TVIR: Building Deep Research Agents Towards Text–Visual Interleaved Report Generation

Anonymous ACL submission

Abstract

Deep Research Agents have shown strong capability in multi-step information retrieval, reasoning, and long-form report generation, but existing benchmarks and systems remain predominantly text-centric, with limited evaluation of whether visual elements are factually reliable and well aligned with the surrounding analysis. To address this gap, we introduce **TVIR** (**T**ext–**V**isual **I**nterleaved **R**eport **G**eneration), which includes **TVIR-BENCH**, a benchmark of 100 expert-curated multimodal deep research tasks that require visual elements to serve specific analytical sub-goals, and **TVIR-AGENT**, a hierarchical multi-agent framework for constructing structured outlines, retrieving images, generating charts with traceable sources, and composing reports through context-aware sequential writing. We further develop a dual-path evaluation framework that combines **Textual Assessment** and **Visual Assessment**. Experiments across nine deep research systems show that TVIR-AGENT achieves strong overall performance, underscoring the importance of explicit multimodal design and evaluation for evidence-driven report generation.

1 Introduction

Recent advances in large language models (LLMs) (OpenAI et al., 2024; DeepSeek-AI et al., 2025) have catalyzed the emergence of Deep Research Agents (DRAs) (Zheng et al., 2025; Coelho et al., 2025; Cai et al., 2026; Tongyi DeepResearch Team et al., 2025), which aim to autonomously conduct multi-step information retrieval, reasoning, and generation to produce comprehensive research reports. These systems have shown promising capability in long-context planning, citation grounding, and analytical writing, positioning them as potential assistants for complex professional decision-making in domains such as policy analysis, finance, and scientific research. As a result, a growing body of benchmarks and systems (Shi et al., 2025) has focused on evaluating end-to-end deep research workflows rather than isolated questions.

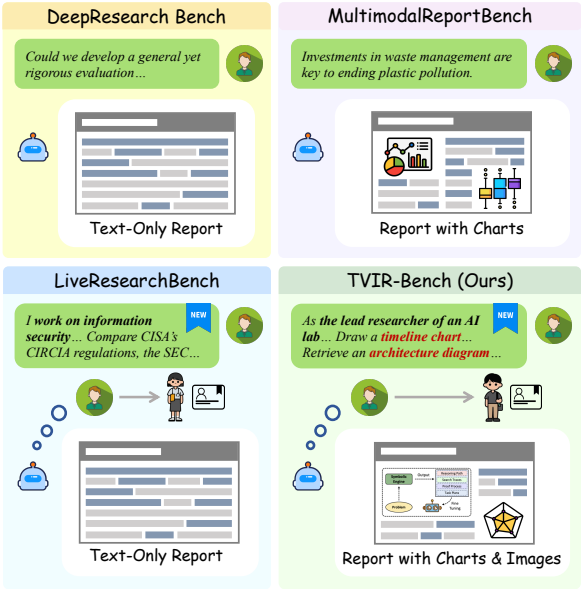

Figure 1: Comparison of representative deep research benchmarks. Existing benchmarks mainly focus on text-only or weakly multimodal reports, whereas TVIR-BENCH requires text–visual interleaved reports with semantically grounded charts and retrieved images.

Despite this progress, as shown in Figure 1, existing deep research paradigms remain predominantly text-centric. Most benchmarks and agent frameworks evaluate success based on textual coherence, depth, and citation support, while overlooking a critical characteristic of real-world professional reports: the integration of visual evidence. In practice, high-quality research outputs rarely rely on text alone. Instead, they interleave narrative analysis with charts, diagrams, and images that serve as evidential artifacts—supporting claims, revealing trends, and enabling rapid sense-making. When visual elements are present in current systems, they are often treated as decorative supplements rather than first-class reasoning components, with limited

evaluation of their fidelity, provenance, or alignment with the surrounding text. This gap leads to a fundamental mismatch between existing deep research benchmarks and the demands of real-world analytical work. A research agent that produces fluent text but inaccurate, misleading, or semantically disconnected visuals cannot be considered reliable for high-stakes decision-making. Moreover, without explicit evaluation protocols, models tend to optimize for superficial visual inclusion—prioritizing aesthetic completeness over evidential rigor. Addressing this limitation requires rethinking deep research not as a purely textual task, but as a multimodal synthesis problem in which text and visuals must be jointly generated, and evaluated.

To this end, we introduce **TVIR** (**T**ext–**V**isual **I**nterleaved **R**eport **g**eneration), a unified benchmark and the corresponding agentic framework designed to advance deep research towards multimodal, evidence-driven report generation. First, we present **TVIR-BENCH**, a multimodal deep research benchmark consisting of 100 expert-curated tasks spanning diverse domains and complexity levels. Unlike prior benchmarks, it enforces strict design principles that require visual elements—both retrieved images and code-generated charts—to be semantically grounded in specific analytical subgoals, rather than appended post hoc.

Second, we propose **TVIR-AGENT**, a hierarchical multi-agent framework that enables text–visual interleaved report generation. The agent decomposes user tasks into structured plans with explicit multimodal constraints, instantiates images and charts with traceable sources, and generates long-form reports through context-aware sequential writing. By explicitly modeling visual evidence throughout the planning and writing stages, TVIR-AGENT treats visuals as integral components of reasoning rather than optional embellishments.

Finally, we introduce a comprehensive evaluation suite that jointly audits textual and visual quality. Our framework combines **Textual Assessment** (focusing on citation grounding, logical consistency, and analytical depth) with **Visual Assessment**, which measures figure quality, chart fidelity, and cross-modal alignment between text and visuals. Through extensive experiments across multiple deep research systems, we highlight the insufficiency of existing paradigms and underscore the need for multimodal deep research agent designs.

In summary, our contributions are threefold :

- We introduce TVIR-BENCH, the first comprehensive benchmark specifically designed to evaluate the end-to-end generation of long-context multimodal research reports, which establishes a new foundation for developing and evaluating deep research agents.
- We propose TVIR-AGENT, an autonomous framework designed to handle the text–visual interleaved report generation. The system consists of a Planner, a Visual Asset Instantiation module, a Writer and a Polisher.
- We develop a rigorous dual-path evaluation framework that assesses reports through Textual Assessment and Visual Assessment. Our extensive experiments with nine representative deep research systems provide the critical insight that while current LLMs excel at textual fluency, they frequently prioritize "decorative" visuals over "evidential" ones, highlighting a significant gap in existing paradigms for evidence-based multimodal reasoning.

2 Related Work

Deep Research Agent Deep Research Agents (DRAs) have become an important paradigm for long-horizon retrieval, reasoning, and report generation. Existing DRAs, such as WebThinker (Li et al., 2025a) and WebWeaver (Li et al., 2025b), are mostly text-centric. Although some recent work explores multimodal settings, such as Multimodal DeepResearcher (Yang et al., 2025) and FinSight (Jin et al., 2025), visuals are still typically treated as auxiliary outputs and mainly rely on charts. In contrast, TVIR-AGENT formulates deep research as text–visual interleaved report generation, integrating retrieved images and code-generated charts into the full research workflow.

Deep Research Benchmark Existing deep research benchmarks mainly evaluate text-only report generation, such as DeepResearch Bench (Du et al., 2025) and DeepResearch Bench II (Li et al., 2026). In multimodal settings, MultimodalReport-Bench (Yang et al., 2025) considers reports with code-generated charts, but its evaluation remains coarse-grained. MMDeepResearch-Bench (Huang et al., 2026) further studies image–text inputs, where report visuals are grounded in task-provided images. In contrast, TVIR-BENCH evaluates end-to-end text-visual interleaved report generation with fine-grained textual and visual assessment.

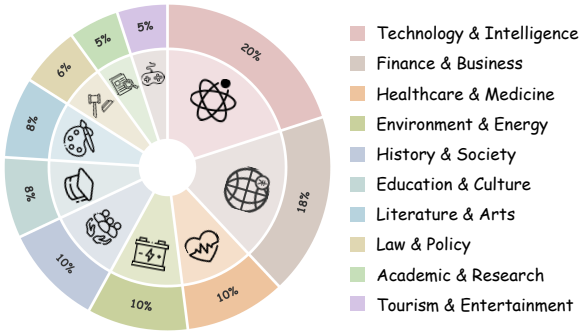

Figure 2: Domain taxonomy of TVIR-BENCH.

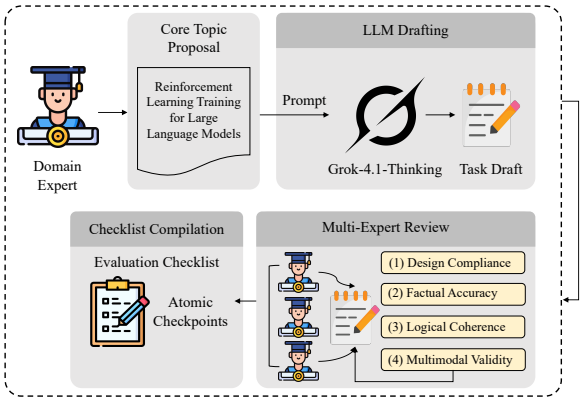

Figure 3: Overview of the Data Construction Process of TVIR-BENCH.

3 TVIR-BENCH

3.1 Data Design

Domain Coverage To ensure comprehensive and representative domain coverage, we draw on the domain taxonomies and task distributions of several existing deep research benchmarks, including DeepResearch Bench (Du et al., 2025), LiveResearchBench (Wang et al., 2025), and DeepResearchEval (Wang et al., 2026), and develop the domain taxonomy shown in Figure 2. While preserving coverage of the humanities and basic sciences, we place greater emphasis on *Technology & Intelligence* and *Finance & Business* to better align the needs of high-stakes decision-making.

Task Design and Complexity Task construction is guided by five core design principles: **role-driven**, **demand-oriented**, **deep research**, **frontier-focused**, and **multimodal integration**. These principles ensure that tasks are grounded in realistic user needs, require substantive analytical synthesis rather than simple information retrieval, and incorporate explicit multimodal elements to better reflect real-world deep research workflows. To systematically evaluate model performance under varying cognitive demands, we further organize tasks into three complexity levels, corresponding to low, medium, and high requirements for multimodal integration and instruction following. Detailed definitions of the design principles and complexity levels are provided in Appendix A.1.

3.2 Data Construction

Our dataset is constructed through a systematic expert-driven workflow, consisting of expert topic proposal, LLM-based task drafting, multi-expert review, and checklist compilation.

Expert Topic Proposal For each task, a domain expert with sufficient professional knowledge and judgment proposes a core topic. To satisfy the frontier-focused principle defined above, the topic must be reasonably novel and timely.

LLM-Based Task Drafting Based on the core topic proposed by the domain expert, we use Grok-4.1-Thinking to draft the task. We guide the model to produce tasks that are logically coherent, researchable, and practically meaningful. The detailed prompt is provided in the Appendix F.1.

Multi-Expert Review Each draft is then examined by three domain experts and revised accordingly. The review focuses on four aspects: **(1) Design Compliance**, which checks whether the task satisfies the predefined task design principles; **(2) Factual Accuracy**, which verifies the correctness of entities, concepts, claims, and contextual details involved in the task; **(3) Logical Coherence**, which evaluates whether the task is clearly formulated and whether its sub-questions form a meaningful and well-connected whole; and **(4) Multimodal Validity**, which ensures that required multimodal elements are practically obtainable, including publicly retrievable images and code-generated charts based on real and accessible data.

Checklist Compilation For each accepted task, we compile a corresponding evaluation checklist. The checklist converts the task into a set of verifiable items for systematically assessing whether a generated report fully and accurately addresses its requirements. Each item is formulated as a small number of clear and actionable atomic checkpoints, without going beyond the task itself.

3.3 Dataset Statistics

TVIR-BENCH comprises 100 high-quality multimodal deep research tasks, including 50 Chinese tasks and 50 English tasks. They span 10 major domains and are proportionally balanced across the three predefined complexity levels. The subquestions within these tasks cover eight high-level functional types, such as trend prediction, mechanism explanation, and comparative analysis, with a roughly balanced distribution across the dataset.

3.4 Evaluation Framework

We propose a multi-dimensional evaluation framework for end-to-end auditing of generated research reports. It consists of two complementary components: **Textual Assessment (TA)** and **Visual Assessment (VA)**. TA and VA are computed as the arithmetic mean of their respective fine-grained metric scores. Unless otherwise specified, all metrics are evaluated using an LLM-as-a-Judge, with scores normalized to a scale of 1–100. The computation methods for all fine-grained metrics are detailed in Appendix B, and the complete evaluation prompts are provided in Appendix F.3.

3.4.1 Report Preprocessing

To support downstream evaluation, we preprocess each generated report to extract structured information using the judge LLM. Specifically, we extract (i) *reference entries*, including their reference indices and associated URLs, (ii) *fact–citation pairs* for textual assessment, where factual statements are linked to citation indices, and (iii) *figure elements* for visual assessment, together with their captions, base64-encoded visual content, surrounding context, and associated citation indices. We then use the extracted URLs to retrieve the corresponding webpage text via the Serper API. The extraction prompts are provided in Appendix F.2.

3.4.2 Textual Assessment (TA)

Citation Support (CS) measures whether factual statements are supported by their cited references. Using the extracted fact–citation pairs and retrieved reference contents, the judge LLM assigns each fact a score of 1, 0.5, or 0 for *supported*, *partially supported*, and *unsupported*, respectively.

Instruction Alignment (IA) evaluates whether the report satisfies the task checklist. For each checklist item, the judge LLM assigns a score of 1, 0.5, or 0 based on how completely and specifically the report addresses it.

Writing Quality (WQ), **Analytical Depth & Breadth (ADB)**, and **Factual & Logical Consistency (FLC)** are assessed at the report level. WQ evaluates coherence and organization, clarity and readability, conciseness, and stylistic consistency. ADB assesses whether the report goes beyond surface-level description through explanatory reasoning, sustained analysis, critical evaluation, forward-looking insight, and broad thematic coverage. FLC measures self-consistency by detecting factual or logical contradictions and mapping the resulting issue count to a discrete score.

3.4.3 Visual Assessment (VA)

Multimodal Composition (MC) provides a report-level assessment of how effectively figure elements are organized across the report, considering their layout, quantity, variety, and richness.

Figure Quality (FQ) captures the intrinsic visual quality of figures. It combines **Image Quality**, derived from CV-based measurements of resolution, aspect ratio, sharpness, and contrast together with a duplication penalty, and **Chart Quality**, assessed by the judge LLM using binary checklist evaluations of layout integrity, readability, and conciseness. The final FQ score is computed as a count-weighted average of the two component scores.

Figure Caption Quality (FCQ), **Figure–Context Integration (FCI)**, and **Chart–Source Consistency (CSC)** are assessed for each extracted figure element. FCQ evaluates whether a caption accurately describes the figure, provides sufficient interpretive information, and remains clear and readable. FCI assesses how well each figure relates to its surrounding text, is incorporated into the narrative flow, and contributes information beyond what text alone can effectively convey. CSC measures the consistency of each chart with its cited sources by identifying contradictions and mapping the resulting issue count to a discrete score.

4 TVIR-AGENT

We develop TVIR-AGENT based on MiroThinker (MiroMind Team et al., 2025) and adapt it for text–visual interleaved report generation. Given a user task $\mathcal{T}$, TVIR-AGENT produces a high-quality research report $\mathcal{R}$ through four stages: planning, visual asset instantiation, writing, and polishing. Figure 4 illustrates the overall architecture.

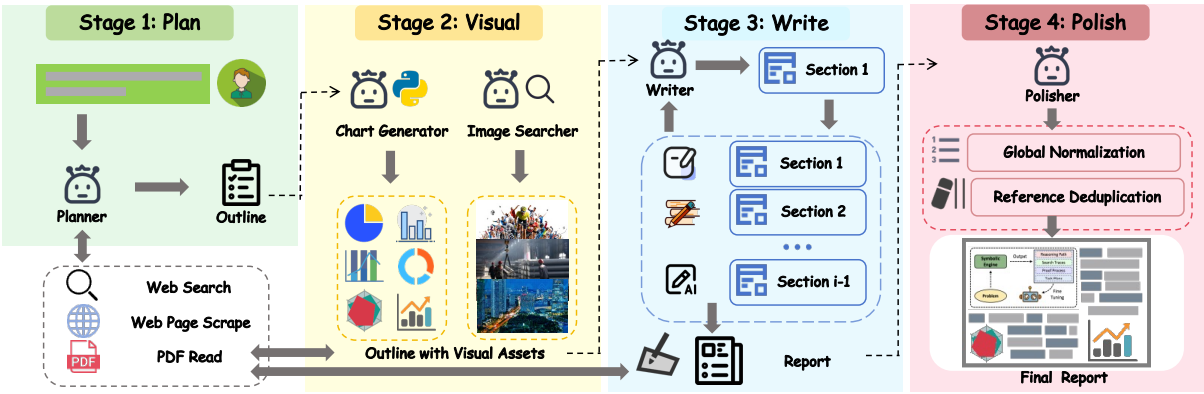

Figure 4: Overview of the proposed multi-stage framework for report generation. The pipeline consists of four stages: (1) Planning, where a structured outline is generated from user queries and external sources (e.g., web pages and PDFs); (2) Visualization, where the outline is enriched with charts and images; (3) Writing, where a draft report is produced based on the enriched outline; and (4) Polishing, where the report is refined through normalization and reference deduplication to produce the final output.

4.1 Research-Grounded Planning

As the system's initialization module, the **Planner** parses the user task $\mathcal{T}$ and iteratively invokes external tools such as Google Search and web scraping to retrieve and verify relevant information, which it organizes into a structured outline $\mathcal{O} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_N\}$. Each section $\mathcal{S}_i$ contains a title, a brief summary, planned visual requirements $\mathcal{V}_i$, and section-level research notes $\mathcal{N}_i$.

Each note $n \in \mathcal{N}_i$ records a citation, source URL, and key findings. These research notes provide factual grounding for subsequent stages, improving both credibility and traceability while serving as reusable knowledge references.

4.2 Visual Asset Instantiation

Given the planned outline $\mathcal{O}$, this stage instantiates the visual requirements of each section and produces an augmented outline $\mathcal{O}^{\text{vis}}$. To address different visual needs, we employ two specialized agents. The **Image Searcher** handles visual concepts such as portraits, scenes, or architecture diagrams by retrieving candidate images through Google Image Search, filtering low-quality results with heuristic rules, and using a VQA tool for relevance verification before selecting the most suitable one. The **Chart Generator** handles content involving data distributions or relationships by retrieving relevant data through search and web scraping tools, verifying authenticity and cross-source consistency, generating Python plotting code, and executing it in a sandbox to produce charts.

After this stage, each section $\mathcal{S}_i$ is associated with instantiated visual assets $\mathcal{V}_i^{\text{vis}}$, together with captions, descriptions, and source provenance. For generated charts, the original data-source URLs are preserved; for retrieved images, the corresponding source webpage URLs are retained.

4.3 Context-Aware Sequential Writing

The **Writer** generates the report section by section. To maintain coherence across sections and reduce redundancy, it conditions on both the section specification $\mathcal{S}_i$ and a dynamically updated global context $\mathcal{C}_{i-1}$, consisting of the titles, summaries, and subsection structures of previously generated sections. It also uses the associated research notes $\mathcal{N}_i$ as supporting evidence and, when they are insufficient, invokes search and web scraping tools to gather additional verifiable information.

During writing, the Writer determines appropriate insertion points for the visual assets $\mathcal{V}_i^{\text{vis}}$ based on their descriptions and then inserts them into the Markdown section content, using local paths for charts and online URLs for retrieved images. Each section maintains its own figure and citation numbering, both starting from 1.

4.4 Global Index Polishing

In the final stage, the **Polisher** processes references and figures at the report level. It first removes uncited references from each section, then deduplicates the remaining references globally by URL and renumbers them into a unified reference list, while updating in-text citation markers accordingly. Finally, it renumbers figures across sections, reassigns figure IDs and labels in sequential order, and updates in-text figure references to match.

Model	Textual Assessment					Visual Assessment					Overall		
	CS	IA	WQ	ADB	FLC	FQ	MC	FCQ	FCI	CSC	**TA**	**VA**	**Avg.**
Commercial Closed-Source Systems													
Gemini-3-Pro Deep Research	14.96	58.31	66.88	63.94	88.50	-	-	-	-	-	58.52	-	-
Grok-4.1-Thinking DeepSearch	17.72	60.65	67.68	57.58	89.20	80.43	52.15	47.04	46.76	5.75	58.56	46.43	52.49
Claude-4.5-Sonnet w/Search	47.53	79.32	69.37	70.52	84.00	90.24	63.85	61.47	53.43	67.49	70.15	67.30	68.72
Perplexity Deep Research	44.60	81.03	67.60	70.64	80.90	73.62	62.90	59.02	63.41	8.35	68.95	53.46	61.20
Genspark Deep Research	35.27	**83.71**	69.28	70.64	84.60	**92.87**	70.70	63.58	59.00	40.28	68.70	65.29	66.99
Manus-1.6	45.57	74.12	**72.15**	62.84	**92.40**	86.27	70.75	66.14	71.03	56.02	69.42	70.04	69.73
TVIR-AGENT Variants (Ours)													
TVIR-Agent (Qwen3-Max)	53.68	76.69	69.30	67.48	83.00	91.71	67.80	72.44	74.56	**78.63**	70.03	77.03	73.53
TVIR-Agent (GLM-4.7)	**68.64**	71.98	69.20	68.16	80.20	84.61	62.55	70.13	73.39	77.35	**71.64**	73.61	72.62
TVIR-Agent (Claude-4.5-Sonnet)	51.20	81.09	69.88	**72.22**	76.20	87.17	**77.80**	**74.49**	**76.75**	77.58	70.12	**78.76**	**74.44**

Table 1: Evaluation results of deep research systems on TVIR-BENCH. **Bold** and underlined values indicate the best and second-best performance on each dimension, respectively. Overall is computed as the mean of TA and VA.

5 Experiments

5.1 Experimental Setup

We evaluate a total of nine deep research systems, including six commercial closed-source systems and three TVIR-AGENT variants built with different backbone LLMs. The commercial systems consist of one text-only report generation system, Gemini-3-Pro Deep Research (Google, 2025a), and five text–visual interleaved report generation systems: Grok-4.1-Thinking DeepSearch (xAI, 2025), Claude-4.5-Sonnet w/Search (Anthropic, 2025), Perplexity Deep Research (Perplexity AI, 2025), Genspark Deep Research (Genspark AI, 2025), and Manus-1.6 (Manus AI, 2025). The three TVIR-AGENT variants use Qwen3-Max (Qwen-Team, 2025), GLM-4.7 (Zhipu AI, 2025), and Claude-4.5-Sonnet as backbone LLMs. Further details on report collection are provided in Appendix D. Report preprocessing and all LLM-based evaluation are conducted using GPT-5.2 (OpenAI, 2025) as the judge LLM.

5.2 Main Results

Overall performance Table 1 presents the main results on TVIR-BENCH. Overall, TVIR-AGENT variants achieve the strongest aggregate performance among all evaluated systems. In particular, TVIR-Agent (Claude-4.5-Sonnet) obtains the best Overall score, followed by TVIR-Agent (Qwen3-Max) and TVIR-Agent (GLM-4.7), while Manus-1.6 is the strongest commercial system. The aggregate metrics also reveal different system strengths: TVIR-Agent (GLM-4.7) achieves the best TA score, showing strong textual capabilities, whereas TVIR-Agent (Claude-4.5-Sonnet) achieves the best VA score by a clear margin, demonstrating superior visual grounding and cross-modal alignment. These results indicate that our framework is particularly effective for deep research tasks requiring both textual synthesis and visual integration.

Key insights The benchmark also reveals several important insights into current deep research systems. First, no single model consistently performs best on every fine-grained dimension, suggesting that strong aggregate performance does not necessarily translate into consistent strength across all aspects of text–visual research. Second, several commercial systems remain competitive on textual assessment metrics, indicating that high-quality textual synthesis is already a relative strength of existing deep research products. Finally, Gemini-3-Pro Deep Research cannot be evaluated on VA and Overall because it generates text-only reports, which further highlights the importance of native multimodal support in TVIR-AGENT.

5.3 Further Analysis

Fine-Grained Strengths and Limitations Compared with commercial systems, TVIR-AGENT variants show clearer advantages in CS and most VA dimensions. On **CS**, TVIR-Agent (GLM-4.7) achieves 68.64, outperforming the best commercial system Claude-4.5-Sonnet w/Search (47.53) by **21.11** points. For **CSC**, TVIR-Agent (Claude-4.5-Sonnet) scores 74.49, exceeding Manus-1.6 (66.14) by **8.35** points. These results indicate better evidence grounding and more reliable cross-modal alignment in TVIR-AGENT. By contrast, **FLC** remains relatively weaker for several strong systems, likely because longer and more detailed reports are inherently harder to maintain consistently. This suggests that long-form factual and logical consis-

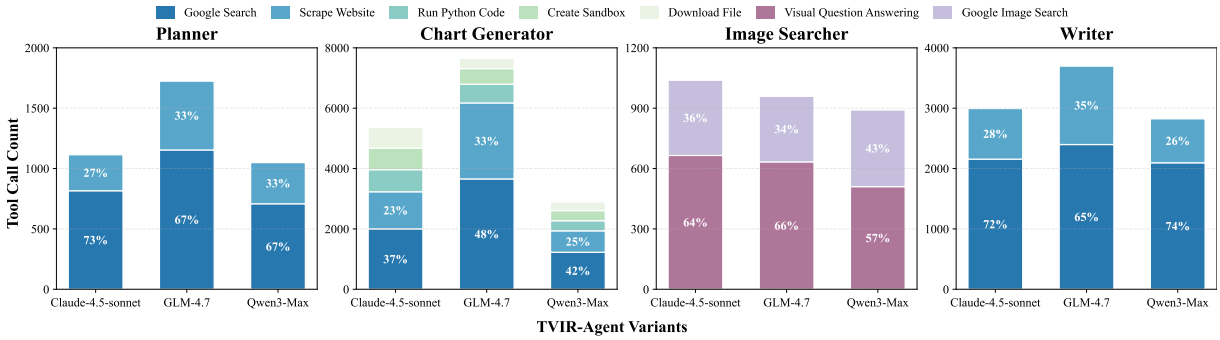

Figure 5: Tool usage distribution of TVIR-AGENT variants across major components.

tency is still a shared challenge.

Performance across Task Complexity Levels Figure 6 shows a consistent trend across TVIR-AGENT variants and leading commercial systems. As task complexity increases, **IA** generally declines, while **ADB** tends to improve. This pattern suggests that more complex tasks place heavier demands on multimodal coordination, fine-grained instruction tracking, and long-horizon reasoning, making it more difficult for research systems to fully satisfy task requirements. At the same time, such tasks appear to encourage more comprehensive and exploratory responses, leading to greater explanatory depth, broader thematic coverage, and more sustained analysis.

Tool Usage Analysis within TVIR-AGENT Figure 5 shows that TVIR-AGENT variants exhibit distinct tool usage patterns across the main components. Inspired by DeepResearch Bench (Du et al., 2025), we additionally report **Average Effective Citations (AEC)** in Figure 8, with its formal definition provided in Appendix E.1. TVIR-Agent (GLM-4.7) shows the heaviest use of search and scraping tools, especially in the Planner and Chart Generator, which is consistent with its highest AEC of **102.41**. However, under a limited agent-turn budget, overly extensive retrieval may come at the expense of chart asset instantiation: despite planning an average of 8.66 charts per task, it generates only 3.33, yielding a chart fulfillment rate of 38.45%. By contrast, TVIR-Agent (Qwen3-Max) plans only 3.68 charts per task on average, suggesting weaker visual planning capability in complex multimodal tasks. TVIR-Agent (Claude-4.5-Sonnet) adopts a more balanced tool usage profile, maintaining a relatively high AEC of 86.14 while achieving the highest chart fulfillment rate of **94.61%**. Overall, these results suggest that the relative strengths of TVIR-AGENT variants depend not only on backbone model capability, but also on how effectively tool usage is allocated between retrieval and chart generation.

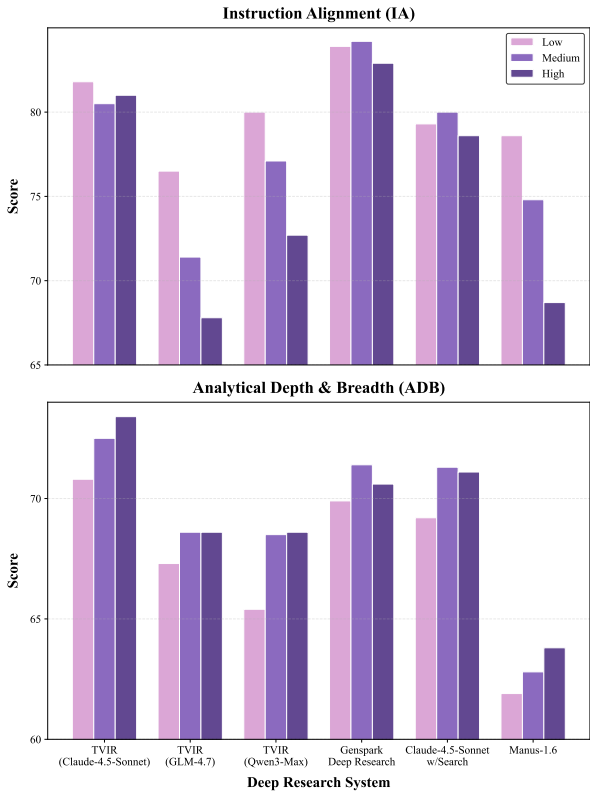

Figure 6: Performance comparison of IA and ADB across task complexity levels for TVIR-AGENT variants and leading commercial systems.

Analysis on the Reliability of Judge LLMs We randomly sample 20 tasks from TVIR-BENCH and collect the corresponding reports generated by six evaluated systems. We then compare the judgments produced by two judge LLMs, **GPT-5.2** and **Gemini-2.5-Pro** (Google et al., 2025b). Following DeepResearch Bench (Du et al., 2025), we adopt **Pairwise Agreement Rate (PAR)** and **Over-**

all Pearson Correlation (OPC) as the primary metrics, and further introduce **Overall Spearman Correlation (OSC)** to capture rank-based consistency. Detailed metric definitions and computation procedures are provided in Appendix E.2. As shown in Table 2, the two judge LLMs exhibit high consistency, suggesting that our automated evaluation framework is stable and reliable.

Dimension	PAR	OPC	OSC
TA	80.00	92.36	82.86
VA	80.33	96.91	100.0
Overall	78.33	95.58	94.29

Table 2: Agreement between the two judge LLMs.

Structural Error Analysis To further examine the reliability of generated reports, we analyze three types of structural errors: **Traceability Error**, **Consistency Error**, and **Completeness Error**. Traceability Error refers to facts, data, or figures presented without explicit and accessible sources; Consistency Error reflects failures in the internal indexing system, including missing numbers or duplicated entries; Completeness Error captures unusable or incomplete components referenced in the text, such as broken figures or absent captions. As shown in Figure 7, TVIR-AGENT variants produce substantially fewer structural errors than commercial systems overall, suggesting better end-to-end control over citation management and multimodal report assembly, with TVIR-Agent (GLM-4.7) achieving the lowest total error count. Nevertheless, traceability remains a persistent challenge across systems, and even TVIR-AGENT exhibits a nontrivial number of errors.

5.4 Ablation Studies

To examine the contribution of key components in TVIR-AGENT, we conduct ablation studies using TVIR-Agent (Claude-4.5-Sonnet) as the reference system. Specifically, we remove **research notes**, the **Image Searcher**, and the **Chart Generator**, respectively, while keeping the rest of the pipeline unchanged. As shown in Table 3, removing any component leads to a drop in overall performance. Among them, removing the Chart Generator has the largest effect, reducing VA from 78.62 to 60.91 and Overall from 73.92 to 63.84, which highlights its central role in visual synthesis and cross-modal alignment. Removing the Image Searcher leads to a clear decline across all metrics, whereas the effect of removing the research notes is relatively small. Overall, these results show that the performance gains of TVIR-AGENT come from the complementary contributions of research-grounded planning and specialized visual capabilities.

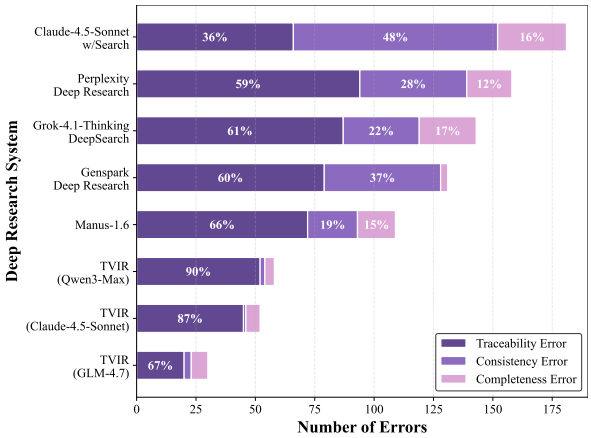

Figure 7: Distribution of structural errors across deep research systems.

System Variant	TA	VA	Overall
TVIR-Agent	69.23	78.62	73.92
w/o research notes	68.63	78.42	73.52
w/o Image Searcher	67.82	77.23	72.53
w/o Chart Generator	66.77	60.91	63.84

Table 3: Ablation results for TVIR-Agent (Claude-4.5-Sonnet) on 20 randomly sampled tasks.

6 Conclusion

In this work, we introduce TVIR, a unified benchmark and agentic framework for text–visual interleaved report generation. We present TVIR-BENCH, an expert-curated benchmark with a dual-path evaluation framework covering both textual and visual assessment, and TVIR-AGENT, a hierarchical multi-agent framework that explicitly models visual evidence throughout planning and writing. Extensive experiments show that TVIR-AGENT achieves strong overall performance and improves evidence grounding and cross-modal alignment over existing systems, while also revealing a key limitation of current deep research systems: they remain much stronger at textual synthesis than at integrating visual assets. We hope TVIR will provide a foundation for future work on trustworthy multimodal deep research agents.

Limitations

Benchmark and Evaluation Despite the strong agreement between judge LLMs, our evaluation framework still inherits the limitations of automated assessment. First, most metrics rely on LLM-as-a-Judge and may remain sensitive to judge preferences, prompting variations, and model updates. Second, citation-based evaluation depends on external webpages, whose content may change over time or become inaccessible, which can affect evaluation stability. Third, report preprocessing requires automatically extracting references, fact–citation pairs, and figure elements from long multimodal reports; such extraction inevitably introduces omissions that may propagate to downstream scoring. In addition, some report-level visual properties remain difficult to assess automatically. For example, metrics such as MC rely mainly on textual cues around figures rather than direct visual perception, which means that report-level multimodal composition, cross-figure coordination, and narrative coherence are only partially captured. In addition, more subjective qualities, such as figure aesthetics and professional visual design, are not explicitly evaluated, although they are often important in real-world analytical writing.

Agent Framework Although TVIR-AGENT provides an effective hierarchical framework for text–visual interleaved report generation, it remains a relatively basic design with substantial room for improvement. The current system is still largely stage-wise, and efficiency may be further improved through stronger parallelization of retrieval, visual asset instantiation, and writing across sections. Its memory management and context compression mechanisms are also relatively simple for long-horizon multimodal tasks with diverse sources and intermediate artifacts. In addition, while TVIR-AGENT already supports iterative search, this iteration is still concentrated before downstream planning and writing. Stronger designs may more tightly interleave search with planning and writing, allowing the agent to dynamically revise outlines, update visual requirements, and retrieve targeted evidence as the report evolves. Finally, the current polishing stage mainly handles report-level indexing and formatting, while more advanced refinement could further improve figure quality, citation quality, and local text–visual alignment.

References

Anthropic. 2025. Introducing claude sonnet 4.5.

Yuxuan Cai, Xinyi Lai, Peng Yuan, Weiting Liu, Huajian Li, Mingda Li, Xinghua Wang, Shengxie Zheng, Yanchao Hao, Yuyang Yin, and Zheng Wei. 2026. Yunque deepresearch technical report. *Preprint*, arXiv:2601.19578.

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.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 176 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

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.

Genspark AI. 2025. Genspark deep desearch.

Google. 2025a. Gemini deep research.

Google, Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, and 3417 others. 2025b. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Peizhou Huang, Zixuan Zhong, Zhongwei Wan, Donghao Zhou, Samiul Alam, Xin Wang, Zexin Li, Zhihao Dou, Li Zhu, Jing Xiong, Chaofan Tao, Yan Xu, Dimitrios Dimitriadis, Tuo Zhang, and Mi Zhang. 2026. Mmdeepresearch-bench: A benchmark for multimodal deep research agents. *Preprint*, arXiv:2601.12346.

Jiajie Jin, Yuyao Zhang, Yimeng Xu, Hongjin Qian, Yutao Zhu, and Zhicheng Dou. 2025. Finsight: Towards real-world financial deep research. *Preprint*, arXiv:2510.16844.

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. *Preprint*, arXiv:2601.08536.

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

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

Manus AI. 2025. Introducing manus 1.6: Max performance, mobile dev, and design view.

MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, Wenhan Dou, Yue Deng, Yunjie Fu, Junqi Ge, Chenxia Han, Tammy Huang, Zhenhang Huang, Jerry Jiao, Shilei Jiang, and 36 others. 2025. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *Preprint*, arXiv:2511.11793.

OpenAI. 2025. Introducing gpt-5.2.

OpenAI, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, and 243 others. 2024. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Perplexity AI. 2025. Introducing perplexity deep research.

QwenTeam. 2025. Qwen3-max: Just scale it.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, Qiujie Xie, Xinyu Guo, Qu Yang, Jiayi Wu, Jujia Zhao, Xiaqiang Tang, Xinbei Ma, Cunxiang Wang, Jiaxin Mao, and 7 others. 2025. Deep research: A systematic survey. *Preprint*, arXiv:2512.02038.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, Kuan Li, Liangcai Su, Litu Ou, Liwen Zhang, Pengjun Xie, Rui Ye, Wenbiao Yin, Xinmiao Yu, Xinyu Wang, and 38 others. 2025. Tongyi deepresearch technical report. *Preprint*, arXiv:2510.24701.

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.

xAI. 2025. Grok deepsearch.

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.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *Preprint*, arXiv:2504.03160.

Zhipu AI. 2025. Glm-4.7: Advancing the coding capability.

A Data Design and Construction Details

A.1 Definitions of Task Design Principles

To ensure that tasks in TVIR-BENCH reflect realistic, challenging, and evaluation-friendly deep research scenarios, we formulate five task design principles: role-driven, demand-oriented, deep research, frontier-focused, and multimodal integration. These principles jointly regulate who the task is for, what the task asks, how much reasoning it requires, how up-to-date the content should be, and how multimodal evidence should be incorporated.

Role-Driven Each task should be centered on a clearly identifiable user with a specific professional identity or functional role, such as a policymaker, clinician, investment analyst, researcher, engineer, or student. The role should be concrete and representative of a real user group that may plausibly raise such a research need in practice, rather than a vague or generic subject such as "a normal user" or "a researcher". To strengthen realism and task specificity, the task context should also describe the user's objective, constraints, or decision scenario. For example, instead of asking a general question about a biomedical trend, a task may be framed from the perspective of a head of R&D at a biopharmaceutical company who needs to assess the clinical and commercial prospects of a new therapeutic direction. This design principle ensures that tasks are anchored in authentic use cases rather than abstract knowledge queries.

Demand-Oriented Each task should express a clear and structured research demand within a specific domain or topic. A task may include multiple sub-questions, but they should remain logically connected and confined to the same thematic scope rather than being artificially assembled from unrelated domains. Each sub-question should correspond to a well-defined information goal, avoiding open-ended prompts such as "give your opinion" without substantive constraints. We encourage tasks to be presented in a structured manner, for example by explicitly listing the aspects that need to be covered, so that the model has a clear pathway for organizing its response and the evaluator has a clear basis for judging completeness. While complex tasks are allowed, the overall instruction should remain coherent, bounded, and operational, without ambiguity, redundancy, or internal contradiction.

Deep Research Tasks should require substantially more than surface-level information aggregation. They should encourage the model to synthesize evidence from multiple sources, construct causal or explanatory chains, compare competing claims, and produce conclusions or recommendations that are explicitly grounded in prior analysis. Relevant evidence may include academic findings, statistical data, case materials, visual evidence, or expert viewpoints, and the task should implicitly or explicitly require the model to explain how such evidence supports its claims. In addition, tasks should encourage critical analysis by asking the model to recognize conflicts across sources, limitations in data, uncertainty in interpretation, or weaknesses in methods and assumptions. A well-designed deep research task should therefore form a closed analytical loop: evidence gathering, reasoning and critique, and then conclusion or actionable recommendation.

Frontier-Focused Tasks should target recent developments, emerging challenges, and open questions across domains, so as to evaluate whether models can engage with up-to-date knowledge rather than relying mainly on static background information. In general, topics are expected to focus on new technologies, policy changes, market shifts, scientific breakthroughs, or other developments that have become salient within the past two to three years. Tasks are encouraged to rely on recent and authoritative sources, especially materials published in or after 2023, such as top-tier conference papers, industry white papers, government releases, company filings, or official statistical reports. When a task involves trend analysis or quantitative comparison, it should call for the latest available data, ideally from 2024–2025 when accessible, with clear attention to timestamps. For frontier topics where no stable consensus yet exists, the task may ask the model to present and compare different institutional or scholarly viewpoints and analyze the basis of their disagreement.

Multimodal Integration Tasks should explicitly require the integration and presentation of multimodal information, so that visual elements become a meaningful part of the research process rather than decorative additions. Such elements may include publicly retrievable images from the web, such as model architecture diagrams, algorithm workflow illustrations, historical portraits, environmental photographs, or scene images, as well

as code-generated charts based on real, publicly accessible data, such as line charts, radar charts, heatmaps, or pie charts. In task wording, multimodal needs should usually be expressed naturally through verbs and phrases such as "visualize", "plot", "show", "illustrate", or "provide a figure with explanation", without always explicitly specifying whether the model must retrieve an existing image or generate a chart. Instead, the intended visual type should be implied by the request itself, for example by asking for a radar chart to compare alternatives or an architecture diagram to explain a technical system. All multimodal elements must directly serve the core analytical objective and should be added only when they improve interpretation, explanation, or decision support. For example, a heatmap may be used to show regional differences in policy outcomes, an architecture diagram to clarify the innovation of a model design, or a time-series plot to support claims about market growth. To ensure practical feasibility, any required retrieved image must be publicly available online, and any required generated chart must rely on real and publicly accessible data sources.

A.2 Definitions of Task Complexity Levels

To evaluate model performance under different levels of cognitive and multimodal demand, we organize tasks in TVIR-BENCH into three complexity levels: **low**, **medium**, and **high**. The complexity label is determined jointly by task length, structural organization, and the extent of multimodal requirements. Here, multimodal requirements refer to distinct requested visual components or visual goals rather than the number of final images or charts. For example, asking for "typical architecture diagrams for different stages of a technology" counts as one multimodal requirement even if multiple figures may ultimately be included.

Low Complexity Low-complexity tasks are relatively concise, typically around 200 Chinese characters or about 130 English words. They usually contain **one to three multimodal requirements** and are primarily intended to assess whether a system can handle focused deep research requests with limited structural and coordination burden. Such tasks may be written in a compact paragraph or brief outline form, and explicit bullet-point decomposition is not required. Although the scope is narrower than that of more complex tasks, the task should still satisfy the core design principles of realism, coherence, and research depth.

Medium Complexity Medium-complexity tasks are more elaborate, typically around 400 Chinese characters or about 260 English words, and contain **two to four multimodal requirements**. They are expected to involve a clearer internal structure and usually benefit from being organized into **three to four major points**, each with **one to three sub-questions**. Compared with low-complexity tasks, they impose greater demands on instruction following, cross-source synthesis, and the coordinated use of text and visual evidence. At this level, the task should require the model not only to answer multiple related questions, but also to maintain logical continuity across them.

High Complexity High-complexity tasks are the most demanding, typically around 600 Chinese characters or about 390 English words, and contain **three to five multimodal requirements**. They must be presented in an explicitly structured format, usually with **four to five major points**, each containing **two to four sub-questions**. These tasks place substantial demands on long-horizon planning, detailed instruction tracking, and multimodal integration across multiple analytical dimensions. In addition to requiring broader coverage and deeper synthesis, high-complexity tasks often involve more intricate dependencies among sub-questions, making it necessary for the model to coordinate evidence, reasoning, and visualization in a highly organized manner.

A.3 Definitions of the Eight High-Level Functional Types

To characterize the substantive objectives covered by task sub-questions, we summarize them into eight high-level functional types. These categories describe the primary analytical function a sub-question is intended to serve, and together they capture the major forms of reasoning and inquiry represented in TVIR-BENCH.

Status Characterization This functional type refers to questions that comprehensively describe the current state of an object, issue, or system, including its composition, major characteristics, distribution, or present-stage structure. The goal is to build an accurate and organized picture of what the situation currently is.

Comparative Analysis This functional type refers to questions that compare two or more enti-

ties in order to identify similarities, differences, relative strengths and weaknesses, or rankings. Such questions are often used to support selection, benchmarking, or comparative evaluation across methods, markets, policies, or institutions.

Mechanism Explanation This functional type refers to questions that clarify the causes, mechanisms, or underlying principles behind a phenomenon, outcome, or behavior. The emphasis is on explaining why or how something happens rather than merely describing that it happens.

Evolution Tracing This functional type refers to questions that examine how an object, idea, technology, or institution emerges and changes over time. These questions focus on origins, developmental trajectories, stage transitions, and patterns of inheritance or transformation.

Impact Assessment This functional type refers to questions that evaluate the actual effects, consequences, benefits, risks, or limitations of a policy, technology, event, or behavioral intervention. The emphasis is on judging what influence something has produced or is likely to produce in practice.

Trend Prediction This functional type refers to questions that infer likely future developments or outcomes on the basis of current data, patterns, or models. These questions typically involve projection, scenario analysis, or forward-looking judgment under uncertainty.

Foundational Exploration This functional type refers to questions centered on fundamental or frontier issues, especially those aimed at probing unknown mechanisms, deeper principles, or the essence of a problem. Compared with other functional types, this category places greater emphasis on open-ended scientific or conceptual inquiry.

Solution Construction This functional type refers to questions that require the design of a framework, model, system, strategy, or actionable plan to solve a specific problem or achieve a stated objective. Such questions move beyond analysis toward synthesis and practical proposal generation.

B Fine-Grained Evaluation Metrics Computation Methods

B.1 Textual Assessment Metrics

Let a report contain M extracted fact–citation pairs, and let K denote the total number of checklist items. For report-level rubric-based metrics, let D denote the number of scoring dimensions in the corresponding rubric.

Citation Support For each extracted fact–citation pair, the judge assigns a discrete score

$$s_i^{\mathrm{CS}} \in \{1,\ 0.5,\ 0\},$$

corresponding to *supported*, *partially supported*, and *unsupported*, respectively. The normalized Citation Support score is computed as the arithmetic mean over all scored fact–citation pairs:

$$\mathrm{CS} = \frac{1}{M_{\mathrm{scored}}} \sum_{i=1}^{M_{\mathrm{scored}}} s_i^{\mathrm{CS}},$$

where M_{scored} is the number of successfully scored pairs. If no fact–citation pair is available, the score is set to 0.

Instruction Alignment For each checklist item, the judge assigns a discrete score

$$s_j^{\mathrm{IA}} \in \{1,\ 0.5,\ 0\},$$

corresponding to *fully satisfied*, *partially satisfied*, and *not satisfied*, respectively. The normalized Instruction Alignment score is computed as

$$\mathrm{IA} = \frac{1}{K_{\mathrm{scored}}} \sum_{j=1}^{K_{\mathrm{scored}}} s_j^{\mathrm{IA}},$$

where K_{scored} is the number of checklist items that receive valid scores.

Writing Quality Writing Quality is assessed using a report-level rubric with $D_{\mathrm{WQ}} = 4$ dimensions. Let

$$r_d^{\mathrm{WQ}} \in \{1, 2, \ldots, 10\}$$

be the judge score for dimension d. The normalized Writing Quality score is

$$\mathrm{WQ} = \frac{1}{10 D_{\mathrm{WQ}}} \sum_{d=1}^{D_{\mathrm{WQ}}} r_d^{\mathrm{WQ}}.$$

In our implementation, this becomes

$$\mathrm{WQ} = \frac{1}{40} \sum_{d=1}^{4} r_d^{\mathrm{WQ}}.$$

Analytical Depth & Breadth Analytical Depth & Breadth is assessed using a report-level rubric with $D_{\mathrm{ADB}} = 5$ dimensions. Let

$$r_d^{\mathrm{ADB}} \in \{1, 2, \ldots, 10\}$$

be the judge score for dimension d. The normalized score is computed as

$$\mathrm{ADB} = \frac{1}{10 D_{\mathrm{ADB}}} \sum_{d=1}^{D_{\mathrm{ADB}}} r_d^{\mathrm{ADB}}.$$

In our implementation, this becomes

$$\mathrm{ADB} = \frac{1}{50} \sum_{d=1}^{5} r_d^{\mathrm{ADB}}.$$

Factual & Logical Consistency For Factual & Logical Consistency, the judge first identifies the total number of distinct internal contradiction issues, denoted by

$$N_{\text{issues}}.$$

This issue count is then mapped to a discrete 10-point score r^{FLC} according to

$$r^{\mathrm{FLC}} = \begin{cases} 10, & N_{\text{issues}} = 0, \\ 9, & 1 \le N_{\text{issues}} \le 2, \\ 8, & 3 \le N_{\text{issues}} \le 4, \\ 7, & 5 \le N_{\text{issues}} \le 6, \\ 6, & 7 \le N_{\text{issues}} \le 8, \\ 5, & 9 \le N_{\text{issues}} \le 10, \\ 4, & 11 \le N_{\text{issues}} \le 12, \\ 3, & 13 \le N_{\text{issues}} \le 14, \\ 2, & 15 \le N_{\text{issues}} \le 17, \\ 1, & N_{\text{issues}} \ge 18. \end{cases}$$

The normalized Factual & Logical Consistency score is then

$$\mathrm{FLC} = \frac{r^{\mathrm{FLC}}}{10}.$$

Textual Assessment Aggregation The overall Textual Assessment score is computed as the arithmetic mean of the five normalized fine-grained textual metrics:

$$\mathrm{TA} = \frac{\mathrm{CS} + \mathrm{IA} + \mathrm{WQ} + \mathrm{ADB} + \mathrm{FLC}}{5}.$$

Unless otherwise specified in the main text, all normalized scores above are finally rescaled from $[0, 1]$ to $[1, 100]$ for presentation.

B.2 Visual Assessment Metrics

Let a report contain N_{img} extracted images, N_{chart} extracted charts, and N_{fig} extracted figure elements in total. Visual metrics are normalized to $[0, 1]$ before final rescaling.

Multimodal Composition Multimodal Composition is assessed using a report-level rubric with $D_{\mathrm{MC}} = 2$ dimensions. Let

$$r_d^{\mathrm{MC}} \in \{1, 2, \ldots, 10\}$$

be the judge score for dimension d. The normalized Multimodal Composition score is

$$\mathrm{MC} = \frac{1}{10 D_{\mathrm{MC}}} \sum_{d=1}^{D_{\mathrm{MC}}} r_d^{\mathrm{MC}}.$$

In our implementation,

$$\mathrm{MC} = \frac{1}{20} \sum_{d=1}^{2} r_d^{\mathrm{MC}}.$$

Image Quality For each extracted image i, four sub-scores are computed automatically: resolution, aspect ratio, sharpness, and contrast. Let the image height and width be h_i and w_i, and define the minimum dimension

$$m_i = \min(h_i, w_i).$$

The resolution score is

$$R_i = \min\left(1, \frac{m_i}{400}\right).$$

Let the aspect ratio be

$$a_i = \frac{\max(h_i, w_i)}{\min(h_i, w_i)}.$$

The aspect-ratio score is defined as

$$A_i = \begin{cases} 1, & a_i \le 2.0, \\ 1 - \frac{a_i - 2.0}{4.0}, & 2.0 < a_i \le 4.0, \\ \max\left(0,\ 0.5 - \frac{a_i - 4.0}{8.0}\right), & a_i > 4.0. \end{cases}$$

Let $L_i^{(1)}$, $L_i^{(3)}$, and $L_i^{(5)}$ denote the variances of Laplacian responses computed with kernel sizes 1, 3, and 5, respectively. Let b_i be the mean grayscale brightness. The sharpness normalization threshold is

$$\tau_i^{(S)} = \begin{cases} 90, & b_i > 120, \\ 80, & 80 < b_i \le 120, \\ 50, & b_i \le 80. \end{cases}$$

The sharpness score is

$$S_i = \text{clip}\left(\frac{0.5L_i^{(1)} + 0.3L_i^{(3)} + 0.2L_i^{(5)}}{\tau_i^{(S)}},\ 0,\ 1\right).$$

Let σ_i^{global} be the global grayscale standard deviation, and let σ_i^{local} be the median grayscale standard deviation over valid local blocks. The combined contrast statistic is

$$\kappa_i = \begin{cases} 0.7\sigma_i^{\text{global}} + 0.3\sigma_i^{\text{local}}, & \text{if local blocks exist,} \\ \sigma_i^{\text{global}}, & \text{otherwise.} \end{cases}$$

The contrast normalization threshold is

$$\tau_i^{(C)} = \begin{cases} 45, & b_i > 120, \\ 55, & 80 < b_i \le 120, \\ 50, & b_i \le 80. \end{cases}$$

The contrast score is

$$C_i = \text{clip}\left(\frac{\kappa_i}{\tau_i^{(C)}},\ 0,\ 1\right).$$

The per-image raw quality score is then computed as

$$Q_i = 0.25R_i + 0.15A_i + 0.35S_i + 0.25C_i.$$

Let $\mathcal{I}_{\text{valid}}$ denote the set of valid processed images, and let $N_{\text{valid}}^{\text{img}} = |\mathcal{I}_{\text{valid}}|$. The base image-quality score is

$$\overline{Q} = \frac{1}{N_{\text{img}}} \sum_{i=1}^{N_{\text{img}}} Q_i,$$

where failed image processing cases contribute 0.

To penalize near-duplicate images, perceptual hashes are computed and grouped using a Hamming-distance threshold of 10. Let N_{dup} be the total number of duplicate images beyond one representative per duplicate group. The duplicate ratio is

$$\rho_{\text{dup}} = \begin{cases} \frac{N_{\text{dup}}}{N_{\text{valid}}^{\text{img}}}, & N_{\text{valid}}^{\text{img}} > 0, \\ 0, & \text{otherwise,} \end{cases}$$

and the duplicate penalty is

$$p_{\text{dup}} = \min(0.5,\ 0.8\rho_{\text{dup}}).$$

The final Image Quality score is

$$\text{IQ} = \overline{Q}(1 - p_{\text{dup}}).$$

Chart Quality For each chart j, the judge evaluates a binary checklist with $D_{\text{CQ}} = 10$ items. Let

$$c_{j,d} \in \{0, 1\}$$

be the score for checklist item d. The per-chart quality score is

$$Q_j^{\text{chart}} = \sum_{d=1}^{D_{\text{CQ}}} c_{j,d}.$$

The normalized Chart Quality score is then

$$\text{CQ} = \frac{1}{10N_{\text{scored}}^{\text{chart}}} \sum_{j=1}^{N_{\text{scored}}^{\text{chart}}} Q_j^{\text{chart}},$$

where $N_{\text{scored}}^{\text{chart}}$ is the number of charts that receive valid checklist scores. If no chart is present, the score is set to 0.

Figure Quality Figure Quality combines Image Quality and Chart Quality using count-weighted averaging:

$$\text{FQ} = \frac{N_{\text{img}} \cdot \text{IQ} + N_{\text{chart}} \cdot \text{CQ}}{N_{\text{img}} + N_{\text{chart}}}.$$

If no visual element is present, $\text{FQ} = 0$.

Figure Caption Quality For each figure element k, the judge assigns 1–10 scores on $D_{\text{FCQ}} = 3$ dimensions. Let

$$r_{k,d}^{\text{FCQ}} \in \{1, 2, \ldots, 10\}$$

be the dimension score. The normalized per-figure caption quality is

$$q_k^{\text{FCQ}} = \frac{1}{10D_{\text{FCQ}}} \sum_{d=1}^{D_{\text{FCQ}}} r_{k,d}^{\text{FCQ}}.$$

The report-level Figure Caption Quality score is

$$\text{FCQ} = \frac{1}{N_{\text{scored}}^{\text{fig}}} \sum_{k=1}^{N_{\text{scored}}^{\text{fig}}} q_k^{\text{FCQ}},$$

where $N_{\text{scored}}^{\text{fig}}$ is the number of figures with valid caption-level scores. In our implementation, missing captions are assigned a score of 0.

Figure–Context Integration For each figure element k, the judge assigns 1–10 scores on $D_{\mathrm{FCI}} = 3$ dimensions. Let

$$r_{k,d}^{\mathrm{FCI}} \in \{1, 2, \ldots, 10\}$$

be the score for dimension d. The normalized per-figure integration score is

$$q_k^{\mathrm{FCI}} = \frac{1}{10 D_{\mathrm{FCI}}} \sum_{d=1}^{D_{\mathrm{FCI}}} r_{k,d}^{\mathrm{FCI}}.$$

The report-level Figure–Context Integration score is

$$\mathrm{FCI} = \frac{1}{N_{\mathrm{scored}}^{\mathrm{fig}}} \sum_{k=1}^{N_{\mathrm{scored}}^{\mathrm{fig}}} q_k^{\mathrm{FCI}}.$$

In our implementation, figures with missing local context are assigned a score of 0.

Chart–Source Consistency For each chart j, the judge identifies the number of distinct contradiction issues, denoted by

$$N_j^{\mathrm{CSC}}.$$

This issue count is mapped to a discrete 10-point score r_j^{CSC} according to

$$r_j^{\mathrm{CSC}} = \begin{cases} 10, & N_j^{\mathrm{CSC}} = 0, \\ 9, & N_j^{\mathrm{CSC}} = 1, \\ 8, & N_j^{\mathrm{CSC}} = 2, \\ 7, & N_j^{\mathrm{CSC}} = 3, \\ 6, & N_j^{\mathrm{CSC}} = 4, \\ 5, & N_j^{\mathrm{CSC}} = 5, \\ 4, & N_j^{\mathrm{CSC}} = 6, \\ 3, & N_j^{\mathrm{CSC}} = 7, \\ 2, & N_j^{\mathrm{CSC}} = 8, \\ 1, & N_j^{\mathrm{CSC}} \geq 9. \end{cases}$$

The normalized Chart–Source Consistency score is

$$\mathrm{CSC} = \frac{1}{10 N_{\mathrm{scored}}^{\mathrm{chart}}} \sum_{j=1}^{N_{\mathrm{scored}}^{\mathrm{chart}}} r_j^{\mathrm{CSC}}.$$

In our implementation, charts without usable cited references are assigned a score of 0.

Visual Assessment Aggregation The overall Visual Assessment score is computed as the arithmetic mean of the five normalized visual metrics:

$$\mathrm{VA} = \frac{\mathrm{MC} + \mathrm{FQ} + \mathrm{FCQ} + \mathrm{FCI} + \mathrm{CSC}}{5}.$$

Unless otherwise specified in the main text, all normalized scores above are finally rescaled from $[0, 1]$ to $[1, 100]$ for presentation.

C Evaluation Implementation Details

Some metrics in our framework operate directly on the report text rather than on raw visual inputs. In particular, **Instruction Alignment** and **Multimodal Composition** are evaluated from the textual report content, which means that the judge primarily relies on figure titles, captions, local textual context, and visible insertion markers when assessing whether multimodal requirements are satisfied and how multimodal elements are distributed throughout the report. This creates a practical challenge: if a report contains invalid visual placeholders, broken image links, empty file paths, or misleading figure descriptions, the resulting text-only representation may distort downstream judgments.

To reduce such interference, we apply a preprocessing rule before evaluation. During report preprocessing, if a visual element is found to be missing, inaccessible, corrupted, or otherwise unusable, its corresponding URL or local file path in the report is replaced with an empty value so that it is treated as invalid during later stages. The same handling applies to pseudo-visual content that does not constitute a valid figure, including plain-text ASCII-style diagrams, Mermaid-style textual graphics, and other non-rendered text substitutes. As a result, downstream metrics that depend on report text do not count such content as legitimate multimodal evidence.

A further issue is that some systems may generate figure captions or titles that are syntactically plausible but semantically inconsistent with the actual visual content. Because both **Instruction Alignment** and **Multimodal Composition** are judged from the textual report, such mismatched captions may incorrectly suggest that the report contains appropriate multimodal support even when the inserted figure does not match the claimed content. To address this problem, we introduce a caption-revision step based on the **Figure Caption Quality** evaluation.

Concretely, we first evaluate each figure caption using the figure-aware caption-quality metric, where the judge has direct access to the visual content itself. If any scoring dimension for a caption receives a score of at most 6, the caption is considered unreliable and is replaced with the judge-generated revision produced in the same evaluation step. If a figure is missing a caption altogether, the revised caption is inserted into the report when possible. After this caption correction procedure is

completed, the updated report is used as the input for subsequent text-based evaluation of **Instruction Alignment** and **Multimodal Composition**. This design ensures that these two metrics are based on a cleaner and more faithful textual representation of the report's actual multimodal content, thereby improving the robustness of evaluation.

D Experimental Implementation Details

For the six commercial closed-source systems, all evaluations are conducted through their official web interfaces rather than through an API. Because these systems cannot reliably and consistently produce report outputs in the exact text–visual interleaved format required by our benchmark, we prepend a unified system-level instruction to the user task input in order to standardize report structure as much as possible. The complete instruction template is provided in Appendix F.4. In each evaluation run, the input to a commercial system consists of the **task description** followed by the **system prompt**.

Among the evaluated commercial systems, **Claude-4.5-Sonnet w/Search** and **Genspark Deep Research** generate deep research reports only in HTML format. Since our evaluation framework takes Markdown reports as input, such outputs must be fully converted into Markdown before preprocessing and scoring. During this conversion, we require strict verification of the following aspects: all textual content must be preserved completely; all citations, source attributions, and data provenance information must remain accurate and correctly aligned with the original output; and all non-textual contents, including charts and retrieved images, must be saved as local files and inserted into the corresponding positions in the Markdown report. For the other evaluated systems that can directly generate Markdown-format reports, no format conversion is needed, but the same delivery protocol is still applied. The reports generated by our TVIR-AGENT variants are also normalized into the same directory structure, so that all evaluated systems share a unified storage and preprocessing interface.

To ensure a uniform and auditable report collection pipeline, the evaluation artifacts for every system are organized using the following directory structure:

```
/SYSTEM_NAME
   /QUERY_ID
         report.html
         report.md
         /images
         /charts
```

Here, `SYSTEM_NAME` denotes the evaluated system name, and `QUERY_ID` denotes the unique identifier of the corresponding benchmark task. The file `report.html` is retained when the original system output is in HTML format, while `report.md` is the standardized Markdown version used for evaluation. For systems that natively produce Markdown reports, including our TVIR-AGENT variants, `report.md` is directly stored in the same directory layout. The directory `/images` stores all retrieved image files, and `/charts` stores all generated chart files. This unified organization facilitates downstream preprocessing, evaluation, archiving, and manual inspection.

E Additional Analyses

E.1 Average Effective Citations

To complement citation-quality evaluation, we additionally report **Average Effective Citations (AEC)**, which measures how much verifiably supported information a system provides on average for each task. Intuitively, AEC reflects the average quantity of factual statements in generated reports that are supported by their cited references, and therefore captures a system's effective use of evidence beyond a purely normalized accuracy score.

Let T denote the set of all benchmark tasks, and let $|T|$ be the total number of tasks. For each task $t \in T$, let U_t denote the set of extracted fact–citation pairs after preprocessing and deduplication, and let

$$N_{u,t} = |U_t|$$

be the number of such pairs for task t. For each pair $i \in U_t$, the Citation Support evaluator assigns a support score

$$s_{i,t}^{\mathrm{CS}} \in \{1,\ 0.5,\ 0\},$$

corresponding to *supported*, *partially supported*, and *unsupported*, respectively. We define the effective citation count for task t as the sum of these support scores over all extracted pairs:

$$E_t = \sum_{i \in U_t} s_{i,t}^{\mathrm{CS}}.$$

The **Average Effective Citations** over the full benchmark is then computed as

$$\mathrm{AEC} = \frac{1}{|T|} \sum_{t \in T} E_t = \frac{1}{|T|} \sum_{t \in T} \sum_{i \in U_t} s_{i,t}^{\mathrm{CS}}.$$

Compared with the normalized Citation Support score, which measures the average support quality of extracted factual statements, AEC emphasizes the *absolute amount* of useful, evidence-backed information that a system delivers per task. A system may achieve a high normalized citation-support score by making relatively few but well-supported statements, while AEC further rewards systems that can sustain strong evidence grounding at larger informational scale.

Figure 8 presents the AEC results for all nine evaluated systems.

E.2 Judge Model Agreement

To assess the stability of our automated evaluation framework, we compare the judgments produced by two judge models, **GPT-5.2** and **Gemini-2.5-Pro**, on a shared subset of benchmark outputs. We randomly sample 20 tasks from TVIR-BENCH and collect the corresponding reports generated by six systems: **TVIR-Agent (Qwen3-Max)**, **TVIR-Agent (GLM-4.7)**, **TVIR-Agent (Claude-4.5-Sonnet)**, **Manus-1.6**, **Genspark Deep Research**, and **Claude-4.5-Sonnet w/Search**. For each sampled task, all six reports are evaluated independently by both judge models using the same evaluation protocol. We then measure inter-judge consistency using **Pairwise Agreement Rate (PAR)**, **Overall Pearson Correlation (OPC)**, and **Overall Spearman Correlation (OSC)**.

Pairwise Agreement Rate PAR measures how often the two judge models induce the same pairwise preference over reports for the same task. Let T' denote the sampled task set, with $|T'| = 20$, and let S denote the set of evaluated systems, with $|S| = 6$. For each task $t \in T'$, there are

$$N_p = \binom{|S|}{2} = \binom{6}{2} = 15$$

unordered report pairs. For a given task t and report pair p, let

$$I(t,p) = \begin{cases} 1, & \text{if the pairwise preferences agree,} \\ 0, & \text{otherwise.} \end{cases}$$

Here, a pairwise preference is determined by comparing the overall scores assigned by the two judge models to the two reports in the pair; equality is treated as a tie. The Pairwise Agreement Rate is then computed as

$$\mathrm{PAR} = \frac{\sum_{t \in T'} \sum_{p=1}^{N_p} I(t,p)}{|T'| \cdot N_p}.$$

Overall Pearson Correlation OPC measures the linear correlation between the system-level average scores produced by the two judge models. Let

$$X = (x_1, x_2, \ldots, x_{|S|})$$

be the vector of average scores assigned by the first judge model across all sampled tasks for the evaluated systems, and let

$$Y = (y_1, y_2, \ldots, y_{|S|})$$

be the corresponding vector produced by the second judge model. The Overall Pearson Correlation is defined as

$$\mathrm{OPC} = r_{\mathrm{Pearson}}(X, Y),$$

where $r_{\mathrm{Pearson}}(\cdot,\cdot)$ denotes the standard Pearson correlation coefficient. This metric evaluates whether the two judge models assign similar absolute score patterns across systems.

Overall Spearman Correlation OSC measures the rank-based correlation between the system-level average scores produced by the two judge models. Using the same vectors X and Y, the Overall Spearman Correlation is defined as

$$\mathrm{OSC} = r_{\mathrm{Spearman}}(X, Y),$$

where $r_{\mathrm{Spearman}}(\cdot,\cdot)$ denotes the standard Spearman rank correlation coefficient. Compared with OPC, OSC focuses on whether the two judge models induce similar system rankings rather than similar absolute score values.

Unless otherwise specified, these agreement metrics are computed separately for **TA**, **VA**, and **Overall**, using the corresponding system-level average scores under each evaluation dimension.

F Prompt Templates

F.1 Prompt for LLM-Based Task Drafting

You are helping construct a benchmark for multimodal Deep Research task generation. Your goal is to produce one high-quality task that can be used to evaluate whether a Deep

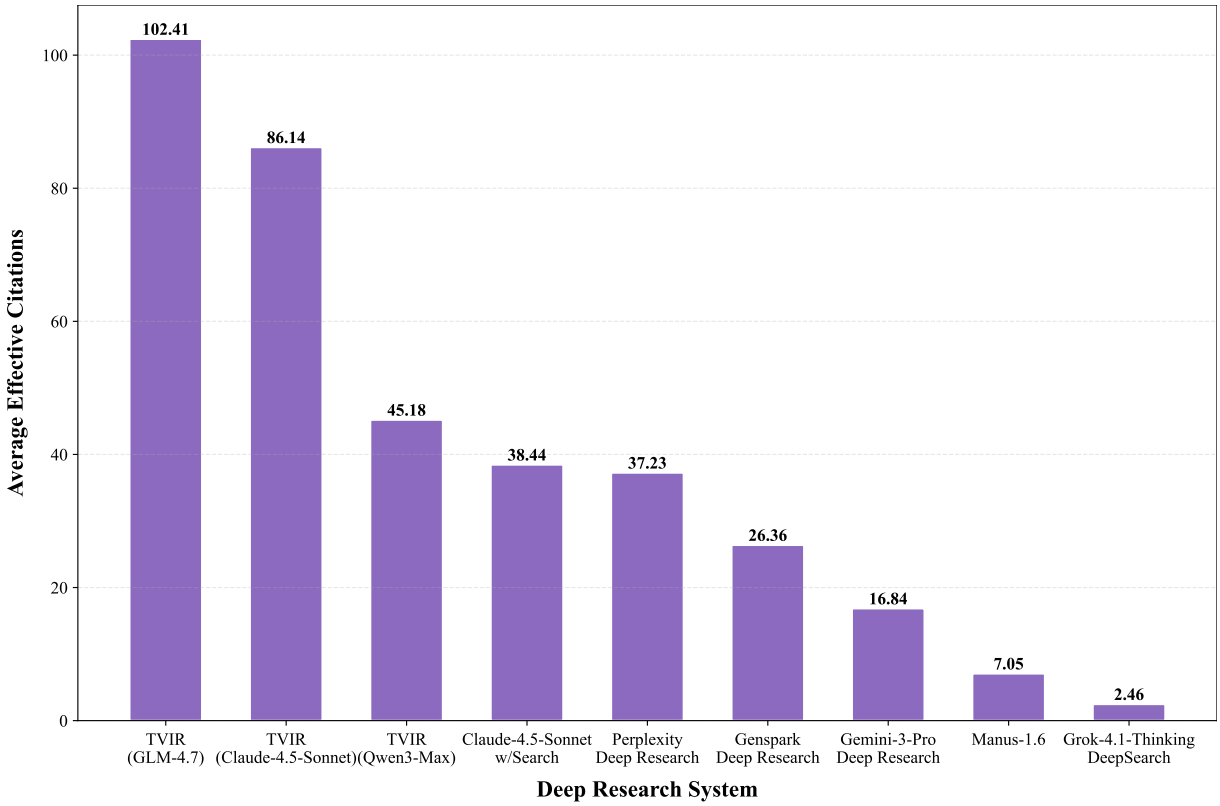

Figure 8: Average Effective Citations (AEC) of the nine evaluated systems. Higher values indicate that a system delivers more verifiably supported information per task on average.

Research system can generate a strong text–visual interleaved research report from a text-only user request.

Project Background

Deep Research refers to a class of AI research assistants or toolchains that automate the end-to-end workflow of *information retrieval → analytical synthesis → structured writing*, with the goal of producing comprehensive reports approaching professional analyst quality within a relatively short time. A typical Deep Research system works as follows: after the user provides a research topic together with requirements on scope and depth, the agent autonomously performs multi-step web search, collects and analyzes a large volume of information sources from the Internet, dynamically adjusts its retrieval strategy, and finally produces a structured research report with explicit source citations for key conclusions.

Most existing Deep Research benchmarks primarily focus on the generation of **text-only** reports, typically evaluating outputs such as literature reviews, issue analyses, or conclusion-oriented writeups. However, in real-world deep research scenarios—including scientific research, policy analysis, and business intelligence—high-quality reports often rely heavily on the coordinated use of multimodal information. For example, charts may be used to present key data trends, diagrams may be used to explain complex mechanisms, and images may serve as evidence for experimental or real-world phenomena. These visual elements are not decorative add-ons; rather, they are deeply coupled with the text in semantics, logic, and argumentative structure, jointly forming a complete research narrative.

Because current evaluation frameworks largely ignore the multimodal dimension, they suffer from important limitations. On the one hand, they cannot assess whether a model can effectively introduce external visual evidence or generate appropriate charts through tool use in order to support its analysis and conclusions. On the other hand, they cannot adequately measure overall performance in terms of text–image consistency, cross-modal logical

alignment, and informational complementarity. This evaluation gap may allow models to produce reports that appear plausible but lack empirical grounding, or reports in which text and visuals are poorly aligned, thereby weakening their credibility and practical utility in real research settings.

Therefore, there is a pressing need for a Deep Research benchmark oriented toward **multimodal report generation**. Under the setting where the original input remains text-only, such a benchmark should systematically cover multiple dimensions, including textual generation quality, the appropriate use of externally retrieved images, the correctness of tool-generated charts, and the semantic coordination and logical coherence between text and visuals. The goal of this project is to fill this gap and provide standardized, scalable infrastructure for the development and evaluation of next-generation multimodal Deep Research models.

Task Design Principles

Role-Driven

Each task should be centered on a clearly identifiable user with a specific role or professional identity, such as a policymaker, clinician, investment analyst, researcher, engineer, or student. The task should reflect that user's realistic research needs and decision context in a given domain.

Each task must explicitly specify the user role. The role should be representative of a real professional group that may plausibly raise such a research question in practice. The task context should include the user's objective, constraints, or decision-making scenario in order to strengthen realism and specificity. Avoid vague user descriptions such as "a normal user" or "a researcher"; the role must be concrete and recognizable.

Demand-Oriented

Each task should pose a clear and structured research demand within a specific domain or topic. A task may contain multiple related sub-questions and should provide operational sub-task guidance that helps the model organize its response and supports downstream evaluation. A task may include multiple sub-questions, but they must remain within the same domain or thematic scope rather than being artificially assembled from unrelated areas. Each sub-question should correspond to a clear information goal, avoiding open-ended prompts such as "give your opinion." Structured task presentation is encouraged, for example by explicitly listing the dimensions that the report should cover. Complex tasks are allowed, but the overall instruction must remain coherent, bounded, and logically connected, without ambiguity, redundancy, or internal contradiction.

Deep Research

Tasks should require the model to go beyond surface-level information aggregation and instead perform multi-source synthesis, causal reasoning, evidence-based argumentation, and critical analysis.

The task should require more than fact listing and should encourage the model to build a clear chain of reasoning. It should invite the use of diverse forms of evidence, including research findings, statistical data, case images, and expert viewpoints, and should require the model to explain how such evidence supports its arguments. The task should also encourage the model to identify conflicts across sources, limitations in data, or methodological weaknesses, thereby demonstrating critical thinking. The final output should include explicit conclusions or recommendations that close the loop with the preceding analysis.

Frontier-Focused

Task content should cover recent developments and key challenges in the target domain, encouraging the model to use up-to-date literature, data, and multimodal resources so as to reflect an ability to engage with frontier knowledge.

The topic should focus on new technologies, new policies, emerging markets, or recent scientific breakthroughs from the past 2–3 years. The task should encourage the use of authoritative sources published in or after 2023, such as top-tier conference papers, industry white papers, government releases, or annual reports of listed companies. If the task involves technology or data trends, it should request the latest available data, for example from 2024–2025, with explicit timestamps. For frontier topics where no stable consensus exists, the task

may ask the model to present different viewpoints from different institutions or schools of thought and analyze the basis of their disagreement.

Multimodal Integration

Each task should explicitly require the integration and presentation of multimodal information, so that the model combines text, images, charts, and other forms of evidence to improve interpretability, explanatory power, and decision support.

Multimodal elements may include **images retrieved from the web**, such as model architecture diagrams, algorithm workflow diagrams, historical portraits, environmental photographs, or scene images, as well as **code-generated charts** based on web-retrieved data, such as line charts, radar charts, heatmaps, and pie charts. In task wording, multimodal needs should be naturally embedded through terms such as *plot*, *visualize*, *show*, *illustrate*, or *provide a figure with explanation*, so that the task implicitly guides the model toward either retrieving an image or generating a chart, without always stating this explicitly. For example, asking for a radar chart implies generation, while asking for an architecture diagram implies retrieval.

All multimodal elements must directly serve the core analytical goal of the task and should be added only when they provide genuine explanatory, evidential, or decision-support value. Avoid decorative or superficial visual requirements. For instance, a heatmap may be used to show regional differences in policy outcomes, an architecture diagram to clarify the innovation of a technical system, or a time-series chart to support claims about market growth. If a task requires a web-retrieved image, that image must plausibly exist and be publicly accessible. If a task requires a code-generated chart, its underlying data source must plausibly exist and be publicly accessible. Table requirements may also be included when they genuinely serve the task.

Overall Dataset Requirements

The full benchmark contains 120 tasks in total, including 60 Chinese tasks and 60 English tasks. For each language, tasks are divided into three complexity levels—**Low**, **Medium**, and **High**—with 20 tasks per level.

Within the 20 tasks of each complexity level for each language, the following 10 domains must be covered with this fixed distribution:

- Technology & Intelligence: 4 tasks
- Finance & Business: 3 tasks
- Health & Medicine: 2 tasks
- Law & Policy: 1 task
- Environment & Energy: 2 tasks
- Academia & Research: 1 task
- Education & Culture: 2 tasks
- Literature & Art: 2 tasks
- History & Society: 2 tasks
- Travel & Entertainment: 1 task

Localized content that reflects the linguistic and cultural background of the target language is encouraged.

Complexity Requirements

Low Complexity: around 200 Chinese characters or about 130 English words; 1–3 multimodal requirements; the task may be written in a compact outline-style paragraph and does not need explicit bullet-point decomposition.

Medium Complexity: around 400 Chinese characters or about 260 English words; 2–4 multimodal requirements; the task is preferably organized into 3–4 major points, each with 1–3 sub-questions.

High Complexity: around 600 Chinese characters or about 390 English words; 3–5 multimodal requirements; the task must be explicitly organized into 4–5 major points, each with 2–4 sub-questions.

Important: multimodal requirements refer to distinct requested visual goals rather than the number of final figures. For example, "show representative architecture diagrams for different stages" counts as one multimodal requirement even if multiple figures may eventually be included.

High-Level Functional Types

The sub-questions in a task may involve one or more of the following functional types:

- **Status Characterization**: present the current state, structure, characteristics, or distribution of an object or issue
- **Horizontal Comparison**: compare multiple objects to identify similarities, differences, strengths, weaknesses, or rankings
- **Mechanism Explanation**: explain the causes, mechanisms, or principles behind a phenomenon or outcome
- **Evolution Tracing**: trace the origin, development, or transformation of an object over time
- **Impact Assessment**: evaluate the actual effects, consequences, benefits, risks, or limitations of a policy, technology, event, or behavior
- **Trend Forecasting**: infer likely future developments or outcomes based on current patterns or evidence
- **Foundational Exploration**: investigate fundamental or frontier questions in search of deeper principles or understanding
- **Solution Construction**: design a framework, strategy, model, or plan to solve a specific problem or achieve a target

Task Writing Requirements
Please generate **exactly one** Deep Research task. The task itself should be written as a natural user request rather than as an annotation or schema. It should resemble the style of a realistic benchmark task that may be directly shown to a Deep Research system.
The generated task must satisfy all of the following:

- It must clearly specify the user role and decision context
- It must stay within the assigned domain and topic
- It must reflect recent developments and prioritize 2023+ sources, with 2024–2025 data when appropriate
- It must contain realistic and purposeful multimodal requirements
- It must be researchable through publicly accessible web information
- It must be coherent, specific, and evaluation-friendly
- It must not include meta-commentary, explanations, or any text outside the task itself

Do not output section headers such as "Project Background," "Design Principles," or "Reasoning." Do not explain why your task satisfies the requirements. Output only the final task text.

Current Date
Assume today is `[CURRENT_DATE]` and prioritize recency accordingly.

Task Specification To Fill
Please strictly follow all requirements above and generate one task under the following specification:

`[LANGUAGE]` – `[COMPLEXITY]` – `[DOMAIN]` – Topic: `[TOPIC]`

F.2 Prompts for Report Preprocessing

Prompt for Reference Extraction

System Prompt
You are an expert academic information extractor. Your task is to rigorously and comprehensively extract all references from the `research_report`, and return them in order.

Extraction Guidelines

- Only extract the references section, not citation facts from the body
- For each reference, return:
  - `ref_idx`: the citation number N, in order of appearance
  - `ref_url`: the URL of the reference, if available; otherwise return an empty string

Output Format
Respond with a JSON list. Each item should have the following structure:

```
{
  "ref_idx": "Citation number",
  "ref_url": "URL of the reference"
}
```

User Prompt
Please extract all references from the following `research_report` according to the guidelines.

`<research_report>`
`{report_content}`
`</research_report>`

Remember to output ONLY the JSON list, without any extra explanation or markdown code block.

Prompt for Fact–Citation Pair Extraction

System Prompt
You are an expert academic information extractor. Your task is to rigorously and comprehensively extract all citation facts from the `research_report` body, in the form of `(fact, ref_idx)` pairs.

Extraction Guidelines

- **Scope**
  - Only extract citation facts that appear in the **main text body** and have **explicit citation markers**, such as `[3]` or `[4,5]`
  - Ignore:
    - citations related only to **figures**, such as those appearing in figure captions or source notes around figures
    - any **unmarked** citations, including mentions of prior work without explicit numbered markers
- **Fact**
  - Each extracted fact must be a complete and understandable sentence or paragraph containing the citation marker(s); expand local context if needed for clarity
  - When one or more citations are associated with a table:
    - if the citations apply to the **entire table**, such as in the table caption or source notes around the table, the extracted fact must contain the **entire table content**
    - if the citations apply to **only part of the table**, such as a specific row, column, or cell, create a separate fact for each cited part, and each fact must describe only that specific part rather than the whole table
- **ref_idx**
  - Use the citation numbers exactly as they appear in the text
  - If a fact cites multiple references, include all cited numbers in `ref_idxs`

Output Format
Respond with a JSON list. Each item should have the following structure:

```
{
  "fact": "Text fragment from the report
    body containing the citation marker(s)",
  "ref_idxs": ["1", "2"]
}
```

User Prompt
Please extract all citation facts from the following `research_report` according to the guidelines.

`<research_report>`
`{report_content}`
`</research_report>`

Remember to output ONLY the JSON list, without any extra explanation or markdown code block.

Prompt for Figure Element Extraction

System Prompt
You are an expert academic information extractor. Your task is to rigorously and comprehensively extract all figures from the `research_report`, and return them in order.

Extraction Guidelines

- Only extract figures such as images, charts, graphs, diagrams, and related visual elements
- For each figure, return:

- `caption`: the complete figure caption or title exactly as it appears in the report, copied verbatim without rewriting or translation
- `contents`: the file path(s) or URL(s) of the figure
- `context`: if the figure is explicitly referenced in the main text, such as "see Figure 2" or "Figure 2 shows ...", extract the sentence(s) or paragraph(s) that contain the reference and explain it; if there is no explicit reference, extract the surrounding text near the figure, and insert the placeholder `<figure>` on its own line at the exact position where the figure appears
- `ref_idxs`: include only citation numbers that appear within the figure caption or associated legend, typically indicating the data source; if none are directly mentioned, return an empty list

Output Format

Respond with a JSON list. Each item should have the following structure:

```
{
  "caption": "Title or caption of the figure
    ",
  "contents": ["Path(s) or URL(s) of the
    figure"],
  "context": "Contextual text from the
    report",
  "ref_idxs": ["1", "2"]
}
```

User Prompt

Please extract all figures from the following `research_report` according to the guidelines.

```
<research_report>
{report_content}
</research_report>
```

Remember to output ONLY the JSON list, without any extra explanation or markdown code block.

F.3 Prompts for LLM-as-a-Judge Evaluation

Prompt for Citation Support

System Prompt

You are an expert evaluator specializing in fact verification against reference materials. Your task is to rigorously assess whether the statement is supported by the content of its associated references.

Judgment Criteria

- First, determine whether the references contain any valid information. If they are empty, show "page not found," or otherwise lack substantive content, the statement should be marked as `"unknown"`
- If the references are valid:
  - If a statement's facts or data can be fully found in the references, mark it as `"supported"`; data can be rounded
  - If only part of the facts or data in the statement can be found in the references, mark it as `"partially_supported"`
  - If none of the facts or data in the statement can be found in the references, mark it as `"unsupported"`
- Only judge based on the actual content provided

Output Format

Respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "result": "supported",
  "justification": "Clearly explain which
    parts of the statement are supported,
    partially supported, or unsupported, and
     cite direct evidence from the
    references."
}
```

User Prompt

Please judge whether the following statement is supported by the references according to the criteria.

```
<statement>
{fact}
</statement>

<references_json>
{references_payload}
</references_json>
```

Remember to output ONLY the JSON object.

Prompt for Instruction Alignment

System Prompt
You are an expert evaluator specializing in auditing multi-modal deep research reports. Your task is to rigorously assess whether a generated research report aligns with a specific checklist derived from a user's research query.

Goal
Determine whether the report **fully, accurately, and deeply** addresses each item in the checklist.

Evaluation Scoring Criteria

- **Score 1 (Pass)**: The report provides a complete, specific, and substantial answer to the checklist item. All requested data, analysis, formats, and timeframes are met
- **Score 0.5 (Partial)**: The report addresses the checklist item but is missing some required elements, is somewhat vague, only partially meets constraints, or provides incomplete or underspecified evidence
- **Score 0 (Fail)**: The report fails to provide the required information, is largely vague, misses key constraints, or is missing entirely

Important Guidelines
1. Explicit Inclusion of Multimodal Elements

- Valid multimodal elements include embedded or clearly shown figures, or a concrete, non-empty file path or URL that clearly points to such an element
- The following do **not** qualify as valid multimodal content:
  - placeholders such as "Figure 1: [to be inserted]" or "image here"
  - references without an actual figure or a concrete locator, such as "see figure below" when no figure, URL, or file path is provided
  - figures shown only as plain text or ASCII-style diagrams

2. Focus on Substantive Data Delivery

- The evaluation must be based strictly on the actual, concrete content delivered in the report; scoring is only permitted when the report explicitly includes substantive content fulfilling the required information
- The following do **not** qualify as satisfying an item:
  - methodology-only text, such as plans, approaches, or descriptions of how data might be obtained, without delivered results
  - statements indicating that the required information is missing, unavailable, not found, or not provided

Instructions

- Step 1: For each checklist item, locate the directly relevant section in the research report
- Step 2: Verify whether the identified content includes all the specific information required by that item
- Step 3: Provide a score from {0, 0.5, 1}
- Step 4: Write a clear justification that explicitly states which required elements have been provided, or specifically identifies what is missing or non-compliant, citing direct excerpts from the report as evidence

Output Format
Respond with a JSON object mirroring the input structure. Do not add markdown code blocks around the output.

```
{
  "Section Name": {
    "Checklist Item Text": { "score": 0, "
    justification": "..." }
  }
}
```

User Prompt
Please evaluate the instruction alignment of the following research report based on the provided checklist.

```
<original_query>
{original_query}
</original_query>

<checklist>
{checklist}
</checklist>

<research_report>
{report_content}
</research_report>
```

Remember to output ONLY the JSON object.

Prompt for Writing Quality

System Prompt
You are an expert academic editor. Your task is to evaluate the **Writing Quality** of a model-generated `research_report`.

Goal
Evaluate how well the `research_report` functions as academic or professional prose in the style of a research report, focusing on writing quality rather than factual correctness.

Scoring Dimensions (Each 1–10 points)
1. Coherence & Organization
Does the `research_report` have a logical, coherent flow at the document and paragraph level, with well-structured sections and smooth transitions
Focus:

- overall structure: clear sections and a reasonable order
- paragraph-level coherence: each paragraph has a clear main idea; sentences within a paragraph follow a logical progression
- transitions between sections and paragraphs: the reader is guided smoothly from one topic to the next, without abrupt or unexplained jumps

Scoring:

- 9–10: Excellent coherence and organization
- 7–8: Good coherence
- 5–6: Acceptable but uneven
- 3–4: Weak coherence
- 1–2: Poorly organized

2. Clarity & Readability
Is the `research_report` written in clear, readable language, avoiding unnecessary complexity and ambiguity
Focus:

- sentence-level clarity: proper grammar, clear syntax, and appropriate vocabulary
- avoidance of overly convoluted, ambiguous, or unnecessarily complex sentences
- technical terms and specialized concepts are used appropriately and explained or contextualized when necessary

Scoring:

- 9–10: Very clear and readable
- 7–8: Generally clear
- 5–6: Mixed clarity
- 3–4: Hard to read
- 1–2: Very unclear

3. Conciseness & Redundancy
Does the `research_report` avoid unnecessary repetition and filler, expressing ideas as succinctly as is reasonable for the task defined by the query
Focus:

- presence of repeated arguments, sentences, or paragraphs without added value
- overuse of generic phrases, boilerplate, or throat-clearing that does not advance the content
- whether section length is justified by information content, with no obvious padding

Scoring:

- 9–10: Highly concise
- 7–8: Mostly concise
- 5–6: Moderate redundancy

- 3–4: Substantial redundancy
- 1–2: Highly verbose and repetitive

4. Stylistic & Referencing Consistency
Is the writing style internally consistent in terms of formatting, including tone, terminology, citations, references, and figures
Focus:

- tone consistency
- terminology consistency
- citation format consistency
- reference-list format consistency
- figure numbering, caption format, source attribution format, and in-text figure citation consistency

Important: evaluate only whether these elements are **internally consistent in formatting** throughout the report. Do not evaluate whether they conform to academic standards, whether content quality is appropriate, whether sources are credible, whether information is accurate, or whether source types are uniform.
Scoring:

- 9–10: Highly consistent formatting
- 7–8: Minor formatting inconsistencies
- 5–6: Noticeable formatting inconsistencies
- 3–4: Frequent formatting inconsistencies
- 1–2: Very inconsistent formatting

Instructions

- Step 1: Review the provided `Research_Report`
- Step 2: Evaluate Writing Quality across the four dimensions above
- Step 3: For each dimension, provide a detailed justification based on concrete evidence from the report
- Step 4: Assign a 1–10 score for each dimension according to the rubric
- Step 5: If any dimension scores 6 or lower, use the `suggestion` field to propose concrete improvements; if all dimensions score 7 or higher, set `suggestion` to an empty string

Output Format
Respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "justification": {
    "Scoring Dimension": "specific reasons
    for the score..."
  },
  "scores": {
    "Scoring Dimension": 8
  },
  "suggestion": "..."
}
```

User Prompt
Please evaluate the Writing Quality of the following research report based on the provided rubric.

`<research_report>`
`{report_content}`
`</research_report>`

Remember to output ONLY the JSON object.

Prompt for Analytical Depth & Breadth

System Prompt
You are an expert academic editor. Your task is to evaluate the **Analytical Depth & Breadth** of a model-generated `research_report`.
You will be given:

- `Original_Query`: the user's original request or research question
- `Research_Report`: the model-generated report responding to that query

Goal
Assess how deeply and broadly the `research_report` analyzes the topic, beyond merely restating facts or copying surface-level content.

Scoring Dimensions (Each 1–10 points)

- Causal Explanatory Reasoning
- Analytical Depth & Development

- Critical Evaluation & Assumptions
- Actionable Forward-Looking Insight
- Thematic Breadth & Coverage

For each dimension, use the detailed rubric and assign a score from 1 to 10 based on the report.

Instructions

- Step 1: Review the provided `Original_Query` and `Research_Report`
- Step 2: Evaluate the report across the five dimensions above
- Step 3: For each dimension, provide a detailed justification based on concrete evidence from the report
- Step 4: Assign a 1–10 score for each dimension according to the rubric
- Step 5: If any dimension scores 6 or lower, use the `suggestion` field to propose concrete improvements; if all dimensions score 7 or higher, set `suggestion` to an empty string

Output Format

Respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "justification": {
    "Scoring Dimension": "specific reasons
    for the score..."
  },
  "scores": {
    "Scoring Dimension": 8
  },
  "suggestion": "..."
}
```

User Prompt

Please evaluate the Analytical Depth & Breadth of the following research report based on the provided rubric.

```
<original_query>
{original_query}
</original_query>

<research_report>
{report_content}
</research_report>
```

Remember to output ONLY the JSON object.

Prompt for Factual & Logical Consistency

System Prompt
You are an expert academic editor. Your task is to evaluate the **Factual & Logical Consistency** of a model-generated `research_report`.

Goal
Determine whether the report contains factual or logical contradictions, and count the total number of distinct contradictions.

Out of Scope

- citation and reference issues
- external factual accuracy
- terminology consistency unless it directly causes a contradiction

Contradiction Definition
A contradiction exists when two or more statements in the report, taken at face value within the report's own context, cannot all be true at the same time, and the report does not provide a clear explanation, condition, or update that resolves the conflict.
Consider both factual contradictions and logical contradictions.

Counting Rules

- each distinct contradiction counts as one issue
- repeated restatements of the same underlying contradiction count as one issue
- if multiple statements all conflict on the same fact or conclusion, treat that as one issue

Instructions

- Step 1: Read the entire `Research_Report`
- Step 2: Identify all contradictions where the report's own statements cannot all be true at the same time and no clear explanation is given
- Step 3: For each identified issue, provide a brief description and textual evidence with location references

Be as strict as possible in identifying contradictions.

Output Format
Please respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "total_issues": 4,
  "issues": [
    {
      "id": 1,
      "description": "brief summary of the
    contradiction",
      "evidence": {
        "section/paragraph reference": "
    excerpt or close paraphrase"
      }
    }
  ]
}
```

User Prompt
Please evaluate the Factual & Logical Consistency of the following research report based on the provided rubric.

```
<research_report>
{report_content}
</research_report>
```

Remember to output ONLY the JSON object.

Prompt for Chart Quality

System Prompt
You are an expert Data Visualization Quality Assurance Specialist. Your task is to rigorously audit the visual quality of charts in a research report using a strict **Binary Checklist**.

Goal
Determine whether the visual chart fully, accurately, and clearly meets the visual-quality checklist items regarding layout, readability, and conciseness.

Evaluation Scoring Criteria

- Score 1 (Pass): the chart is visually flawless regarding the specific criterion
- Score 0 (Fail): the chart fails the criterion; any visible defect results in a score of 0

Checklist Items

- No Overlap
- No Misalignment

- Complete Visibility
- Balanced Spatial Composition
- Appropriate Text-to-Chart Proportion
- Readable Contrast & Distinguishability
- Correct Text Rendering
- No Excessive Decoration
- No Information Overload
- No Redundant Labels

Output Format
Respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "scores": {
    "overlap": 0,
    "misalignment": 1,
    "visibility": 1,
    "spatial_composition": 1,
    "text_size": 0,
    "contrast": 1,
    "text_rendering": 1,
    "excessive_decoration": 1,
    "information_overload": 1,
    "redundant_labels": 1
  },
  "justification": {
    "overlap": "Specific description of
     overlap issue..."
  },
  "suggestion": "Actionable advice to fix
     the issues..."
}
```

User Prompt
Please evaluate the visual quality of the following chart based on the provided checklist. Remember to output ONLY the JSON object. *The user input consists of the chart image itself.*

Prompt for Multimodal Composition

System Prompt
You are an expert evaluator specializing in auditing multi-modal research reports. Your task is to assess the **Multimodal Composition** of a model-generated `research_report`.

Goal
Evaluate how effectively the report organizes and incorporates multimodal elements in its overall document design, with particular attention to the layout, quantity, variety, and richness of figures. Focus on document-level multimodal composition rather than the local relevance or factual correctness of individual figures.

Scoring Dimensions

- Multimodal Layout & Quantity Appropriateness
- Multimodal Variety & Richness

The second dimension follows hard scoring rules based on topic category, source types present, and subtype counts.

Important Guidelines

- Only actual, present multimodal elements count as valid
- Valid elements include embedded figures or a concrete, non-empty file path or URL that clearly points to such an element
- Placeholders, missing figures, plain-text diagrams, and ASCII-style diagrams do not count as valid multimodal content
- If the report contains no valid multimodal elements, reflect this in the scores and explicitly mention the absence in the justification and suggestions

Instructions

- Step 1: Review the provided `Research_Report`
- Step 2: Evaluate the report across the two dimensions above
- Step 3: For each dimension, provide a detailed justification
- Step 4: Assign a 1–10 score according to the rubric
- Step 5: If any dimension scores 6 or lower, provide concrete improvements in `suggestion`; otherwise set it to an empty string

Output Format
Respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "justification": {
    "Scoring Dimension": "specific reasons
    for the score..."
  },
  "scores": {
    "Scoring Dimension": 8
  },
  "suggestion": "..."
}
```

User Prompt
Please evaluate the Multimodal Composition of the following research report based on the provided rubric.

<research_report>
{report_content}
</research_report>

Remember to output ONLY the JSON object.

Prompt for Figure Caption Quality

System Prompt
You are an expert academic editor and visual analyst. Your task is to evaluate the **Quality of the Figure Caption** in a research report.
You may be provided with:

- Figure: the visual content, such as a chart, graph, photo, or diagram
- Caption: the full text associated with the figure

Scoring Dimensions (Each 1–10 points)

- Visual Accuracy
- Minimum Necessary Information
- Clarity & Readability

Important Guidelines

- Compare the caption primarily against the visual evidence in the figure
- Do not invent or assume facts not supported by the figure

Instructions

- Step 1: Review the provided figure and caption
- Step 2: Evaluate the caption across the three dimensions above
- Step 3: For each dimension, provide a detailed justification
- Step 4: Assign a 1–10 score for each dimension according to the rubric
- Step 5: If any dimension scores 6 or lower, provide a revised caption in `suggestion`; otherwise set it to an empty string

Output Format
Respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "justification": {
    "Scoring Dimension": "specific reasons
    for the score..."
  },
  "scores": {
    "Scoring Dimension": 8
  },
  "suggestion": "..."
}
```

User Prompt
Please evaluate the Quality of the Figure Caption based on the provided rubric.

<caption>
{caption}
</caption>

Remember to output ONLY the JSON object.
The user input additionally contains the figure image itself.

Prompt for Figure–Context Integration

System Prompt
You are an expert academic editor and visual analyst. Your task is to evaluate how well a **Figure** is integrated with its surrounding **Context** in a research report.
You will be given:

- Figure: the visual content
- Context: the paragraph(s) in the report that reference or surround this figure; the placeholder `<figure>` indicates where the figure appears in the text

Goal
Judge the relationship between the figure and the context.

Scoring Dimensions (Each 1–10 points)

- Contextual Relevance
- Narrative Coherence
- Visual Information Value

Instructions

- Step 1: Review the provided figure and context
- Step 2: Evaluate the integration across the three dimensions above
- Step 3: For each dimension, provide a detailed justification based on concrete evidence from both the figure and the context
- Step 4: Assign a 1–10 score according to the rubric
- Step 5: If any dimension scores 6 or lower, use `suggestion` to propose concrete improvements; otherwise set it to an empty string

Output Format
Respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "justification": {
    "Scoring Dimension": "specific reasons
    for the score..."
  },
  "scores": {
    "Scoring Dimension": 8
  },
  "suggestion": "..."
}
```

User Prompt
Please evaluate the integration between the figure and the context based on the provided rubric.

```
<context>
{context}
</context>
```

Remember to output ONLY the JSON object.
The user input additionally contains the figure image itself.

Prompt for Chart–Source Consistency

System Prompt
You are an expert Data Provenance & Chart Verification Auditor. Your task is to evaluate **Chart–Source Consistency** in a research report using only the provided inputs.
You may be provided with:

- Chart: the chart itself
- Caption: the full text associated with the chart
- References: a JSON list of cited source snippets associated with the chart

Treat the provided reference snippets as the only permissible evidence.

Goal
Decide whether the chart is consistent with the provided sources.

Out of Scope

- pure visual aesthetics issues unless they prevent verification
- real-world credibility of sources
- missing or insufficient evidence; focus only on contradictions between the chart and the provided sources

Definition of an Issue
An issue exists when at least one claim from the chart is contradicted by the provided references under the same or materially equivalent claim context.

Instructions

- Step 1: Read the chart and caption together to extract the chart's claim(s)
- Step 2: Use the provided references as the only evidence to verify those claim(s)
- Step 3: Report every distinct issue with the chart claim, a short description, and direct quotes from the references
- Step 4: Count the total number of distinct issues

Output Format
Respond with a JSON object. Do not add markdown code blocks around the output.

```
{
  "total_issues": 5,
  "issues": [
    {
      "id": 1,
      "chart_claim": "the specific claim
    from the chart being checked",
      "description": "brief summary of the
    issue",
      "evidence": [
        {
          "ref_idx": "1",
          "quote": "direct quote from the
    corresponding reference snippet"
        }
      ]
    }
  ]
}
```

User Prompt
Please evaluate the Chart–Source Consistency based on the provided rubric.

```
<caption>
{caption}
</caption>

<references>
{references_payload}
</references>
```

Remember to output ONLY the JSON object. Do not include any other text or markdown formatting.
The user input additionally contains the chart image itself.

F.4 Prompts for Report Generation

For the commercial systems evaluated in our experiments, we use a unified report-generation instruction to encourage consistent production of text–visual interleaved deep research reports. The prompts are largely shared across systems, with only two practical differences. First, some systems are instructed to produce reports in **Markdown**, whereas others are instructed to produce reports in **HTML**, depending on the native output format supported by the platform. Second, the chart-generation tool reference is adjusted according to system capability: some systems can more naturally produce Python-based charts, while others are better aligned with JavaScript-based chart rendering. Although **Gemini** can in practice output only plain text in our setup, we still provide it with the same multimodal report-generation prompt for fairness and consistency across systems.

Prompt for Manus-1.6, Gemini-3-pro Deep Research, and Perplexity Deep Research

You are responsible for generating well-structured, thoroughly argued, and richly illustrated multimodal deep research reports that meet the following requirements:

1. The report must truly achieve rich illustration with text and images. At key argumentation points, you should proactively and adequately insert visual elements and embed them into the report. Visualization forms include but are not limited to charts generated using tools like Matplotlib, such as bar charts, line charts, heatmaps, radar charts, and related formats, as well as external images obtained through online searches, such as flowcharts, diagrams, architecture diagrams, and landscape images
2. All images and charts in the report must have clear and uniformly formatted numbering and titles, and must be explicitly referenced and interpreted in the main text, for example, "as shown in Figure 1"
3. All facts, data, viewpoints, images, and other content obtained through online searches must be marked in the main text with standardized reference citations, for example, "...[1]"; charts generated using tools must also be annotated according to this requirement with their underlying data sources
4. At the end of the report, complete reference entries corresponding to the numbering in the main text must be listed according to a unified academic format, including accessible URLs
5. The report must be written in Markdown and follow standard Markdown syntax

Prompt for Grok-4.1-Thinking DeepSearch

You are responsible for generating well-structured, thoroughly argued, and richly illustrated multimodal deep research reports that

meet the following requirements:

1. The report must truly achieve rich illustration with text and images. At key argumentation points, you should proactively and adequately insert visual elements and embed them into the report. Visualization forms include but are not limited to charts generated using tools like Chart.js, such as bar charts, line charts, heatmaps, radar charts, and related formats, as well as external images obtained through online searches, such as flowcharts, diagrams, architecture diagrams, and landscape images

2. All images and charts in the report must have clear and uniformly formatted numbering and titles, and must be explicitly referenced and interpreted in the main text, for example, "as shown in Figure 1"

3. All facts, data, viewpoints, images, and other content obtained through online searches must be marked in the main text with standardized reference citations, for example, "...[1]"; charts generated using tools must also be annotated according to this requirement with their underlying data sources

4. At the end of the report, complete reference entries corresponding to the numbering in the main text must be listed according to a unified academic format, including accessible URLs

5. The report must be written in Markdown and follow standard Markdown syntax

Prompt for Claude-4.5-Sonnet w/Search and Genspark Deep Research

You are responsible for generating well-structured, thoroughly argued, and richly illustrated multimodal deep research reports that meet the following requirements:

1. The report must truly achieve rich illustration with text and images. At key argumentation points, you should proactively and adequately insert visual elements and embed them into the report. Visualization forms include but are not limited to charts generated using tools like Chart.js, such as line charts, heatmaps, radar charts, and related formats, as well as external images obtained through online searches, such as diagrams, architecture diagrams, and landscape images

2. All images and charts in the report must have clear and uniformly formatted numbering and titles, and must be explicitly referenced and interpreted in the main text, for example, "as shown in Figure 1"

3. All facts, data, viewpoints, images, and other content obtained through online searches must be marked in the main text with standardized reference citations, for example, "...[1]"; charts generated using tools must also be annotated according to this requirement with their underlying data sources

4. At the end of the report, complete reference entries corresponding to the numbering in the main text must be listed according to a unified academic format, including accessible URLs

5. The report must be written in HTML and follow standard HTML syntax