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 that serves as a strong baseline for constructing 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.

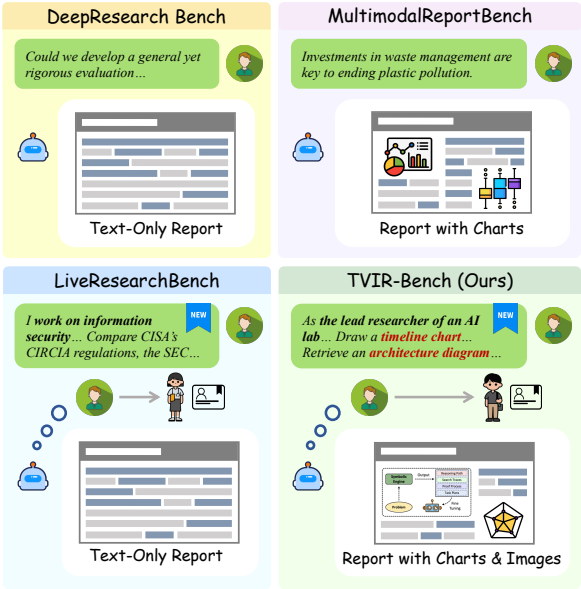

Figure 1: Comparison of representative deep research benchmarks. Existing benchmarks mainly focus on text-only or weakly multimodal reports, whereas TVIR-BENCH requires text–visual interleaved reports with semantically grounded charts and retrieved images.

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 baseline framework for 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.

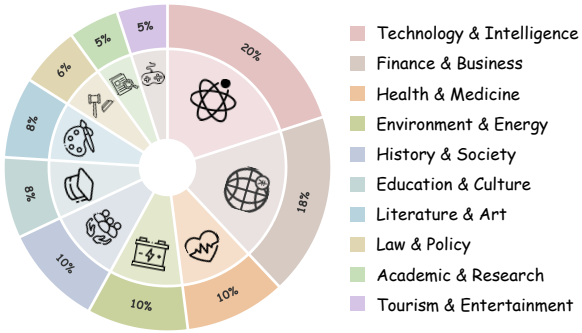

Figure 2: Domain taxonomy of TVIR-BENCH.

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. Further details are provided in Appendix A.1.

3.2 Data Construction

As shown in Figure 3(a), dataset construction follows an expert-driven workflow with four stages.

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 expert-proposed core topic, we guide Grok-4.1-Thinking to produce a task draft that is logically coherent, researchable, and practically meaningful. The detailed prompt is provided in Appendix G.1.

Multi-Expert Review and Revision Each draft is then examined by three 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 sub-questions 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

Figure 3(b) illustrates our 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.

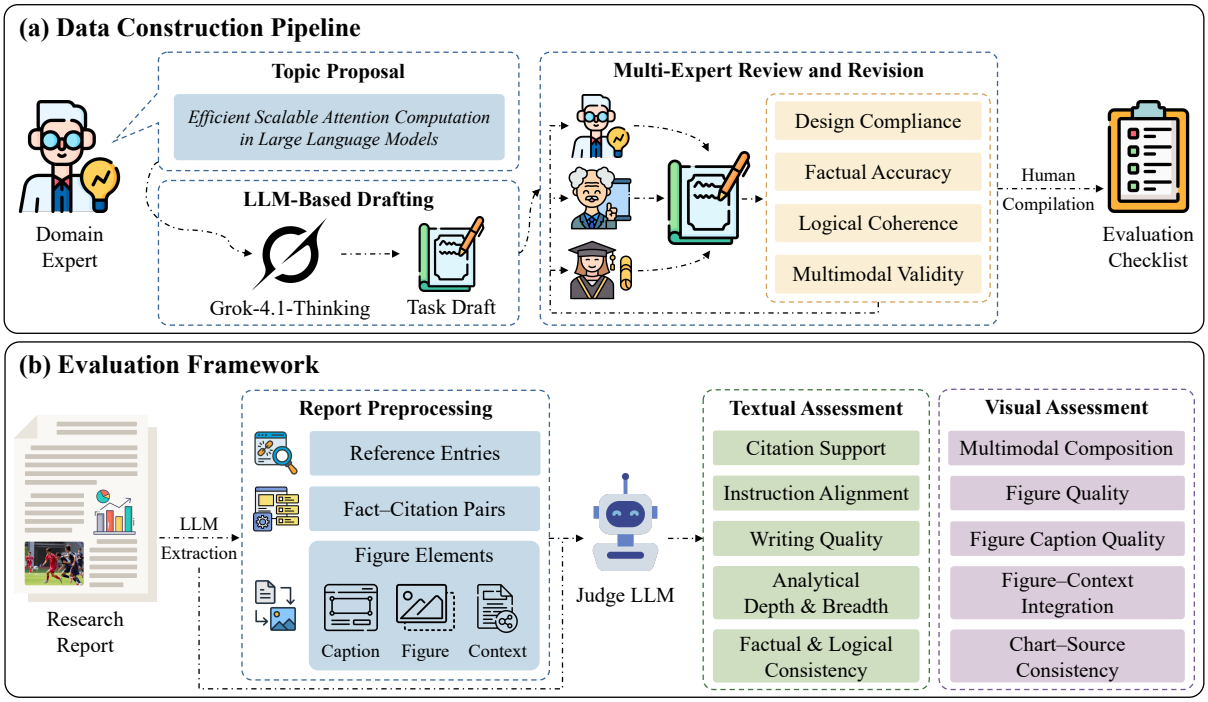

Figure 3: Overview of TVIR-BENCH, including data construction pipeline and evaluation framework.

See Appendix B for metric computation details and Appendix G.3 for evaluation prompts.

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 G.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**

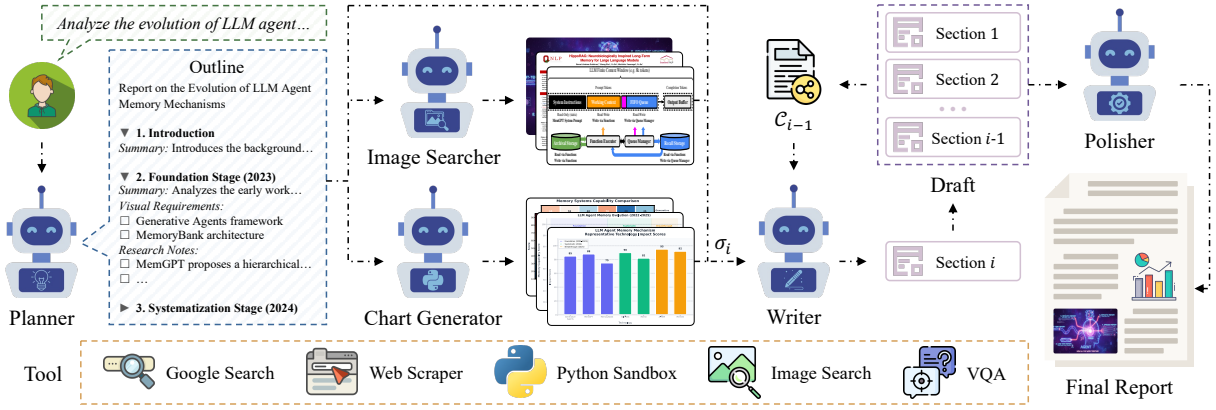

Figure 4: Overview of the proposed multi-stage framework for report generation.

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) for text–visual interleaved report generation. Figure 4 illustrates its overall architecture. Given a user task $\mathcal{T}$, TVIR-AGENT produces a high-quality research report $\mathcal{R}$ through four stages, which we introduce below.

4.1 Research-Grounded Planning

As the initialization module, the **Planner** parses the user task $\mathcal{T}$ and iteratively invokes external tools such as Google Search and web scraping to retrieve relevant information. It then synthesizes the collected information into a structured outline $\mathcal{O} = \{\sigma_1, \sigma_2, \ldots, \sigma_N\}$, where each outline unit σ_i contains a section title, a brief summary, planned visual requirements $\mathcal{V}_i^{\text{req}}$, and section-level research notes $\mathcal{N}_i$. Each note $n \in \mathcal{N}_i$ records a citation, source URL, and key findings. These notes provide factual grounding for subsequent stages and improve both credibility and traceability.

4.2 Visual Asset Instantiation

Given the planned outline $\mathcal{O}$, this stage instantiates the visual requirements of each section. To address different visual needs, we employ two specialized agents. The **Image Searcher** handles visual concepts such as portraits, scenes, and architecture diagrams by retrieving candidate images through Google Image Search, filtering low-quality results with heuristic rules, and using a visual question answering (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 outline unit σ_i is updated with instantiated visual assets $\mathcal{V}_i^{\text{inst}}$, together with captions, descriptions, and source provenance, yielding the augmented outline $\mathcal{O}^{\text{vis}}$. For code-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 the current outline unit σ_i and a dynamically updated global context $\mathcal{C}_{i-1}$, which consists 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 generation, the Writer determines insertion points for the instantiated visual assets in σ_i based on their descriptions and composes Markdown content with interleaved text and visual ele-

Model	Textual Assessment					Visual Assessment					Aggregate		
	CS	IA	WQ	ADB	FLC	FQ	MC	FCQ	FCI	CSC	**TA**	**VA**	**Overall**
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: Main results on TVIR-BENCH, averaged over three independent runs. **Bold** and underlined indicate the best and second-best performance on each dimension, respectively. Overall is computed as the mean of TA and VA.

ments, using local paths for code-generated 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 normalized content, 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.

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 (QwenTeam, 2025), GLM-4.7 (Zhipu AI, 2025), and Claude-4.5-Sonnet as backbone LLMs. More 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, with temperature set to 0.

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.

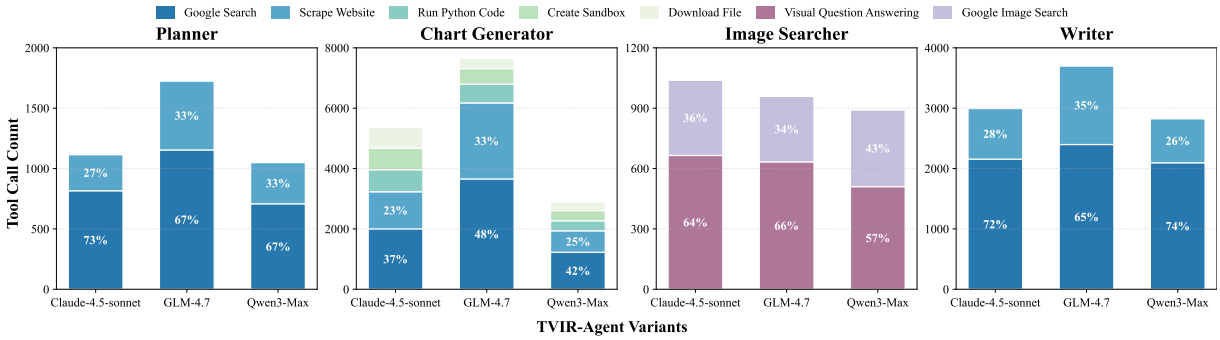

Figure 5: Tool usage distribution of TVIR-AGENT variants across major components.

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.

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 suggests that more complex tasks place heavier demands on multimodal coordination, 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 definition provided in Appendix E.2. 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**.

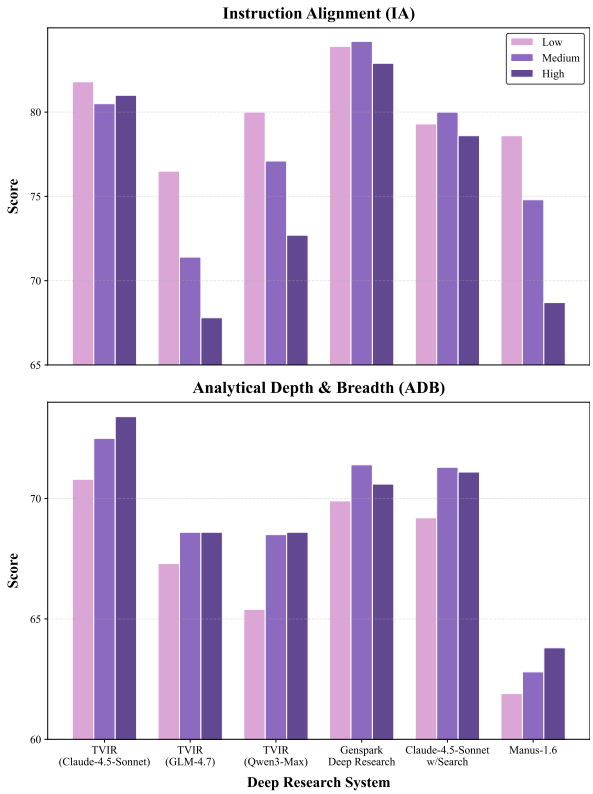

Figure 6: Performance comparison of IA and ADB across task complexity levels for TVIR-AGENT variants and leading commercial systems.

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 (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%**. 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.

Analysis on Human Alignment To assess the reliability of our automated evaluation, we conduct a large-scale human alignment study. We recruit 20 annotators with Master's degrees and relevant domain expertise, and each report is independently evaluated by three annotators. Following DeepResearch Bench (Du et al., 2025), we adopt **Pairwise Agreement Rate (PAR)** and **Overall Pearson Correlation (OPC)** as the primary metrics, and further report **Overall Spearman Correlation (OSC)** to capture rank-based consistency. Full details are provided in Appendix F. Table 10 shows strong agreement between automated evaluation and human judgments. We also report human inter-annotator agreement, with an Overall PAR of **74.20**.

Dimension	PAR	OPC	OSC
TA	70.00	99.12	97.62
VA	81.07	99.42	100.00
Overall	78.39	99.73	97.62

Table 2: Agreement between automated evaluation and human judgments.

Structural Error Analysis To further examine the reliability of 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. 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

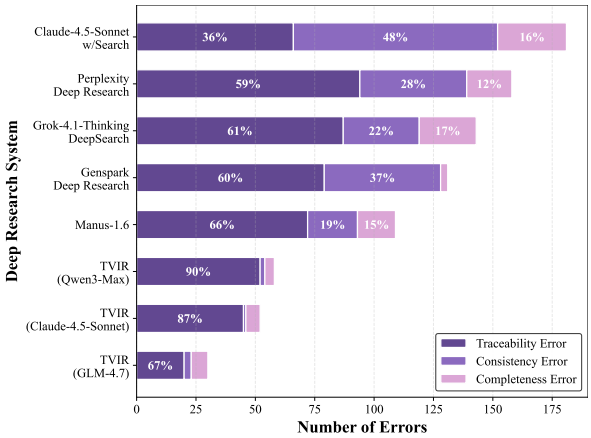

Figure 7: Distribution of structural errors across deep research systems.

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.

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

We introduce TVIR, a unified benchmark and agentic framework for text–visual interleaved report generation. It includes TVIR-BENCH, an expert-curated benchmark with textual and visual evaluation, and TVIR-AGENT, a hierarchical multi-agent framework that models visual evidence during planning and writing. Experiments show that TVIR-AGENT improves evidence grounding and cross-modal alignment, while revealing that current deep research systems remain much stronger at textual synthesis than at visual integration.

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.

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 Details

All fine-grained metrics are first computed on a normalized scale in $[0, 1]$ and then linearly rescaled to $[0, 100]$ for presentation.

B.1 Textual Assessment Metrics

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

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

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

corresponding to *supported*, *partially supported*, and *unsupported*, respectively. The Citation Support score is

$$\mathrm{CS} = \frac{1}{M} \sum_{i=1}^{M} s_i^{\mathrm{CS}}.$$

If no fact–citation pair is available, the score is set to 0.

Instruction Alignment For each checklist item, the judge assigns

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

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

$$\mathrm{IA} = \frac{1}{K} \sum_{j=1}^{K} s_j^{\mathrm{IA}}.$$

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 score is

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

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

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

Factual & Logical Consistency Let N_{issues} denote the number of distinct internal contradiction issues identified by the judge. This issue count is mapped to a discrete 10-point score r^{FLC}, which is then normalized as

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

N_{issues}	r^{FLC}
0	10
1–2	9
3–4	8
5–6	7
7–8	6
9–10	5
11–12	4
13–14	3
15–17	2
≥ 18	1

Table 4: Mapping from contradiction count to the FLC score.

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.

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

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

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

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

If no figure element is present, the score is set to 0.

Figure Quality Figure Quality combines **Image Quality (IQ)** and **Chart Quality (CQ)**.

For each image i, we compute four normalized sub-scores: resolution R_i, aspect ratio A_i, sharpness S_i, and contrast C_i. The raw image quality score is defined as

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

The weights are human-calibrated rather than manually assigned. Expert annotators rated extracted images on the same four sub-dimensions and an overall image quality score, each ranging from 1 to 10. We then fit a linear regression model from the four sub-scores to the overall rating and normalized the fitted coefficients to obtain the final weights. To penalize near-duplicate images, we compute a duplicate penalty p_{dup} based on perceptual-hash similarity. The Image Quality score is

$$\text{IQ} = \left(\frac{1}{N_{\text{img}}} \sum_{i=1}^{N_{\text{img}}} Q_i \right) (1 - p_{\text{dup}}).$$

If no image is present, the score is set to 0.

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 Chart Quality score is

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

If no chart is present, the score is set to 0.

The final Figure Quality score is the count-weighted average of IQ and CQ:

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

If no figure element is present, the score is set to 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 score is

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

The report-level Figure Caption Quality score is

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

If no figure element is present, the score is set to 0. 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 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

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

If no figure element is present, the score is set to 0. In our implementation, figures with missing context are assigned a score of 0.

Chart–Source Consistency For each chart j, let N_j^{CSC} denote the number of distinct contradiction issues identified by the judge. This issue count is mapped to a discrete 10-point score r_j^{CSC}, which is then normalized as

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

If no chart is present, the score is set to 0. In our implementation, charts without usable cited references are assigned a score of 0.

N_j^{CSC}	r_j^{CSC}
0	10
1	9
2	8
3	7
4	6
5	5
6	4
7	3
8	2
≥ 9	1

Table 5: Mapping from contradiction count to the CSC score.

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 G.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 Experimental Results

E.1 Variance Across Independent Runs

To assess the stability of system performance, we repeat each experiment three times under the same setup and report the standard deviation for all metrics. The detailed results are shown in Table 6. The standard deviations are consistently small across all evaluated systems, indicating strong run-to-run stability. In particular, the aggregate metrics TA, VA, and Overall exhibit very limited variation. These results suggest that the performance differences reported in the main paper are robust and unlikely to be driven by random variation across runs.

E.2 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 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. 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

Model	Textual Assessment					Visual Assessment					Aggregate		
	CS	IA	WQ	ADB	FLC	FQ	MC	FCQ	FCI	CSC	**TA**	**VA**	**Overall**
Commercial Closed-Source Systems													
Gemini-3-Pro Deep Research	0.002	0.002	0.003	0.004	0.003	-	-	-	-	-	0.002	-	-
Grok-4.1-Thinking DeepSearch	0.006	0.002	0.002	0.004	0.000	0.001	0.007	0.005	0.004	0.000	0.002	0.002	0.000
Claude-4.5-Sonnet w/Search	0.002	0.011	0.002	0.008	0.014	0.003	0.020	0.003	0.004	0.007	0.002	0.006	0.002
Perplexity Deep Research	0.002	0.003	0.007	0.006	0.006	0.000	0.005	0.001	0.004	0.003	0.005	0.001	0.003
Genspark Deep Research	0.003	0.010	0.001	0.006	0.003	0.002	0.008	0.004	0.002	0.014	0.001	0.004	0.002
Manus-1.6	0.006	0.005	0.002	0.004	0.005	0.005	0.007	0.002	0.001	0.001	0.003	0.000	0.001
TVIR-AGENT Variants (Ours)													
TVIR-Agent (Qwen3-Max)	0.002	0.015	0.001	0.003	0.011	0.002	0.005	0.008	0.005	0.018	0.004	0.003	0.001
TVIR-Agent (GLM-4.7)	0.003	0.006	0.002	0.006	0.009	0.001	0.011	0.004	0.006	0.010	0.002	0.004	0.003
TVIR-Agent (Claude-4.5-Sonnet)	0.002	0.001	0.005	0.002	0.000	0.006	0.004	0.003	0.001	0.004	0.001	0.001	0.001

Table 6: Standard deviations of the main results on TVIR-BENCH over three independent runs.

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.3 Performance by Language

We further analyze system performance on the Chinese and English subsets of TVIR-BENCH. The detailed results are reported in Table 7.

Overall, the results suggest a mild advantage on the Chinese subset. This trend is most consistent for textual assessment: all evaluated systems obtain higher TA scores on Chinese tasks than on English tasks. For VA and Overall, most systems also perform slightly better on the Chinese subset, although a few systems, such as Grok-4.1-Thinking DeepSearch, Perplexity Deep Research, and Genspark Deep Research, achieve comparable or stronger results on English tasks. Despite these differences, the performance gap is generally modest, and the relative ranking of systems remains broadly stable across the two languages.

One possible reason is that the Chinese and English tasks in TVIR-BENCH are not direct translations of each other. Instead, they are constructed with consideration of the cultural and real-world context associated with each language. As a result, the two subsets should be viewed as parallel benchmark slices rather than strictly matched pairs. The observed differences may therefore reflect not only language effects, but also mild differences in task composition and information environment. Importantly, the three TVIR-AGENT variants remain among the strongest systems on both subsets, indicating that the proposed framework generalizes well across the bilingual benchmark.

E.4 Performance by Domain

We also conduct a domain-stratified analysis on TVIR-BENCH. As shown in Table **??**, the TVIR-AGENT variants remain consistently strong across domains and achieve leading or near-leading performance in most cases, suggesting that the proposed framework generalizes well across a diverse set of long-form multimodal research tasks.

To further characterize domain-level difficulty, we use a ranking-based analysis rather than directly averaging scores, as ranking is less sensitive to unusually high or low scores from individual systems. Specifically, for each system, we rank all domains according to their Overall scores, and then compute, for each domain, the average of its ranks across all systems. A smaller average rank indicates that a domain tends to receive higher Overall scores across systems, while a larger average rank suggests that the domain is more challenging overall.

Table 9 shows that *History & Society* and *Ed-*

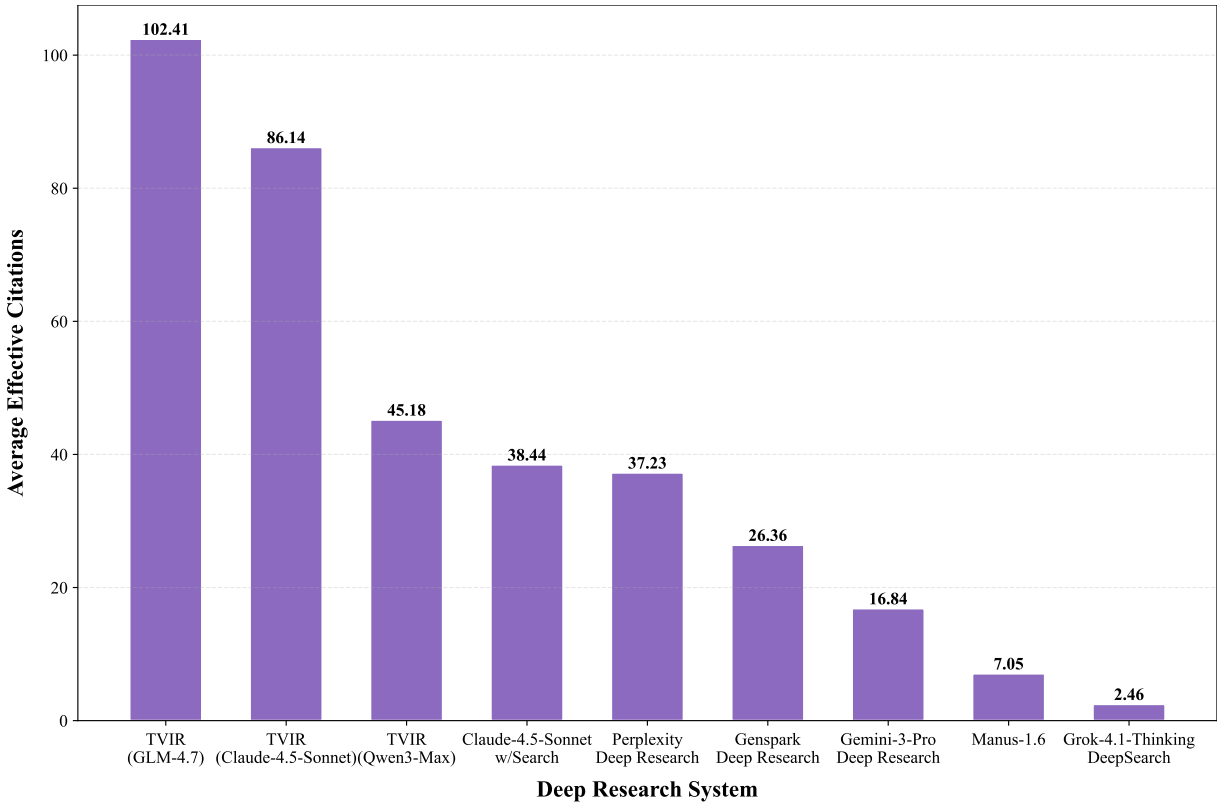

Figure 8: Average Effective Citations (AEC) of the nine evaluated systems.

ucation & Culture are relatively easier domains for current systems, whereas *Technology & Intelligence*, *Finance & Business*, and *Law & Policy* appear more challenging. This pattern is intuitively plausible: domains such as technology, finance, and policy often involve denser terminology, more rapidly evolving factual content, and stricter requirements for precise citation grounding, all of which increase the difficulty of long-form synthesis and text–visual interleaved report generation. By contrast, domains such as history, culture, and tourism often permit broader narrative organization and may place comparatively less pressure on highly technical grounding. We regard these findings as descriptive rather than causal, but they nevertheless provide useful insight into the cross-domain behavior of current deep research systems.

F Validation of the Evaluation Framework

F.1 Human Evaluation Protocol

To assess the reliability of our model-based evaluation, we conduct a large-scale human alignment study. Specifically, we evaluate reports from 8 systems on the full TVIR-BENCH benchmark of 100 tasks, excluding Gemini-3-Pro Deep Research because it generates text-only reports, making VA inapplicable and Overall unavailable. Each report is independently assessed by 3 expert annotators, and we recruit a total of 20 annotators with Master's degrees and relevant domain expertise. Because the reports are long, citation-grounded, and multimodal, the annotation process is highly time-intensive: assessing the reports from all 8 systems for a single query requires approximately one annotator-day. The full study takes about two weeks to complete.

F.2 Agreement Metrics

Pairwise Agreement Rate Pairwise Agreement Rate (PAR) measures how often two evaluators induce the same pairwise preference over reports for the same task. Let T denote the set of tasks, and let S denote the set of evaluated systems. For each task $t \in T$, there are

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

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

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

Model	TA (ZH)	TA (EN)	VA (ZH)	VA (EN)	Overall (ZH)	Overall (EN)
Commercial Closed-Source Systems						
Gemini-3-Pro Deep Research	59.64	57.39	–	–	–	–
Grok-4.1-Thinking DeepSearch	58.84	58.29	38.65	54.20	48.75	56.24
Claude-4.5-Sonnet w/Search	70.84	69.46	69.61	64.98	70.23	67.22
Perplexity Deep Research	69.13	68.78	52.90	54.02	61.01	61.40
Genspark Deep Research	69.29	68.11	64.20	66.37	66.75	67.24
Manus-1.6	71.10	67.74	71.82	68.27	71.46	68.00
TVIR-AGENT Variants (Ours)						
TVIR-Agent (Qwen3-Max)	71.01	69.04	77.24	76.82	74.13	72.93
TVIR-Agent (GLM-4.7)	72.91	70.36	74.41	72.81	73.66	71.59
TVIR-Agent (Claude-4.5-Sonnet)	70.60	69.63	79.64	77.88	75.12	73.76

Table 7: Performance on the Chinese and English subsets of TVIR-BENCH.

Model	Tech. & Intell.	Fin. & Bus.	Health & Med.	Law & Policy	Env. & Energy	Acad. & Res.	Edu. & Cult.	Lit. & Art	Hist. & Soc.	Tour. & Ent.
Commercial Closed-Source Systems										
Grok-4.1-Thinking DeepSearch	51.40	52.78	53.93	54.68	52.55	56.44	50.86	53.37	55.64	41.26
Claude-4.5-Sonnet w/Search	67.09	68.87	69.91	69.36	67.87	65.93	69.59	69.32	70.75	69.66
Perplexity Deep Research	62.07	59.95	59.03	61.80	61.21	61.58	62.45	60.41	62.99	61.23
Genspark Deep Research	66.21	66.39	66.97	65.88	68.83	67.46	68.35	65.51	67.34	69.05
Manus-1.6	68.48	70.72	68.51	65.68	72.06	73.58	63.13	70.60	74.10	70.39
TVIR-AGENT Variants (Ours)										
TVIR-Agent (Qwen3-Max)	73.38	71.70	73.55	74.34	73.83	71.57	75.63	75.09	73.52	75.20
TVIR-Agent (GLM-4.7)	73.44	70.86	74.28	72.11	74.82	67.66	74.33	75.27	69.42	72.96
TVIR-Agent (Claude-4.5-Sonnet)	73.16	75.27	73.73	73.45	74.34	75.40	75.63	74.95	74.90	74.70

Table 8: Domain-stratified Overall results on TVIR-BENCH.

Domain	Avg. Rank
History & Society	3.88
Education & Culture	4.12
Tourism & Entertainment	5.00
Environment & Energy	5.12
Literature & Art	5.25
Academic & Research	5.50
Health & Medicine	5.88
Law & Policy	6.25
Finance & Business	6.62
Technology & Intelligence	7.38

Table 9: Average domain ranks computed from Overall scores across all evaluated systems.

Here, a pairwise preference is determined by comparing the scores assigned by the two evaluators 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 Overall Pearson Correlation (OPC) measures the linear correlation between the system-level average scores produced by two evaluators. Let

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

be the vector of average scores assigned by the first evaluator 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 evaluator. 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 evaluators assign similar absolute score patterns across systems.

Overall Spearman Correlation Overall Spearman Correlation (OSC) measures the rank-based correlation between the system-level average scores produced by two evaluators. 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 evaluators induce similar system rankings rather than similar absolute score values.

F.3 Cross-Judge Consistency

To further examine the stability of the evaluation framework, we compare the results produced by two judge LLMs, GPT-5.2 and Gemini-2.5-Pro. As shown in table 10, the agreement remains consistently high across all three dimensions. In particular, both OPC and OSC are near-perfect for TA, VA, and Overall, while PAR also reaches a high level. These results provide strong evidence that the proposed evaluation framework is stable across judge LLM choices, not only for the aggregate Overall score but also for the textual and visual assessment dimensions separately.

Dimension	PAR	OPC	OSC
TA	81.96	99.42	92.86
VA	88.04	99.26	100.00
Overall	86.07	99.44	97.62

Table 10: Cross-judge consistency between GPT-5.2 and Gemini-2.5-Pro on 8 evaluated systems over the full TVIR-BENCH benchmark of 100 tasks.

F.4 Reliability of LLM-based Information Extraction

The preprocessing stage is a necessary component of our evaluation pipeline, as it extracts structured information for downstream assessment. In practice, we find that LLM-based extraction is substantially more robust than rule-based parsing for this task, since report formats vary widely across systems and rule-based methods are brittle under formatting inconsistency.

To validate the reliability of this stage, we manually annotate ground truth for 90 reports in total, including 10 reports per system across 9 systems. The annotation covers three extraction targets: (i) reference entries; (ii) fact–citation pairs; and (iii) figure elements. Table 11 reports the average extraction quality. The results show that LLM-based extraction is highly accurate across all three targets, with near-perfect precision, recall, and F1. This suggests that the preprocessing stage introduces negligible noise into downstream evaluation.

In addition, webpage retrieval is required because metrics such as Citation Support and Chart–Source Consistency depend on the actual content of cited webpages rather than URLs alone. We retrieve webpage text via the Serper API. Across the full evaluation set of 9×100 reports, the overall retrieval success rate is 96.53%, including cases where the source webpage itself is no longer available. In practice, this has only a limited impact on evaluation coverage.

Extraction Target	Recall	Precision	F1
Reference entries	100.00	100.00	100.00
Fact–citation pairs	99.20	99.55	99.35
Figure elements	100.00	100.00	100.00

Table 11: Extraction quality of the LLM-based preprocessing stage, evaluated on 90 manually annotated reports.

G Prompt Templates

G.1 Prompt for LLM-Based Task Drafting

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 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 weak-

nesses, 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]`

G.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.

G.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/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/paragraphs: the reader is guided smoothly from one topic to the next, without abrupt or unexplained jumps.

Scoring:

- **9–10 – Excellent coherence and organization**: The report's structure is clear and well-designed; sections and paragraphs are logically ordered. Transitions are smooth, the progression of ideas is easy to follow, and the overall reading experience feels cohesive from start to finish.
- **7–8 – Good coherence**: The overall structure is clear and sensible. There may be minor rough transitions or slightly uneven sections, but the progression of topics or ideas is consistently understandable and largely well ordered.
- **5–6 – Acceptable but uneven**: A recognizable structure exists, but some paragraphs or sections feel disjointed, misplaced, or weakly connected. Transitions are occasionally abrupt, and the reader sometimes has to work to reconstruct how one part leads to the next.
- **3–4 – Weak coherence**: Organization is confusing or inconsistent. Sections or paragraphs appear in a suboptimal order; topic shifts are abrupt; transitions are often missing. The overall flow is unclear and hard to follow.
- **1–2 – Poorly organized**: The report is largely incoherent in its sequencing. Top-

ics are mixed together without clear structure, transitions are absent, and the reader struggles to understand how the text is organized or how different parts relate to each other.

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**: Sentences are well-constructed and easy to understand. Technical terms are used appropriately and explained when necessary. The text is pleasant to read and accessible to the intended audience.
- **7–8 – Generally clear**: The text is understandable with only minor awkward phrasing or occasional dense passages. Overall readability is good.
- **5–6 – Mixed clarity**: The main ideas can be understood, but there are noticeable awkward sentences, occasional ambiguities, or sections that are harder to parse. The reader needs extra effort at times.
- **3–4 – Hard to read**: Frequent grammar or syntax issues, long and tangled sentences, or confusing wording. Clarity problems significantly hinder comprehension.
- **1–2 – Very unclear**: Serious language problems or extremely convoluted writing make the report difficult to understand or even follow.

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; no obvious padding or "water".

Scoring:

- **9–10 – Highly concise**: The report is tight and efficient: almost no redundancy or filler. Each paragraph contributes new information or perspective relevant to the query.
- **7–8 – Mostly concise**: Some minor repetition or slightly wordy passages, but overall the report is reasonably succinct and does not feel bloated.
- **5–6 – Moderate redundancy**: The report contains noticeable repeated ideas or verbose phrasing. While still usable, it could be significantly tightened without loss of content.
- **3–4 – Substantial redundancy**: Many points are repeated, or large portions add little beyond what has already been stated. The report feels padded or overly long relative to its substantive content.
- **1–2 – Highly verbose and repetitive**: Extensive redundancy and filler severely dilute the message. The key ideas are buried under unnecessary text.

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

- Tone: the level of formality is stable; the text does not jump needlessly between very informal and highly formal styles.
- Terminology: key terms, labels, and names for the same concept, variable, or method are used consistently.
- Citations: one coherent citation format is followed uniformly.
- References: the format of reference list entries is consistent.
- Figures:
  - Numbering format is consistent.
  - Caption format is consistent.
  - Source attribution format is consistent across all figures.
  - In-text citation format for figures is consistent.

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, whether the number of sources is uniform, or whether source types are uniform. Focus solely on consistency of formatting: Are the same things formatted the same way throughout?
Scoring:

- **9–10 – Highly consistent formatting**: Tone level, terminology spelling/capitalization, citation format, reference list format, and figure formatting are uniform throughout. All formatting choices follow a single consistent pattern.
- **7–8 – Minor formatting inconsistencies**: Overall formatting is consistent, with only occasional small deviations that do not significantly affect the professional appearance.
- **5–6 – Noticeable formatting inconsistencies**: There are several visible inconsistencies in formatting that reduce visual polish but do not prevent understanding.
- **3–4 – Frequent formatting inconsistencies**: The report often switches between different formatting styles for the same elements. Citation formats are mixed, terminology formatting is unstable, and figure formatting varies significantly. This is visually distracting and appears unprofessional.
- **1–2 – Very inconsistent formatting**: Formatting is chaotic with no clear standard. The same terms are formatted differently throughout, citation styles change arbitrarily, reference entries have no consistent format, and figure numbering/captions follow no pattern. The report appears visually disorganized and unprofessional.

INSTRUCTIONS

- **Step 1**: Review the provided **Research_Report**.
- **Step 2**: Evaluate the **Writing Quality** of the Research_Report across the four dimensions: **Coherence & Organization**, **Clarity & Readability**, **Conciseness & Redundancy**, and **Stylistic & Referencing Consistency**.
- **Step 3**: For each dimension, provide a **detailed justification** that examines concrete evidence from the report and explains the reasoning leading to the score.
- **Step 4**: Based on the justification, assign a **1–10 score** for each dimension according to the detailed scoring criteria.
- **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 the **suggestion** field as an empty string `""`.

OUTPUT FORMAT
Respond with a JSON object. Do not add markdown code blocks (like “‘`json`) 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:

1. **Original_Query**: The user's original request or research question.
2. **Research_Report**: The model-generated report responding to that query.

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

SCORING DIMENSIONS (Each 1–10 points)
1. Causal Explanatory Reasoning
To what extent does the report go beyond description and provide clear, causal or mechanism-oriented explanations for its main claims?
Focus:

- Does the report move beyond simply stating "what happened" to explain "why" and "how it happens"?
- Are there clear causal or mechanism links (e.g., linking causes, conditions, and outcomes), rather than a loose list of factors?
- Does it specify key causal pathways and the conditions under which they hold, rather than implying causality?

Scoring:

- **9–10 – Excellent causal explanatory reasoning**: The report consistently explains why and how key phenomena occur via explicit causal pathways and mechanisms, clarifying conditions/boundaries for major claims with minimal unexplained jumps.
- **7–8 – Strong but slightly uneven reasoning**: Most important points include plausible causal/mechanism explanations, though some claims lack clear pathways or boundary conditions.
- **5–6 – Moderate explanatory reasoning**: Some causal/mechanism explanation is present, but several key claims rely on generic "drivers" language without specifying how the mechanism operates or when it applies.
- **3–4 – Limited or shallow reasoning**: Explanations are mostly vague ("due to", "driven by") with little mechanism detail or conditionality; causal links are often implied rather than articulated.
- **1–2 – Very poor or purely descriptive**: The report is largely descriptive, with almost no attempts to explain why or how; causal/mechanism thinking is essentially absent.

2. Analytical Depth & Development
To what extent does the report develop its key points into sustained analysis rather than brief, list-like statements?
Focus:

- Are key claims unpacked with multi-step reasoning rather than asserted in a single step?
- Are the most important points developed into coherent, connected paragraphs instead of fragmented one-line bullets?

- Does the report prioritize depth on central points, with added length being substantive rather than repetitive?

Scoring:

- **9–10 – Excellent development**: Central claims are unpacked with clear multi-step reasoning; key sections show sustained, connected analysis with strong internal structure. Depth is concentrated on the most important points; added length is substantively analytical with minimal filler or repetition.
- **7–8 – Strong development with minor thin spots**: Most key points are developed beyond surface summary and include some multi-step reasoning. Structure is generally coherent, but a few sections remain somewhat compressed, list-like, or stop short of fully unpacking implications/logic.
- **5–6 – Mixed/uneven development**: The report alternates between some developed passages and many thin statements. Several important claims are only partially unpacked, with reasoning that is short, generic, or stops after one step. List-like presentation is common for core points.
- **3–4 – Shallow/list-like development**: Predominantly brief bullets or high-level summaries. Claims are often asserted with little to no reasoning, and important points lack sustained development or connected explanation.
- **1–2 – Very poor development**: Fragmentary and minimally elaborated. The report mostly consists of headings or one-liners with little analytical content and no meaningful development of key points.

3. Critical Evaluation & Assumptions
To what extent does the report critically examine its assumptions, limitations, and alternative viewpoints instead of treating its own perspective as unquestioned?
Focus:

- Does the report identify and discuss key assumptions underlying its analysis, models, or narratives?
- Does it acknowledge limitations in data, methods, or perspective, and consider how these might affect the reliability or scope of its conclusions?
- Does it engage with relevant alternative explanations, interpretations, or strategies, rather than presenting a single viewpoint as obviously correct?

Scoring:

- **9–10 – Excellent critical evaluation**: The report explicitly surfaces major assumptions and examines their plausibility. It discusses important limitations of data and methods and thoughtfully considers how these constrain the conclusions. Alternative viewpoints or explanations are actively engaged and weighed, showing strong self-critical awareness.
- **7–8 – Solid but incomplete critical reflection**: The report shows clear critical thinking: it notes some key assumptions, limitations, and alternative perspectives. However, not all important issues are explored in depth, and parts of the analysis still rely on relatively unexamined premises.
- **5–6 – Moderate critical evaluation**: There is some acknowledgement of assumptions or limitations, and the report may briefly mention alternative views. However, this is sporadic and not integrated into the core reasoning; most arguments still proceed as if their premises are mostly unquestioned.
- **3–4 – Limited critical reflection**: Assumptions and limitations are largely implicit or mentioned only superficially. Alternative explanations or viewpoints are barely considered, and the report mostly treats its own framing and data as straightforward and unproblematic.
- **1–2 – Very poor or absent critical evaluation**: The report presents its analysis as if it is fully self-evident, with virtually no discussion of assumptions, limitations, or alternatives. It shows little awareness

of potential biases, blind spots, or uncertainties.

4. Actionable Forward-Looking Insight
Does the report translate its analysis into meaningful, forward-looking implications or recommendations that help the reader understand what to do or watch for next?
Focus:

- Does the report derive concrete implications for action, decision-making, strategy, policy, or further work from its analysis?
- Are recommendations or takeaways specific, directional, and conditional (where appropriate), rather than vague or purely formulaic statements?
- Does the report consider future scenarios, risks, opportunities, or monitoring points, instead of stopping at a description of the current state?

Scoring:

- **9–10 – Excellent actionable and forward-looking insight**: The report offers clear, specific, and well-justified implications or recommendations that follow from the analysis. It highlights relevant future scenarios, risks, or opportunities and gives the reader a strong sense of "what to do or pay attention to next."
- **7–8 – Strong but somewhat limited insight**: There are meaningful, reasonably specific implications or recommendations, and some forward-looking reflection. However, coverage is not fully systematic: certain conclusions are left without actionable follow-through or future-oriented consideration.
- **5–6 – Moderate actionable insight**: The report contains some useful takeaways, but they are either unevenly developed, somewhat generic, or only loosely tied to the preceding analysis. Forward-looking elements are present but not deeply integrated.
- **3–4 – Limited or generic implications**: Implications and recommendations, if present, are mostly high-level, vague, or boilerplate with little specificity or prioritization. There is little sense of future scenarios or concrete next steps.
- **1–2 – Very poor or absent forward-looking content**: The report mostly describes or analyzes the current situation with almost no attempt to derive practical implications or future-oriented insights.

5. Thematic Breadth & Coverage
How comprehensively does the report cover the key aspects of its topic, showing a sufficiently broad and well-chosen thematic scope without obvious blind spots?
Focus:

- Does the report address the main dimensions that are clearly relevant to the core topic, rather than focusing narrowly on a single aspect?
- Does it reasonably extend beyond only the explicitly mentioned points in the original_query to include other clearly important, closely related angles?
- While remaining focused, does the report avoid obvious gaps where a major, relevant facet of the topic is missing or severely underdeveloped?

Scoring:

- **9–10 – Excellent thematic breadth and coverage**: The report covers the topic in a well-balanced, comprehensive way, addressing the major relevant dimensions with appropriate depth. It not only responds to the explicitly stated points in the original_query but also proactively includes other clearly important, closely related angles. This extension beyond the original_query feels natural and necessary rather than off-topic. There are no clear blind spots.
- **7–8 – Strong but slightly uneven coverage**: Most important aspects of the topic are covered, and the scope feels suitably

broad. The report does move beyond the exact wording of the original_query to touch on some additional, relevant facets, but one or two significant dimensions are treated briefly or somewhat underdeveloped. The overall breadth is good, with only minor gaps.

- **5–6 – Moderate thematic breadth**: The report addresses several key facets of the topic, but coverage is selective or uneven. It may mostly stick to the points explicitly mentioned in the original_query, with only limited extension to other relevant angles, or it may treat some dimensions in depth while leaving others only hinted at. Overall breadth is acceptable but leaves noticeable room for a more complete picture.
- **3–4 – Narrow or imbalanced coverage**: The report focuses heavily on a limited subset of relevant aspects and largely stays within the most obvious or explicitly stated parts of the original_query. It rarely brings in additional, closely related angles that would help complete the thematic picture. As a result, the scope feels constrained or skewed, and important dimensions of the topic are missing or severely underdeveloped.
- **1–2 – Very poor or severely narrow coverage**: The report's treatment of the topic is extremely narrow, fragmentary, or misaligned. It may restrict itself almost entirely to a small portion of what the original_query mentions, ignoring other clearly relevant aspects, and it does not attempt to go beyond the original_query's explicit wording in any meaningful way. The reader is left without a coherent overall understanding of the subject.

INSTRUCTIONS

- **Step 1**: Review the provided **Original_Query** and **Research_Report**.
- **Step 2**: Evaluate the **Analytical Depth & Breadth** of the Research_Report across the five dimensions: **Causal Explanatory Reasoning**, **Analytical Depth & Development**, **Critical Evaluation & Assumptions**, **Actionable Forward-Looking Insight**, and **Thematic Breadth & Coverage**.
- **Step 3**: For each dimension, provide a **detailed justification** that examines concrete evidence from the report and explains the reasoning leading to the score.
- **Step 4**: Based on the justification, assign a **1–10 score** for each dimension according to the detailed scoring criteria.
- **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 the **suggestion** field as an empty string `""`.

OUTPUT FORMAT

Respond with a JSON object. Do not add markdown code blocks (like "'`json`) 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

1. **Citation and reference issues**: Do NOT evaluate any problems related to citations or references. This includes inconsistencies in attribution, source identification, or provenance of methods/concepts/-data that arise from how citation markers are used or organized. **If a contradiction exists only because of citation or reference-list issues, it is OUT OF SCOPE.** Only identify contradictions that would still exist if all citation markers and reference entries were removed from the report.

2. **External factual accuracy**: Do NOT evaluate whether the report's statements are true in the real world; only check whether the report contradicts itself internally.

3. **Terminology consistency**: Do NOT evaluate whether key terms, labels, or names for the same concept, variable, or method are used consistently throughout the report; only when such terminological differences lead to a direct contradiction about the same underlying object or result should it be counted as an issue.

CONTRADICTION DEFINITION
An **contradiction (issue)** 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 this conflict.
You must consider both:

1. **Factual contradictions**

   Examples include, but are not limited to:

   - The same quantity (e.g., sample size, number of companies, growth rate, percentage, etc.) is given conflicting values with no explanation (e.g., no mention of different subsamples, filtering, or time points).
   - The same time period, region/market, or population/scope is described differently in different places for what is clearly the same analysis or dataset.
   - The same dataset's source or status (complete vs missing/bias, etc.) is described in mutually conflicting ways for the same stage of analysis.
   - The same result (same metric, same time period) is described with opposite trends or directions in different parts of the text (e.g., "increasing" vs "decreasing") without clarification.

2. **Logical contradictions**

   Examples include, but are not limited to:

   - The report states an explicit premise, assumption, or applicability condition, and later reasoning or conclusions clearly violate that premise under the same context.
   - The report gives a certain judgment/evaluation of a model, strategy, or view in one place and later uses an opposite characterization of the same item in the same context, without new evidence or conditions.
   - The conclusions or recommendations directly state the opposite of the report's own presented results (e.g., results show "unlikely to succeed," conclusion claims "almost certain to succeed").
   - The report explicitly labels its results as non-generalizable or restricted, but later presents them as broadly applicable without any additional justification.

COUNTING RULES

- Each distinct contradiction (factual or logical) 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 of the contradiction.
  - Provide textual evidence: quote or closely paraphrase the relevant conflicting statements with location references (section/paragraph labels if available, or your own brief locator such as "Introduction, paragraph 2").

REMEMBER
Be as strict as possible in identifying contradictions. Scrutinize the report thoroughly to find any potential logical or factual inconsistencies.

OUTPUT FORMAT
Please respond with a JSON object. Do not add markdown code blocks (like "`` `json ``) 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 if 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. No issues are visible.
- **Score 0 (Fail)**: The chart fails the criterion. Any visible defect (e.g., text overlap, unreadable font, broken encoding) results in a score of 0.

CHECKLIST ITEMS TO EVALUATE
A. Layout & Structure

1. **No Overlap**

   All text elements (titles, legends, axis labels, tick labels, annotations, data labels) and graphic elements do not overlap with each other or with the core data visualization components.

2. **No Misalignment**

   Elements that should align (axis lines, tick marks with tick labels, legend markers with legend text, subplot panels in a grid) are consistently aligned, without clear, unintended offset or irregular positioning.

3. **Complete Visibility**

All chart content is fully visible within the canvas boundaries, with no elements cut off, clipped, or extending beyond the viewable area.

4. **Balanced Spatial Composition**

   The chart's overall composition is balanced; the plot area, titles, legends, and annotations occupy appropriate proportions of the canvas; whitespace is distributed reasonably, appearing neither cramped nor empty.

B. Readability

5. **Appropriate Text-to-Chart Proportion**

   Text elements (titles, legends, axis labels, tick labels, data labels) are proportionate to the chart's visual elements: text is neither dominating the plot area nor so small relative to marks/axes that it harms readability.

6. **Readable Contrast & Distinguishability**

   Text is clearly readable against the background, and different series/categories are visually distinguishable. Fail if low contrast or overly similar colors make text or categories hard to read/identify.

7. **Correct Text Rendering**

   All characters, symbols, and numbers are displayed correctly, without encoding errors, missing glyphs, or rendering issues.

C. Conciseness

8. **No Excessive Decoration**

   The chart does not contain distracting or excessive decorative elements; all visual effects are restrained and purposeful, enhancing rather than obscuring data readability.

9. **No Information Overload**

   The number of visual elements in the chart is not excessive; the chart avoids being too dense or overcrowded, which would cause difficulty in interpretation.

10. **No Redundant Labels**

    Labels are used efficiently; the same information is not repeated or presented redundantly in ways that create visual clutter.

OUTPUT FORMAT

Respond with a JSON object. Do not add markdown code blocks (like "'`json`) 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...",
    "text_size": "Specific description of
     font size 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 additionally contains 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 `research_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 (Each 1–10 points)

1. Multimodal Layout & Quantity Appropriateness

Are multimodal elements placed, distributed, and used in an appropriate amount across the report's stages, supporting a smooth reading flow without obvious gaps or overload?

Focus:

- Figures are placed near the sections where they are discussed, not arbitrarily clustered or detached.
- The overall visual distribution supports a natural reading rhythm, avoiding both long stretches with no visuals despite complex content and sudden dense clusters that disrupt the flow.
- Important complex content (e.g., structures, workflows, data comparisons) has at least some visual support, rather than relying solely on dense text.
- The total number of figures feels proportionate to the length and complexity of the report (neither too few nor excessive).
- Multimodal elements are reasonably spread across key parts, rather than confined to only one section while other equally complex sections have none.

Scoring:

- **9–10 – Excellent layout and quantity appropriateness**: Multimodal elements are consistently placed near relevant text and are well distributed across the report's stages. Coverage is well targeted (no clear gaps), and the quantity feels proportionate with no obvious redundancy. Overall, the layout and amount clearly support a smooth, coherent reading experience.
- **7–8 – Good layout and quantity with minor issues**: Placement and distribution are mostly sensible with occasional issues (a figure could be closer to its discussion, slight clustering, or a small coverage gap). Quantity is generally appropriate, with only minor under/over-use that does not strongly affect readability.
- **5–6 – Acceptable but uneven**: The report remains readable, but layout and/or quantity feel insufficiently planned. There may be noticeable stretches lacking visuals despite complexity, awkward groupings, uneven coverage across sections, or a quantity that is somewhat misaligned (a bit sparse or somewhat cluttered).
- **3–4 – Weak layout/quantity choices**: Figures are frequently detached from relevant text or awkwardly grouped; some key sections lack needed visuals while others feel cluttered. Coverage and quantity choices noticeably hinder reading flow and understanding.
- **1–2 – Very poor layout and quantity appropriateness**: Figures (if any) are placed with little regard to structure and discussion points, and the report is either severely under-illustrated despite complexity or overloaded with visuals. The multimodal design seriously disrupts or confuses the reading experience.

2. Multimodal Variety & Richness

Does the report demonstrate appropriate variety and richness in its multimodal elements based on the report's topic and nature?

Focus:

- **Source-type variety**:
  - **Retrieved external images**: e.g., architecture diagrams, system schematics, flowcharts, sequence diagrams, environment photos, illustrations, etc.
  - **Code-generated charts**: e.g., line charts, bar charts, scatter plots, heatmaps, radar charts, box plots, etc.
- **Intra-type variety**: Within each source type present, the report uses multiple appropriate sub-types.
  - **For charts**: Different chart types count as different sub-types. (e.g., line chart, bar chart, scatter plot are 3 distinct sub-types).

- **For images**: Different purposes count as different sub-types (e.g., offshore wind power architecture diagram, reinforcement learning algorithm architecture diagram are 2 distinct sub-types).

- **Topic-appropriate multimodal strategy**: The choice and balance of multimodal elements should align with the report's subject matter.
  - **Data/analysis-heavy topics** (e.g., financial analysis, statistical studies, performance benchmarks): Naturally emphasize code-generated charts with rich variety.
  - **Technical/architectural topics** (e.g., system design, software architecture, infrastructure): Naturally emphasize retrieved images with rich variety.
  - **Balanced topics** (e.g., comprehensive surveys, product analyses): Should include both types with reasonable variety in each.

Hard Constraints

1. You MUST compute these intermediate variables:
   - topic_category: string in {"data_heavy", "technical_architectural", "balanced"}
   - source_types_present: integer in {0,1,2}
   - image_subtypes_count: integer (0 if no valid images)
   - chart_subtypes_count: integer (0 if no valid charts)
   - total_subtypes: integer (image_subtypes_count + chart_subtypes_count)
   - dominant_subtypes_count: integer (chart_subtypes_count for data_heavy, image_subtypes_count for technical_architectural)
2. ABSOLUTE SCORING RULES – NO EXCEPTIONS:
   - Once the numerical thresholds are met, the score range is LOCKED and cannot be changed.
   - You MUST NOT apply any additional subjective criteria beyond the defined rules, including but not limited to:
     - Subjective quality judgments (e.g., "not especially diverse", "basic/common types")
     - Suggestions for additional content (e.g., "could include more advanced visualizations", "lacks X type")
     - Personal preferences about specific chart or image types
   - The score range is determined EXCLUSIVELY by these objective factors:
     - The COUNT of subtypes (as computed above)
     - Whether both source types are present (for balanced topics only)
     - The topic category classification
   - Within the determined score range (e.g., 9–10 or 7–8), you may select the specific score based on the minor execution quality differences. BUT you CANNOT move to a different score range under any circumstances.
3. You MUST determine the score by STRICTLY following the mapping below. The score ranges are **hard boundaries** that you CANNOT violate.

Score Mapping (EVALUATE IN ORDER, STOP AT FIRST MATCH):

- If source_types_present == 0:
  - → score MUST be 1 or 2. STOP.
- If topic_category == "balanced":
  - If source_types_present == 2:
    - If image_subtypes_count >= 3 AND chart_subtypes_count >= 3:
      - → score MUST be 9 or 10. STOP.

        - Else if image_subtypes_count >= 2 AND chart_subtypes_count >= 2:
            - → score MUST be 7 or 8. STOP.
        - Else:
            - → score MUST be 5 or 6. STOP.
    - If source_types_present == 1:
        - If max(image_subtypes_count, chart_subtypes_count) >= 3:
            - → score MUST be 5 or 6. STOP.
        - Else:
            - → score MUST be 3 or 4. STOP.

- If topic_category in {"data_heavy", "technical_architectural"}:
    - If dominant_subtypes_count >= 4:
        - → score MUST be 9 or 10. STOP.
    - Else if dominant_subtypes_count >= 3:
        - → score MUST be 7 or 8. STOP.
    - Else if total_subtypes >= 2:
        - → score MUST be 5 or 6. STOP.
    - Else:
        - → score MUST be 3 or 4. STOP.

IMPORTANT GUIDELINES

- Only **actual, present multimodal elements** are considered valid for this evaluation.
- Valid 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** count as valid multimodal content:
    - Placeholders (e.g., "Figure 1: [to be inserted]", "image here").
    - Any references without an actual figure or a concrete locator (e.g., "see figure below" but no figure, URL, or file path is provided).
    - Any figures shown only as plain text or ASCII-style diagrams.
- If the report contains **no valid multimodal elements** under these rules, reflect this in your scores (e.g., very weak layout, quantity and variety) and explicitly mention the absence of real multimodal content in your justification and suggestions.

INSTRUCTIONS

- **Step 1**: Review the provided **Research_Report**.
- **Step 2**: Evaluate the **Multimodal Composition** of the Research_Report across the two dimensions: **Multimodal Layout & Quantity Appropriateness**, and **Multimodal Variety & Richness**.
- **Step 3**: For each dimension, provide a **detailed justification** that examines concrete evidence from the report and explains the reasoning leading to the score.
- **Step 4**: Based on the justification, assign a **1–10 score** for each dimension according to the detailed scoring criteria.
- **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 the **suggestion** field as an empty string `""`.

OUTPUT FORMAT

Respond with a JSON object. Do not add markdown code blocks (like "`'json`) 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:

1. **Figure**: The visual content (such as a chart, graph, photo, diagram, etc.).
2. **Caption**: The full text associated with the figure.

SCORING DIMENSIONS (Each 1–10 points):
1. Visual Accuracy
Does the caption correctly identify and describe what is actually shown?
Focus:

- Alignment between caption and figure.
- Correct identification of subject, variables, and visible relationships.

Scoring:

- **9–10** – Perfect match. The caption correctly identifies what the figure is (e.g., chart/diagram/photo), names the key entities/components/variables, and accurately describes the visible content. All claims are consistent with the figure.
- **7–8** – High accuracy. The main identification and description are correct; minor visible elements are omitted or slightly under-specified, but nothing is wrong or misleading.
- **5–6** – Partial match. The caption is too generic or incomplete: it roughly matches the topic, but misses key visible elements needed to understand what is shown.
- **3–4** – Low accuracy. The caption refers to related but materially different content (e.g., wrong type of figure, wrong components/variables, or mischaracterized relationship/comparison).
- **1–2** – Contradiction. The caption describes content not present in the figure, or asserts relationships/structures/process steps that are clearly contradicted by what is shown.

2. Minimum Necessary Information
*Using only the figure and this caption, can a reader correctly interpret what is shown? Does the caption provide the **minimum necessary** information given the figure's complexity (no less, no more)?*
Focus:

- If a reader can understand the figure's main point using **the figure itself + this caption** (without the main text), the caption is sufficient.
- Provide only the **minimal guidance needed to read this figure**. **Do NOT require** the caption to restate or list details already visible in the figure.
- A caption can be short and still score high if the figure is self-explanatory. A longer caption scores high only if the added details are necessary to remove ambiguity for this specific figure.

Scoring:

- **9–10** – Minimally sufficient and perfectly calibrated. The caption provides just enough information for correct interpretation (given the figure), with little or no ambiguity, and without redundant listing or filler content.
- **7–8** – Sufficient with minor inefficiency. Interpretation is possible without the main text; one small clarification is missing **or** there is a small amount of unnecessary detail, but overall still clear.
- **5–6** – Borderline. The caption misses 1–2 key clarifications needed to interpret

the figure **or** includes noticeable non-essential detail that does not improve interpretability; a reader may need the main text.

- **3–4** – Insufficient or bloated. Multiple necessary clarifications are missing **and/or** the caption is mostly verbose filler, leading to confusion or ambiguity.
- **1–2** – Not usable. The caption is too vague/empty to interpret the figure, or is largely irrelevant to what is shown, making independent interpretation impossible.

3. Clarity & Readability
Is the caption written in clear and accessible language that is easy to understand?
Focus:

- Clear sentence structure and word choice.
- Avoids unnecessary jargon or defines technical terms when necessary.
- Easy to parse and understand on first reading.

Scoring:

- **9–10** – Highly clear. The caption uses straightforward, direct language that is immediately understandable. Technical terms are either avoided or clearly defined. No ambiguous phrasing.
- **7–8** – Clear. Generally easy to understand with minor issues (e.g., one slightly complex sentence or undefined term that doesn't significantly hinder comprehension).
- **5–6** – Moderately clear. Contains some unclear phrasing, unnecessary complexity, or undefined jargon that requires rereading or may confuse readers.
- **3–4** – Unclear. Multiple confusing sentences, heavy jargon without explanation, or overly complex language that obscures the main point.
- **1–2** – Incomprehensible. The caption is so poorly written, convoluted, or jargon-heavy that readers would struggle to understand what is being described.

IMPORTANT GUIDELINES

- You must primarily compare the **Caption** against the **Visual Evidence** in the **Figure**.
- Do **not** invent or assume facts that are not supported by the figure.

INSTRUCTIONS

- **Step 1**: Review the provided **Figure** and **Caption**.
- **Step 2**: Evaluate the **Quality of the Figure Caption** across the three dimensions: **Visual Accuracy**, **Minimum Necessary Information**, and **Clarity & Readability**.
- **Step 3**: For each dimension, provide a **detailed justification** that examines concrete evidence from the figure and explains the reasoning leading to the score.
- **Step 4**: Based on the justification, assign a **1–10 score** for each dimension according to the detailed scoring criteria.
- **Step 5**: If any dimension scores **6 or lower**, provide a **revised caption** in the **suggestion** field, using only information supported by the figure. If **all** dimensions score **7 or higher**, set the **suggestion** field as an empty string `""`.

OUTPUT FORMAT
Respond with a JSON object. Do not add markdown code blocks (like "'`json`) 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:

1. **Figure**: The visual content (such as a chart, graph, photo, diagram, etc.).
2. **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):
1. Contextual Relevance
How well does the figure's content match the specific topic and claims in the context?
Focus:

- **Semantic alignment** between what the figure shows and what the context discusses.
- Whether the figure depicts the same topic, variables, phenomena, or cases mentioned in the surrounding text.

Scoring:

- **9–10 – Highly Relevant**: The figure directly depicts or represents the specific subject matter discussed in the immediate context. A reader can clearly see why this particular figure was chosen for this passage.
- **7–8 – Relevant**: The figure clearly depicts the specific subtopic or subject in this passage, though it may not capture every detail mentioned in the text. The figure choice is appropriate for this context.
- **5–6 – Moderately Relevant**: The figure belongs to the same broader theme or research area and is not out of place, but the connection between what the figure shows and what this passage specifically discusses is loose.
- **3–4 – Weak Relevance**: The connection to the context is vague or forced. It only shares a very general domain, without depicting what this passage is specifically about.
- **1–2 – Irrelevant**: The figure does not depict anything meaningfully related to the local context; a typical reader would be confused why this particular figure is placed here.

2. Narrative Coherence
How smoothly is the figure integrated into the flow of the text?
Focus:

- Whether the context **explicitly or clearly implicitly** points the reader to this figure.
- How natural the transition is between text and the figure, and whether the discussion around it is sufficient.

Scoring:

- **9–10 – Seamless Integration**: The text explicitly references the figure (e.g., "As shown in Figure 3…"), and it meaningfully discusses key elements of the figure, so the reader is guided to look at and interpret it.
- **7–8 – Good Integration**: There is a clear reference to the figure, but the transition is slightly abrupt, or the discussion is a bit brief or incomplete.
- **5–6 – Weak Link**: The figure is mentioned (e.g., only "(see Figure 2)" at the

end of a sentence), but the surrounding text does not really explain or interpret it.

- **3–4 – Disconnected**: No explicit reference; the figure appears between paragraphs or at the side, and the relation has to be guessed by the reader.
- **1–2 – Disruptive**: The placement of the figure breaks the reading flow, feels random, or even contradicts the narrative sequence.

3. Visual Information Value
Does the figure provide visual information that text alone cannot effectively convey?
Focus:

- Whether the figure offers **unique visual value** that would be difficult, inefficient, or impossible to communicate through text alone.
- Consider how much understanding would be lost if the figure were removed. Would readers struggle to grasp the information from the text alone?

Scoring:

- **9–10 – Visually Essential**: The figure conveys information that is extremely difficult or impractical to describe in text. Without the visual, readers would struggle to grasp the information even with lengthy description.
- **7–8 – High Visual Value**: The figure significantly enhances understanding by providing visual information that would require extensive text to approximate. Text alone would be notably less efficient or clear.
- **5–6 – Moderate Visual Value**: The figure provides some visual convenience, but most information could be reasonably conveyed in a few sentences. The visual format is helpful but not particularly necessary.
- **3–4 – Low Visual Value**: The figure adds minimal visual information beyond what text easily provides. The visual medium is barely justified.
- **1–2 – No Visual Necessity**: The figure provides no meaningful visual information that text cannot easily convey. Using a visual medium here is unnecessary regardless of content relevance.

INSTRUCTIONS

- **Step 1**: Review the provided **Figure** and **Context**.
- **Step 2**: Evaluate the **Integration between the Figure and the Context** across the three dimensions: **Contextual Relevance**, **Narrative Coherence**, and **Visual Information Value**.
- **Step 3**: For each dimension, provide a **detailed justification** that examines concrete evidence from the figure and context, **then** explains the reasoning leading to the score.
- **Step 4**: Based on the justification, assign a **1–10 score** for each dimension according to the detailed scoring criteria.
- **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 the **suggestion** field as an empty string `""`.

OUTPUT FORMAT
Respond with a JSON object. Do not add markdown code blocks (like “‘`json`) 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:

1. **Chart**: The chart itself.
2. **Caption**: The full text associated with the chart.
3. **References**: A JSON list of cited source snippets associated with the chart. Treat this list as the only permissible evidence; do not assume access to external links or any information not contained in the provided snippets.

GOAL
Decide whether the chart is **consistent** with the provided sources.

OUT OF SCOPE

1. Pure visual aesthetics issues (layout, fonts, colors) unless they prevent verification.
2. Real-world credibility of sources; only whether the provided snippets support the chart as presented.
3. Missing or insufficient evidence. Only focus 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 References under the same (or materially equivalent) claim context.

COUNTING RULES

- Each distinct underlying problem counts as ONE issue.
- Multiple symptoms from the same root cause count as ONE issue.
- If there are multiple unrelated contradictions, count them as separate issues.

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 you find, each with:
  - the chart claim,
  - a short description,
  - direct quote(s) from References with ref_idx.
- **Step 4**: Count the total number of distinct issues.

OUTPUT FORMAT
Respond with a JSON object. Do not add markdown code blocks (like "'`json`) 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}
</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.

G.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