MMDEEPRESEARCH-BENCH: A BENCHMARK FOR MULTIMODAL DEEP RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep Research Agents (DRAs) generate citation-rich reports via multi-step search and synthesis, yet assessing faithful use of textual and visual evidence in open-ended reports remains challenging. We introduce **MMDeepResearch-Bench (MMDR-Bench)**, a benchmark of 140 researcher-curated tasks across 19 domains, where each task provides an image–text bundle to evaluate multimodal understanding and citation-grounded report generation. Compared to prior setups, MMDR-Bench emphasizes report-style synthesis with explicit evidence use, where models must connect visual artifacts to sourced claims and maintain consistency across narrative, citations, and visual references. We further propose a unified, interpretable evaluation pipeline: **F**ormula–LLM **A**daptive **E**valuation (**FLAE**) for report quality, **T**rustworthy **R**etrieval-**A**ligned **C**itation **E**valuation (**TRACE**) for citation-grounded evidence alignment, and **M**ultimodal **S**upport-**A**ligned **I**ntegrity **C**heck (**MOSAIC**) for text–visual integrity, each producing fine-grained signals for aggregate failure-mode analysis beyond a single overall score. Experiments across 28 state-of-the-art models and agents reveal systematic trade-offs between generation quality, citation discipline, and multimodal grounding, highlighting that strong prose alone does not guarantee faithful evidence use and that multimodal integrity remains a key bottleneck for deep research agents. Anonymous dataset: `https://huggingface.co/datasets/anomyousMMDR/MMdeepresearch`. Anonymous code: `https://anonymous.4open.science/r/MMDeepResearch-Bench-01CE`.

1 INTRODUCTION

Recent advancements in foundation models have driven a shift from language-centric systems to large multimodal models (LMMs) that jointly process text and visual inputs (Bommasani, 2021). Enabled by vision–language pretraining and instruction tuning (Li et al., 2023; Dai et al., 2023; Zhu et al., 2023), modern LMMs can reason over structured visual artifacts such as charts and documents, forming the basis of current vision–language benchmarks (Bai et al., 2025; Masry et al., 2022; Mathew et al., 2021; Tang et al., 2025).

Yet static models remain limited by fixed parametric memory, motivating retrieval-augmented generation (Lewis et al., 2020) and tool-using agents that browse and collect external evidence (Wong et al., 2025; Xi et al., 2025; Lan et al., 2025). Building on this paradigm, Deep Research Agents (DRAs) target open-ended, long-horizon tasks by iteratively retrieving sources, reconciling hypotheses, and producing research-style reports (Shi et al., 2025; Shao et al., 2024; Du et al., 2025). Since real research is rarely text-only, DRAs must also align textual claims with figures, charts, and diagrams, motivating multimodal deep research (Du et al., 2025; Patel et al., 2025; Venkit et al., 2025).

As DRAs proliferate (Patel et al., 2025; Venkit et al., 2025), evaluation becomes crucial but difficult: intermediate reasoning and retrieval are opaque, and open-ended questions seldom admit a single gold answer, making the final cited report the primary evaluation interface (Venkit et al., 2025). Existing benchmarks either isolate web or retrieval competence (Wong et al., 2025; Xi et al., 2025), focus on text-only deep research reports (Du et al., 2025; Patel et al., 2025), or emphasize short-horizon multimodal perception (Zou et al., 2025; Tang et al., 2025). Live retrieval further complicates evaluation via issues such as search-time data contamination (Han et al., 2025). Recent benchmarks cover complementary targets: visual–textual search and answer correctness in VDR-Bench (Zeng

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2026), expert-derived report rubrics in DeepResearch Bench II (Li et al., 2026), and synthesis, factuality, and process audits in MiroEval (Ye et al., 2026). MMDR-Bench jointly evaluates report quality, citation support, input-image fidelity, and report-image alignment.

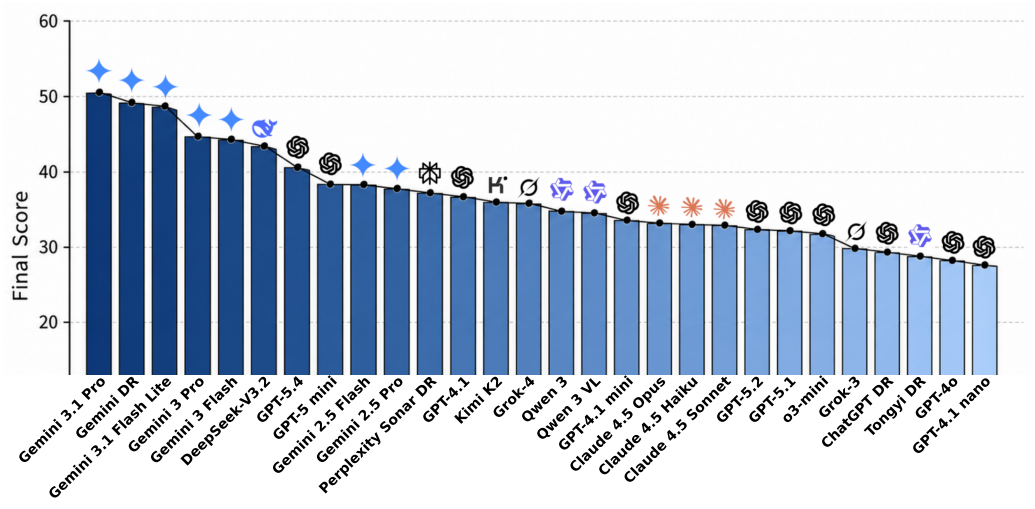

Figure 1: Overall scores for 28 systems on MMDR-Bench (higher is better). Gemini DR denotes Gemini Deep Research (Gemini 3 Pro).

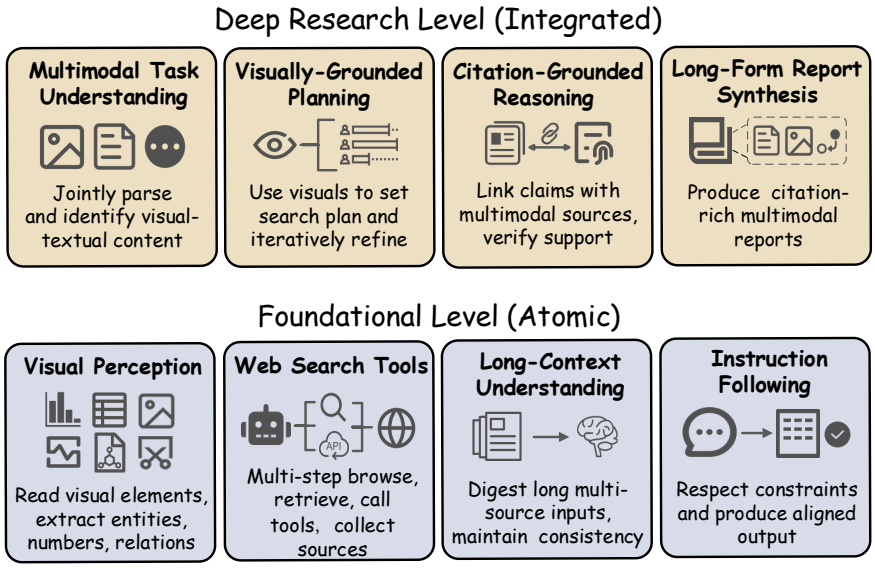

Figure 2: Multimodal research capabilities motivating MMDR-Bench.

We introduce **MMDeepResearch-Bench (MMDR-Bench)**, comprising 140 image–text tasks that elicit citation-grounded reports, integrating multimodal understanding, evidence-grounded reasoning, and long-form synthesis (Figure 2). MMDR-Bench includes two complementary regimes, *Daily* and *Research*, reflecting lightweight everyday usage and analysis-heavy research settings. All tasks are iteratively refined by researchers with relevant disciplinary training to ensure multimodal necessity and verifiability.

Building on MMDR-Bench, we develop an evaluation framework along three aspects: **F**ormula–**L**LM **A**daptive **E**valuation (**FLAE**) for long-form report quality, **T**rustworthy **R**etrieval–**A**ligned

Citation **E**valuation (**TRACE**) for citation-grounded support and source quality, and **M**ultimodal **S**upport–**A**ligned **I**ntegrity **C**heck (**MOSAIC**) for text–image consistency of the report. Figure 1 summarizes overall scores; Section 5 reports selected diagnostic dimensions.

- **A Multimodal Research Benchmark.** MMDR-Bench contains 140 researcher-curated tasks across 19 domains under Daily and Research regimes, with image–text inputs requiring multimodal understanding and evidence-grounded report synthesis.
- **A Unified Evaluation Framework.** FLAE, TRACE, and MOSAIC jointly assess report quality, citation-grounded reasoning, and consistency between textual claims and visual artifacts.
- **Visual Evidence Fidelity.** VEF checks task-specific textualized visual requirements with a strict PASS/FAIL rule, penalizing misinterpretation of critical input-image evidence.
- **Systematic Evaluation and Open Resources.** Results across 28 models and agentic systems expose trade-offs between writing quality, citation discipline, and multimodal grounding. We release the dataset, evaluation code, and detailed metrics.

2 Related Work

Deep Search and Agentic Reasoning. Agentic search frameworks decompose queries and gather evidence through sequential tool use (Shao et al., 2024; Deng et al., 2023). Search-R1 (Jin et al., 2025) and DeepDive (Lu et al., 2025) use reinforcement learning to improve retrieval and reasoning. These systems motivate evaluating not only final answers but also how reports substantiate their claims.

Multimodal Search and Reasoning. MMSearch (Jiang et al., 2024) and DeepEyes (Zheng et al., 2025) study visual information seeking. Vision-DeepResearch Benchmark (Zeng et al., 2026) targets realistic visual–textual search with 2,000 VQA instances and a cropped-search workflow. Our focus is instead long-form reports, including whether input-image facts and report images support the accompanying claims.

Benchmarks and Evaluation. BrowseComp (Wei et al., 2025) and BrowseComp-Plus (Chen et al., 2025) are text-based search benchmarks; the latter provides a fixed-corpus setting. DeepResearch Bench (Du et al., 2025) and DeepScholar (Patel et al., 2025) evaluate report synthesis. Concurrent DeepResearch Bench II (Li et al., 2026) derives fine-grained rubrics from expert reports, while MiroEval (Ye et al., 2026) combines synthesis, factuality, and process-centric evaluation. MMDR-Bench complements these efforts with image–text task bundles and separate checks for input-image fidelity, citation support, and report-image alignment.

3 Dataset Collection

We define multimodal deep research as tasks that require multi-round web browsing, evidence gathering, and report synthesis while explicitly interpreting and using provided images. Following this definition, MMDR-Bench targets realistic long-horizon cases that require multimodal understanding and evidence-grounded report writing. Each task is an image–text bundle: a textual query paired with a small set of images that must be interpreted and integrated into a research report. This design jointly evaluates (i) multimodal question understanding and (ii) evidence-grounded multimodal report generation with citations.

From a collected pool of 98k+ real-world queries, author-team researchers with relevant disciplinary training curate 140 tasks across 19 domains (Figure 3). The final tasks comprise two regimes: *Daily* (40 tasks) with casual visuals such as screenshots and UI captures, and *Research* (100 tasks) with structured, information-dense figures such as charts, tables, and diagrams that require deeper synthesis. Curation emphasizes multimodal necessity, answerability, public evidence, and auditability. Two example tasks are shown in Figure 4; further construction details are in Appendix A.1.

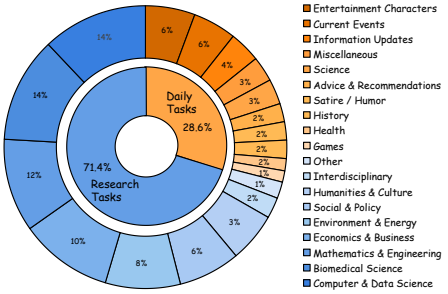

Figure 3: Task distribution of MMDR-Bench.

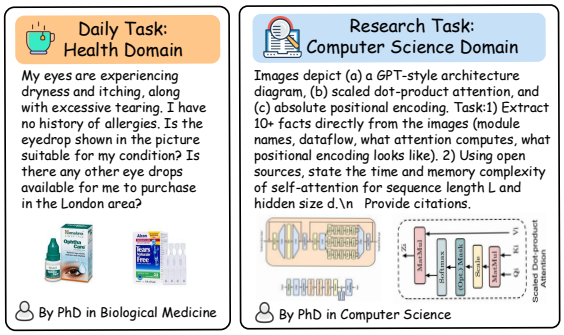

Figure 4: Two example tasks from MMDR-Bench.

4 Evaluation Methodology

Our evaluation framework jointly assesses research-report quality, citation-grounded evidence use, and multimodal integrity. It comprises FLAE, TRACE, and MOSAIC (Figure 5).

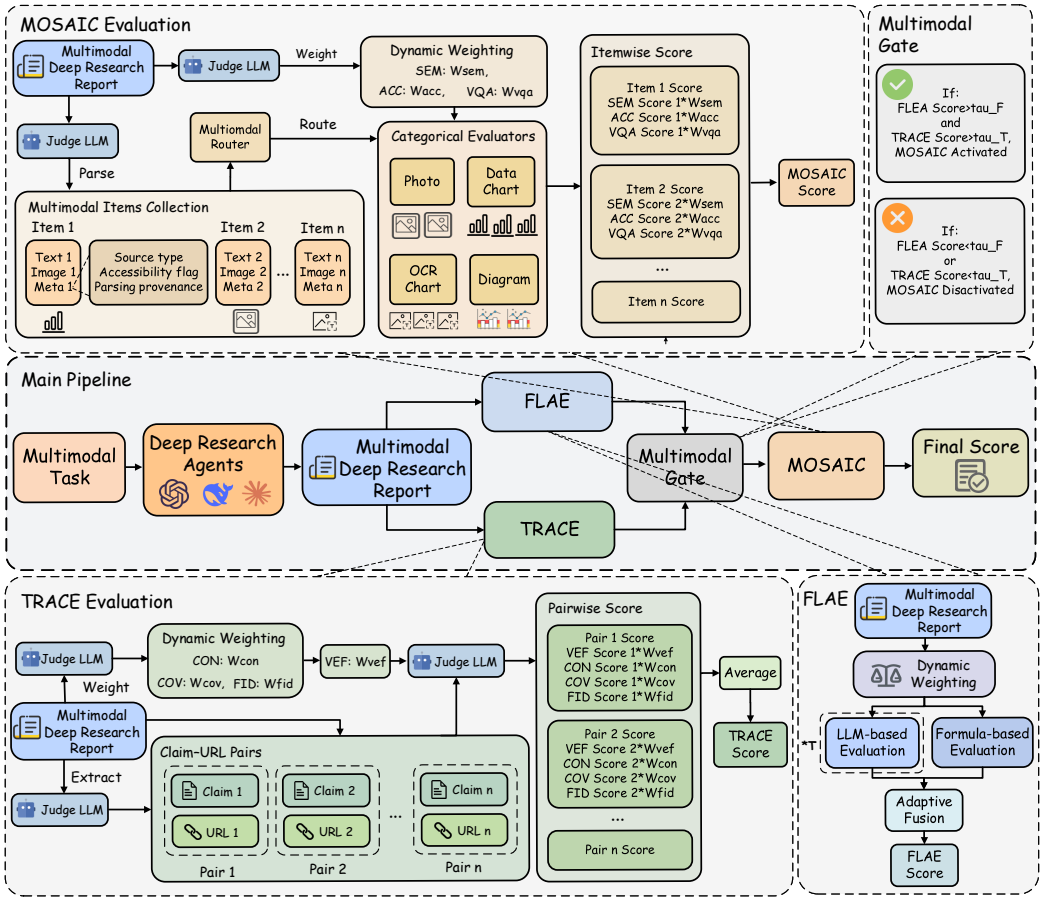

Figure 5: MMDR-Bench evaluation pipeline, integrating report-quality assessment, citation and visual-evidence verification, and gated visual-support evaluation.

FLAE assesses report quality, TRACE verifies citation support and fidelity to task-provided visual evidence, and MOSAIC checks the images used in the generated report. Positive report-quality and

text-evidence scores gate MOSAIC evaluation. Together, these modules connect writing quality to the evidence supporting a report's claims. Execution and aggregation details are provided in Appendix A.5.

4.1 FLAE: FORMULA-LLM ADAPTIVE EVALUATION

Writing requirements vary across research tasks. **FLAE** (**F**ormula–**LL**M **A**daptive **E**valuation) combines fixed feature transforms with holistic judge scores on four dimensions: Readability (READ.), Insightfulness (INSH.), Coherence (COH.), and Structural Completeness (STRU.).

Formula-Based Channel. The feature vector $\phi(R)$ combines directly measured text statistics with judge-estimated attributes. Fixed transforms produce scores in $[0, 1]$:

$$s_d^{\text{form}}(R) = f_d\big(\phi(R)\big), \qquad \mathcal{D} = \{\text{READ.}, \text{INSH.}, \text{COH.}, \text{STRU.}\}. \tag{1}$$

The formula channel captures explicit report characteristics, complementing the judge's holistic assessment. Feature definitions and transforms are provided in Appendix A.3.2.

Judge Channel and Fusion. A judge estimates holistic scores $s_d^{\text{judge}}(t, R)$. We fuse the channels as $s_d = \alpha\, s_d^{\text{form}} + (1 - \alpha) s_d^{\text{judge}}$. The fusion coefficient α is set by report type, while task-adaptive dimension weights account for differences in writing requirements. This combination balances explicit features with holistic assessment; detailed settings are provided in Appendix A.3.4.

$$\text{FLAE}(t, R) = 100 \sum_{d \in \mathcal{D}} W_d(t, R) s_d(R), \qquad \sum_d W_d = 1. \tag{2}$$

4.2 TRACE: TRUSTWORTHY RETRIEVAL-ALIGNED CITATION EVALUATION

TRACE (**T**rustworthy **R**etrieval-**A**ligned **C**itation **E**valuation) assesses citation support and task-image fidelity. We map citation indices to URLs, extract atomic Claim–URL pairs, retrieve source content, and record accessibility. A judge assesses support, hallucination, causal inversion, and reasoning-chain soundness. Text-evidence aggregation uses four dimensions in $[0, 1]$: Consistency (CON.), Coverage (COV.), Textual Fidelity (FID.), and Source Diversity (DIV.).

Visual Evidence Fidelity (VEF.) checks whether the report faithfully uses the task-provided images. It compares the report against task-specific textualized visual ground truth and applies a strict binary criterion to critical visual facts, including entity identities, object presence, and numerical details. This makes fidelity to the input images an explicit part of evidence evaluation. Ground-truth construction and the decision rule are detailed in Appendix A.1.2.

With $\mathcal{K} = \{\text{CON.}, \text{COV.}, \text{FID.}, \text{DIV.}\}$, TRACE is

$$\text{TRACE}(t, R) = 100 \left[\lambda_{\text{VEF.}}\, y(t, R) + (1 - \lambda_{\text{VEF.}}) \sum_{k \in \mathcal{K}} W_k(t, R) E_k(R) \right], \tag{3}$$

where $y(t, R)$ is the binary VEF outcome, $\lambda_{\text{VEF.}} = 0.4$, and $\sum_k W_k = 1$. The citation weights use a fixed prior, task-controller adjustments, and document diagnostics (Appendix A.3.5).

4.3 MOSAIC: MULTIMODAL SUPPORT-ALIGNED INTEGRITY CHECK

MOSAIC (**M**ultim**o**dal **S**upport-**A**ligned **I**ntegrity **C**heck) evaluates whether image-referenced statements match the images embedded or linked in a report. It extracts multimodal items and routes them into photo, data-chart, OCR-chart, or diagram types for modality-appropriate checks.

Multimodal Support Scoring. Applicable diagnostics include Visual-Semantic Alignment (SEM.), Visual Data Interpretation Accuracy (ACC.), and Visual Claim Support (VQA). VQA uses visual question answering to assess whether an image supports its associated statement. This targets the specific claim made in the report, complementing overall visual-semantic alignment. Item scores use fixed type-specific weights, a formatting/integration factor f_i, and a diagram hallucination factor h_i:

$$s_i = f_i \sum_k \omega_{\text{type}(i),k}\, m_{i,k}, \qquad \sum_k \omega_{\text{type}(i),k} = 1. \tag{4}$$

For diagrams, m_i includes $1 - h_i$; other types use applicable SEM., ACC., and VQA scores. Appendix A.3.6 gives the weights and missing-value conventions.

4.4 DESIGN RATIONALE: ARTIFACT-GROUNDED EVALUATION

MMDR-Bench evaluates three complementary aspects of multimodal deep research: faithful use of the provided visual inputs (VEF.), grounding of textual claims in retrieved evidence (TRACE), and consistency between a report's statements and its visual artifacts (MOSAIC). Open-ended research permits multiple valid evidence paths, while live sources and time-sensitive information continue to evolve. Rather than prescribing a single gold report or search route, our framework checks the evidence underlying the generated report. This accommodates diverse research strategies while making failures in visual interpretation, citation support, and multimodal synthesis visible through separate diagnostic dimensions.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Benchmark and Protocol. We evaluate generated reports on MMDR-Bench using Gemini-2.5-Pro as the primary judge. Overall scores combine FLAE, TRACE, and MOSAIC with weights of 20/50/30%, emphasizing evidence-grounded research quality. Model access information and testing windows appear in Appendix A.4; judge settings, coverage, and missing-output handling are detailed in Appendix A.5.

Evaluated Report Models. We evaluate a diverse set of systems spanning four groups: (i) Single-modal LLM baselines without web search function, (ii) multimodal LLM baselines without web search function, (iii) Multimodal LLM baselines with web search function and (iv) specific deep research agents. For each model, we report the overall MMDR-Bench score and metric dimensions for FLAE, TRACE and MOSAIC.

5.2 MAIN RESULTS AND FINDINGS

Table 1 reports Overall and selected diagnostics. Gemini 3.1 Pro achieves the highest Overall score (50.59), followed by Gemini Deep Research (49.41). Gemini Deep Research leads several displayed FLAE/TRACE dimensions; GPT-5 mini has the highest MOSAIC accuracy score. GPT-5.4 has the largest displayed VEF value (52.45).

Finding 1: Vision is beneficial only when it is reliable as evidence. Comparisons within the same model family such as Qwen 3 235B (A22B) vs. Qwen 3 VL 235B (A22B) show that adding vision is not a monotonic win. Although their multimodal variants improve visual grounding, the unified error analysis in Figure 6 reveals increased detail-level extraction failures from mis-reading fine-grained literals such as numerals, dates, labels and table cells. The comparison illustrates that stronger visual alignment can coexist with less reliable extraction of task-critical details. The detailed failure case analysis is in Appendix A.5.4.

Finding 2: Multimodal alignment and citation grounding can diverge. Stronger multimodal alignment or prompt-following does not guarantee more reliable citation grounding. Compared with Gemini 3 Pro, its backbone, Gemini Deep Research has higher citation coverage, yet Figure 6 shows more entity-level failures in its final reports. This coexistence indicates that broader evidence coverage need not ensure correct entity attribution.

Finding 3: Agentic scaffolding helps, but base-model progress can outweigh it. Gemini Deep Research (Gemini 3 Pro) scores 49.41 versus 44.68 for Gemini 3 Pro, with particularly large gains in citation consistency and coverage. The later Gemini 3.1 Pro reaches 50.59, surpassing the previous-generation deep-research agent.

Table 1: Results for 28 systems on MMDR-Bench, with Overall scores and selected diagnostic dimensions (0–100). Highlights indicate the highest scores in each column. MOSAIC submetrics average applicable task-level summaries, with system-dependent coverage (ACC: 6 tasks for o3-mini; 45 for Kimi K2). Aggregation and coverage details are provided in Appendix A.5.

Model	Overall	FLAE			TRACE				MOSAIC		
		READ.	INSH.	STRU.	VEF.	CON.	COV.	FID.	SEM.	ACC.	VQA
Single-Modal, w/o Search											
OpenAI o3-mini	31.96	51.32	52.65	37.11	13.57	28.45	33.74	**48.35**	15.47	90.00	12.60
DeepSeek-V3.2	43.71	59.06	87.82	58.16	19.29	33.34	45.48	18.77	42.19	83.85	12.88
Kimi K2 (Thinking)	36.91	55.60	77.27	47.34	17.14	23.54	24.62	27.20	42.00	90.00	9.50
Qwen 3 235B (A22B)	36.04	57.77	85.74	54.05	17.14	35.60	45.73	22.98	20.43	53.09	4.95
Multimodal, w/o Search											
Qwen 3 VL 235B (A22B)	35.08	58.47	86.48	52.21	43.57	18.34	15.25	10.68	30.58	93.52	16.98
GPT-4o	28.62	54.38	68.41	40.90	10.04	10.94	4.61	11.89	24.10	71.43	18.72
GPT-4.1	36.95	58.29	89.04	53.00	39.29	15.90	10.06	5.61	29.66	80.56	19.92
GPT-4.1 mini	34.23	57.61	83.62	49.60	12.86	24.20	25.44	12.33	32.62	89.91	13.21
GPT-4.1 nano	28.07	51.74	64.82	37.28	10.79	18.99	19.86	24.42	27.02	76.30	13.04
GPT-5 mini	38.49	51.57	81.73	47.18	39.29	20.02	26.64	32.61	33.90	**94.23**	15.60
GPT-5.1	32.69	56.41	83.57	51.80	35.71	13.22	2.30	13.67	22.03	84.29	14.32
GPT-5.2	32.76	59.34	83.92	54.31	46.43	14.00	1.43	5.30	12.83	50.00	9.16
GPT-5.4	40.84	59.17	76.44	49.26	**52.45**	14.56	6.75	5.75	19.56	69.23	12.60
Grok-3	29.89	56.14	86.13	52.24	20.00	12.57	5.79	2.80	22.18	68.39	13.89
Grok-4 (Fast Reasoning)	36.10	57.48	80.49	52.99	36.43	17.30	14.62	6.12	28.46	87.45	19.34
Multimodal, w/ Search											
Claude 4.5 Haiku	33.67	58.34	81.80	53.22	28.57	17.90	14.10	18.56	25.98	76.90	11.70
Claude 4.5 Sonnet	33.61	57.14	82.31	51.65	32.14	14.36	15.09	16.11	20.73	70.13	14.41
Claude 4.5 Opus	33.84	55.73	83.86	50.70	35.00	30.64	41.14	21.97	21.30	77.21	14.75
Gemini 2.5 Flash	38.40	55.44	68.58	46.91	32.86	25.35	27.77	38.30	40.67	75.96	25.49
Gemini 2.5 Pro	38.04	56.06	85.94	51.44	38.57	30.18	28.77	14.98	19.47	92.86	12.50
Gemini 3 Flash	44.43	56.59	**90.22**	52.00	45.71	31.95	35.07	15.42	36.61	87.31	18.99
Gemini 3 Pro	44.68	57.38	75.39	49.85	46.43	37.98	41.85	6.46	40.69	80.44	23.15
Gemini 3.1 Flash Lite	49.08	56.56	86.11	48.88	30.89	12.28	11.65	18.52	**53.04**	87.86	28.91
Gemini 3.1 Pro	**50.59**	59.49	87.59	52.82	35.01	13.99	12.49	11.54	46.54	83.53	**31.54**
Deep Research Agent											
Tongyi Deep Research (30B-A3B)	29.02	54.62	62.67	40.07	12.86	25.99	30.87	24.25	20.39	93.33	7.52
Perplexity Sonar Deep Research	37.55	53.64	64.35	47.80	27.86	33.12	41.51	16.68	50.79	87.75	21.22
ChatGPT Deep Research (o3-mini)	29.50	53.72	63.61	37.30	29.29	10.19	4.16	11.07	27.32	73.44	21.75
Gemini Deep Research (Gemini 3 Pro)	49.41	**75.19**	89.56	**70.86**	35.71	**56.17**	**52.84**	31.29	41.29	87.54	28.45

These results highlight the importance of both agent design and continued backbone improvement: a stronger base model can exceed an agent built on an earlier generation. Tongyi Deep Research (30B-A3B) also scores below several larger systems, and some offline models outperform web-enabled systems on citation coverage, showing that effective evidence use depends on more than tool availability alone.

5.3 Fine-Grained Domains Analysis

Figure 7 illustrates Research-task scores for five selected systems, grouped by domain. Substantial variation within each system shows how a single overall score can conceal task-specific weaknesses. This view complements the aggregate comparison by highlighting the uneven performance of current systems across research tasks.

5.4 Human Consistency Check

We assess evaluator–human agreement on reports from all 140 tasks using pairwise preference agreement (PAR) and Pearson correlation of system-level mean scores (OPC). The annotation protocol, annotator composition, and sampling procedures are detailed in Appendix A.6.

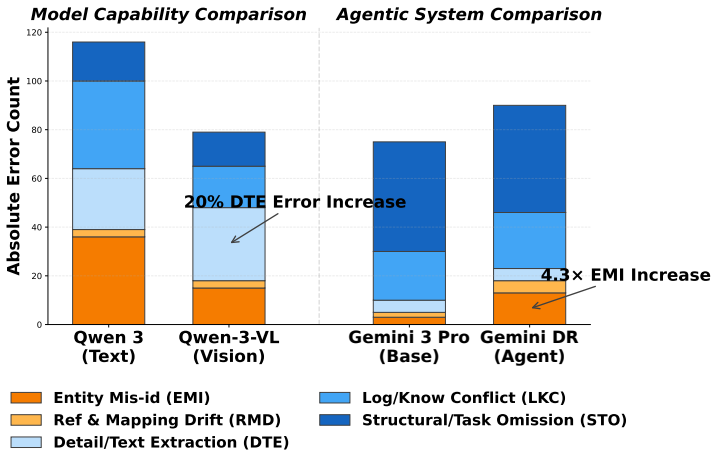

Figure 6: Visual-evidence failure modes for text-only versus vision-enabled backbones (left) and base versus agentic systems (right).



Figure 7: An illustrative view of Research-task scores for five selected systems, grouped by domain. Higher scores indicate better performance.

Table 2: Evaluator–human agreement on MMDR-Bench. PAR: agreement with majority human preferences; OPC: Pearson correlation of system-level average scores.

Method	PAR↑	OPC↑
Vanilla Prompt Judge	61.2	93.0
MMDR-Bench-Eval (Full)	**73.5**	**96.4**
w/o VEF.	68.0	95.2
w/o MOSAIC	70.1	95.8
Human Inter-Annotator Agreement	69.8	–

Table 2 shows that the full evaluator achieves 73.5% PAR and 96.4% OPC, compared with 61.2% and 93.0% for the vanilla prompt judge. Removing VEF. or MOSAIC reduces agreement, supporting their contribution to evaluating multimodal research reports. Additional blinded calibration checks cover report preferences and visual judgments; the two-rater MOSAIC audit yields quadratic-weighted $\kappa = 0.887$. Full results and protocols are provided in Appendix A.6.

5.5 Robustness of Judge Model and Weights

Cross-judge scoring. We rescore a fixed set of Gemini-generated reports with Gemini-2.5-Pro and GPT-5.2 under the same parsing, retrieval, and aggregation pipeline. Table 3 (left) shows module-level differences: GPT-5.2 gives lower FLAE and VEF scores and higher textual TRACE scores, while MOSAIC remains close (29.44 vs. 29.53). Opposing module shifts yield aggregate means of 36.76 and 37.06, a difference of 0.30 points (about 0.8%).

Table 3: **Left:** Mean scores under two judge backbones on the same report set. **Right:** Repeated-scoring ICC(3,1) with Gemini-2.5-Pro as judge; report generators are Gemini-3.1-Pro (Gemini) and GPT-5.4 (GPT). Full consistency statistics appear in Appendix A.7.1.

Metric	Gem-2.5-Pro	GPT-5.2
FLAE	61.89	45.82
TRACE (w/o VEF.)	28.39	39.87
VEF.	38.57	26.42
MOSAIC	29.44	29.53
MMDR (w/ VEF.)	36.76	37.06
MMDR (w/o VEF.)	38.04	35.23

Module	Gemini	GPT
FLAE	0.295	0.153
TRACE (w/o VEF.)	0.884	0.702
VEF.	0.273	0.506
MOSAIC	0.970	0.983
Overall MMDR	0.709	0.614

Controlled cross-judge rankings. We further hold reports, evidence artifacts, prompts, routing, and aggregation fixed on 120 byte-identical reports from six systems and twenty shared tasks (14 Research, 6 Daily). Gemini-2.5-Pro, GPT-5.4, and Grok-4.3 preserve the same winner, with pairwise Spearman correlations of 0.714–0.943. A 10,000-replicate task-paired bootstrap preserves the winner in 98.4%, 93.3%, and 99.8% of resamples, respectively. These results characterize ranking stability on the matched subset; middle-rank intervals are wider. Selection criteria and full uncertainty estimates are in Appendix A.7.

Aggregation-weight sensitivity. Six local perturbations around $(w_F, w_T, w_M) = (0.2, 0.5, 0.3)$ preserve the winner and Top-3, with mean Spearman correlation 0.978 against the official ranking. This supplementary analysis anchors each system at its historical Overall score and applies reweighting deltas from a later independently restarted component run. Separately, 2,000 perturbations of task-adaptive FLAE dimension weights by up to $\pm 30\%$ yield mean/minimum Spearman correlations of 0.999/0.995. The component-run provenance, estimated-component exclusion check, and alternative aggregation results are detailed in Appendix A.7.

Repeated-judge consistency. We score each fixed report five times with Gemini-2.5-Pro (default temperature 0.2; VQA 0.0). Gemini-3.1-Pro reports cover 18 tasks with three generations each, yielding 270 ratings; GPT-5.4 reports cover 17 tasks with three generations each, yielding 255 ratings. Table 3 (right) reports overall ICC(3,1) of 0.709 and 0.614. MOSAIC is particularly repeatable (0.970/0.983), while FLAE and VEF show greater variability. Binary VEF judgments on the Gemini reports have 73.3% mean pairwise agreement. Appendix A.7.1 gives the complete protocol, Krippendorff's α, and coefficients of variation.

6 CONCLUSION

In this work we introduce MMDR-Bench, a benchmark for multimodal evidence use in long-form research reports. Building on the benchmark dataset comprising 140 tasks across 19 domains under Daily and Research regimes, we further propose a unified evaluation pipeline that jointly measures report quality, citation-grounded faithfulness, and text–visual evidence consistency. Results across 28 state-of-the-art LLMs and DRAs reveal persistent trade-offs between writing quality, citation discipline, and multimodal grounding.

7 LIMITATIONS

MMDR-Bench currently contains 140 tasks, predominantly in English and Chinese. Its evaluation focuses on groundedness in retrieved sources and visual evidence, with coverage and score stability affected by inaccessible sources, changing hosted models, and judge variability. Broader language coverage and continued benchmark maintenance are important directions for future work. Further methodological and implementation details are discussed in Appendix A.8.

AI USE STATEMENT

The evaluated generative models produced research reports. Gemini-2.5-Pro served as the primary evaluation judge; GPT-5.2, GPT-5.4, and Grok-4.3 were used in the reported cross-judge studies. OpenAI Codex assisted with LaTeX formatting, manuscript organization and wording, translation for author review, and literature and citation checks. Claude assisted with manuscript review and revision suggestions.

ETHICS STATEMENT

Human involvement comprises researcher-led task construction, author-group report assessment, and supplementary blinded audits with non-author raters (Appendix A.6); results are reported in aggregate. Task images and cited materials include third-party content, whose reuse remains subject to the respective source terms and licenses. Health- and finance-related reports are evaluation artifacts and require appropriate expert review before use in consequential decisions.

REPRODUCIBILITY STATEMENT

The anonymous dataset and evaluation-code links are provided in the abstract. Dataset construction and visual ground-truth preparation are described in Appendix A.1; evaluator formulas, weights, prompts, and fallback behavior are documented in Appendices A.2 and A.3. Model access information, evaluation coverage, human assessment, and robustness protocols appear in Appendices A.4, A.5, A.6, and A.7, respectively. We distinguish frozen-report rescoring from regeneration through hosted APIs. The availability of backend snapshot identifiers and original human annotation records is documented alongside the relevant protocols.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Perplexity AI. Perplexity ai: Answer engine and sonar models. `https://www.perplexity.ai/`, 2025. Official Perplexity AI website and Sonar model family overview.

Anthropic. Claude: Constitutional ai models from anthropic. `https://www.anthropic.com/claude`, 2024. Official description of the Claude model family.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Rishi Bommasani. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, et al. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Erhang Li, Fangqi Zhou, Fangyun Lin, Fucong Dai, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Hao Li, Haofen Liang, Haoran Wei, Haowei Zhang, Haowen Luo, Haozhe Ji, Honghui Ding, Hongxuan Tang, Huanqi Cao, Huazuo Gao, Hui Qu, Hui Zeng, Jialiang Huang, Jiashi Li, Jiaxin Xu, Jiewen Hu, Jingchang Chen, Jingting Xiang, Jingyang Yuan, Jingyuan Cheng, Jinhua Zhu, Jun Ran, Junguang Jiang, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kaige Gao, Kang Guan, Kexin Huang, Kexing Zhou, Kezhao Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Wang, Liang Zhao, Liangsheng Yin, Lihua Guo, Lingxiao Luo, Linwang Ma, Litong Wang, Liyue Zhang, M. S. Di, M. Y Xu, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Panpan Huang, Peixin Cong, Peiyi Wang, Qiancheng Wang, Qihao Zhu, Qingyang Li, Qinyu Chen, Qiushi Du, Ruiling Xu, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runqiu Yin, Runxin Xu, Ruomeng Shen, Ruoyu Zhang, S. H. Liu, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaofei Cai, Shaoyuan Chen, Shengding Hu, Shengyu Liu, Shiqiang Hu, Shirong Ma, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, Songyang Zhou, Tao Ni, Tao Yun, Tian Pei, Tian Ye, Tianyuan Yue, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjie Pang, Wenjing Luo, Wenjun Gao, Wentao Zhang, Xi Gao, Xiangwen Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaokang Zhang, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xingyou Li, Xinyu Yang, Xinyuan Li, Xu Chen, Xuecheng Su, Xuehai Pan, Xuheng Lin, Xuwei Fu, Y. Q. Wang, Yang Zhang, Yanhong Xu, Yanru Ma, Yao Li, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Qian, Yi Yu, Yichao Zhang, Yifan Ding, Yifan Shi, Yiliang Xiong, Ying He, Ying Zhou, Yinmin Zhong, Yishi Piao, Yisong Wang, Yixiao Chen, Yixuan Tan, Yixuan Wei, Yiyang Ma, Yiyuan Liu, Yonglun Yang, Yongqiang Guo, Yongtong Wu, Yu Wu, Yuan Cheng, Yuan Ou, Yuanfan Xu, Yuduan Wang, Yue Gong, Yuhan Wu, Yuheng Zou, Yukun Li, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehua Zhao, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhixian Huang, Zhiyu Wu, Zhuoshu Li, Zhuping Zhang, Zian Xu, Zihao Wang, Zihui Gu, Zijia Zhu, Zilin Li, Zipeng Zhang, Ziwei Xie, Ziyi Gao, Zizheng Pan, Zongqing Yao, Bei Feng, Hui Li, J. L. Cai, Jiaqi Ni, Lei Xu, Meng Li, Ning Tian, R. J. Chen, R. L. Jin, S. S. Li, Shuang Zhou, Tianyu Sun, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xinnan Song, Xinyi Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, Dongjie Ji, Jian Liang, Jianzhong Guo, Jin Chen, Leyi Xia, Miaojun Wang, Mingming Li, Peng Zhang, Ruyi Chen, Shangmian Sun, Shaoqing Wu, Shengfeng Ye, T. Wang, W. L. Xiao, Wei An, Xianzu Wang, Xiaowen Sun, Xiaoxiang Wang, Ying Tang, Yukun Zha, Zekai Zhang, Zhe Ju, Zhen Zhang, and Zihua Qu. Deepseek-v3.2: Pushing the frontier of open large language models, 2025. URL https://arxiv.org/abs/2512.02556.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36:28091–28114, 2023.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. Deepresearch bench: A comprehensive benchmark for deep research agents, 2025. URL https://arxiv.org/abs/2506.11763.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Google. Gemini deep research documentation. https://ai.google.dev/gemini-api/docs/deep-research, 2025. Official documentation for Gemini Deep Research endpoints.

Ziwen Han, Meher Mankikar, Julian Michael, and Zifan Wang. Search-time data contamination. *arXiv preprint arXiv:2508.13180*, 2025.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, Peng Gao, Yu Liu, Chunyuan Li, and Hongsheng Li. Mmsearch: Benchmarking the potential of large models as multi-modal search engines, 2024. URL https://arxiv.org/abs/2409.12959.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Tian Lan, Bin Zhu, Qianghuai Jia, Junyang Ren, Haijun Li, Longyue Wang, Zhao Xu, Weihua Luo, and Kaifu Zhang. Deepwidesearch: Benchmarking depth and width in agentic information seeking. *arXiv preprint arXiv:2510.20168*, 2025.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474, 2020.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

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

Rui Lu, Zhenyu Hou, Zihan Wang, Hanchen Zhang, Xiao Liu, Yujiang Li, Shi Feng, Jie Tang, and Yuxiao Dong. Deepdive: Advancing deep search agents with knowledge graphs and multi-turn rl. *arXiv preprint arXiv:2509.10446*, 2025.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2200–2209, 2021.

OpenAI. Gpt-5 system card. `https://cdn.openai.com/gpt-5-system-card.pdf`, 2025a. System card describing the GPT-5 model family.

OpenAI. o3 / o4-mini system card. `https://openai.com/index/o3-o4-mini-system-card/`, 2025b. Official OpenAI system card for o3 and o4-mini models.

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

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

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, Qingyao Ai, Jen-Tse Huang, Wenxuan Wang, Yue Zhang, Yiming Yang, Zhaopeng Tu, and Zhaochun Ren. Deep research: A systematic survey, 2025. URL `https://arxiv.org/abs/2512.02038`.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Leo Liu, Zayne Sprague, et al. Chartmuseum: Testing visual reasoning capabilities of large vision-language models. *arXiv preprint arXiv:2505.13444*, 2025.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

Tongyi DeepResearch Team. Tongyi deepresearch: A new era of open-source ai researchers. https://github.com/Alibaba-NLP/DeepResearch, 2025.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Kung-Hsiang Huang, Yixin Mao, and Chien-Sheng Wu. Deeptrace: Auditing deep research ai systems for tracking reliability across citations and evidence. *arXiv preprint arXiv:2509.04499*, 2025.

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

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, et al. Widesearch: Benchmarking agentic broad info-seeking. *arXiv preprint arXiv:2508.07999*, 2025.

xAI. Grok: xai's multimodal reasoning model. https://x.ai/blog/grok, 2024. Official description of the Grok model family.

Yunjia Xi, Jianghao Lin, Menghui Zhu, Yongzhao Xiao, Zhuoying Ou, Jiaqi Liu, Tong Wan, Bo Chen, Weiwen Liu, Yasheng Wang, et al. Infodeepseek: Benchmarking agentic information seeking for retrieval-augmented generation. *arXiv preprint arXiv:2505.15872*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Fangda Ye, Yuxin Hu, Pengxiang Zhu, Yibo Li, Ziqi Jin, Yao Xiao, Yibo Wang, Lei Wang, Zhen Zhang, Lu Wang, Yue Deng, Bin Wang, Yifan Zhang, Liangcai Su, Xinyu Wang, He Zhao, Chen Wei, Qiang Ren, Bryan Hooi, An Bo, Shuicheng Yan, and Lidong Bing. Miroeval: Benchmarking multimodal deep research agents in process and outcome. *arXiv preprint arXiv:2603.28407*, 2026.

Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, Shiting Huang, Yiming Zhao, Xu Tang, Yao Hu, Philip Torr, Wanli Ouyang, and Shaosheng Cao. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Kai Zou, Ziqi Huang, Yuhao Dong, Shulin Tian, Dian Zheng, Hongbo Liu, Jingwen He, Bin Liu, Yu Qiao, and Ziwei Liu. Uni-mmmu: A massive multi-discipline multimodal unified benchmark. *arXiv preprint arXiv:2510.13759*, 2025.

A APPENDIX

A.1 ADDITIONAL DATASET CONSTRUCTION DETAILS

Two Task Regimes. MMDR-Bench contains 140 researcher-curated tasks spanning 19 domains, which are organized into two complementary regimes: Daily (40 tasks across 10 domains) and Research (100 tasks across 9 domains). Daily tasks reflect everyday deep-research workflows driven by loosely structured visual inputs, such as screenshots, photographs, and user-facing interfaces, and require lightweight yet verifiable evidence use. In contrast, Research tasks span scientific and social-science domains and emphasize analysis-heavy settings with information-dense visuals, including charts, diagrams, and tables, where models must integrate multi-source evidence and synthesize it into coherent, citation-grounded reports under stricter fidelity constraints.

Researcher-Driven Task Proposal and Refinement. Researchers with relevant disciplinary training propose candidate tasks, which we iteratively refine with clarity checks, multimodal-necessity checks (tasks must be image-dependent), and evidence-grounding checks (reports must be verifiable via citations). This process yields the final 140-task benchmark. The source pool contains 98k+ real-world queries and supplies material for task curation. The pool size describes the source collection rather than a selection ratio. The benchmark pairs 140 curated tasks with multi-stage evidence assessment, averaging 35.48 evaluation calls per report (Appendix A.5.2); task-paired bootstrap estimates for the matched study are reported in Appendix A.7.

Multimodal Packaging, Difficulty, and Multilinguality. Each task is packaged as an image–text bundle with a variable number of images. We annotate instance difficulty (easy, hard, complex) and record the task language. The benchmark is multilingual, dominated by English and Chinese, with additional languages in the long tail.

A.1.1 STANDARDIZED REPORT GENERATION PROTOCOL

We prompt models to generate reports grounded in Claim–URL verifiability and inline visual support. Citations should appear immediately after factual claims, and each citation index must map to exactly one URL in a `References:` block. For image-dependent claims, the report should embed the referenced image inline with a short caption carrying the same citation index, and task-provided input images should be embedded before drawing image-dependent conclusions. We encourage diverse sources (Daily about 6 and Research about 10 when feasible) while avoiding social media and question-answering forums, and we note that Daily inputs can be noisy or partial.

A.1.2 TEXTUALIZED VISUAL GROUND TRUTH FOR VEF.

VEF. uses a task-specific textualized visual ground truth (visual GT) that records directly observable facts from the task-provided images. Concretely, when a task author prepares a benchmark task, they additionally write a concise visual GT by describing the key visual facts in the image bundle as-is. Depending on the visual type, the GT may include: salient entities/objects and their identities; explicit numbers, labels, and table entries; chart titles, axes, units, legends, and trend/ordering relations; and screenshot/UI states such as selected options, warnings, and on-screen text. This GT is intended to be minimal yet sufficient to distinguish correct visual interpretation from visually grounded hallucinations such as mis-identifying the central entity in an image.

Decision Rule and Threshold. The judge returns a 0–10 score and a raw PASS/FAIL label. In the released worker, the numerical endpoint is $100\,\mathbf{1}[s_{\text{judge}} \geq 6]$; the numerical score determines the endpoint, and the raw PASS/FAIL label is retained for diagnostics. The judging instructions classify wrong identities and false presence as failures. The binary endpoint gives critical visual errors an explicit penalty within VEF, while the 0–10 score and raw label provide diagnostic detail. Historical score provenance is documented in Appendix A.5.1.

Future Versioning and Drift Management. Building on retained task/GT hashes for the audited subsets, future maintenance will extend immutable version identifiers, release snapshots, correction records, and canary regression checks to the benchmark.

Table 4: Weighting configuration. Adaptive dimension weights and fixed channel/module shares are distinct design choices.

Scope	Rule	Weighted quantities
Overall	Fixed	FLAE/TRACE/MOSAIC shares of 0.20/0.50/0.30.
FLAE dimensions	Task-adaptive	Readability, insightfulness, coherence, and structure; normalized weights W_d.
FLAE fusion	Type-fixed	Coefficient α combines fixed-transform and holistic scores; see Appendix A.3.4.
TRACE visual fidelity	Fixed	$\lambda_{\text{VEF.}} = 0.4$, giving VEF an Overall share of 0.20.
TRACE evidence	Task-adaptive	Consistency, coverage, fidelity, and source diversity; normalized weights W_k.
MOSAIC items	Type-fixed	Routed item weights in Table 5.

VEF. Score Semantics. For the set $\mathcal{D}_s$ with retained final VEF outcomes, let $y(t, R) \in \{0, 1\}$ denote that binary outcome. We report

$$\text{VEF.}(\mathcal{D}_s) = \frac{100}{|\mathcal{D}_s|} \sum_{t \in \mathcal{D}_s} y(t, R). \tag{5}$$

This defines the binary endpoint; historical aggregation and coverage are documented in Appendix A.5.1.

Quality Control and Calibration. Task authors inspect the input images to construct the textualized visual GT. As shown in Figure 8, the VEF judge receives the task text, visual GT, and report, rather than the images themselves. Human spot-checks on a small development subset inform refinements for incorrect entity identity, swapped labels, and incorrect numeric readings. Repeated-judging results appear in Appendix A.7.1.

A.2 Evaluation Metrics

We score each generated report with three modules: FLAE (generation quality), TRACE (evidence and task faithfulness), and MOSAIC (image-grounded integrity). Figure 5 summarizes the conceptual dependencies.

A.2.1 Weighting Hyperparameters

Table 4 summarizes the weighting hyperparameters used by the evaluator, including task-adaptive weights and fixed shares.

Rationale. Task-adaptive dimension weights are intended to accommodate heterogeneous tasks. Fixed module shares encode the benchmark's emphasis on evidence support. Sensitivity checks characterize ranking changes under the specified alternative configurations; the weights are design settings.

A.3 Module Implementation Details

A.3.1 Measured and Judge-Estimated Features

The directly measured features include type–token ratio TTR, heading density H, and long-sentence ratio L. The judge supplies E, p_{analysis}, $\rho_{\text{connective}}$, $\rho_{\text{synthesis}}$, U, O, T, κ, θ, and section-presence flags, together with holistic dimension scores. Here E represents errors; p_{analysis}, $\rho_{\text{connective}}$, and $\rho_{\text{synthesis}}$ describe analytical/synthetic content; U, O, T are coherence sub-scores; κ is the conflict term; and θ is the structural-adequacy factor. Insightfulness operationalizes analytical and synthetic content through these judge-estimated attributes. Retained feature values permit deterministic recomputation of the transforms.

A.3.2 FIXED FORMULA CHANNEL

Let $c(x) = \min(1, \max(0, x))$, $\sigma(x) = (1 + e^{-x})^{-1}$, and H_m be the harmonic mean with a 10^{-9} numerical floor. The retained evaluator computes

$$\begin{aligned} R_{\text{form}} &= c\big(\sigma(2TTR + 1.2H - 2L - 3E)\big), \\ I_{\text{form}} &= c\left[\big(\max(10^{-9}, p_{\text{analysis}})\ \max(10^{-9}, \rho_{\text{connective}})\ \max(10^{-9}, \rho_{\text{synthesis}})\big)^{1/3}\right], \\ C_{\text{form}} &= c\big(H_m(U, O, T)\max(0, 1 - 2\kappa)\big), \\ S_{\text{form}} &= c\big(P_{\text{sections}}\min(1, n_{\text{img}}/n^*_{\text{img}})c(\theta)\big). \end{aligned} \tag{6}$$

Here n_{img} counts unique HTTP(S) Markdown image URLs, and $n^*_{\text{img}} = \max(1, n_{\text{words}}/350)$ for evaluator label `research`, or $\max(1, n_{\text{words}}/700)$ otherwise. P_{sections} is the product of the implementation's required-section indicators. The research list comprises Executive Summary, Background, Methodology, Analysis & Evidence, Visual Integration, and Conclusion; other labels use the first, fourth, and sixth entries.

Evaluator Version and Structural Scoring. The reported results use the same evaluator version across systems: its lowercase detected-section keys are looked up with title-case required keys, yielding $P_{\text{sections}} = S_{\text{form}} = 0$, so structural scoring retains the weighted judge-channel contribution and the section-based $g_S+ = 1$ adjustment always applies.

A.3.3 JUDGE INPUTS AND OUTPUT SCHEMA

The feature/scoring call receives the task, evaluator report-type hint, and the first 1,400/14,000 characters of the rules/report, returning strict JSON with the feature fields, section-presence map, and optional $[0, 1]$ holistic fields `llm_R`, `llm_I`, `llm_C`, and `llm_S`. Missing holistic fields use the corresponding formula scores; failed structured calls use the implementation's default feature/score record.

A.3.4 TASK-ADAPTIVE WEIGHTS AND TYPE-FIXED FUSION

Fusion uses the retained per-report evaluator label and coefficient; evaluator labels may differ from the benchmark's canonical task regimes. The coefficient follows a fixed lookup:

$$\alpha = \begin{cases} 0.60, & \texttt{research}, \\ 0.40, & \texttt{daily_hard}, \\ 0.30, & \texttt{daily_easy}, \\ 0.35, & \text{otherwise}. \end{cases} \qquad s_d = \alpha s_d^{\text{form}} + (1 - \alpha) s_d^{\text{judge}}. \tag{7}$$

This is distinct from adaptive dimension weighting. A separate task-controller call receives the question and report-type hint and returns four-dimensional adjustment vectors `v_general` and `v_evidence`. For FLAE,

$$W_d = \text{softmax}_d\big(\log(\pi_d + 10^{-9}) + v_d + g_d\big), \qquad \pi = (0.25, 0.25, 0.25, 0.25). \tag{8}$$

In (R, I, C, S) order, g starts at zero. The code adds 1 to g_S if fused $S < 0.5$ or $P_{\text{sections}} = 0$; adds 0.8 to g_R if $E > 0.10$ or $L > 0.35$; and adds 0.4 to both g_I and g_C if $H > 0.12$ and $S > 0.7$. These rules define the implemented weighting configuration.

A.3.5 TRACE: EVIDENCE AND TASK FAITHFULNESS

TRACE evaluates citation fidelity and multimodal task faithfulness, aggregating as Eq. (3). We parse citation markers and the references: block to build a one-to-one index-to-URL map, extract atomic claims, and align each claim to its cited URL(s) to form claim-URL pairs. For accessible pages, a judge verifies whether each claim is supported and aggregates citation fidelity into $\mathcal{K}$: CON., COV., FID., and DIV.. We additionally compute VEF. by matching the report to a textualized visual ground truth for the task, then threshold it at $\tau_{\text{VEF.}}$.

SYSTEM
You are a STRICT QA Judge for VEF.. Use the task, the provided visual ground truth, and the report. Return **only** valid JSON.
USER
Segment:
{SEGMENT}
Question:
{TASK_TEXT}
Visual ground truth (text-form requirements):
{VEF_GT}
Report:
{REPORT_TEXT}

Rules: any wrong visual identity is FAIL; any false presence is FAIL; missing details allowed only if no wrong identities; score below 6 must be FAIL.
Output JSON with keys: score, reason, verdict (PASS or FAIL).

Figure 8: TRACE prompt template for VEF..

Table 5: Default MOSAIC item-score weights by routed visual type.

Visual type	Item score s_i (normalized)
DATACHART / OCRCHART	$s_i = f_i \cdot \left(0.9\,\text{ACC.}_i + 0.1\,\text{VQA}_i\right)$
PHOTO	$s_i = f_i \cdot \left(0.5\,\text{SEM.}_i + 0.5\,\text{VQA}_i\right)$
DIAGRAM	$s_i = f_i \cdot \left(0.5\,\text{VQA}_i + 0.5\,(1 - h_i)\right)$

Text-Evidence Components and Weights. Consistency is a relevance-weighted support mean multiplied by one minus the prohibited-source fraction. Coverage is a soft supported fraction over extracted key textual claims, with the denominator restricted to extracted key textual claims and image-dependent claims excluded. Fidelity is the harmonic mean of mean non-hallucination, non-inversion, and chain-soundness scores. Diversity is normalized domain entropy multiplied by $\min(1, (m_{\text{auth}} + 0.5(m_{\text{acc}} - m_{\text{auth}}))/3)$, where the counts denote authoritative and non-prohibited domains.

Evidence weights are $W = \text{softmax}(\log(\pi + 10^{-9}) + v + h)$ with prior $\pi = (0.40, 0.25, 0.25, 0.10)$ in (Con., Cov., Fid., Div.) order and task-controller vector v. Starting from zero, the diagnostic adjustment h adds 1 to Con. and subtracts 0.8 from Div. when the prohibited-source fraction exceeds 0.05; adds 0.8 to Cov. when at least three key claims have coverage below 0.4; and adds 0.8 to Fid. when mean hallucination exceeds 0.25 or inversion exceeds 0.15. Empty text-evidence sets receive a zero aggregate. For non-prohibited sources whose fetch fails, the retained implementation floors support at 0.18 (excluding direct-image/local items); this heuristic fallback distinguishes an unverified citation from a contradicted one without assigning full support.

VEF. Thresholding. The threshold and diagnostic fields are defined in Appendix A.1.2; repeated-judging consistency is reported in Appendix A.7.1.

VEF. Judge Prompt Template. Figure 8 shows the prompt used to score VEF. against the task-specific textualized visual ground truth and to output a PASS/FAIL verdict.

A.3.6 MOSAIC: IMAGE-GROUNDED INTEGRITY

MOSAIC checks whether image-grounded statements match the referenced visuals and aggregates by Eq. (4). It extracts figure-linked items and routes them into visual types (photo, datachart, ocrchart, diagram). Each item is scored with a formatting and integration factor $f_i \in [0, 1]$; for diagrams we additionally use a hallucination factor $h_i \in [0, 1]$.

Visual Claim Support and Temperature. The Gemini VQA module asks whether the image supports its associated statement at temperature 0.0 (other primary-judge calls: 0.2), using up to five first-token log-probability candidates to compute positive mass (yes/true/correct) divided by positive plus negative mass (no/false/incorrect/not). It falls back to returned yes/no text when log-probabilities are unavailable, and to zero for insufficient recognized mass or API exceptions.

A.4 Evaluated Model Details

Table 6 lists the 28 evaluated systems, access platforms or model families, usage settings, and testing windows. The original systems were tested in December 2025; GPT-5.4, Gemini 3.1 Pro, and Gemini 3.1 Flash-Lite were added in April 2026, using Azure for GPT-5.4 and the Gemini API for the two Gemini models. Exact backend snapshot identifiers are controlled by the serving platforms and are unavailable in our retained records. We report model names, access platforms, and testing windows as the available identifiers. Exact regeneration is constrained by platform-managed backend changes; frozen-report rescoring holds the generated reports fixed. The retained label "ChatGPT Deep Research (o3-mini)" denotes the historical Azure-based setup; its provenance is recorded at the platform and deployment-label level. We group systems into three tiers to reflect progressively stronger tool access and orchestration capability. Tier 1 covers single-shot LLM/LMM baselines with external browsing disabled, corresponding to the first two groups in Table 1. For the text-only baselines (o3-mini, DeepSeek-V3.2, Kimi K2, and Qwen 3 235B), retained task inputs contain query text and textual image-file references; these references do not convey image content, and the systems serve as no-vision reference points. Tier 2 includes web-enabled report generators with built-in browsing, representing mainstream tool-using LMM deployments. Tier 3 contains dedicated Deep Research Agents that explicitly orchestrate multi-step retrieval and synthesis, and thus reflect agentic search-and-write behavior beyond a single response.

A.5 Coverage, Reliability, and Missing Data

A.5.1 Additional Experimental Setup and Reliability Handling

We report reliability-related settings that affect score stability and effective coverage.

Primary Judge Configuration. Gemini-2.5-Pro is the primary judge. General judging calls default to temperature 0.2, while the VQA yes/no module uses 0.0 to reduce sampling variability. Its token-level scoring and fallbacks are specified in Appendix A.3.6.

TRACE Weighting. With the benchmark-level module weight $w_T = 0.5$, we set $\lambda_{\text{VEF.}} = 0.4$ so that VEF. contributes $0.5 \times 0.4 = 0.2$ of the overall score. The remaining 0.3 overall share from TRACE is allocated to citation-fidelity metrics (CON., COV., FID., DIV.) using task-adaptive weights.

MOSAIC Gate and Missing Outputs. MOSAIC is gated on valid positive report-quality and text-evidence scores. A missing or unusable MOSAIC output contributes zero within a scored task; this differs from excluding a task for which no complete pipeline score is available. No reason-aware validity penalty is applied to Table 1.

Score Provenance and Coverage. Table 1 preserves the original reported values. Historical VEF verdicts were also evaluated separately and incorporated offline. Reconciliation of the main-table values with separately retained run-level aggregation records and per-task denominator manifests remains incomplete; results from distinct runs are kept separate. The benchmark contains 140 tasks, with repeated provider refusals for some image-based requests. MOSAIC submetrics are conditional means over applicable task-level dimension summaries with retained scores, so denominators can differ across systems. The later-run components and estimated values used for weight sensitivity are identified in Appendix A.7.

Interpretation of Tables and Domain Plots. Column highlights in Table 1 mark descriptive maxima among the displayed means. Figure 7 illustrates actual Research-task scores for five selected systems across unequal-sized domains. Cross-generation comparisons involve the evaluated backbones, tool configurations, testing windows, and effective coverage.

Table 6: Evaluated systems, access information, and testing windows. Identifiers comprise model names, access platforms, and testing windows; exact platform-managed backend snapshots are unavailable.

Tier	System (paper)	Modality/Setting	Testing window	Provider / access
Tier 1: Single-shot LLM/LMM baselines				
T1	OpenAI o3-mini	Text/Reasoning	December 2025	OpenAI models[a]
T1	DeepSeek-V3.2 (Base)	Text	December 2025	DeepSeek API[b]
T1	Kimi K2 (Thinking)	Text/Reasoning	December 2025	Kimi API[c]
T1	Qwen 3 235B (A22B)	Text	December 2025	Qwen API[d]
T1	Qwen 3 VL 235B (A22B)	Multimodal	December 2025	Qwen-VL[d]
T1	GPT-4o	Multimodal	December 2025	OpenAI models[a]
T1	GPT-4.1	Multimodal	December 2025	OpenAI models[a]
T1	GPT-4.1 mini	Multimodal	December 2025	OpenAI models[a]
T1	GPT-4.1 nano	Multimodal	December 2025	OpenAI models[a]
T1	GPT-5 mini	Multimodal	December 2025	OpenAI models[a]
T1	GPT-5.1	Multimodal	December 2025	OpenAI models[a]
T1	GPT-5.2	Multimodal	December 2025	OpenAI models[a]
T1	GPT-5.4	Multimodal	April 2026	Azure
T1	Grok-3	No web browsing	December 2025	xAI API[e]
T1	Grok-4 (Fast Reasoning)	Reasoning	December 2025	xAI API[e]
Tier 2: Web-enabled report generators				
T2	Claude 4.5 Haiku	Web-enabled	December 2025	Anthropic[f]
T2	Claude 4.5 Sonnet	Web-enabled	December 2025	Anthropic[f]
T2	Claude 4.5 Opus	Web-enabled	December 2025	Anthropic[f]
T2	Gemini 2.5 Flash	Web-enabled	December 2025	Gemini API[g]
T2	Gemini 2.5 Pro	Web-enabled	December 2025	Gemini API[g]
T2	Gemini 3 Flash	Web-enabled	December 2025	Gemini API[g]
T2	Gemini 3 Pro	Web-enabled	December 2025	Gemini API[g]
T2	Gemini 3.1 Pro	Web-enabled	April 2026	Gemini API
T2	Gemini 3.1 Flash-Lite	Web-enabled	April 2026	Gemini API
Tier 3: Deep Research Agents (DRA)				
T3	Gemini Deep Research (Gemini 3 Pro)	Agentic DR	December 2025	Google DR[h]
T3	ChatGPT Deep Research (o3-mini)	Agentic DR	December 2025	Azure (historical label)
T3	Tongyi Deep Research (30B-A3B)	Agentic/IR	December 2025	Tongyi DR[i]
T3	Perplexity Sonar Deep Research	Agentic/IR	December 2025	Perplexity[j]

[a] OpenAI model-family and system documentation: (OpenAI, 2025b;a; Achiam et al., 2023).
[b] DeepSeek documentation: (DeepSeek-AI et al., 2025).
[c] Kimi documentation: (Team et al., 2025).
[d] Qwen/Qwen-VL documentation: (Yang et al., 2025; Bai et al., 2025).
[e] xAI Grok documentation: (xAI, 2024).
[f] Anthropic Claude documentation: (Anthropic, 2024).
[g] Gemini model-family documentation: (Gemini Team, 2025).
[h] Google Gemini Deep Research: (Google, 2025; Gemini Team, 2025).
[i] Tongyi Deep Research: (Team, 2025).
[j] Perplexity Sonar Deep Research: (AI, 2025).

A.5.2 Evaluation Cost and Timing

A full end-to-end evaluation from raw reports requires 35.48 model calls per report on average (median 35, p90 56, max 70). The mean call decomposition is 2.00 calls for FLAE, 14.92 for TRACE, 1.00 for Vef., and 17.56 for MOSAIC. Evaluating one system on all 140 tasks therefore requires roughly 4,967 evaluation calls; extrapolating gives about 139,000 calls for one 28-system scoring cycle.

These figures cover evaluation only. Across 3,439 reports with retained timing instrumentation, introduced partway through the runs, median end-to-end scoring time was 163 seconds per report (mean 185, p90 356). Under the Gemini-2.5-Pro configuration, scoring cost about USD 0.36 per report, or approximately USD 50 to evaluate an additional system on 140 tasks.

A.5.3 Failure-Reason Taxonomy

Missing or unscorable outputs may arise from unusable model responses, parser/schema failures, provider refusals or infrastructure errors, and inaccessible evidence assets. These categories support

diagnostic interpretation of the retained logs. Provider and access failures indicate operational constraints; reason tags support diagnosis.

A.5.4 FAILURE CASE ANALYSIS

For qualitative analysis of multimodal task understanding, we select reports marked FAIL by the Visual Evidence Fidelity (VEF.) check in TRACE. We examine visual-evidence issues in this judge-selected report set using the five categories below; other metrics provide complementary diagnostics.

We group VEF.-level failures into five categories: EMI (entity mis-identification), RMD (reference or mapping drift), DTE (detail or symbol extraction errors), LKC (logical or knowledge conflicts), and STO (structural or task-level omissions). Figure 6 summarizes the distribution of these failure modes, comparing (left) a text-only backbone versus a vision-enabled backbone and (right) a base model versus its agentic deep-research system.

On the left, the vision-enabled model shows a clear increase in DTE, consistent with mis-reading fine-grained literals such as small numerals, axis labels, timestamps, or table cells. These local errors leave some report claims unsupported.

On the right, the agentic system exhibits more EMI failures in final reports, including evidence bound to an incorrect referent. Attributing these final-report errors to intermediate retrieval or reasoning steps requires process logs.

A.6 HUMAN EVALUATION PROTOCOL

The original author-group study assesses evaluator–human agreement; two supplementary blinded audits below examine report preferences and visual judgments on frozen artifacts. We use the full 140 tasks from both Daily and Research regimes and collect reports from all evaluated models. For each task, we form balanced system pairs by stratifying on overall score and tier (to avoid trivial comparisons), and sample a fixed number of pairs per task. Twelve annotators with relevant disciplinary background independently assess report pairs: for each pair, three annotators provide an overall preference and a coarse score on the same rubric, and we aggregate by majority vote. Evaluator–human agreement is measured by pairwise agreement (PAR) on preferences and score correlation (OPC), where OPC is the Pearson correlation between system-level mean scores computed by averaging over tasks and sampled pairs. We also manually spot-check extracted Claim–URL units and supporting snippets in a subset of pairs. These checks provide a qualitative audit of the sampled evidence units. Only aggregate results remain available for the original study; exact pair counts, individual votes, tie handling, and the computation underlying the reported 69.8% inter-annotator agreement cannot be reconstructed from them.

We additionally conduct two small audits on frozen artifacts with samples and endpoints distinct from the original study. In the preference pilot, two non-author raters blindly assess twenty score-independently selected report pairs with system identities and evaluator scores hidden. They agree on 13/20 A/B/Tie labels (Cohen's $\kappa = 0.466$); an author who had not previously read the paired reports adjudicates the seven disagreements using the same hidden-identity interface. In this small three-class pilot (A/B/Tie), pooled strict agreement is 40.0% at the default and 38.3–45.0% across six local weight settings; the differences are statistically inconclusive.

In the visual audit, two blinded raters assess 48 self-contained MOSAIC visual judgments, yielding human–human quadratic-weighted $\kappa = 0.887$. All 13 VQA items classified as supported by MOSAIC receive the maximum human rating. This positive-item agreement concerns the audited evaluator-positive items; the audit does not estimate recall or false-negative rates.

A.7 ROBUSTNESS OF JUDGE MODEL AND WEIGHTS

We test robustness to judge backbones and aggregation weights in two ways. First, we retain the cross-judge comparison from the original experiment, where a fixed set of Gemini-generated reports is re-scored by two judge backbones under the same parsing, retrieval, and aggregation pipeline. Second, to isolate judge choice more strictly, we conduct a controlled same-report study on 120

byte-identical reports from six systems and twenty shared tasks, with reports, evidence artifacts, prompts, routing, score ranges, and aggregation held fixed.

Table 3 (left) shows module-level differences on the fixed report set: GPT-5.2 assigns lower mean FLAE and VEF. scores, while mean MOSAIC scores are similar.

Despite per-stage shifts, the overall MMDR score changes by only 0.30 points (36.76 to 37.06, about 0.8%), with opposing module shifts contributing to aggregate stability. Repeated-scoring consistency is assessed separately below.

Controlled same-report judge study. The controlled subset contains six systems and twenty common tasks (14 Research, 6 Daily), selected before new judging using only exact artifact availability, provider/model-family diversity, rank strata, system category, and task metadata. Rank stratification used previously available scores, and selection was finalized before new cross-judge scores and disagreement outcomes were computed. We compare retained exact-match Gemini-2.5-Pro evaluations with new GPT-5.4 and Grok-4.3 rescoring. All three judge backbones preserve the winner. Pairwise system-ranking correlations are descriptive because $n = 6$, and characterize ordering across providers: Spearman/Kendall are 0.714/0.467 for Gemini vs. GPT-5.4, 0.943/0.867 for Gemini vs. Grok-4.3, and 0.771/0.600 for GPT-5.4 vs. Grok-4.3. Top-3 overlap ranges from 2/3 to 3/3 and Top-5 overlap from 4/5 to 5/5. A 10,000-replicate task-paired bootstrap further preserves the winner in 98.4%, 93.3%, and 99.8% of resamples under Gemini-2.5-Pro, GPT-5.4, and Grok-4.3, respectively. Winner-rank 95% bootstrap intervals are $[1, 1]$, $[1, 2]$, and $[1, 1]$, with mean resampled-ranking Spearman correlations of 0.803, 0.954, and 0.888 against the official ordering restricted to the matched systems. In the 14-task Research stratum, winner preservation is 84.5%, 97.0%, and 97.4%; the six-task Daily stratum is too small for stable stratified estimates. Middle-rank intervals are wide.

Same-family score differentials. For reports from a judge's own model family, we compare that judge's score with the mean score from the other two judges; negative values indicate lower own-family scores. The mean differences are -3.41 points for Gemini, -2.53 for GPT, and +1.61 for Grok. On this matched subset, the Gemini differential does not show an own-family score uplift; the mixed signs do not support a uniform self-preference pattern across judges.

Holistic judge control and component interpretation. On the same 120 frozen reports, plain holistic LLM-as-a-judge controls produce a stable scalar ordering. The modular pipeline additionally provides module- and item-level diagnostics for visual fidelity, citation support, and report-image alignment. Separately, the original human study in Table 2 reports higher preference agreement for the full evaluator than for its vanilla-prompt baseline; that baseline and protocol differ from the frozen-report holistic control.

Aggregation-weight sensitivity. We perturb the official top-level weights around $(w_F, w_T, w_M) = (0.2, 0.5, 0.3)$ using six local settings, each transferring 0.10 between one component pair while preserving the published internal TRACE coefficient $\lambda_{\text{VEF.}} = 0.4$. The six (w_F, w_T, w_M) settings are $(0.1, 0.5, 0.4)$, $(0.1, 0.6, 0.3)$, $(0.2, 0.4, 0.4)$, $(0.2, 0.6, 0.2)$, $(0.3, 0.4, 0.3)$, and $(0.3, 0.5, 0.2)$.

The component summaries and displayed Overall scores come from separate runs: the former were obtained in a later independently restarted run. Each system is anchored at its historical Overall and receives only a counterfactual reweighting delta from those later components. One system, Gemini 2.5 Pro, uses partly estimated components constrained by displayed submetrics and Overall; the retained exclusion check gives the same qualitative conclusion. This is an aggregate-level reweighting analysis, not a full-pipeline rerun.

Across these six local perturbations, the winner and full Top-3 set are preserved; mean Spearman with the official 28-system leaderboard is 0.978 (minimum 0.964), and mean Top-5 overlap is 4.5/5.

Alternative aggregation on the available components is subject to the same provenance limitation. Geometric, harmonic, and min-gated aggregation preserve the winner and Top-3 set (Spearman 0.938, 0.871, and 0.755). Equal weighting has Spearman 0.967 and swaps only the top two systems, which are separated by 1.18 points in the official table. Less compensatory aggregation gives weak modules greater influence; the lower min-gated correlation illustrates the sensitivity of ordering to this aggregation choice.

FLAE internal-weight sensitivity and component complementarity. We perturb each task-adaptive FLAE dimension weight independently by up to $\pm 30\%$, renormalize the weights per task, and

recompute system rankings over 2,000 randomized perturbations. Mean Spearman correlation with the default FLAE ranking is 0.999 and the minimum is 0.995; uniform and global-mean weighting are similarly stable. Across report-level module scores, the largest absolute pairwise correlation is 0.313. Under the official judge, leave-one-component-out rankings (with remaining top-level weights renormalized) have Spearman correlations of 0.943, 0.829, and 0.771 with the full ranking when removing FLAE, TRACE, and MOSAIC, respectively.

A.7.1 Intra-Judge Consistency: Full Results

We assess repeated-scoring consistency separately from cross-judge and aggregation-weight sensitivity. Gemini-3.1-Pro reports cover 18 tasks with three generation runs each; GPT-5.4 reports cover 17 tasks with three runs each, excluding Q115 because generation failed. Each fixed report is judged five times by Gemini-2.5-Pro with default temperature 0.2 (VQA uses 0.0), yielding 270 and 255 ratings, respectively.

Table 7: Repeated-judge consistency for two report generators. ICC denotes ICC(3,1), α is Krippendorff's agreement coefficient, and CV is coefficient of variation.

	Gemini-3.1-Pro reports			GPT-5.4 reports		
Module	ICC	α	CV	ICC	α	CV
FLAE	0.295	0.283	0.161	0.153	0.148	0.180
TRACE (w/o VEF.)	0.884	0.883	0.210	0.702	0.699	0.213
VEF.	0.273	0.271	0.558	0.506	0.494	0.462
MOSAIC	0.970	0.970	0.081	0.983	0.983	0.097
Overall MMDR	0.709	0.705	0.160	0.614	0.602	0.158

Decision-level repeatability and interpretation. On the 54 Gemini report instances with five repeated ratings each, VEF.'s binary PASS/FAIL judgments have 73.3% mean pairwise agreement. We report this alongside ICC because the thresholded endpoint differs from a continuous score. With each report held fixed, these repeated ratings measure judge variability. Overall agreement is higher than FLAE and VEF agreement. These report-level repeatability results are distinct from system scores averaged across tasks; the task-paired bootstrap above separately assesses system-ranking stability.

A.8 Extended Limitations and Mitigations

Evaluation scope and coverage. MMDR-Bench evaluates support relationships between final reports, images, and cited sources. Independent factual verification and causal attribution to retrieval or reasoning stages require additional evidence and protocols. Coverage is defined over extracted key textual claims, and source diversity describes the cited-source distribution. The benchmark is predominantly English and Chinese. Separate module diagnostics, supplementary blinded audits, and documented evaluator-version and coverage conventions support interpretation of these scores (Appendices A.3.2, A.5, and A.6).

Dynamic services and public tasks. Retained reports and evidence enable frozen-artifact rescoring. Regeneration also depends on live web access and platform-managed model backends. Public task exposure introduces a risk of memorized task framings; held-out tasks, expanded language coverage, and release versioning are planned maintenance directions.

Judge variability. Appendix A.7.1 reports distinct repeatability levels for MOSAIC, FLAE, and VEF. VEF's hard threshold is sensitive to borderline judgments. System-ranking sensitivity and individual-report repeatability characterize complementary evaluation levels; repeated judging can quantify uncertainty at additional evaluation cost.

Figures 9 and 10 provide annotated report examples from Grok-4 (Fast Reasoning) and Gemini-2.5-Pro; Figure 11 adds a retained VEF FAIL case illustrating image-reference misalignment.

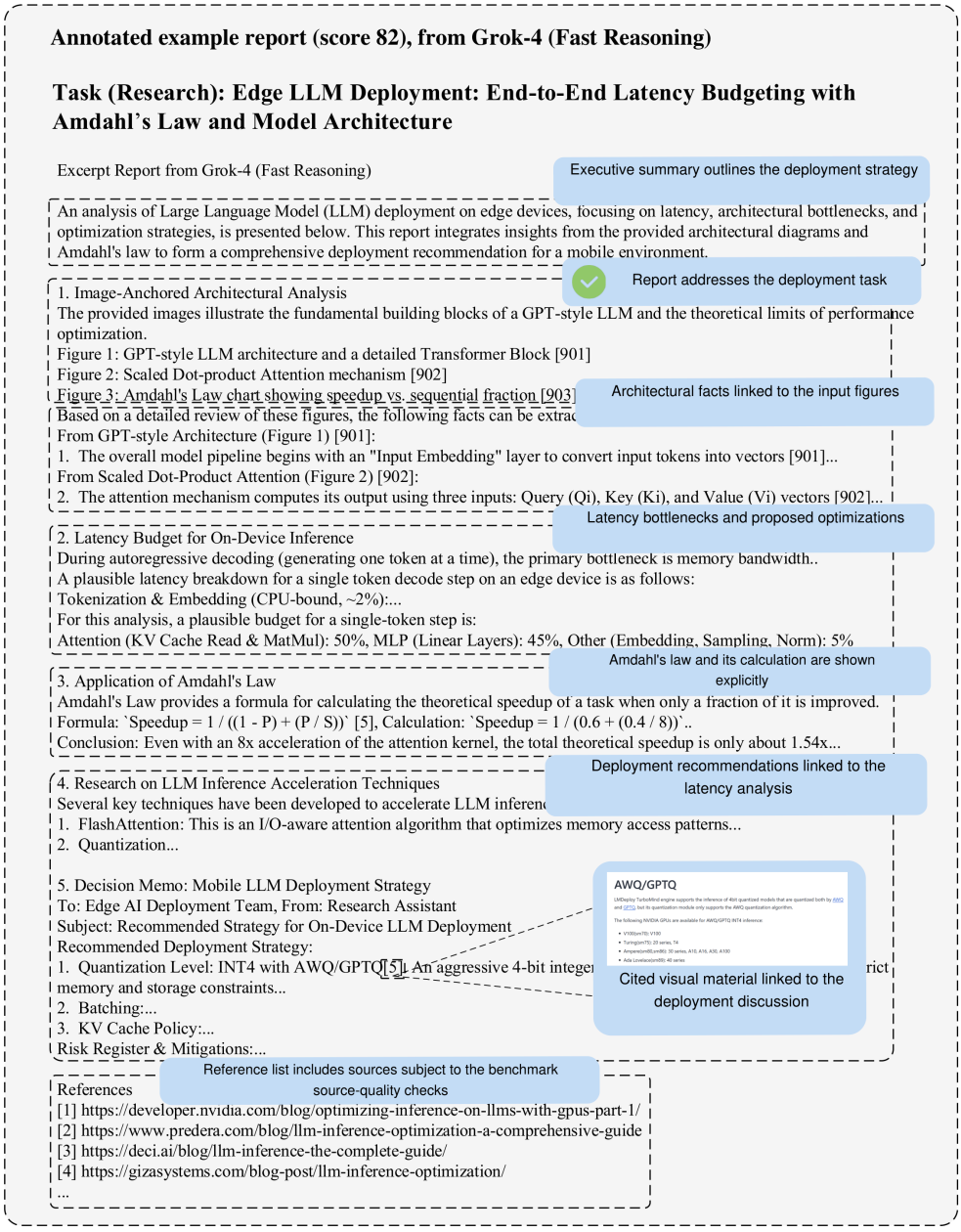

Figure 9: Example scored report generated by Grok-4 (Fast Reasoning).

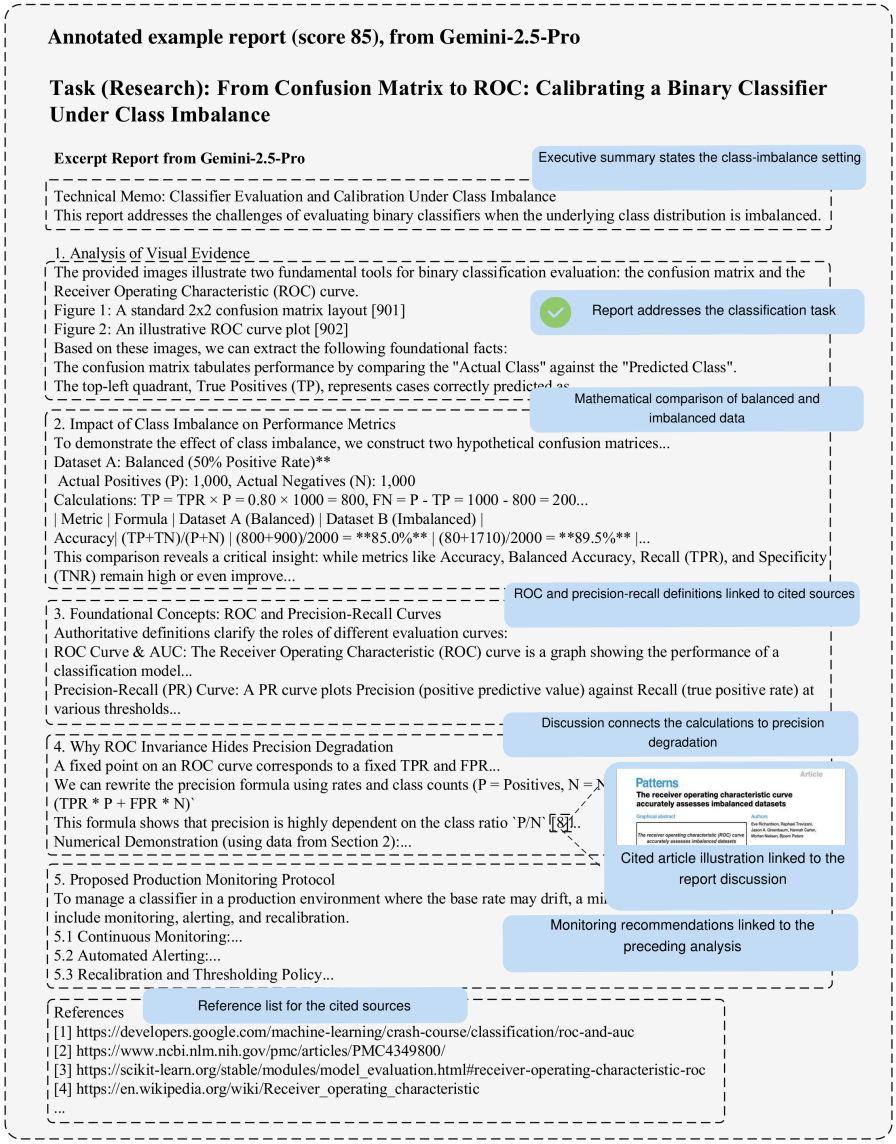

Figure 10: Example scored report generated by Gemini-2.5-Pro.

Retained VEF FAIL example: Qwen3-VL, R0012
Task visuals. A population table, an unemployment plot, and an asset/output plot for Japan.
Textual visual GT (translated summary). Image 00: population data for 2000–2015; Image 01: long-term unemployment rates; Image 02: asset/output indices.
Report headings (translated from Chinese). "Image 00: Japan's unemployment rate"; "Image 01: Assets/output"; "Image 02: Population changes."
Retained judge result. FAIL; diagnostic score 1/10 under Gemini-2.5-Pro. The report cyclically assigns all three image numbers to the wrong visual content.
Failure category. Reference or mapping drift (RMD): statements are attributed to the wrong task images.
Selection. This report is retained in the matched audit; the example uses its existing evaluation record.

Figure 11: A retained VEF failure from Qwen3-VL on task R0012. The visual GT and report headings are translated for presentation; the FAIL label and diagnostic score come from the retained evaluation record.