, precise interpretation, where five additional queries are rejected.

3.2 Dataset Statistics

DR^3-Eval comprises 100 independent tasks with an even split between English and Chinese samples. In Figure 3(a), the tasks cover technology, economy, and humanities, subdivided into 13 atomic domains such as computer science, healthcare, and policy. Regarding input modalities, Figure 3(b) presents a distribution of 45.98% documents, 27.68% images, and 13.84% videos, alongside data sheets, audio, and HTML files, with specific format breakdowns detailed in Appendix B. Figure 3(c) further plots the file density per task, which averages 2.24 user files and reaches a maximum of 6. Besides, the sandbox corpus introduces massive external noise, containing an average of 465.5 web pages per task under the 512k token configuration.

4 DR^3-Agent

4.1 Framework Construction

Current open-source deep research frameworks (ByteDance, 2025; DeepResearch Team, Tongyi Lab, 2025; CAMEL-AI, 2026) cannot directly handle the sandbox corpus and the cross-reading of multimodal files tasks proposed in DR^3-Eval. Therefore, we develop DR^3-Agent, an LLM-driven system based on the MiroFlow framework (MiroMind Team et al., 2025). In Figure 2, we integrate perception tools into the main agent to effectively handle multimodal user files such as audio and video. Besides, we employ two sub-agents that share the model backbone and return only condensed summaries to the main agent. Specifically, the RAG search sub-agent interacts with the static sandbox corpus. Unlike conventional RAG, our RAG sub-agent performs autonomous multi-step retrieval with iterative query refinement within the sandbox corpus. This requires the agent to evaluate incomplete or conflicting evidence and revise its search direction across iterations, making the search process functionally analogous to heuristic exploration over hyperlink graphs. Meanwhile, the file reader sub-agent specializes in parsing long-text user files, utilizing tools to execute fine-grained queries and retrieve content by page numbers.

4.2 Evaluation Metrics

DR^3-Eval comprises five metrics, categorized into two dimensions: *Information Seeking*, which assesses the quality of gathered evidence, and *Report Generation*, which evaluates the final output. We utilize Φ (GPT-5.1) as the evaluator, with specific prompts detailed in Appendices E and F. A report example using DR^3-Agent is shown in Appendix C, with evaluation details in Appendix D.

4.2.1 Information Seeking

Information Recall (IR) We employ Gemini-2.5-Flash to extract insight sets $\mathcal{I}_{\text{UF}}$ and $\mathcal{I}_{\text{SC}}$ (Pradeep et al., 2025) from user files and the sandbox corpus, respectively. All extracted insights are manually verified to ensure accuracy. Subsequently, we use the evaluator model Φ to assess the report R's coverage of each insight i, assigning a score $\text{cov}(i, R) \in \{1, 0.5, 0\}$ (Patel et al., 2025).

$$\text{IR}_{\text{UF}}(R, \mathcal{I}_{\text{UF}}) = \frac{1}{|\mathcal{I}_{\text{UF}}|} \sum_{i \in \mathcal{I}_{\text{UF}}} \mathbb{1}[\text{cov}(i, R) = 1] \tag{1}$$

$$\text{IR}_{\text{SC}}(R, \mathcal{I}_{\text{SC}}) = \frac{1}{|\mathcal{I}_{\text{SC}}|} \sum_{i \in \mathcal{I}_{\text{SC}}} \mathbb{1}[\text{cov}(i, R) = 1] \tag{2}$$

Citation Coverage (CC) Inspired by DeepScholar-Bench (Patel et al., 2025), we establish the ground truth set $\mathcal{D}_{\text{req}}$. Let $\mathcal{D}_{\text{cited}}$ denote the documents explicitly cited in R. The coverage is defined as:

$$\text{CC}(R, \mathcal{D}_{\text{req}}) = \frac{|\mathcal{D}_{\text{req}} \cap \mathcal{D}_{\text{cited}}|}{|\mathcal{D}_{\text{req}}|} \tag{3}$$

4.2.2 Report Generation

Factual Accuracy (FA) We extract the set of all claim-source pairs $\mathcal{C}$ from the generated report R and employ Φ to evaluate textual claims, while utilizing Gemini-2.5-Pro to verify claims grounded in video or audio content. For each pair $(c, s) \in \mathcal{C}$, we define an entailment function $\mathbb{V}(c, s)$ that equals 1 if s supports c, and 0 otherwise:

$$\text{FA}(R) = \frac{1}{|\mathcal{C}|} \sum_{(c,s) \in \mathcal{C}} \mathbb{V}(c, s) \tag{4}$$

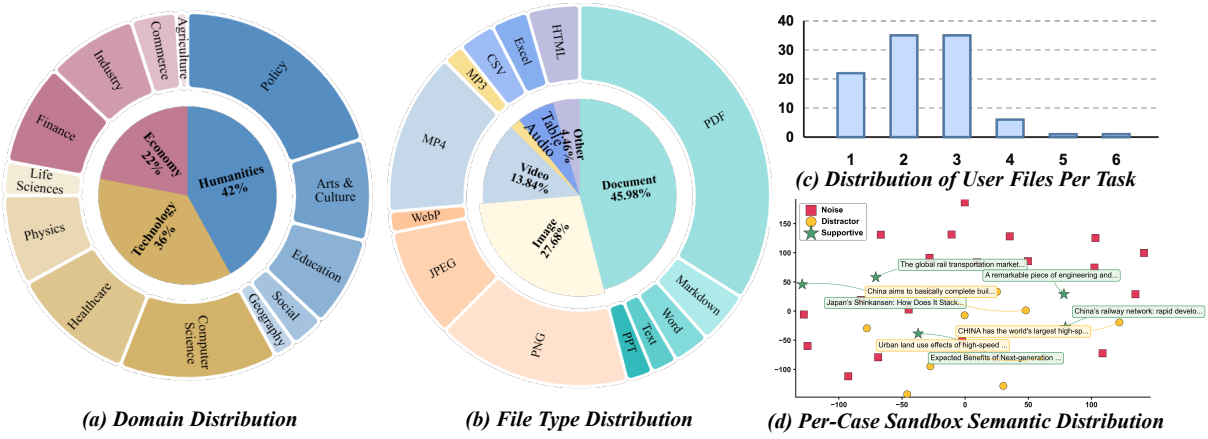


Figure 3: Dataset statistics. (a) Domain coverage spanning Technology, Economy, and Humanities, comprising 13 atomic sub-domains. (b) Distribution of file types. (c) Distribution of user files per task. (d) t-SNE visualization of the semantic distribution in the Sandbox Corpus.

Instruction Following (IF) We utilize Φ to generate a checklist $\mathcal{L}$ based on the task query, covering aspects such as content, evidence, and analysis (Sharma et al., 2025; Wang et al., 2025), and checklists are manually verified. We then evaluate whether R satisfies each requirement $l \in \mathcal{L}$, indicated by a binary score $\mathbb{S}(l, R) \in \{1, 0\}$:

$$\mathrm{IF}(R, \mathcal{L}) = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \mathbb{S}(l, R) \tag{5}$$

Depth Quality (DQ) We employ the model Φ to evaluate the analytical substance and logical rigor of R. The quality score is assigned conditioned on the query Q and a predefined rubric $\mathcal{P}$:

$$\mathrm{DQ}(R, Q) = \Phi(R, Q \mid \mathcal{P}) \tag{6}$$

5 Experiments

5.1 Experimental Settings

For DR^3-Agent, the maximum interaction turns for the main agent is set to 10, while the sub-agents for RAG and file reading are limited to five and three turns, respectively. We utilize OpenAI's text-embedding-3-small for vectorization. For baselines, we use DR^3-Agent as the outer framework and evaluate GPT-4.1 (OpenAI, 2025), Claude Sonnet 4 (Anthropic, 2025), Gemini 2.5 Pro (Google DeepMind, 2025), Qwen3-235B-A22B (Tongyi Lab, 2025), Qwen3-30B-A3B (Tongyi Lab, 2025), Qwen3-32B (Tongyi Lab, 2025), GLM-4.6 (Zhipu AI, 2025), GLM-4.7 (Zhipu AI, 2025), and additional backbone candidates including DeepSeek-V4-Pro (DeepSeek-AI, 2026), GLM-5.1 (Zhipu AI, 2026), Gemini-2.5-Flash (Google DeepMind, 2025), Qwen3-8B (Tongyi Lab, 2025), Qwen3.6-Max-Preview (Qwen Team, 2026), and GPT-5.5 (OpenAI, 2026) by replacing only the backbone model. In evaluation, we introduce GPT-5.1 (OpenAI, 2025) as a model-anonymized blind judge, and use Gemini-2.5-Pro as an assistant judge. To ensure the evaluation is deterministic, the temperature for all judge models is set to 0. Runtime and API cost statistics are provided in Appendix G.

5.2 Main Results

In Table 2 and Figure 4, we have the following observations: (1) **DR^3-Eval is very challenging.** GPT-5.5 achieves the highest overall performance across all context lengths, while GLM-5.1, Claude Sonnet 4, and DeepSeek-V4-Pro form the next strongest group. Besides, within the same model family (i.e., Qwen), the scaling law is still a key factor in complex tasks. (2) **Longer contexts lead to lower performance.** As the size of the sandbox corpus grows from 64k to 512k, a general drop in performance is seen across all models. We hypothesize that longer contexts introduce more noise and irrelevant information, which makes it more difficult to obtain valuable insights. (3) **Better instruction following does not indicate higher factual accuracy.** For example, some models (e.g., Qwen3-235B-A22B and GPT-4.1) achieve relatively good results in Instruction Following(IF), but obtain very low factual accuracies. This suggests these models cannot accurately obtain sufficient information from the given materials: they tend to create a report that "looks" complete and satisfies the query, at the high cost of Factual Accuracy (FA). We further provide a modality-level breakdown in Ap-

Models	Information Seeking									Report Generation									Total Score		
	IR_{UF}			IR_{SC}			CC			FA			IF			DQ			Avg.		
	64k	128k	512k	64k	128k	512k	64k	128k	512k	64k	128k	512k	64k	128k	512k	64k	128k	512k	64k	128k	512k
GPT-5.5	**73.8**	**72.5**	**71.5**	55.2	51.2	46.0	**66.8**	**60.4**	**51.6**	82.7	**83.2**	80.5	**95.1**	**95.2**	**95.4**	**80.5**	**80.3**	**80.7**	**75.7**	**73.8**	**71.0**
GLM-5.1	64.5	65.7	68.3	54.8	**53.2**	**48.8**	65.1	58.7	47.2	75.1	71.5	69.8	86.2	84.7	88.2	72.4	72.2	74.6	69.7	67.7	66.2
Claude Sonnet 4	58.8	60.4	60.8	55.3	46.6	41.8	64.7	54.8	48.5	**87.0**	82.7	82.1	87.4	89.2	88.5	70.7	71.5	72.0	70.7	67.5	65.6
DeepSeek-V4-Pro	69.0	63.0	60.5	**58.8**	50.1	47.1	63.0	54.2	46.7	78.4	70.0	72.5	90.0	85.5	88.3	76.4	73.9	74.9	72.6	66.1	65.0
GLM-4.7	55.7	55.0	57.1	53.1	47.6	42.1	65.4	55.9	45.3	84.5	82.1	80.3	88.8	89.3	88.1	71.1	71.8	72.1	69.8	66.9	64.1
Qwen3.6-Max-Preview	58.8	58.0	58.3	51.2	48.2	42.9	62.5	57.3	46.3	82.2	77.7	74.0	88.5	87.0	87.0	73.1	72.3	72.7	69.4	66.7	63.5
GLM-4.6	53.4	52.6	50.3	49.5	43.9	39.8	58.2	52.0	44.0	84.0	82.3	**82.9**	85.6	87.2	86.4	70.1	69.3	70.6	66.8	64.5	62.3
Gemini-2.5-Pro	52.7	53.7	50.4	44.8	42.4	37.6	54.3	49.5	36.6	81.3	80.7	80.0	84.9	84.5	84.5	67.1	68.3	67.4	64.2	63.2	59.4
Gemini-2.5-Flash	50.7	52.7	51.8	42.9	37.0	34.8	48.8	44.8	38.6	86.1	82.3	81.2	78.8	81.0	82.2	66.1	67.7	67.8	62.2	60.9	59.4
GPT-4.1	40.7	42.5	41.3	30.9	29.4	29.2	37.2	35.6	30.0	56.4	54.2	58.8	83.0	83.3	82.7	63.1	63.2	63.4	51.9	51.3	50.9
Qwen3-235B-A22B	37.4	36.0	39.7	35.7	29.8	28.8	40.6	36.6	31.8	52.5	53.6	49.8	78.0	78.6	80.2	62.1	62.8	61.9	51.1	49.6	48.7
Qwen3-32B	33.2	36.6	35.4	26.5	25.3	24.7	34.2	32.3	26.1	49.4	52.2	51.5	73.5	74.2	74.3	58.8	59.9	59.3	45.9	46.7	45.2
Qwen3-8B	27.7	32.9	29.9	27.8	25.1	22.7	36.2	38.2	29.7	57.9	44.3	60.0	69.0	73.6	67.6	56.3	58.1	57.4	45.8	45.4	44.6
Qwen3-30B-A3B	30.9	38.2	34.1	23.2	25.7	23.5	26.6	25.3	21.5	41.9	46.8	45.2	73.2	71.8	74.7	57.6	58.2	58.0	42.2	44.3	42.8

Table 2: Evaluation results on DR^3-Agent. The best and second-best performances are highlighted in **bold** and underlined, respectively. **Key**: IR_{UF}/IR_{SC} = Information Recall from user files/sandbox corpus; **CC** = Citation Coverage; **FA** = Factual Accuracy; **IF** = Instruction Following; **DQ** = Depth Quality; **Avg.** = Average.

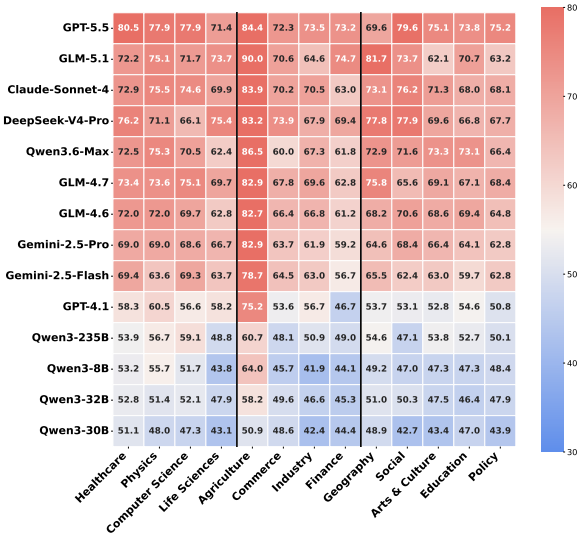


Figure 4: Performance on different domains.

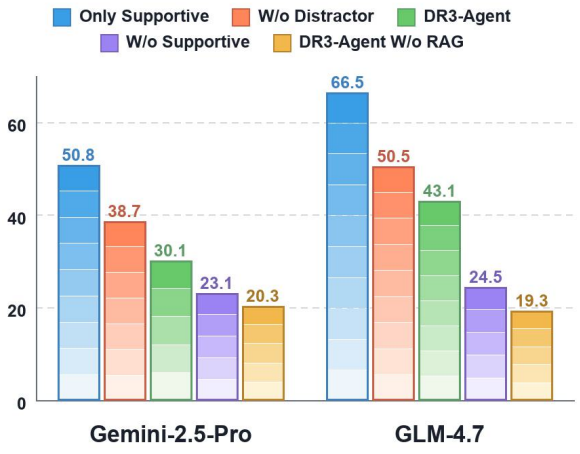


Figure 5: Analysis on the effect of sandbox corpus.

pendix H, which compares multimodal and text-only tasks and shows that multimodal inputs generally introduce additional difficulty, especially for information recall. (4) **Performance exhibits significant variance across different domains and different models**. In Figure 4, we observe that results from several domains (e.g., GLM 4.7 achieves the best on "Industry" domains while Claude Sonnet 4 achieves the best on "Physics").

5.3 Further Analysis

Correlation between sandbox corpus and real-world web corpus. We keep the agent architecture, and vary the external information source. In the real-time web search setting, the sandbox corpus is removed and the agent instead uses real-time web search to acquire evidence. We conduct experiments on all English tasks using Qwen3-235B-A22B and Gemini-2.5-Pro to verify whether the sandbox corpus can approximate information acquisition in real-world web environments. In Table 3, the overall performance remains close across the two settings, with Citation Coverage showing only a slight decrease because open-web retrieval can return alternative valid sources outside our predefined ground-truth citation set. This indicates that the core evidence chains ultimately relied upon by the models in the live web setting highly overlap with the supportive documents predefined in the sandbox, and the sandbox preserves the main information difficulty that determines task performance while enabling deterministic citation evaluation.

Effect of different sizes of sandbox corpus. We conduct evaluations on five sandbox corpora (i.e., 32k, 64k, 128k, 256k, and 512k tokens). As shown in Figure 6, as the size of the sandbox corpus increases, the overall performance (Avg.), IR_{SC}, and CC of all models show a clear downward trend. This indicates that the models find it harder to identify effective answers among the increasing noise

	Qwen3-235B-A22B			Gemini-2.5-Pro		
Metric	Baseline	w/ Web	Change (Δ)	Baseline	w/ Web	Change (Δ)
IR_{SC}	33.2	38.5	(+5.3)	40.4	41.9	(+1.5)
IR_{UF}	23.9	20.2	(-3.7)	27.4	25.4	(-2.0)
CC	36.3	28.0	(-8.3)	50.4	49.0	(-1.4)
FA	59.0	60.3	(+1.3)	76.3	75.9	(-0.4)
IF	73.6	79.2	(+5.6)	80.1	84.1	(+4.0)
DQ	63.8	62.0	(-1.8)	67.8	70.4	(+2.6)
Avg.	48.3	48.0	(-0.3)	57.1	57.8	(+0.7)

Table 3: Analysis on the correlation between sandbox corpus and real-world web corpus.

and distracting information. However, FA metric shows relative stability. We believe this mainly measures the model's inherent performance.

Comparison of different framework architectures. We compare DR^3-Agent with DeerFlow (ByteDance, 2025). Considering that DeerFlow's native retrieval mechanism cannot directly process DR^3-Eval, we transplant the Agentic RAG component from DR^3-Agent, without removing any of its original functions. In Figure 7, both frameworks exhibit converging capabilities in information acquisition within the sandbox corpus, and DR^3-Agent demonstrates distinct architectural advantages when tackling complex deep research tasks.

Effect of sandbox corpus. To verify the reasonableness of our sandbox corpus design, we systematically analyze the impact of different document components on model performance using English subset of the datasets. The experiments are mainly based on the 128k-sized corpus (except for the only supportive setting). In Figure 5, after removing the distractor web pages, the performance of all models improved significantly. This shows that our distractor documents effectively increased the task difficulty. Furthermore, we observe that the agent's performance is nearly identical when the sandbox corpus is provided without supportive documents relative to the no-corpus setting. This demonstrates that, apart from the designated supportive documents, our sandbox corpus contains no other effective information that the agent can exploit to complete the task. When only supportive web pages are present in the sandbox corpus, it establishes the model's performance upper bound under perfect information retrieval.

Correlation between LLM-as-judge and human evaluation. We conducted a correlation study on 50 reports randomly sampled across all domains, which were independently reviewed by four experts. Consistency was measured using Pearson and Spearman correlation coefficients (Han et al., 2026) alongside pairwise agreement (Du et al., 2025) , with details provided in Appendix I. As shown in Table 4, our automated scoring exhibits strong concurrence with expert evaluations. Furthermore, we evaluated the consistency between our automated claim extraction for factual accuracy and human annotations, achieving a Precision of 0.924 and a Recall of 0.960.

Method	r	ρ	**Agr.**
DR^3-Eval (Ours)	0.88	0.83	0.89
Inter-Human	0.93	0.86	0.92

Table 4: LLM-as-judge vs. human evaluation.

Model	**OpenAI-Emb**	**Qwen-Emb**	**BM25**
GLM-4.7	**56.58**	**53.61**	**50.71**
GPT-4.1	36.15	35.64	22.60
Gemini-2.5-Pro	49.51	37.16	31.25

Table 5: Effectiveness of different retrievers.

Effect of different judge LLMs. We select Claude Sonnet 4, Gemini-2.5-Pro, and Qwen-Max as alternatives to GPT-5.1 to re-rank a representative subset of six models. Relative to the rankings from GPT-5.1, the rankings derived from their scores are almost identical, with a mean Spearman's ρ of 0.924. Details are shown in Appendix J. Regarding the multimodal assistant judge LLM, Gemini-2.5-Pro, we replace it with Qwen3-VL-Plus and Moonshot Kimi-k2, yielding a mean Spearman's ρ of 0.864. The impact on the final scores is not significant, with an average difference of less than two points ($p > 0.05$)

Effect of maximum iteration turns in Agentic-RAG. Considering that our RAG sub-agent employs an Agentic-RAG, we set the maximum iteration turns to one, three, five, and seven. In Table 6, as the allowed iteration turns increase, the overall performance of different models exhibits a clear upward trend, particularly in IR and CC. However, we observe that once the iteration turns increase to a certain extent, the model's performance not only peaks but even experiences a slight decline.

Effect of different retrievers. To compare different retrieval strategies, we select three representative models under the 128k-token sandbox corpus. By default, we use OpenAI text-embedding-3-

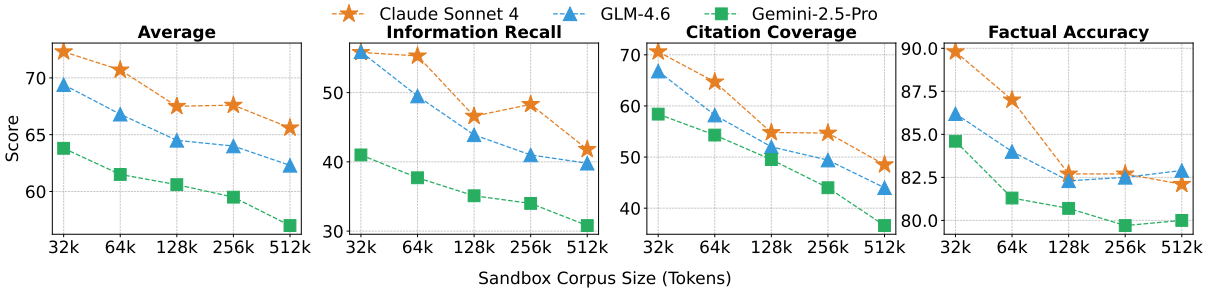


Figure 6: Analysis on the performance of different sizes of sandbox corpus.

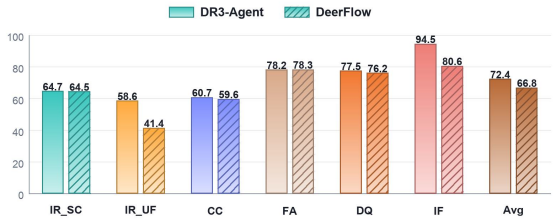


Figure 7: Comparison of framework architectures.

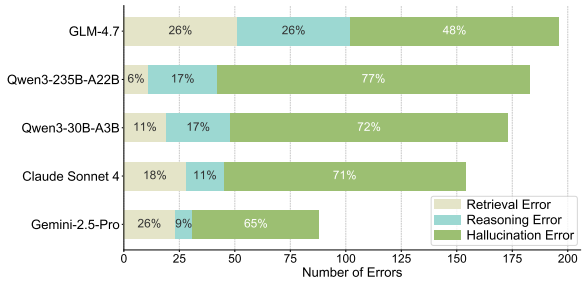


Figure 8: Error type analysis across LLMs.

	Qwen3-235B-A22B		Gemini-2.5-Pro	
Turns	**IR**	**CC**	**IR**	**CC**
1	27.2	14.8	32.4	21.0
3	34.7	27.1	39.6	47.6
5	33.9	27.1	**44.6**	**51.0**
7	**44.0**	**32.9**	38.1	48.1

Table 6: Effect of different maximum RAG turns.

small (OpenAI, 2024) and compare its results with Qwen-text-embedding-v2 (Alibaba Cloud, 2025) and BM25 (Robertson and Zaragoza, 2009). In Table 5, the text-embedding-3-small achieves the best performance on all models, while Qwen-text-embedding-v2 performs slightly lower and BM25 performs significantly worse.

Case Study. We conduct an error attribution analysis on five selected models, based on 100 reports per model, and classify the root causes into three categories: (1) **Retrieval Error**, denoting where the agent fails to locate or omits key information required to answer the question during the retrieval stage; (2) **Reasoning Error**, denoting where the agent, despite obtaining relevant information, makes mistakes in information integration, logical inference, or detail processing; and (3) **Hallucination Error**, denoting where the model's generated response is not based on the provided context but is instead fabricated from its parametric knowledge. In Figure 8, hallucination remains the primary cause of failure for most models. In contrast, the distributions of retrieval and reasoning errors vary across models.

Evaluation stability and significance. A 10,000-iteration bootstrap analysis shows no overlap in the 95% confidence intervals between two representative models selected for the pairwise comparison, with a Wilcoxon test ($p = 0.0046$) confirming a significant difference in their scores. Furthermore, the average total score (Avg.) variance across repeated evaluations is only 0.874, and the Kendall's τ and Spearman's ρ for model rankings under resampling reach 0.969 and 0.991, respectively. Taking Claude Sonnet 4, GLM-4.7, and GLM-4.6 as examples, the standard deviations across three repeated runs for each model remain exceptionally low at 0.83, 0.85, and 1.33, respectively.

6 Conclusion

We introduces DR^3-Eval, a benchmark designed to address key limitations in the evaluation of deep research agents. By grounding tasks in authentic user research scenarios, constructing controlled yet web-like sandbox corpus, and eliminating evaluation ambiguity through reverse task construction, DR^3-Eval provides a principled testbed for assessing long-horizon research capabilities. Our experimental results show that DR^3-Eval poses substantial challenges for state-of-the-art LLMs and reveals systematic failure modes of themr.

Limitations

The benchmark does not evaluate long-term maintenance issues such as corpus refresh schedules, version migration, storage costs, or indexing updates. These engineering factors are important for deployed evaluation platforms, but they are orthogonal to the core assessment of agent performance.

The task collection cannot exhaust the full range of user intents, languages, domains, expertise levels, or document formats encountered in real deployments. Privacy-oriented sanitization and manual curation may also remove ambiguity, noise, or contextual details that would appear in unconstrained user-provided materials.

Automatic evaluation of long-form reports remains imperfect, especially for analytical depth, nuanced factual support, and instruction satisfaction. Although human checks can reduce obvious errors, judge-model bias and residual annotation subjectivity may still affect scores and rankings.

Ethical Considerations

This paper presents DR^3-Eval, a benchmark designed to advance the evaluation of Deep Research Agents. Our work has several broader impacts and ethical considerations:

Data Privacy and Human Subjects: The dataset construction involved collecting authentic files from human participants. Participation was voluntary, contributors were compensated for their contributions, and consent and compensation records were stored separately from the benchmark data. We instructed contributors not to submit unnecessary personal, proprietary, or sensitive information, and implemented a two-stage sanitization protocol comprising automated redaction and manual cross-validation to remove or mask personally identifiable information (PII) and sensitive proprietary data before inclusion in the benchmark. Additional details on volunteer and annotator instructions are provided in Appendix K.

Societal Implications of Deep Research Agents: The advancement of autonomous research agents holds significant potential to increase productivity in knowledge-intensive fields. However, we acknowledge the risks associated with this technology, such as the potential for generating convincing hallucinations, over-reliance on unsupported claims, or the misuse of automated information gathering for malicious purposes. By introducing metrics specifically focused on *Factual Accuracy* and *Citation Coverage*, and by providing a static, reproducible sandbox environment, our work aims to steer the field towards developing safer, more reliable, and verifiable systems, rather than those that merely optimize for persuasiveness.

Environmental Impact: Furthermore, by utilizing a static sandbox corpus rather than relying on repetitive live-web crawling for every evaluation run, our framework promotes more computationally efficient and environmentally sustainable benchmarking practices.

References

Amirhossein Abaskohi, Tianyi Chen, Miguel Muñoz-Mármol, Curtis Fox, Amrutha Varshini Ramesh, Étienne Marcotte, Xing Han Lù, Nicolas Chapados, Spandana Gella, Christopher Pal, Alexandre Drouin, and Issam H. Laradji. 2025. Drbench: A realistic benchmark for enterprise deep research. *Preprint*, arXiv:2510.00172.

Perplexity AI. 2024. Perplexity pro research: From search to synthesis.

Alibaba Cloud. 2025. Text embedding. https://help.aliyun.com/zh/model-studio/embedding.

Anthropic. 2025. Claude Sonnet 4 Model Card. https://huggingface.co/CometAPI/Claude_Sonnet4.

ByteDance. 2025. Deerflow: An open-source long-horizon research agent framework. https://github.com/bytedance/deer-flow. Official GitHub repository, accessed April 15, 2026.

CAMEL-AI. 2026. Workforce. https://docs.camel-ai.org/key_modules/workforce. Official documentation page, accessed April 15, 2026.

Zhuo Chen, Xinyu Geng, Xinyu Wang, Yong Jiang, Zhen Zhang, Pengjun Xie, and Kewei Tu. 2026. Efficient multimodal planning agent for visual question-answering. *Preprint*, arXiv:2601.20676.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. 2025. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. *Preprint*, arXiv:2508.06600.

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

DeepResearch Team, Tongyi Lab. 2025. A new era of open-source ai researchers. https://tongyi-agent.github.io/blog/introducing-tongyi-deep-research/. Official blog post, accessed April 15, 2026.

DeepSeek-AI. 2026. DeepSeek-V4-Pro Model Card. https://huggingface.co/deepseek-ai/DeepSeek-V4-Pro.

Design Council. 2005. The eleven lessons: Managing design in eleven global brands. Technical report, Design Council. Introducing the Double Diamond design process model.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *Preprint*, arXiv:2506.11763.

Google. 2025. Gemini deep research: Advanced information synthesis and planning.

Google DeepMind. 2025a. Gemini 2.5 Flash Model Card. https://modelcards.withgoogle.com/assets/documents/gemini-2.5-flash.pdf.

Google DeepMind. 2025b. Gemini 2.5 Pro Model Card. https://huggingface.co/CometAPI/gemini2.5_pro_preview.

Janghoon Han, Heegyu Kim, Changho Lee, Dahm Lee, Min Hyung Park, Hosung Song, Stanley Jungkyu Choi, Moontae Lee, and Honglak Lee. 2025. Deer: A comprehensive and reliable benchmark for deep-research expert reports. *arXiv preprint arXiv:2512.17776*.

Janghoon Han, Heegyu Kim, Changho Lee, Dahm Lee, Min Hyung Park, Hosung Song, Stanley Jungkyu Choi, Moontae Lee, and Honglak Lee. 2026. Deer: A benchmark for evaluating deep research agents on expert report generation. *Preprint*, arXiv:2512.17776.

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

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. 2025. Webthinker: Empowering large reasoning models with deep research capability. *Preprint*, arXiv:2504.21776.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, Pan Zhang, Liangming Pan, Yu-Gang Jiang, Jiaqi Wang, Yixin Cao, and Aixin Sun. 2024. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Preprint*, arXiv:2407.01523.

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.

MiroMind Team and 1 others. 2025. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *Preprint*, arXiv:2511.11793.

MiroMindAI. 2025. MiroFlow: An open-source multi-agent framework for deep research. Available at https://github.com/MiroMindAI/MiroFlow.

OpenAI. 2024. New embedding models and api updates. https://openai.com/index/new-embedding-models-and-api-updates/.

OpenAI. 2025a. GPT-4.1 Model Card. https://openai.com/index/gpt-4-1/.

OpenAI. 2025b. GPT-5.1 Model Card. https://openai.com/zh-Hans-CN/gpt-5/.

OpenAI. 2025. Introducing deep research: A new paradigm for long-horizon reasoning.

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

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. *Preprint*, arXiv:2508.20033.

Long Phan, Alice Gatti, and 1 others. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Ronak Pradeep, Nandan Thakur, Shivani Upadhyay, Daniel Campos, Nick Craswell, Ian Soboroff, Hoa Trang Dang, and Jimmy Lin. 2025. The great nugget recall: Automating fact extraction and rag evaluation with large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 180–190.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, Rui Min, Minpeng Liao, Yong Jiang, Pengjun Xie, Fei Huang, and Jingren Zhou. 2025. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *Preprint*, arXiv:2509.13309.

Qwen Team. 2026. Qwen3.6-Max-Preview: Smarter, Sharper, Still Evolving. https://qwen.ai/blog?id=qwen3.6-max-preview.

Stephen Robertson and Hugo Zaragoza. 2009. *The Probabilistic Relevance Framework: BM25 and Beyond*, volume 4. Now Publishers Inc.

Jie Ruan, Inderjeet Nair, Shuyang Cao, Amy Liu, Sheza Munir, Micah Pollens-Dempsey, Tiffany Chiang, Lucy Kates, Nicholas David, Sihan Chen, and 1 others. 2025. Expertlongbench: Benchmarking language models on expert-level long-form generation tasks with structured checklists. *arXiv preprint arXiv:2506.01241*.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. 2025. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *Preprint*, arXiv:2511.07685.

Qwen Team. 2025. Qwen-deepresearch: Scaling reasoning for complex research tasks.

Tongyi Lab. 2025a. Qwen3-235B-A22B Model Card. https://huggingface.co/Qwen/Qwen3-235B-A22B.

Tongyi Lab. 2025b. Qwen3-30B-A3B Model Card. https://huggingface.co/Qwen/Qwen3-30B-A3B.

Tongyi Lab. 2025c. Qwen3-32B Model Card. https://huggingface.co/Qwen/Qwen3-32B.

Tongyi Lab. 2025d. Qwen3-8B Model Card. https://huggingface.co/Qwen/Qwen3-8B.

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. *Preprint*, arXiv:2510.14240.

Yibo Wang, Lei Wang, Yue Deng, Keming Wu, Yao Xiao, Huanjin Yao, Liwei Kang, Hai Ye, Yongcheng Jing, and Lidong Bing. 2026. Deepresearcheval: An automated framework for deep research task construction and agentic evaluation. *Preprint*, arXiv:2601.09688.

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. *arXiv preprint arXiv:2504.12516*.

xAI Team. 2024. Grok-1: An open large language model from xai. https://github.com/xai-org/grok-1.

Renjun Xu and Jingwen Peng. 2025. A comprehensive survey of deep research: Systems, methodologies, and applications. *Preprint*, arXiv:2506.12594.

Zhaorui Yang, Bo Pan, Han Wang, Yiyao Wang, Xingyu Liu, Luoxuan Weng, Yingchaojie Feng, Haozhe Feng, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025. Multimodal deepresearcher: Generating text-chart interleaved reports from scratch with agentic framework. *Preprint*, arXiv:2506.02454.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L McManus, Yuan Katsman, Junxian Chen, and Kelvin Guu. 2023. Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Zhipu AI. 2025a. GLM-4.6 Model Card. https://huggingface.co/zai-org/GLM-4.6.

Zhipu AI. 2025b. GLM-4.7. Model Card. https://huggingface.co/zai-org/GLM-4.7.

Zhipu AI. 2026. GLM-5.1 Model Card. https://huggingface.co/zai-org/GLM-5.1.

Anni Zou, Wenhao Yu, Hongming Zhang, Kaixin Ma, Deng Cai, Zhuosheng Zhang, Hai Zhao, and Dong Yu. 2024. Docbench: A benchmark for evaluating llm-based document reading systems. *Preprint*, arXiv:2407.10701.

A Prompts for Data Construction

This section outlines the prompts used to synthesize our dataset.

A.1 Search Terms Generation

Prompt for Search Terms Generation

You are a professional research assistant. Based on the following user files, generate search terms for web search.
User Files
files_summary
Task Requirements
Generate 10 Search Terms
Based on the content of the user files, generate 10 **English** search terms for web search. These search terms should:

- Be able to retrieve supplementary information related to the topic of the user files
- Cover different angles and aspects (such as background knowledge, case studies, policies and regulations, technical details, industry trends, etc.)
- **Each search term should be 3-5 words** - combining 2-3 related concepts naturally
- **Use common, widely-searched phrases** that are likely to return 100+ search results
- Search terms should be able to find **information not in the user files but helpful for understanding the topic**
- **All 10 terms must be ORTHOGONAL** - they should target completely different aspects and NOT overlap in search results

CRITICAL RULES for generating search terms that return many results:

1. **Balance specificity with searchability**
    - If the topic is about a specific country/region (e.g., Japan tax policy), include it
    - If the topic is general, prefer broader terms that apply globally
    - BAD: "Japan income tax bracket structure 2024" (too specific + year)
    - GOOD:"Japan income tax reform" or "income tax policy reform"
2. **Use organization names strategically**
    - Include organization names when they are central to the topic
    - Avoid obscure acronyms that few people search for
    - BAD: "HUD FMR calculation methodology" (obscure acronym)
    - GOOD: "HUD housing affordability" or "federal housing assistance"
3. **Prefer common terms over academic jargon**
    - Use terms that practitioners and general public would search
    - BAD: "arthrofibrosis prophylaxis protocols" (too technical)
    - GOOD: "joint stiffness prevention therapy"
4. **Use common search patterns**:
    - "[Topic] + [Aspect]": "smart agriculture technology"
    - "[Topic] + [Method]": "precision farming data analytics"
    - "[Topic] + [Trend]": "renewable energy market growth"
    - "[Topic] + [Impact]": "climate change economic effects"
    - "[Topic] + [Impact]": "climate change economic effects"

Good Examples (3-5 words, will return 100+ results):

- "smart agriculture technology development"
- "precision farming data analytics"
- "vaccine efficacy clinical research"
- "housing affordability policy reform"

- “pension system sustainability analysis"
- “renewable energy investment trends"
- “supply chain risk management"
- “Japan pension reform challenges" (specific country when relevant)
- “OECD education policy recommendations" (organization when relevant)

Bad Examples (too specific or obscure):

- “Japan corporate tax rate competitiveness analysis 2024" (too long, includes year)
- “Fair Market Rent determination methodology HUD" (obscure + too specific)
- “California DMV licensing procedures step by step" (too specific)
- “pension actuarial valuation discount rate assumptions" (too technical)

Language Requirements
Important: All output content (search_terms, topic_summary) must be in **English**.
Output Format
Please output in JSON format:
“‘json
"search_terms": [
"search term 1",
"search term 2",
"search term 3",
"search term 4",
"search term 5",
"search term 6",
"search term 7",
"search term 8",
"search term 9",
"search term 10"
],
"topic_summary": "Brief summary of the user files topic (1-2 sentences)"
“‘ Only output JSON, no other content.

A.2 Query Construction

Prompt for Query Generation

Task
Generate a natural user query based on user files, search results, and keyword info.
Core Goal: The generated query should naturally guide an AI agent to search for "relevant keywords" without directly exposing them.
User Files
chr(10).join(f’- name’ for name in user_file_names)
Relevant Keywords (what agent should search)
relevant_kw_text
Noise Keywords (what agent should NOT be guided to search)
noise_kw_text
Keyword Classification Reasoning
reasoning
Search Results (len(useful_search) total)
search_text
Query Design Requirements
Core Principles

1. **Concise**: Query must be SHORT (50-100 words), like a real user’s brief question, not verbose
2. **Natural**: Query should be from user’s perspective, like a real person would ask
3. **Guiding**: Query topic should naturally lead agent to search “relevant keywords", but don’t over-hint
4. **No Exposure**: Don’t directly use technical terms from keywords, use simple natural expressions
5. **Brief File Reference**: Query must briefly mention user files, like “based on my xxx file" or “see attachment"
6. **Cover All Results**: Query must be designed so ALL len(useful_search) search results are needed for a complete answer, even if each result is only used a little
7. **Use All Files**: Query must be designed so ALL len(user_file_names) user files

are needed for a complete answer, even if each file is only used a little

Design Approach

1. Analyze the common theme of relevant keywords
2. Design a SHORT natural query (50-100 words), don't over-describe background
3. Query should:
   - Be short and direct, like a casual question
   - Not contain technical jargon or hint-like words
   - Briefly mention user files

Examples If relevant keywords are: - "Three-distance method spatial layout modern pocket park design cases" - "Scattered perspective step-by-step scenery urban micro-renewal"
User file is:
- Suzhou_Garden_Design.pdf
✗BAD Query (too long, over-hinting):
"Based on my Suzhou Garden Design.pdf showing spatial layout techniques, help me research how these traditional garden concepts are applied in contemporary urban small green spaces, please find some recent practice cases"
✗BAD Query (directly exposes keywords):
"Please analyze the application cases of three-distance method in modern pocket parks"
✓GOOD Query (short, natural):
"Based on my Suzhou garden design file, help me find some similar style applications in modern small parks"
✓GOOD Query (short, natural):
"With the attached garden diagram, research how traditional garden techniques are practiced in urban small green spaces"
Output Format
Output JSON only:

"query": "Complete user query (50-100 words, short and natural, mentions user files)",
"file_references": ["User files referenced in query and their content description"],
"subtasks": [

"id": 1,
"description": "Subtask description",
"target_keywords": ["Relevant keywords this subtask should guide agent to search"]

],
"design_rationale":"Why this query guides agent to search relevant keywords",
"avoided_terms": ["Technical terms deliberately avoided in query"],
"natural_expressions": ["Natural expressions used to replace technical terms"]

Note: Output JSON only, no other content. Query must include clear reference to user files!

B Detailed breakdown of file types

Figure 9 illustrates the specific file format breakdown for the document and image categories.

C Report Example

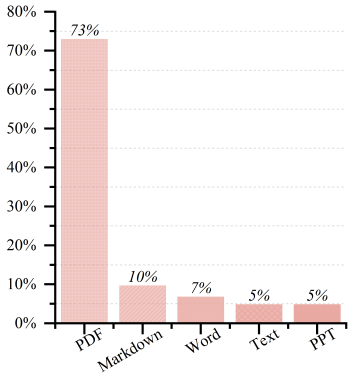
We present a representative task and the corresponding report generated by DR^3-Agent.

User Files:

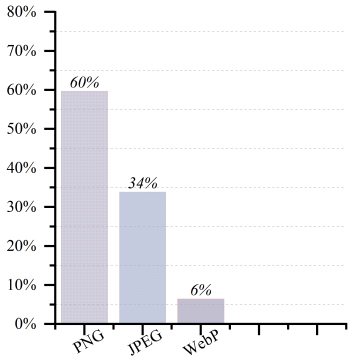
- **China.jpg** - China's high-speed rail network map showing extensive coverage and connectivity
- **Shinkansen.jpg** - Japan's Shinkansen network map displaying route structure and operational lines
- **China's high-speed rail.mp4** - Analysis of China's high-speed rail development narrative and technical achievements

Figure 10 displays the visual components of the user files provided in this case study

Query: I've attached a video about China's high-speed rail, along with maps of China's and Japan's networks. It's impressive how fast and extensive China's system has become. Could you summarize the story behind its rapid development? I'm curious about its evolution, how its core system compares to Japan's Shinkansen, the planning behind its vast



(a) document formats



(b) image formats

Figure 9: Breakdown of specific file formats for documents and images.

network, and what the next big steps in high-speed rail technology look like globally.

Claude Sonnet 4's Report

The Remarkable Story of China's High-Speed Rail Revolution: From Zero to Global Leader Executive Summary

China's high-speed rail development represents one of the most extraordinary infrastructure transformations in modern history. In just over three decades, China evolved from having no high-speed rail capability to operating the world's largest network, fundamentally reshaping both domestic transportation and global rail technology markets. This comprehensive analysis examines the evolution of China's high-speed rail system, compares it with Japan's pioneering Shinkansen network, explores the strategic planning behind China's vast infrastructure, and identifies the next generation of rail technologies that will define the future of high-speed transportation. **The Evolution:**

A Three-Decade Journey from Concept to Global Dominance

- **Early Foundations and Strategic Planning (1990-2007)** China's high-speed rail journey began with visionary planning in the early 1990s. In December 1990, the Ministry of Railways submitted the first proposal to build a high-speed railway between Beijing and Shanghai to the National People's Congress, marking the first time high-speed rail was proposed as an option for intercity transportation in China ["Transportation Planning Casebook/China High-Speed"]. This initial proposal set in motion a carefully orchestrated development strategy that would unfold over the following decades.

  The planning phase was methodical and comprehensive. In 1995, Premier Li Peng announced that preparatory work on the Beijing Shanghai HSR would begin in the 9th Five Year Plan (1996–2000), demonstrating the government's commitment to long-term infrastructure development ["Transportation Planning Casebook/China High-Speed"]. The State Council commissioned a feasibility study for the Beijing-Shanghai line in December 1994, establishing the technical and economic foundation for what would become the backbone of China's high-speed rail network ["Transportation Planning Casebook/China High-Speed"].

  The first operational breakthrough came in 1998 when the Guangzhou-Shenzhen line was electrified and, by

China.jpg

Shinkansen.jpg

China's high-speed rail.mp4

Figure 10: The view of user files.

introducing Swedish-made X 2000 trains, the service speed increased to 200 km/h [“Transportation Planning Casebook/China High-Speed”]. This project served as a crucial testing ground for high-speed rail operations in the Chinese context. On October 12, 2003, the Qinshen Railroad started operation as the first passenger-only railroad line in China, featuring the “China Star” electrified locomotive with a designed speed of 200 km/h and a top speed of 300 km/h [“Transportation Planning Casebook/China High-Speed”].

- **Strategic Technology Acquisition Phase (2004-2008)** Rather than attempting to develop high-speed rail technology entirely from scratch, China pursued a sophisticated strategy of technology acquisition and transfer. In June 2004, the Ministry of Railways accepted bids from Alstom of France, Bombardier Transportation of Germany, and Kawasaki of Japan for 200 high-speed train sets with top speed of 200 km/h, crucially including terms of technology transfer [“Transportation Planning Casebook/China High-Speed”]. This marked the beginning of systematic technology acquisition that would accelerate China’s development timeline by decades.

  The technology transfer relationships were comprehensive and strategic. CSR obtained Japanese high-speed technology starting in 2004 as part of a deal with Kawasaki, with the CRH2 series being directly derived from the E2 Series 1000 Shinkansen and nicknamed “Hairtail” [“Hexie (train)”]. From June to September 2005, the Ministry of Railways purchased train sets with a top speed of 350 km/h, demonstrating the rapid escalation in technical capabilities [“Transportation Planning Casebook/China High-Speed”].

  A pivotal decision came in 2006 when the State Council, in its Mid-to-Long Term Railway Development Plan, adopted conventional track HSR technology over maglev, enabling rapid construction of standard gauge, pas-

senger dedicated HSR lines throughout China ["Transportation Planning Casebook/China High-Speed"]. By April 2007, the high-speed rail system had upgraded to more than 423 km of tracks capable of supporting 250 km/h train service and 3,002 km of tracks capable of supporting 200 km/h as a result of the Speed Up Campaign ["Transportation Planning Casebook/China High-Speed"].

- **The Breakthrough Era (2008-2020)** The transformative moment arrived on August 1, 2008, when the Beijing-Tianjin Inter-city Railway was launched at a top speed of 350 kilometers per hour, marking the beginning of China's high-speed railway era [Video: China's high-speed rail.mp4]. This railway served as China's first independently built high-speed railway operating at 350 kilometers per hour ["On Aug. 1, 2008, the Beijing-Tianjin Inter-city Ra"]. The video analysis reveals that this achievement represented rapid development accomplished in just over 10 years, establishing China as a world leader in high-speed rail technology [Video: China's high-speed rail.mp4].

  The scale and pace of subsequent expansion was unprecedented in infrastructure history. When China's Premier Wen Jiabao boarded the first bullet train to open the service of the new high-speed line between Beijing and Shanghai on June 30, 2011, he described the step as a 'new chapter' in China's railway history ["When China's Premier Wen Jiabao boarded the fi"]. The 1,318km-long high-speed line connected China's two major metropolitan hubs in only four hours and 48 minutes, compared to about ten hours on the old railway ["When China's Premier Wen Jiabao boarded the fi"].

  The passenger adoption was remarkable. By the end of 2017, the total number of passengers who traveled on China's high-speed trains had reached 7 billion ["On Aug. 1, 2008, the Beijing-Tianjin Inter-city Ra"]. By 2020, China's total operation mileage of high-speed railway reached 30,000 kilometers, covering 80 percent of its major cities ["On Aug. 1, 2008, the Beijing-Tianjin Inter-city Ra"]. The World Bank described this as "the biggest rail expansion the world has ever seen" with China being "the only country that has ever attempted such an increase" ["When China's Premier Wen Jiabao boarded the fi"].

Comparative Analysis: China's System vs. Japan's Shinkansen

The comparison between China's high-speed rail network and Japan's Shinkansen reveals dramatic differences in scale, approach, and strategic objectives. The total operating mileage of China's high-speed railway network exceeds 25,000 kilometers, accounting for two thirds of the world's total ["On Aug.1, 2008, the Beijing-Tianjin Inter-city Ra"]. The China high-speed rail network map demonstrates extensive coverage that stretches from the Bohai Sea in the east to the Gobi Desert in the west, from the central plains to the mountains in the southwest, and from the snowfields in the north to the riverside towns in the south [Image: China.jpg].

The network analysis reveals dense connectivity in eastern and central regions, with key routes connecting major cities such as Beijing, Shanghai, Guangzhou, and Wuhan, forming an integrated system linking provincial capitals and major urban centers [Image: China.jpg]. The map shows a well-developed network with multiple intersecting lines and branch routes, demonstrating extensive coverage and connectivity as of November 27, 2009, with continued expansion since then [Image: China.jpg].

In contrast, Japan's Shinkansen network, while pioneering and highly efficient, covers approximately 2,800 km as of March

2025 [Image: Shinkansen.jpg]. The Japanese network shows a highly focused approach along Japan's main urban corridor, connecting major cities like Tokyo, Osaka, Nagoya, and Fukuoka through nine distinct lines: Tokaido, Sanyo, Kyushu, Tohoku, Joetsu, Hokuriku, Hokkaido, Akita, and Yamagata Shinkansen [Image: Shinkansen.jpg]. The coverage pattern is linear, following Japan's geographic constraints and population distribution, with most lines concentrated along the Pacific coast and major population centers [Image: Shinkansen.jpg]. **Strategic Planning Behind China's Vast Network**

China's network planning represents a revolution in infrastructure development methodology, incorporating cutting-edge technologies and comprehensive strategic thinking. Modern railway planning involves strategic oversight of all railway systems and related infrastructure, including track alignment, rail capacity, logistics planning, maintenance scheduling, safety checks, signalling and communication networks, track access, landscaping, and vegetation control ["What is railway planning? A brief overview"]. This comprehensive approach requires synchronizing human resources such as engineers, architects, urban planners, rail construction managers, transport officers, railway contractors and crews, as well as coordinating traffic, passengers, and pedestrian flow ["What is railway planning? A brief overview"]. Leaders of the best railway infrastructure worldwide are increasingly turning towards 3D virtual replicas or digital twins as a planning tool ["What is railway planning? A brief overview"]. Digital twin railway planning software generates intelligent, dynamic models of the rail network, creating a virtual representation of complex 3D environments, accurately mimicking the appearance and behaviour of real-world infrastructure ["What is railway planning? A brief overview"]. This technological approach addresses traditional inefficiencies, as an international survey by BearingPoint revealed that 81% of rail infrastructure managers felt traditional decision-making processes were too time-consuming, 89% believed these processes led to unnecessary duplication of effort, and 86% were overwhelmed by paperwork duties ["What is railway planning? A brief overview"].

Conclusion

China's high-speed rail development represents a remarkable transformation that redefined global transportation infrastructure within a single generation. The journey from initial proposals in 1990 to achieving world leadership status demonstrates the power of strategic planning, systematic technology acquisition, massive investment, and unwavering political commitment. The evolution followed a clear pattern: early planning and feasibility studies (1990-2003), strategic technology acquisition through foreign partnerships (2004-2008), rapid domestic expansion (2008-2020), and emergence as a global technology exporter (2010s-present).

The comparison with Japan's Shinkansen reveals two different but successful approaches to high-speed rail development. Japan pioneered the technology with careful engineering and operational efficiency, creating a reliable and profitable system that serves as a model for precision and safety. China leveraged this foundation to create an unprecedented scale of infrastructure that prioritizes national integration and social benefits alongside transportation efficiency. The strategic planning behind China's vast network demonstrates sophisticated integration of digital technologies, comprehensive infrastructure coordination, and long-term economic development objectives. The network serves not merely as transportation infrastructure but as a tool for national integration, regional development, and international influence. Looking toward the future, the next generation of high-speed rail technology promises even more dramatic advances. Maglev technology offers the

potential for speeds exceeding 600 km/h, while digital integration through IoT, AI, and cloud computing will revolutionize operational efficiency and passenger experience. China's continued investment in ultra-high speed conventional rail, maglev systems, and emerging technologies like hyperloop positions the country to maintain its leadership in this critical infrastructure sector. The global expansion of high-speed rail technology, driven by Chinese companies competing with established European and Japanese manufacturers, suggests that the next decade will see accelerated adoption worldwide. As the video demonstrates with China's CR450 prototype achieving 450 km/h with improved smoothness, the pursuit of "amazing China speed" continues to push the boundaries of what is possible in rail transportation, setting the stage for a new era of global mobility.

References
Video Sources:

- China's high-speed rail.mp4 - Analysis of China's high-speed rail development narrative and technical achievements

Image Sources:

- China.jpg - China's high-speed rail network map showing extensive coverage and connectivity
- Shinkansen.jpg - Japan's Shinkansen network map displaying route structure and operational lines

Database Sources:

- Transportation Planning Casebook/China High-Speed Rail development timeline and milestones
- Historical development of Beijing-Tianjin Inter-city Railway and network expansion
- Technology transfer relationships between China and international partners
- Railway planning methodologies and digital twin technologies
- Maglev technology development and global implementation
- Global high-speed rail market analysis and future projections
- Digital railways and IoT integration in transportation systems
- International expansion projects and technology export initiatives

D Evaluation Example

This section details the calculation of evaluation metrics, beginning with the assessment of insight coverage derived from user files and the sandbox corpus, followed by the computation of the remaining metrics. Tables 7–11 show the intermediate results for this evaluation example.

D.1 Information Recall from User Files and Sandbox Corpus

D.1.1 Information Recall from User Files

Golden insight from user files

1. China's high-speed rail network is dense, especially in the east.
2. The map shows the extensive network as of November 2023.
3. The network includes lines with speeds of 300 km/h or more.
4. Rail lines are color-coded by speed, from $<$ 200 to $\geq$ 300 km/h.
5. Map of Japan's Shinkansen lines as of March 2025.
6. Shows operational, planned, and under-construction routes.
7. A future Linear Chūō Shinkansen (maglev) line is projected.
8. The network connects major cities like Tokyo, Osaka, and Hakata.
9. Developed from non-existent to world-class in just over 10 years.

10. Current trains travel at world-leading speeds of 300-350 km/h.

11. The new CR450 EMU prototype is the world's fastest.

12. CR450 prototype reaches 450 km/h in tests.

D.1.2 Information Recall from Sandbox Corpus

Extracted insight from sandbox corpus:

1. Reducing aerodynamic resistance is crucial for faster trains.

2. Shinkansen's strengths are efficiency and passenger comfort.

3. China has an ambitious 2035 high-speed rail expansion plan.

4. Digital transformation is key to future rail network evolution.

5. Future rail relies on IoT, 5G, and AI technologies.

6. China plans to extend its HSR network to Southeast Asia.

D.2 Citation Coverage

D.3 Factual Accuracy

D.4 Format Compliance

D.5 Depth Quality

Score: 0.70

Justification: The report gives a solid, well-structured narrative of China's HSR evolution. It clearly explains the technology-transfer strategy, compares China's system to Japan's on scale, speed, and philosophy, and touches on planning tools and future technologies like maglev and digital railways.

However, much of the "planning" section drifts into generic global railway-planning and IoT discussion rather than concrete Chinese institutional planning choices (e.g., corridor selection, financing models, governance reforms, safety scandals). Furthermore, the comparison with Shinkansen lacks deeper critical analysis of safety records, cost overruns, demand risk, and long-term economic performance.

The future-tech section is descriptive and somewhat market-report-like, with limited critical assessment of feasibility, trade-offs, and timelines. Consequently, the analysis is adjudged as good but not yet truly deep or nuanced.

E Prompts for Evaluation

We utilize an LLM-as-a-Judge approach to score the model's responses

E.0.1 Information Recall

Prompt for Evaluation of Information Recall

You are an intelligent assistant that labels atomic insights based on whether they can be extracted from a given report.
Search Query/Task
self.query or "No specific query provided"
Insights to Label (Total: len(batch))
insights_list
Report/Passage
result_text
TASK: Label each insight as 1.0 (support), 0.5 (Partial Support), or 0.0 (Not Support) based on whether the insight can be extracted from the report.
TERNARY SCORING SYSTEM
[SCORE 1.0 - FULLY COVERED]
Definition: The report EXPLICITLY and COMPLETELY covers the insight's core meaning.
Requirements for 1.0 (ALL must be met):

- ✓The core information point is clearly stated in the report
- ✓No significant details are missing
- ✓The meaning is unambiguous and complete
- ✓A reader would understand the full insight from the report alone

Examples of 1.0:

- Insight: "In 2023, Shanghai achieved a waste classification coverage rate of 95%"

  Report: "The coverage rate of household waste classification in Shanghai in 2023 has reached over 95%" → 1.0 (exact match)

Number	Status	Evidence
1	Covered	The network analysis reveals dense connectivity in eastern and central regions, with key routes connecting major cities...
2	Half Covered	The map shows a well-developed network ... as of November 27, 2009, with continued expansion since then [Image: China.jpg].
3	Covered	The Beijing-Tianjin Inter-city Railway was launched at a top speed of 350 kilometers per hour." "CRH380A reaching up to 380 km/h...
4	Not Covered	Not Found
5	Covered	Japan's Shinkansen network ... covers approximately 2,800 km as of March 2025 [Image: Shinkansen.jpg].
6	Half Covered	Japan's Shinkansen network ... connecting major cities ... through nine distinct lines...
7	Half Covered	The experimental Japanese L0 Series maglev set a train speed record of 603 km/h (375 mph) in 2015...
8	Covered	connecting major cities like Tokyo, Osaka, Nagoya, and Fukuoka through nine distinct lines...
9	Covered	The video analysis reveals that this achievement represented rapid development accomplished in just over 10 years, establishing China as a world leader in high-speed rail technology.
10	Half Covered	Beijing-Tianjin Inter-city Railway ... top speed of 350 kilometers per hour." "CRH380A reaching up to 380 km/h." "Shinkansen ... up to 320 km/h...
11	Half Covered	the CR450 EMU prototype that reaches 450 km/h in testing, representing a significant advancement in speed capabilities
12	Covered	the CR450 EMU prototype that reaches 450 km/h in testing

Table 7: Evaluation of Information Recall from User Files.

Number	Status	Evidence
1	Not Covered	Not Found
2	Half Covered	Japan... has established a reliable and profitable system, serving as a paradigm of precision and safety.
3	Not Covered	The medium and long-term railway development plan from 2004 to 2020... plans to invest..." and so on, only up to the investment planning for 2025.
4	Covered	Digital twins... as planning tools... address the inefficiencies of traditional decision-making... Digital railways encompass real-time data collection, predictive modeling, automated decision-making, and more.
5	Covered	Leading this evolution are various IoT-related technologies, such as 5G, AI, and cloud computing... which play a pivotal role in digital logistics and transportation.
6	Half Covered	China's high-speed rail technology is going global... projects such as the Ankara-Istanbul high-speed rail, the Jakarta-Bandung high-speed rail, and the Moscow-Kazan high-speed rail.

Table 8: The Calculation of Information Recall from Sandbox Corpus.

- Insight: "EPNP algorithm for camera pose estimation"

  Report: "Using EPNP algorithm for camera pose estimation" → 1.0 (same meaning, different wording)

[SCORE 0.5 - PARTIALLY COVERED] Definition: The report contains RELATED content but is INCOMPLETE or VAGUE. Conditions for 0.5 (any of these):

1. Only part of the core meaning is covered (missing key details)
2. The topic is mentioned but specifics are absent
3. Related concept exists but not the exact point
4. Generalization without the specific insight
5. The connection requires inference (not explicit)

Examples of 0.5:

- Insight: "Shanghai's garbage classification coverage rate will reach 95% by 2023"

  Report: "Shanghai's garbage classification has achieved significant results"

No.	Source Title	Status
Web Page Coverage		
1	Japan's Shinkansen: How Does It Stack Up Worldwide?	**Cited**
2	The global rail transportation market was valued at US$ 724,180 million in 2022 and, by 2029, is pro	**Cited**
3	Expected Benefits of Next-generation High-speed Rail Introduction.	Missed
4	China's railway network: rapid development to support domestic mobility	Missed
5	A remarkable piece of engineering and efficiency, the Shanghai-Kunming line spans from the east coas	Missed
User File Coverage		
6	China.jpg	**Cited**
7	Shinkansen.jpg	**Cited**
8	China's high-speed rail.mp4	**Cited**

Table 9: Evaluation of Citation Coverage.

→ 0.5 (topic covered, but no specific percentage)

- Insight: "Germany adopts a dual track recycling system"

  Report: "Developed countries have mature recycling systems" → 0.5 (generalization, no specific country/system)

- Insight: "AI ethics need to establish accountability mechanisms"

  Report: "The development of AI needs to consider ethical issues" → 0.5 (related topic, but no specific mechanism)

[SCORE 0.0 - NOT COVERED]
Definition: The report does NOT contain the insight's information.
Conditions for 0.0:

- ✗The topic is completely absent from the report
- ✗Only unrelated keywords appear
- ✗The information contradicts the insight
- ✗No semantic connection exists

Examples of 0.0:

- Insight: "Singapore's waste incineration power generation technology"

  Report: (never mentions Singapore) → 0.0

- Insight: "The specific content of the 666 training plan"

  Report: (never mentions "666" or this training plan) → 0.0

THREE-STEP JUDGMENT PROCESS
For EACH insight, follow these steps:
STEP 1: Extract Core Meaning

- Identify the KEY information points (usually 1-3)
- Note any specific details (numbers, names, dates, methods)
- Example: "Shanghai's garbage classification coverage rate will reach 95% by 2023" → Core: [Shanghai, 2023, garbage classification, coverage rate, 95%]

STEP 2: Search Report for Coverage

- Look for exact matches first
- Then look for semantic equivalents
- Note what IS found and what is MISSING

STEP 3: Apply Scoring Rules

- If ALL core points are explicitly covered → 1.0
- If SOME core points are covered OR topic is mentioned without specifics → 0.5

No.	Claim to be verified	Status
1	In December 1990, the Ministry of Railways submitted the first proposal to build a high-speed railway...	✓
2	In 1995, Premier Li Peng announced that preparatory work on the Beijing Shanghai HSR would begin in...	✓
3	The State Council commissioned a feasibility study for the Beijing-Shanghai line in December 1994, e...	✓
4	The first operational breakthrough came in 1998 when the Guangzhou-Shenzhen line was electrified and...	✓
5	On October 12, 2003, the Qinshen Railroad started operation as the first passenger-only railroad line...	✓
6	In June 2004, the Ministry of Railways accepted bids from Alstom of France, Bombardier Transportation...	✓
7	CSR obtained Japanese high-speed technology starting in 2004 as part of a deal with Kawasaki, with t...	✓
8	From June to September 2005, the Ministry of Railways purchased train sets with a top speed of 350 k...	✓
	⋮	
48	Planned investments in 2025 alone total 590 billion yuan (approximately $80.8 billion) to develop an...	✗
49	The expansion of China's rail network benefits regions in the interior of the country, with provinces...	✗
	⋮	
72	As the video demonstrates with China's CR450 prototype achieving 450 km/h with improved smoothness,...	✓

Table 10: The Evaluation of Factual Accuracy.

- If NO core points are found → 0.0

SEMANTIC EQUIVALENCE GUIDELINES
✓The following are considered SEMANTICALLY EQUIVALENT (score 1.0):
a) Synonym Substitution:

- "Law" ≈ "Regulations" ≈ "Regulations" ≈ "Provisions"
- "Significant results" ≈ "Significant effects" ≈ "Success achieved"
- "Diversified technology" ≈ "Multiple technologies" ≈"Rich variety of technologies"

b) Hypernym/Hyponym:

- "EPNP algorithm" → "PnP algorithm"(EPNP is a type of PnP) → 0.5 (too general)
- "Germany, Japan" → "Developed Countries "/" International Experience" → 0.5 (generalization)
- "Gaode Map" → "Navigation Application"(if context is clear) → 0.5 (less specific)

c) Paraphrasing :

- ··Establish AI Ethics Review Committee" ≈ "Establish AI Ethics Committee for Review" → 1.0
- ··Six ethical design details" ≈ "Six major ethical issues" → 1.0 (if all 6 are covered)

d) Information Fusion:

- Insight content is integrated into a larger discussion
- Core information points are preserved with different expression

✗The following are NOT semantically equivalent (score 0.0 or 0.5):
a) Completely Different Topics:

- Insight discusses topic A, report discusses topic B
- No semantic overlap whatsoever

b) Keyword Overlap Only:

- Same words appear but with completely different meanings
- Example: Insight says "Apple company", report says "apple fruit"

c) Missing Information:

No.	Requirement	Satisfied
1	Mention China's high-speed rail rapid development	✓
2	Describe early stages of China's high-speed rail	✓
3	Describe key phases in China's network expansion	✓
4	Explain main drivers of China's rapid rail growth	✓
5	Mention role of national planning in China	✓
6	Describe overall layout of China's high-speed network	✓
7	Explain planning logic behind China's network structure	✓
8	Describe core technical system of China's high-speed rail	✓
9	Describe core technical system of Japan's Shinkansen	✓
10	Compare technical features of China rail and Shinkansen	✓
11	Compare operational characteristics of China rail and Shinkansen	✓
12	Explain key differences between China rail and Shinkansen	✓
13	Mention future global high-speed rail technology trends	✓
14	Describe next big technological steps in high-speed rail	✓
15	Explain global implications of upcoming high-speed rail advances	✓

Table 11: The Evaluation of Format Compliance.

- Core information points from insight are completely absent in report
- Example: Insight mentions specific number "666", report never mentions it

d) Contradictory Information:

- Report information contradicts the insight
- Example: Insight says "successful", report says "failed"

BOUNDARY CASE HANDLING
When uncertain, follow this priority:

1. If >50% of core meaning is covered → 1.0
2. If reasonable semantic connection exists → 1.0
3. If only weak connection or keyword overlap → 0.5
4. If no connection at all → 0.0

Principle: Prefer false positives over false negatives (The goal of recall assessment is to check if information is missing)
RESPONSE FORMAT
Respond ONLY with valid JSON (no markdown, no extra text):
"results": [
"id": 1,
"core_points": ["point1", "point2"],
"found_in_report": "[quote or describe what was found]",
"missing_points": ["what is missing, if any"],
"score": 1.0,
"reasoning": "[brief reason in Chinese, <30 words]"
,
"id": 2,
"core_points": ["point1", "point2"],
"found_in_report": "Related content but incomplete",
"missing_points": ["specific detail X"],
"score": 0.5,
"reasoning": "[brief reason in Chinese, <30 words]"
,
"id": 3,
"core_points": ["point1"],
"found_in_report": "NOT FOUND",
"missing_points": ["all points missing"],
"score": 0.0,
"reasoning": "[brief reason in Chinese, <30 words]"
]
You MUST provide exactly len(batch) results. Use ONLY scores 1.0, 0.5, or 0.0.

E.0.2 Factual Accuracy

Prompt for Evaluation of Factual Accuracy

Task
You are a fact-checking expert. Please verify whether the following statement is **relevant and not contradictory** to the content of the source document.
Important principle: Maintain the stability of the evaluation, avoid fluctuations, and adopt a relatively lenient evaluation standard. Only consider a result as false if there are obvious errors.
Statements to be verified
claims_text
Validation Criteria
Cases where the determination is supported: true
As long as **any one of the following conditions** is met, it is determined to be true:

1. **Any part** of the statement has relevant content in the source document
2. A statement is a **generalization, summary, inference, or extension** of the content of the source document
3. The statements use different wording, but **have similar semantics**
4. The statement contains **implicit information** from the source document
5. For images/videos: The content described may be visually visible or inferable
6. The statement is a **reasonable interpretation** of the content of the source document, even if it is not the only interpretation
7. The source document contains **partially supporting** content for this statement

Situations where it is determined as supported:
false (limited to the following situations)
Only when one of the following conditions is met, it is determined as false:

1. Statements that are **directly contradictory** to the source document (such as significant errors in numbers or completely opposite facts)
2. The source document **completely lacks** any relevant content stated
3. The statement cannot be reasonably inferred from the source document

Judgment principles

- Allow **substantial generalization and inference**
- Allow **wording differences** and **different ways of expression**
- Allow **partially correct** statements (as long as they are not completely wrong)
- For situations that are **ambiguous or uncertain**, they should all be determined as true
- For multimodal content, **subjective description** and **reasonable imagination** are allowed

Example

- Source document: "The company's revenue increased by 25% in 2023"
  - "The company's revenue has seen a significant increase" → true (generalization)
  - "The company's performance is good" → true (inferred)
  - "The company's revenue in 2023 increased by approximately 25%." → true (approximately)
  - "The company's revenue increased by about 55% in 2023" → false (not true)
  - "The company's revenue in 2023 decreased by 25%" → false (direct contradiction)
- Source document: A product image
  - "The product's exterior design is exquisite" → true (subjective description)
  - "The product adopts a modern design style" → true (reasonable inference)

Output format (strict JSON)
"results": ["id": 1, "supported": true, "explanation": "Brief explanation"]
Please start the verification:

E.0.3 Instruction Following

Prompt for Evaluation of Instruction Following

Check if the report satisfies each requirement. Output ONLY true/false for each item.
Requirements (batch_size items)
checklist_list
Report
result_text
Rules

- true: Report clearly addresses this requirement with specific content
- false: Not mentioned, or only briefly/-superficially mentioned

Response Format (JSON only, no explanation needed)
"results": ["", ".join([f'"id": i+1, "satisfied": true/false' for i in range(batch size)])]

E.0.4 Depth Quality

Prompt for Evaluation of Depth Quality

You are an expert Research Report Evaluator.
You are given:

1. A research report
2. A research question that the report attempts to answer

<research_question>
Question
</research_question>
<Report>
result_text
</Report>
Instructions:
ANALYZE THOROUGHLY: Examine the report in detail and identify any issues, even small ones. Look for subtle problems, minor inconsistencies, areas that could be improved, or any shortcomings that might affect the quality.
Evaluate the report according to the **Depth & Quality of Analysis** criterion. Provide:

- A **score between 1 and 10** (must be an integer) using the scale defined below.
- A **detailed justification** (2–3 sentences) in **simple plain English** explaining why you gave that score, including any specific issues or strengths you identified.

General Scoring Scale (1-10, integers only):

- **1** = Completely Unusable: No meaningful content, completely off-topic, or gibberish
- **2** = Severely Deficient: Major fundamental flaws, almost no useful information
- **3** = Very Poor: Significant problems throughout, barely addresses the topic
- **4** = Poor: Multiple major issues, lacks coherence or depth, many gaps
- **5** = Below Average: Notable weaknesses, incomplete coverage, superficial analysis
- **6** = Adequate: Meets minimum requirements but with clear limitations and room for improvement
- **7** = Good: Solid work that addresses the topic well, minor issues present
- **8** = Very Good: High quality with comprehensive coverage, few minor issues
- **9** = Excellent: Outstanding work with deep insights, nearly flawless execution
- **10** = Perfect: Exceptional, publication-ready quality with innovative insights

CRITICAL EVALUATION REQUIREMENTS:

1. **Use the FULL scoring range**: Distribute scores across 1-10 based on actual quality differences. Do NOT cluster scores in a narrow range.
2. **Differentiate clearly**: A mediocre report should score 4-5, a good report 6-7, an excellent report 8-9. Only truly exceptional work deserves 10.
3. **Be discriminating**: Look for specific quality differences between reports. Better analysis, clearer structure, and deeper insights should result in higher scores.
4. **Penalize appropriately**: Minor issues = small deductions (0.5-1 point), major issues = significant deductions (2-3 points).
5. **Reward excellence**: If a report demonstrates exceptional depth, clarity, or insight, give it the high score it deserves.
6. **Compare mentally**: Consider how this report compares to the best and worst possible reports on this topic.

Evaluation Criterion: Depth & Quality of Analysis
Evaluate how thoroughly the report analyzes the research question. **BE HARSH**: Look for superficiality, missing details, lack of evidence, weak reasoning.

- **1-2**: Completely superficial, no real analysis, just lists facts
- **3-4**: Very basic analysis, misses most key factors, no depth at all
- **5**: Basic analysis present but very incomplete, misses many important aspects
- **6**: Covers main points but analysis is shallow, lacks nuance and deeper insights
- **7**: Adequate analysis with some depth, but still missing sophistication or critical thinking
- **8**: Good depth with multiple factors explored, shows some sophisticated reasoning (RARE)
- **9**: Exceptional depth with nuanced understanding and insights (VERY RARE)
- **10**: Perfect analysis with groundbreaking insights and comprehensive understanding (ALMOST NEVER)

Output format:
<evaluation>
<depth_quality>
<score>
1–10 (integer only, based on the scoring scale above)
</score>
<justification>
Give a detailed 2–3 sentence justification for your score in simple plain English, including specific issues or strengths.
</justification>
</depth_quality>
</evaluation>

F Prompts for Evaluation Preparation

F.1 Insights Extraction from User Files

Prompt for Insights Extraction from User Files

Please extract **Key Insights** relevant to the query from the following web search results.
Query Task
query
Important Notes (Web Search Results)
You are processing web search results. These sources contain key information needed to answer the query. Please extract **the core information provided by each source**.
Core Principle
Source Specificity: Extract information that is **unique and specific to each web page**, NOT common knowledge or generic statements.
Extraction Requirements

1. **Source Specificity (MOST IMPORTANT)**:
   - ✓Extract **specific information unique to this source**, such as: specific names, dates, data, events, product names, company names, project names, method names, etc.
   - ✓Extract **unique viewpoints, conclusions, methods, or findings** from this source
   - ✗**DO NOT** extract common knowledge, such as: "AI can improve efficiency", "data analysis is important", "user experience matters"
   - ✗**DO NOT** extract generic descriptions, such as: "this method works well", "performance improved", "solved the problem"
2. **Source Contribution**: Extract the main contribution of each source to answering the query, such as:
   - Methods/techniques/concepts introduced by the source (e.g., "proposes AfME embedding", "uses MCMC optimization")
   - Core topics or problems discussed by the source
   - Key conclusions or findings of the source
   - **Note**: No need to extract precise numbers (e.g., "19.86% improvement"), just extract core concepts (e.g., "significant PSNR improvement")
3. **Relevance**: Information is relevant to the query task
4. **Verifiability**: Can determine whether the report mentions this information (semantic similarity is sufficient, exact match not required)
5. **Atomicity**: Each insight must be atomic, containing only 1-12 words, expressing a simple fact or concept

Examples of Common Knowledge (DO NOT Extract)

- "Machine learning requires large amounts of data" → This is common knowledge
- "User experience is important" → This is common knowledge
- "This method improved accuracy" → Too vague, no specific value or comparison
- "Further research is needed" → This is a generic conclusion

Examples of Source-Specific Content (SHOULD Extract)

- "Uses ResNet-50 as backbone network" → Specific technical choice
- "Achieves 92.3% accuracy on ImageNet" → Specific data
- "Proposed by John Smith in January 2024" → Specific name and time
- "Employs AfME embedding method" → Specific method name

Quantity Control

- **Extract 1-2 core insights per source (web page)**
- Ensure each source has corresponding insights
- Prioritize conceptual information over precise numbers
- If a source lacks specific information, extract fewer insights rather than extracting common knowledge

Language Requirement
Important: Output insights must be in English.
Output Format
Please output insights in JSON array format, each insight should include the source title:
"'json
[

"insight": "Brief source-specific information point (1-12 words)", "source": "Source Title"
]
```
Only output the JSON array, no other content.
content_text

### F.2 Insights Extraction from Sandbox Corpus

**Prompt for Insights Extraction from Sandbox Corpus**

Please extract **Key Insights** relevant to the query from the following web search results.
**Query Task**
query
**Important Notes (Web Search Results)**
You are processing web search results. These sources contain key information needed to answer the query. Please extract **the core information provided by each source**.
**Core Principle**
**Source Specificity**: Extract information that is **unique and specific to each web page**, NOT common knowledge or generic statements.
**Extraction Requirements**

1. **Source Specificity (MOST IMPORTANT)**:
   - ✓Extract **specific information unique to this source**, such as: specific names, dates, data, events, product names, company names, project names, method names, etc.
   - ✓Extract **unique viewpoints, conclusions, methods, or findings** from this source
   - ✗**DO NOT** extract common knowledge, such as: "AI can improve efficiency", "data analysis is important", "user experience matters"
   - ✗**DO NOT** extract generic descriptions, such as: "this method works well", "performance improved", "solved the problem"
2. **Source Contribution**: Extract the main contribution of each source to answering the query, such as:
   - Methods/techniques/concepts introduced by the source (e.g., "proposes AfME embedding", "uses MCMC optimization")
   - Core topics or problems discussed by the source
   - Key conclusions or findings of the source
   - **Note**: No need to extract precise numbers (e.g., "19.86% improvement"), just extract core concepts (e.g., "significant PSNR improvement")
3. **Relevance**: Information is relevant to the query task
4. **Verifiability**: Can determine whether the report mentions this information (semantic similarity is sufficient, exact match not required)
5. **Atomicity**: Each insight must be atomic, containing only 1-12 words, expressing a simple fact or concept

**Examples of Common Knowledge (DO NOT Extract)**

- "Machine learning requires large amounts of data" → This is common knowledge
- "User experience is important" → This is common knowledge
- "This method improved accuracy" → Too vague, no specific value or comparison
- "Further research is needed" → This is a generic conclusion

**Examples of Source-Specific Content (SHOULD Extract)**

- "Uses ResNet-50 as backbone network" → Specific technical choice
```

- “Achieves 92.3% accuracy on ImageNet" → Specific data
- “Proposed by John Smith in January 2024" → Specific name and time
- “Employs AfME embedding method" → Specific method name

Quantity Control

- **Extract 1-2 core insights per source (web page)**
- Ensure each source has corresponding insights
- Prioritize conceptual information over precise numbers
- If a source lacks specific information, extract fewer insights rather than extracting common knowledge

Language Requirement
Important: Output insights must be in English.
Output Format
Please output insights in JSON array format, each insight should include the source title:
“‘json
[“insight": “Brief source-specific information point (1-12 words)",
“source": “Source Title"
]
“‘
Only output the JSON array, no other content.
content_text

F.3 Checklist Generation

Prompt for Checklist Generation

Extract atomic evaluation requirements from the following query. Each requirement must be short, specific, and independently verifiable.
Query:
query_text
Rules:

1. **Atomic Decomposition**: Break down complex requirements into minimal, independent checkpoints
   - Each requirement checks only one specific point
   - Example: “Analyze aspects A, B, and C" → Split into “Mention A", “Mention B", “Mention C"
   - Example: “Compare X and Y" → Split into “Describe X", “Describe Y", “Explain differences"
2. **Short and Clear**: Each requirement should be one sentence, no more than 10 words
   - Bad: “Based on the uploaded innovation theory article and e-commerce comparison chart, analyze Temu’s business model"
   - Good: “Mention innovation theory", “Cite e-commerce data", “Analyze Temu model"
3. **Independently Verifiable**: Each requirement can be judged individually (yes/no)
   - Bad: “Deeply analyze multiple dimensions of the business model"
   - Good: “Mention value proposition", “Mention revenue sources", “Mention cost structure"
4. **Cover Implicit Requirements**: Extract implicit checkpoints from the query
   - If “compare" is mentioned → Need “Describe A", “Describe B", “Explain differences"
   - If “determine whether" is mentioned → Need “Provide evidence", “Draw conclusion"
   - If “refer to article" is mentioned → Need “Cite article content"
5. **Quantity Target**: A query should typically have 8-15 atomic requirements
   - Simple query: 8-10 items
   - Medium query: 10-12 items
   - Complex query: 12-15 items
   - Keep moderate granularity, not too few or too many

6. **Categories**:
   - content: Mention a specific content point
   - evidence: Cite a material/data source
   - analysis: Perform some analysis
   - comparison: Make a comparison
   - conclusion: Draw a conclusion
7. **Format Requirements**:
   - Start with verbs: "Mention", "Cite", "Analyze", "Compare", "Explain", "Describe"
   - Avoid compound sentences, one thing per item
   - Each item no more than 10 words

Output JSON:
"checklist": [
"id": 1, "requirement": "Mention X", "category": "content",
"id": 2, "requirement": "Cite Y data", "category": "evidence",
"id": 3, "requirement": "Analyze Z features", "category": "analysis",
...
]

G Inference and Evaluation Cost

We report the approximate runtime and API cost of the full pipeline.

Setting	Avg. time / task	Approx. API cost / task
Inference	300–400 s	\$0.3–1.0\$
Evaluation	90 s	\$0.10\$

Table 12: Approximate runtime and API cost of the full pipeline.

H Modality Performance Breakdown

Table 13 reports the modality-level performance statistics, comparing factual accuracy and information recall on multimodal and text-only tasks.

Model	FA-M	FA-T	IR-M	IR-T
GLM-4.7	72.5	81.1	0.664	0.689
Gemini-2.5-Pro	86.0	83.3	0.525	0.551
Claude Sonnet 4	89.8	86.8	0.617	0.722

Table 13: Modality performance breakdown. FA-M and FA-T denote factual accuracy on multimodal and text-only tasks, respectively; IR-M and IR-T denote information recall on multimodal and text-only tasks, respectively.

I Human Evaluation

For human evaluation, we randomly sample 50 generated reports across domains and models. Each report is independently reviewed by four expert annotators using the same five evaluation dimensions as the automatic evaluator: Information Recall, Factual Accuracy, Citation Coverage, Instruction Following, and Depth Quality. Annotators are provided with the task query, user-provided files, sandbox evidence, and the generated report, and are instructed to score only based on the provided materials and evaluation rubric.

For each report, we average the four annotators' scores to obtain the human score. We then compare the automatic score and the averaged human score using the following metrics:

- **Pearson's correlation coefficient.**

$$r = \frac{\sum_{i=1}^{n}(A_i - A)(B_i - B)}{\sqrt{\sum_{i=1}^{n}(A_i - A)^2}\sqrt{\sum_{i=1}^{n}(B_i - B)^2}}$$

- **Spearman's correlation coefficient.**

$$s = \frac{\sum_{i=1}^{n}(R(A_i) - R_A)(R(B_i) - R_B)}{\sqrt{\sum_{i=1}^{n}(R(A_i) - R_A)^2}\sqrt{\sum_{i=1}^{n}(R(B_i) - R_B)^2}}$$

- **Pairwise agreement.**

$$PAR = \frac{1}{\binom{n}{2}} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \mathbb{I}_{ij}$$

$$\mathbb{I}_{ij} = \begin{cases} 1 & \text{if } (A_i - A_j) \cdot (B_i - B_j) > 0 \\ 0 & \text{otherwise} \end{cases}$$

where A_i,B_i represent the automated total score and human total score for a single case, respectively.

Table 14 presents the fine-grained scores of the first five cases across various dimensions.

J Model Rankings across Different Judge LLMs

Table 15 compares the leaderboard rankings produced by GPT-5.1 against alternative judges (Gemini-2.5-Pro and Qwen-Max) to quantify the consistency of our evaluator selection.

Case ID	LLM-Judge Score					Human-Judge Score				
	IR_{llm}	FA_{llm}	DQ_{llm}	IF_{llm}	CC_{llm}	IR_{human}	FA_{human}	DQ_{human}	IF_{human}	CC_{human}
001	30.00	46.23	70.00	93.33	66.67	32.50	42.76	72.00	92.42	66.67
002	40.27	50.38	70.00	86.67	66.67	38.68	53.89	68.00	83.25	66.67
003	42.85	52.25	70.00	64.00	16.67	40.78	56.65	66.86	62.38	16.67
004	30.00	43.84	60.00	77.78	50.00	31.50	46.29	58.76	75.94	50.00
005	49.29	60.33	70.00	84.00	71.43	46.73	57.83	73.56	82.37	71.43

Table 14: Comparison of LLM-Judge and Human-Judge scores across the first five cases.

Model	**GPT-5**	**Gemini-2.5-Pro**	**Qwen-Max**
GLM-4.7	1	1	1
Claude Sonnet 4	2	3	2
Gemini-2.5-Pro	3	2	3
GPT-4.1	4	5	5
Qwen3-235B-A22B	5	4	4
Qwen3-30B-A3B	6	6	6
Disagreement (Δ)	–	1	1

Table 15: Comparison of rankings assigned by three different judge models. The "Disagreement" row indicates the number of rank swaps relative to GPT-5.1 .

K Volunteer and Annotator Instructions

This appendix summarizes the instructions given to volunteers who contributed source materials and to annotators who processed the collected data. Volunteers were recruited through voluntary calls among undergraduate and graduate student communities and were informed that their materials would be anonymized and used to construct research-style evaluation tasks, that participation was voluntary, and that they should only submit files for which they had the right to share research-use copies. They were asked to provide realistic material sets that could naturally motivate a deep research request, such as reports, spreadsheets, images, web pages, or other documents from non-sensitive domains. They were explicitly instructed not to include unnecessary personal information, confidential institutional records, private communications, financial identifiers, medical records, or other sensitive content. When a potentially useful file contained incidental personal or proprietary information, volunteers were asked to either remove it before submission or flag it for sanitization.

For each submitted material set, volunteers also provided a short description of the intended research need, the expected user context, and any files that should be treated as essential evidence. This information was used only to guide task construction and was not released as participant metadata. All participants provided consent for this anonymized research use and were compensated for the expected time required to prepare and submit materials. Consent and compensation records were stored separately from the benchmark data.

Annotators then screened the submitted materials before benchmark construction. Their instructions were to verify that each material set supported a meaningful long-form research task, remove or mask personally identifiable information, discard files with unresolved privacy or ownership concerns, and check that the remaining evidence was sufficient for constructing a verifiable answer path. Annotators also marked the modality and file type of each item, identified candidate supportive documents, and flagged ambiguous, duplicated, corrupted, or low-quality files for exclusion. A second annotator cross-checked the sanitized materials before they were used to build the sandbox corpus.

For human evaluation, annotators were asked to score generated reports using the same five dimensions as the automatic evaluator: Information Recall, Factual Accuracy, Citation Coverage, Instruction Following, and Depth Quality. They were instructed to rely only on the provided user files, sandbox evidence, and task specification, rather than their own background knowledge. Disagreements or uncertain cases were recorded for adjudication, and aggregate agreement statistics are reported in Section I.