ApodexEval: Benchmarking Multimodal Deep Research Agents in Process and Outcome

Fangda Ye[1,2*†], Yuxin Hu[1,2*†], Pengxiang Zhu[1,2*†], Yibo Li[1,2*†], Ziqi Jin[1,3†], Yao Xiao[1], Yibo Wang[1†], Lei Wang[1‡], Zhen Zhang[1], Lu Wang[1], Yue Deng[1], Bin Wang[1], Yifan Zhang[1], Liangcai Su[1], Xinyu Wang[1], He Zhao[1], Chen Wei[1], Qiang Ren[1], Bryan Hooi[2], Bo An[1,3], Shuicheng Yan[2], Lidong Bing[1✉]

[1]Apodex [2]National University of Singapore [3]Nanyang Technological University

Abstract

Recent progress in deep research systems has been impressive, yet their evaluation remains limited not only by benchmark coverage but also by methodology. Existing benchmarks predominantly assess final reports using fixed rubrics, failing to evaluate the underlying research process. Most also offer limited multimodal coverage, rely on synthetic tasks that do not reflect real-world query complexity, and cannot be refreshed as knowledge evolves. To address these gaps, we introduce **ApodexEval**, an evaluation framework for deep research systems, accompanied by a 100-task benchmark (70 text-only, 30 multimodal) constructed via a dual-path pipeline that supports periodic updates, enabling a live and evolving setting. The proposed evaluation suite assesses deep research systems along three complementary dimensions: adaptive synthesis quality evaluation with task-specific rubrics, agentic factuality evaluation via active retrieval and reasoning over both web sources and multimodal attachments, and process-centric evaluation that audits how the system searches, reasons, and refines throughout its investigation. Evaluation across 11 systems supports a diagnostic reading, in which deep-research capability is not a single quantity, through three principal findings: the three evaluation dimensions capture complementary aspects of system capability, with each revealing distinct strengths and weaknesses across systems; the research process reveals weaknesses invisible to output-level metrics, most notably that report content often cannot be traced back to the documented research; and multimodal tasks degrade most systems, with declines concentrated in synthesis and process while factuality stays nearly flat. Each component is validated against human judgment, complemented by an adaptive-rubric ablation, a judge-family bias analysis, and three-annotator benchmark verification. ApodexEval provides a holistic diagnostic tool for the next generation of deep research agents.

1 Introduction

The rapid advancement of Large Language Models (LLMs) has driven a pivotal transition from passive text generation to agentic systems capable of autonomous planning and execution [39, 25, 10, 30, 45].

*Equal contribution.

†Work done during internships at Apodex.

‡Project lead.

✉Corresponding author: lyndon@apodex.com.

[1]This work was supported by Apodex, which provided the compute, storage, API access, and engineering infrastructure for the reported experiments.

40th Conference on Neural Information Processing Systems (NeurIPS 2026). Track on Evaluations and Datasets.

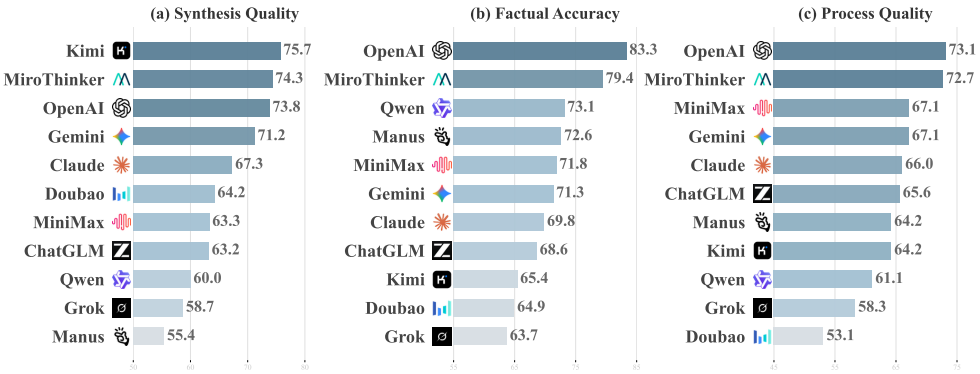


Figure 1: Model performance comparison on 70 text-only deep research tasks.

Deep research, broadly defined as the autonomous, multi-step process of investigating complex information needs through iterative search, evidence gathering, verification, and synthesis [18, 49, 7], has become a prominent agentic paradigm in this transition. Deep research systems [36, 2, 31, 14, 13] operationalize this paradigm by integrating planning, tool use, heterogeneous source interaction, and long-form report generation into a unified workflow. As these systems are increasingly adopted in high-stakes domains such as finance, healthcare, and legal analysis, users demand more than a fluent final report: they need answers that are factually reliable, grounded in thorough and traceable investigation, and capable of incorporating multimodal materials (images, PDFs, spreadsheets) that real-world research queries often involve. Meeting these demands requires continued improvement of deep research systems, which in turn requires reliable ways to measure whether a system truly conducts thorough, factually grounded investigation or merely produces a plausible-looking report.

Existing benchmarks have made valuable progress [1, 22, 24], but limitations persist along both axes. On the methodology side, three problems stand out. First, most benchmarks evaluate only the final report and leave the underlying research process unobserved [5, 22, 32]. A system that produces a polished report through redundant exploration is then indistinguishable from one that follows a disciplined investigation. Second, evaluation rubrics are typically fixed and task-agnostic, applying uniform criteria across heterogeneous queries. A financial market analysis and a clinical treatment comparison are scored under the same general dimensions, even though each demands its own task-specific expertise [9, 46]. Third, evaluation methodologies are predominantly text-centric, reducing multimodal assessment to short-form QA over attachments rather than long-form, attachment-grounded analysis [21, 16, 19, 12]. On the benchmark side, task construction often relies on synthetic or academic queries that do not fully capture the complexity of authentic user needs [51, 32, 38], and static benchmarks risk becoming stale as the information landscape evolves [20, 37].

To address these challenges, we introduce **ApodexEval**, a three-layer evaluation framework for deep research systems (§3), accompanied by a refreshable 100-task benchmark drawn from authentic deep-research usage patterns (§2). The framework comprises three layers, each natively supporting multimodal inputs. **Comprehensive Adaptive Synthesis Quality Evaluation** (§3.1) dynamically generates task-specific rubrics and importance weights to assess the final report, moving beyond fixed criteria to capture domain-specific nuances. **Agentic Factuality Evaluation** (§3.2) decomposes reports into atomic claims and employs an evaluation agent to verify them against both live web sources and multimodal attachments, utilizing a four-way consistency assessment: `RIGHT`, `WRONG`, `CONFLICT`, or `UNKNOWN`. **Process-Centric Evaluation** (§3.3) audits research trajectories across five intrinsic dimensions, including search breadth, analytical depth, progressive refinement, critical thinking, and efficiency. It further measures bidirectional alignment between process findings and the final report (Process→Report and Report→Process) along with contradiction detection. The accompanying benchmark, comprising 70 text-only and 30 multimodal tasks across 12 domains and 10 task types, is built through two complementary paths, both grounded in authentic deep-research demands and both refreshable, though not in the same way: the auto-generated path can be re-run on demand, while the user-derived path refreshes only as new internal-testing data accrues. One path rewrites real user queries under privacy-preserving anonymization. The other distills recurring patterns from user demands and generates new queries by agentically retrieving real-time web trends.

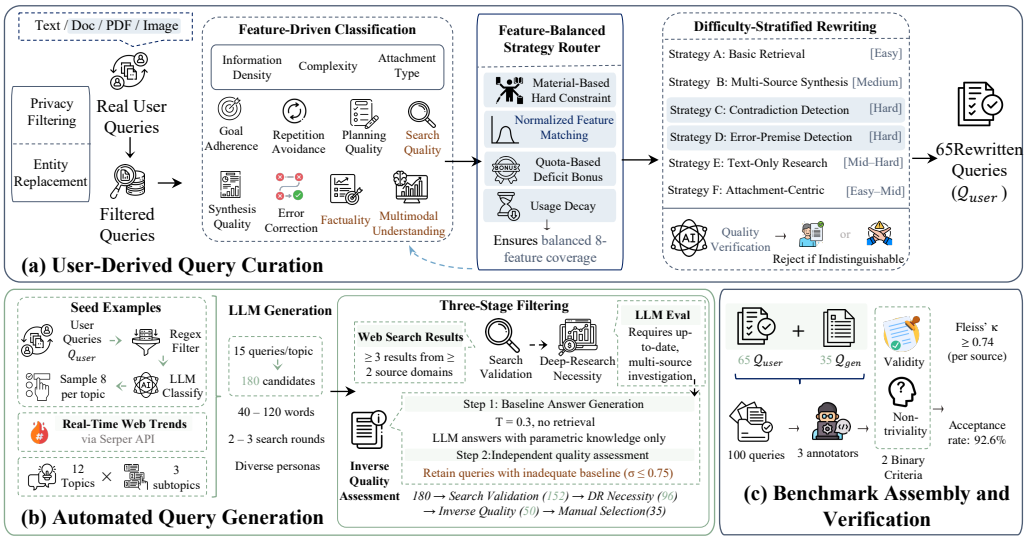


Figure 2: Overview of the dual-path query construction pipeline. The user-derived path (top) curates queries from anonymized real interactions; the auto-generated path (bottom) generates queries from real-time web trends with cascaded filtering. Both feed into a 100-query benchmark with three-annotator verification.

Evaluating 11 leading systems with ApodexEval (§4) supports one takeaway: deep-research capability is not a single quantity, so a system has to be diagnosed along all three dimensions rather than ranked on one number. Three findings instantiate this. First, synthesis and factuality come apart. System rankings on the two dimensions are nearly uncorrelated (Spearman's $\rho = 0.07$): the system with the highest text-only synthesis score ranks near the bottom on factuality, while another with the lowest synthesis score outperforms several stronger writers on factual precision. Second, the research process reveals what the report hides. Process→Report coverage exceeds Report→Process traceability by about 20 points or more for every system, and the asymmetry persists among the systems that expose the most complete traces, so part of report content cannot be traced back to the documented research. Process quality and outcome quality are also strongly rank-correlated across systems ($\rho = 0.88$, 95% CI $[0.73, 0.94]$), a limited-sample association over 11 systems rather than a predictive claim. Third, multimodal tasks degrade most systems, with the loss concentrated in synthesis and process while factuality remains nearly flat. The framework is validated through robustness experiments (per-dimension Kendall's τ of 0.79 to 0.93 under judge-model substitution), a human ranking study (maximum rank shift of 1 position, Spearman's $\rho = 0.98$), and human validation of each component (§4.5). Three expert annotators further verify benchmark query quality at a 92.6% acceptance rate. ApodexEval thus provides both a methodology and a refreshable benchmark for diagnosing deep research systems.

2 Query Collection and Verification

Constructing a benchmark for deep research agents poses three competing demands (authenticity, complexity, and freshness) that no single construction strategy meets at once. Real user queries supply authenticity but face volume and privacy limits, while purely synthetic queries scale but drift toward textbook patterns; existing benchmarks typically choose one source and inherit its bias. We resolve this tension with a dual-path pipeline (Figure 2) that combines both sources under a common privacy and refreshability discipline, yielding **100 queries (70 text-only, 30 multimodal)** spanning 12 domains and 10 task types, evaluated on synthesis quality, factuality, and process integrity (§4).

User-Derived Path (65 queries: 35 text + 30 multimodal). After privacy filtering and entity replacement (named institutions, organizations, and individuals are substituted by realistic same-type counterparts of comparable scale), an LLM annotates each anonymized query with eight evaluation features (Appendix C.2), and a feature-balanced router selects one of six rewrite strategies across three

difficulty tiers, ranging from basic retrieval to contradiction injection and erroneous-premise detection. No original user query appears verbatim in the benchmark; the rewrite preserves methodological complexity while abstracting identity-revealing personas, ensuring authenticity without re-disclosing private content. Contributors are internal staff testing the system; their notice, retrospective written confirmation, and an identifiability audit of all 65 released queries are documented in Appendix F.2.

Auto-Generated Path (35 text-only queries). An LLM seeded with real-time web trends across 12 topics produces 180 candidates, filtered in cascade: search validation (≥ 3 results from ≥ 2 source domains; 152 retained, 84.4%); necessity assessment (rejects queries answerable from pre-2023 knowledge; 96, 63.2%); and inverse quality assessment retaining only queries where a no-retrieval baseline is demonstrably inadequate ($\sigma \leq 0.75 \wedge \ell \neq \texttt{high} \wedge \texttt{requires_search}$; 50, 52.1%). A final manual selection yields 35 queries (19.4% cumulative retention), a stringency that reflects the difficulty of producing queries which genuinely demand external retrieval rather than well-documented parametric knowledge.

All 100 queries are independently verified by three expert annotators on validity and non-triviality (with a concurrent flag-and-correct difficulty-tier review), achieving a **92.6% acceptance rate** weighted by source size (per-source Fleiss' $\kappa \geq 0.74$). Both paths share a parameterized design that mitigates the staleness risk affecting static benchmarks as new events recede into training data, but they refresh differently: the auto-generated path can be re-run at any time by re-fetching trend context, whereas the user-derived path, which supplies all 30 multimodal tasks, refreshes only as new internal-testing data accrues and cannot be regenerated independently by others. Construction details, prompts, and full statistics: Appendices C–G; license, maintenance, and Responsible AI metadata: Appendix E.

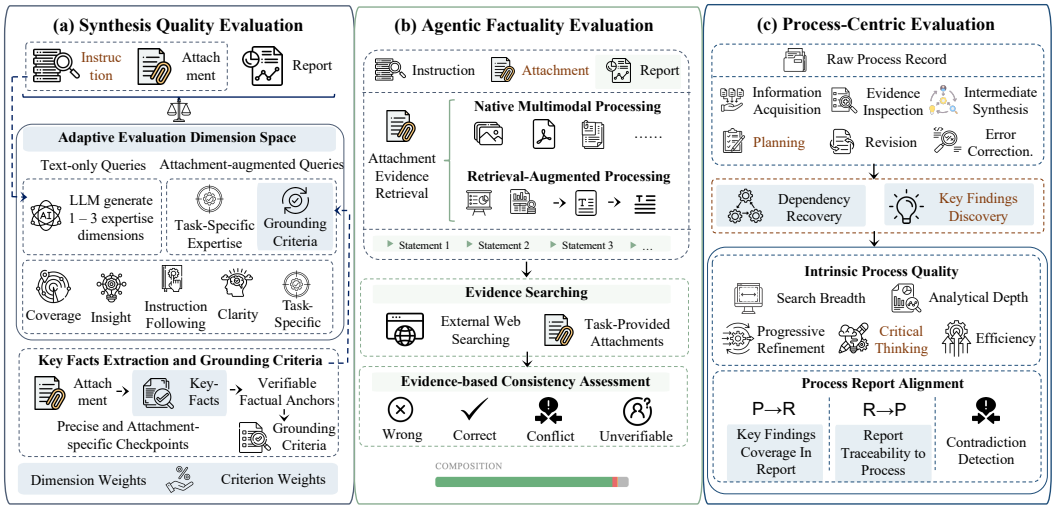


Figure 3: Overview of the evaluation pipeline.

3 Evaluation Methodology

3.1 Adaptive Synthesis Quality Evaluation

Because deep research tasks vary substantially in domain, objectives, and input modality, we tailor evaluation dimensions, criteria, and weights to each task.

Adaptive Evaluation Dimension Space. Let $Q = (I, A)$ denote the input query, where I is the research instruction and A is an optional set of attachments (e.g., images, PDFs, spreadsheets). For each task, the framework constructs a tailored evaluation dimension space $D = D_{\text{fixed}} \cup D_{\text{dynamic}}(Q)$. The fixed component D_{fixed} captures universal aspects of synthesis quality such as Coverage, Insight, Instruction-following, and Clarity. The dynamic component $D_{\text{dynamic}}(Q)$ adapts to the query:

- **Text-only queries** ($A = \emptyset$): the LLM generates 1–3 task-specific expertise dimensions from I (e.g., "Policy Pragmatism").
- **Attachment-augmented queries** ($A \neq \emptyset$): the framework adds a **Grounding** dimension forming composite "Grounding & Task-specific Expertise" dimensions, penalizing superficial referencing.

The four fixed dimensions are a shared backbone, but the criteria inside every dimension and every dimension's weight are generated per query; the dynamic dimensions are reported as *Task-Specific* (TS) in Table 2, and Appendix L.3 ablates the adaptation.

Key Facts Extraction and Grounding Criteria. For attachment-augmented queries, an upstream module extracts condensed **key facts** from the raw attachments, distilling heterogeneous materials (tables, image captions, structured text) into a set of verifiable factual anchors. These key facts transform abstract evaluation requirements into precise, attachment-specific checkpoints. For example, given a task to "analyze the global EV market" with an accompanying sales spreadsheet, a context-free evaluator can only assess general criteria such as whether quantitative analysis is used. With extracted key facts, the evaluator can generate concrete criteria such as whether the report correctly identifies the inflection point where BYD surpassed Tesla in 2023Q3. For text-only queries, criteria are generated directly from I.

Dynamic Weighting and Scoring. The evaluator analyzes Q to derive dimension-level weights W_d and criterion-level weights $w_{d,c}$ with $\sum_d W_d = 1$ and $\sum_c w_{d,c} = 1$). The report R is assessed against each criterion via $s_{d,c} = \text{LLM}_\theta(R, d, c, Q) \in [0, 10]$, and computes

$$S_{\text{quality}} = \sum_{d \in D} W_d \sum_c w_{d,c}\, s_{d,c}. \tag{1}$$

3.2 Agentic Factuality Evaluation

In deep research scenarios, reports often contain numerous verifiable statements that must be validated against heterogeneous evidence from both external web resources and task-provided attachments. These sources may yield conflicting conclusions, making binary fact-checking insufficient.

Building on recent advances in long-form factuality verification with agentic or multi-step reasoning [42, 27, 28, 40, 35], we design an *agentic factuality evaluation* framework that enables an evaluation agent to retrieve and reason over evidence from multiple sources. Given a query $Q = (I, A)$ and report R, the system decomposes the report into a set of verifiable statements $\mathcal{S}(Q, R) = \{s_1, \ldots, s_n\}$. For each statement s, the agent retrieves evidence from two complementary sources, forming an evidence set $E(s) = E_{\text{search}}(s) \cup E_{\text{attach}}(s)$, where E_{search} denotes evidence from external search and E_{attach} from task-provided attachments.

Attachment Evidence Retrieval. To handle diverse attachment formats, the framework adopts a hybrid retrieval strategy.

- **Native Multimodal Processing.** For formats directly interpretable by multimodal LLMs (images, PDFs, plain-text), the attachment is passed alongside the query, allowing the model to reason over visual and structural information without intermediate conversion.
- **Retrieval-Augmented Processing.** For formats not directly ingestible (spreadsheets, slides), the attachment is converted to text, segmented, and retrieved on demand to locate supporting evidence in large documents.

Evidence-based Consistency Assessment. The agent evaluates each statement against its evidence set and assigns a label $y(s) \in \{\texttt{RIGHT}, \texttt{WRONG}, \texttt{CONFLICT}, \texttt{UNKNOWN}\}$. The added `CONFLICT` label captures cases where heterogeneous sources disagree, preserving the disagreement rather than forcing a binary judgment. The factuality score is the per-task fraction of statements labeled `RIGHT`. Since `UNKNOWN` mixes claims the verifier could not reach with unverifiable ones, we also report the ratio excluding `UNKNOWN` and each verdict rate separately (Appendix L.1); verdicts from live web evidence inherit the reliability of the sources retrieved.

3.3 Process-Centric Evaluation

We complement output-level evaluation with a *process-centric* framework that examines how the system conducts the research procedure rather than only the final text.

Process Representation. We transform each raw process record P into a structured representation by decomposing it into a sequence of atomic units corresponding to functionally distinct steps such as information acquisition, evidence inspection, intermediate synthesis, planning, revision, and error correction. We further recover their local dependency structure and extract key findings that emerge during the procedure.

Intrinsic Process Quality. The structured trajectory is evaluated along five complementary dimensions.

- **Search Breadth**. Coverage of sources, perspectives, and sub-topics relevant to the query.
- **Analytical Depth**. Extent of multi-step reasoning, follow-up investigation, and in-depth analysis beyond surface retrieval.
- **Progressive Refinement**. Iterative improvement of understanding as new evidence accumulates.
- **Critical Thinking**. Assessment of source reliability, evidence limitations, and handling of conflicts.
- **Efficiency**. Avoidance of redundancy such as repeated queries, circular paths, and unused retrievals.

Process–Report Alignment. We further check the alignment between process findings and report findings along two directions and one consistency check.

- **Process→Report (P→R)**. Whether major findings established during the process are realized in the report.
- **Report→Process (R→P)**. Whether report conclusions can be linked back to sufficient support in the process.
- **Contradiction Detection (Contr)**. Whether the system identifies and resolves conflicting evidence encountered during research.

The overall process score combines intrinsic quality and alignment.

$$S_{\text{process}} = \alpha\, S_{\text{intrinsic}}(P) + (1-\alpha)\, S_{\text{align}}(P, R), \tag{2}$$

where $S_{\text{intrinsic}}$ averages the five intrinsic dimensions and S_{align} averages the three alignment scores. We set $\alpha = 0.5$ in our experiments. Since only what a system exposes can be evaluated, cross-system comparisons of process levels carry an exposure caveat (Appendix L.5).

4 Evaluation of Deep Research Systems

4.1 Experiment Setup

We conduct evaluations on a range of mainstream deep research systems, including OpenAI Deep Research [31], Gemini-3.1-Pro Deep Research [13], Grok Deep Research [44], Claude-Opus-4.6 Research [3], Manus-1.6-Max Wide Research [2], Doubao Deep Research [8], ChatGLM Agent [50], Kimi-K2.5 Deep Research [36], Qwen-3.5-Plus Deep Research [34], MiniMax-M2.5 Research [29], and MiroThinker-1.7 [4]. For Kimi-K2.5 Deep Research and Doubao Deep Research, we report only text-only results, as these systems do not support multimodal deep research. All systems are evaluated through their official online interfaces using automated pipelines that submit queries, collect final reports, and gather available process traces. Details on the data collection procedure and report statistics are provided in Appendix B. For automatic evaluation, we use GPT-5.1 as the judge model for synthesis quality and GPT-5.2 for process evaluation, and GPT-5-mini for factuality evaluation. Uncertainty is quantified with a task-level bootstrap (10,000 resamples, tasks resampled as clusters shared by all systems, stratified by setting); 95% confidence intervals for every cell of Table 1 are given in Appendix L.7.

4.2 Main Results

Table 1 presents the performance of all evaluated systems across Synthesis quality, Factuality, and Process under both the Text-Only and MultiModal settings.

Table 1: Performance comparison of models with ApodexEval. Each system is scored on Synthesis quality, Factuality, and Process under Text-Only and MultiModal settings. Overall within each setting is the equal-weight mean of all three dimensions, so it includes Process; the last column is the task-weighted average across the two settings (70 Text-Only + 30 MultiModal tasks). 95% confidence intervals for every cell are given in Table 29.

Model	**Text-Only**				**MultiModal**				**Overall**
	Synthesis	Factuality	Process	Overall	Synthesis	Factuality	Process	Overall	
Doubao Deep Research	64.2	64.9	53.1	60.7	–	–	–	–	–
Kimi-K2.5 Deep Research	**75.7**	65.4	64.2	68.4	–	–	–	–	–
Grok Deep Research	58.7	63.7	58.3	60.2	56.3	71.5	53.9	60.5	60.3
Qwen-3.5-Plus Deep Research	60.0	73.1	61.1	64.7	44.6	69.9	53.8	56.1	62.1
Manus-1.6-Max Wide Research	55.4	72.6	64.2	64.0	54.3	70.0	61.8	62.0	63.4
ChatGLM Agent	63.2	68.6	65.6	65.8	61.6	71.6	57.7	63.6	65.1
MiniMax-M2.5 Research	63.3	71.8	67.1	67.4	56.7	71.0	62.2	63.3	66.2
Claude-Opus-4.6 Research	67.3	69.8	66.0	67.7	62.5	70.7	65.9	66.4	67.3
Gemini-3.1-Pro Deep Research	71.2	71.3	67.1	69.9	66.4	73.7	64.1	68.1	69.3
MiroThinker-1.7	74.3	79.4	72.7	75.5	**69.0**	**78.4**	**67.4**	**71.6**	74.3
OpenAI Deep Research	73.8	**83.3**	**73.1**	**76.7**	66.7	77.0	66.8	70.2	**74.8**

Overall Results. In the Text-Only setting, systems separate into roughly three performance tiers. OpenAI Deep Research forms the top tier at 76.7, followed by MiroThinker-1.7 at 75.5. Kimi-K2.5 Deep Research, Gemini-3.1-Pro Deep Research, Claude-Opus-4.6 Research, MiniMax-M2.5 Research, and ChatGLM Agent constitute a middle tier between roughly 65 and 70. A lower tier includes Qwen-3.5-Plus Deep Research, Manus-1.6-Max Wide Research, Doubao Deep Research, and Grok Deep Research, all scoring below 65. A broadly similar ordering holds in the MultiModal setting, though overall scores decrease for most systems and the inter-system gaps narrow. Text-Only 95% CI half-widths are 1.0–3.6 points and the 30-task MultiModal ones 2.3–8.7, so adjacent systems within a tier are not reliably separated; we read the results as orderings between tiers rather than as point magnitudes.

Key Findings. Three findings emerge from the dimension-level comparison. First, ***rankings shift substantially across evaluation dimensions***, and synthesis and factuality rankings are nearly uncorrelated (Spearman's $\rho = 0.07$, task-weighted over both settings). Kimi-K2.5 achieves the highest Text-Only Synthesis score at 75.7, yet its Factuality of 65.4 ranks ninth of eleven, and it stays in the lower half when `UNKNOWN` verdicts are excluded. Conversely, Manus-1.6-Max obtains the lowest Synthesis score at 55.4, yet its Factuality of 72.6 surpasses several systems with much stronger reports, including Gemini-3.1-Pro and MiniMax-M2.5, under both scoring rules. These contrasting cases illustrate that a polished report does not guarantee factual grounding, nor does a factually disciplined system necessarily produce well-structured output. We investigate the sub-metric sources of this divergence in §4.3. Second, ***process quality and outcome quality are aligned***. The top-process systems (OpenAI at 73.1 and the next-best at 72.7) also rank highest on the combined synthesis–factuality outcome, while the weakest process system, Doubao at 53.1, also produces a near-bottom outcome. Overall in Table 1 includes Process, whereas the outcome column of Table 2 excludes it by design. We deepen this observation with quantitative correlation analysis in §4.4. Third, ***multimodal tasks degrade most systems***, with declines concentrated in synthesis and process while factuality stays nearly flat. A detailed cross-setting comparison is provided in Appendix H.2.

4.3 Outcome-Level Analysis

To understand the synthesis–factuality divergence observed in §4.2, we analyze the sub-metric structure of each dimension as shown in Table 2.

First, in synthesis quality, *Task-Specific* (TS) quality is the universal bottleneck, the lowest sub-metric for nearly every system in both settings, while *Insight* is the primary differentiator, with the widest spread across systems. Coverage and Instruction-following are approaching saturation among top systems and no longer separate them. Second, factuality scores reveal a precision–volume trade-off, as shown in Figure 4. Systems with very large claim volumes often achieve high absolute counts

Table 2: Combined evaluation on Synthesis quality and Factuality. Synthesis is assessed across five dimensions: Coverage, Insight, Instruction-following, Clarity, and Task-Specific (TS), the aggregate of the per-query dynamic dimensions (§3.1). Factuality is measured by the average right ratio scaled to $[0, 100]$; Ratio$_{\neg U}$ is the same ratio with UNKNOWN verdicts excluded. Overall is the average of Synthesis Avg and Factuality Ratio; it is an outcome score and deliberately excludes Process. Text-Only comprises 70 tasks and Multimodal comprises 30 tasks.

Model	Synthesis						Factuality						Overall
	Cov.	Insight	Instr.	Clarity	TS	Avg	Right	Wrong	Conf.	Unk.	Ratio	Ratio$_{\neg U}$	
Text-Only (70 Tasks)													
Grok Deep Research	67.3	56.3	74.9	64.7	51.1	58.7	1924	368	–	699	63.7	82.1	61.2
Manus-1.6-Max Wide Research	61.2	54.8	67.9	65.6	48.1	55.4	1972	191	–	459	72.6	88.5	64.0
Doubao Deep Research	72.9	62.7	74.6	67.2	58.2	64.2	3890	780	–	1393	64.9	83.1	64.6
ChatGLM Agent	69.9	62.8	74.5	67.5	57.1	63.2	4096	580	–	981	68.6	85.0	65.9
Qwen-3.5-Plus Deep Research	64.0	64.7	69.9	67.8	52.6	60.0	1706	244	–	380	73.1	87.5	66.5
MiniMax-M2.5 Research	69.8	62.7	74.2	70.6	56.7	63.3	3872	486	–	921	71.8	87.4	67.5
Claude-Opus-4.6 Research	73.3	72.0	73.5	71.2	61.1	67.3	2838	338	–	910	69.8	90.2	68.6
Kimi-K2.5 Deep Research	**80.4**	**79.8**	78.6	76.3	**71.7**	**75.7**	3702	595	–	1256	65.4	85.2	70.6
Gemini-3.1-Pro Deep Research	77.4	76.6	80.0	70.1	64.9	71.2	4039	526	–	1068	71.3	88.1	71.3
MiroThinker-1.7	79.2	74.7	**84.7**	**80.1**	68.4	74.3	3334	181	–	670	79.4	**94.6**	76.9
OpenAI Deep Research	78.2	74.3	81.6	77.1	69.1	73.8	3335	170	–	496	**83.3**	94.5	**78.6**
Multimodal (30 Tasks)													
Qwen-3.5-Plus Deep Research	46.8	46.3	52.9	52.6	30.1	44.6	576	99	19	101	69.9	79.9	57.3
Manus-1.6-Max Wide Research	58.7	50.2	65.0	61.2	40.4	54.3	681	81	32	134	70.0	81.6	62.2
MiniMax-M2.5 Research	63.1	53.3	69.1	62.0	39.2	56.7	1255	184	59	255	71.0	82.4	63.8
Grok Deep Research	61.8	52.5	68.9	60.4	40.5	56.3	734	104	37	163	71.5	83.1	63.9
ChatGLM Agent	67.1	60.2	71.7	65.4	45.1	61.6	1038	144	46	215	71.6	83.0	66.6
Claude-Opus-4.6 Research	68.9	66.8	62.8	59.3	50.0	62.5	964	84	44	243	70.7	87.0	66.6
Gemini-3.1-Pro Deep Research	72.4	**70.8**	72.4	62.5	50.1	66.4	1502	158	94	302	73.7	86.3	70.0
OpenAI Deep Research	70.6	63.9	74.8	70.5	**54.2**	66.7	1062	100	36	157	77.0	89.2	71.8
MiroThinker-1.7	**72.6**	69.2	**78.6**	**75.1**	53.6	**69.0**	1306	103	63	235	**78.4**	**89.8**	**73.7**

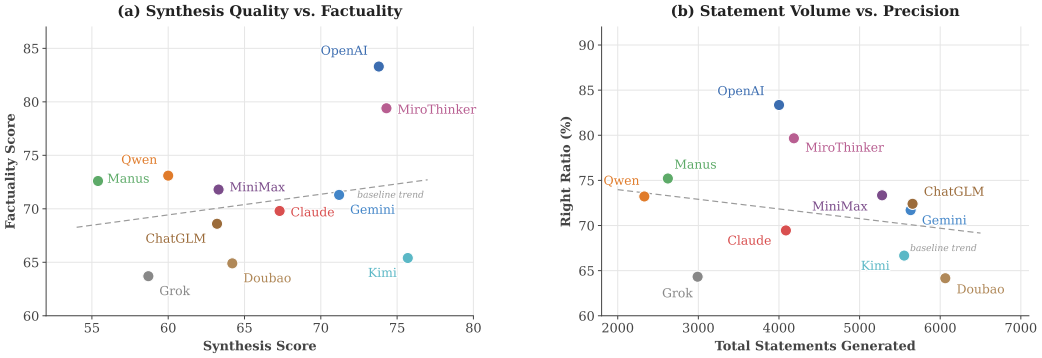


Figure 4: Relationship between synthesis quality, factuality, and statement-level precision across different systems. Left: synthesis quality vs. factuality score. Right: total number of generated statements vs. right ratio. Each point represents a system. The gray dashed lines denote linear regression fits, illustrating a weak positive correlation between synthesis quality and factuality, and a negative correlation between statement volume and precision.

of correct claims, but they also accumulate substantial numbers of wrong claims. For example, ChatGLM Agent and Gemini-3.1-Pro both exceed 4,000 correct claims, yet their hundreds of wrong claims reduce their Factuality Ratios to around 70. In contrast, systems with stricter generation behavior, such as OpenAI Deep Research, produce fewer claims but achieve a substantially higher Ratio of 83.3%. This trade-off represents a salient design axis along which deep research systems can be tuned. Part of the spread, however, reflects verifier coverage rather than error: on Text-Only, UNKNOWN-rate differences alone explain 35% of the cross-system variance of the pooled ratio and WRONG-rate differences 23% (their covariance the rest), and excluding UNKNOWN narrows the range across systems from 63.7–83.3 to 82.1–94.6. The two scoring rules rank systems similarly ($\rho = 0.80$), and our comparisons hold under both (Appendix L.1). Detailed sub-metric breakdowns, the synthesis–factuality misalignment analysis, and per-system case discussions are provided in Appendix I.

Table 3: Process evaluation results. Intrinsic metrics assess the quality of the research process itself across five dimensions: Search Breadth, Analytical Depth, Progressive Refinement, Critical Thinking, and Efficiency. Alignment metrics measure the consistency between the research process and the final report (Process→Report coverage, Report→Process traceability, and Contradiction detection). Overall is the weighted average of Intrinsic Avg and Alignment Avg.

Model	**Intrinsic**						**Alignment**				**Overall**
	Brdth	Depth	Refin	Critl	Effic	Avg	P→R	R→P	Contr	Avg	
Text-Only (70 Tasks)											
Doubao Deep Research	59.3	41.6	59.6	55.7	53.3	53.9	65.7	36.8	54.2	52.2	53.1
Grok Deep Research	50.9	49.4	61.0	54.6	64.7	56.1	74.6	42.2	64.6	60.4	58.3
Qwen-3.5-Plus Deep Research	74.4	64.1	75.0	74.1	63.2	70.2	59.6	39.7	56.9	52.1	61.1
Manus-1.6-Max Wide Research	62.8	58.4	60.6	53.5	**68.8**	60.8	75.1	51.3	76.3	67.6	64.2
Kimi-K2.5 Deep Research	77.5	59.4	71.0	67.6	53.5	65.8	70.7	46.8	70.4	62.6	64.2
ChatGLM Agent	76.2	59.4	67.1	59.3	59.0	64.2	77.1	51.4	72.3	67.0	65.6
Claude-Opus-4.6 Research	**79.1**	58.8	67.2	56.7	62.2	64.8	81.0	47.1	73.5	67.2	66.0
Gemini-3.1-Pro Deep Research	75.4	66.6	75.9	64.1	59.0	68.2	72.9	50.6	74.4	66.0	67.1
MiniMax-M2.5 Research	71.9	62.2	70.1	62.5	63.5	66.0	77.4	53.0	74.3	68.3	67.1
MiroThinker-1.7	74.4	64.4	75.7	71.6	64.6	70.1	**83.7**	**59.4**	**82.5**	**75.2**	72.7
OpenAI Deep Research	77.4	**67.3**	**76.7**	**74.7**	63.7	**72.0**	83.6	59.0	79.9	74.1	**73.1**
Multimodal (30 Tasks)											
Qwen-3.5-Plus Deep Research	57.0	51.3	58.7	57.7	51.3	55.2	61.7	39.3	56.3	52.4	53.8
Grok Deep Research	41.9	44.3	52.4	42.4	59.5	48.1	72.4	41.4	65.2	59.7	53.9
ChatGLM Agent	52.7	52.3	55.7	44.7	54.3	51.9	73.0	47.0	70.7	63.6	57.7
Manus-1.6-Max Wide Research	52.4	57.2	60.7	43.4	**65.9**	55.9	74.5	54.5	74.1	67.7	61.8
MiniMax-M2.5 Research	51.0	59.0	65.0	43.7	63.0	56.3	77.0	52.0	75.0	68.0	62.2
Gemini-3.1-Pro Deep Research	69.7	**65.3**	71.0	58.3	47.0	62.3	75.7	49.0	73.0	65.9	64.1
Claude-Opus-4.6 Research	**75.2**	60.7	69.6	59.3	60.0	65.0	78.9	49.3	72.6	66.9	65.9
MiroThinker-1.7	65.0	57.0	72.0	63.0	57.7	62.9	**80.7**	**58.7**	**76.0**	**71.8**	**67.4**
OpenAI Deep Research	65.5	62.1	**73.8**	**70.0**	54.5	**65.2**	77.2	56.2	72.1	68.5	66.8

4.4 Process-Level Analysis

Table 3 reports Intrinsic and Alignment process metrics. Three structural patterns emerge.

First, *Analytical Depth* and *Efficiency* emerge as the main Intrinsic bottlenecks. Most systems achieve reasonable *Search Breadth*, with Text-Only scores of 71.9 to 79.1 for eight of the eleven systems, but exhibit much larger variation in *Analytical Depth*, ranging from 41.6 to 67.3. This suggests that current systems are generally capable of retrieving information broadly, yet they often fall short in sustained investigation and deeper synthesis. Meanwhile, *Efficiency* is the lowest Intrinsic score for five of the eleven systems, indicating substantial redundancy in research trajectories.

Second, *Process–Report alignment is asymmetric*. Process→Report (P→R) scores are uniformly high, meaning information uncovered during research is generally incorporated into reports. Report→Process (R→P) scores are much lower for every system, by about 20 points or more. Grouping systems into three tiers by observable trace exposure (Appendix L.5), the gap is essentially unchanged across tiers (23.6, 26.3, and 26.5 points) and significant within the most-exposed tier alone (95% CI $[22.6, 24.6]$); the *levels* fall with exposure (R→P from 55.8 to 42.8), so R→P levels are compared with an exposure caveat. Part of report content thus cannot be traced back to the documented research, a traceability gap that undermines auditability.

Third, *process quality and outcome quality are strongly rank-correlated*. The Spearman correlation between Process and the combined synthesis–factuality outcome, both task-weighted over the two settings, is $\rho = 0.88$ (95% CI $[0.73, 0.94]$; leave-one-system-out range $[0.84, 0.93]$; Pearson $r = 0.83$), higher than with either Synthesis ($\rho = 0.62$) or Factuality ($\rho = 0.69$) alone. With 11 systems this is a limited-sample association rather than a predictive model; the value of the process layer is diagnostic, since it reveals weaknesses, such as the traceability gap above, that the report alone hides. Detailed per-dimension breakdowns and contradiction-detection analyses are provided in Appendix J.

4.5 Reliability and Validity of ApodexEval

We assess the framework's stability and validity and the benchmark's stability (Appendices K, L, and H).

Framework Reliability. Re-running the primary GPT judge three times on the MultiModal setting, a targeted diagnostic on three systems, yields Overall standard deviations of 0.3 to 0.7 with identical rankings in every run. Substituting Gemini as an alternative judge shifts absolute scores by 13 to 19 points but preserves rankings closely (Kendall's $\tau = 0.93$, 0.90, and 0.79 for Synthesis, Factuality, and Process; Overall changes only by adjacent swaps); we therefore interpret results as orderings, not point magnitudes. Changing the judge prompt shifts Overall scores within 2.2 points with no rank changes. A holistic study with 10 annotators matches the ApodexEval ranking with a maximum rank shift of one position (Spearman's $\rho = 0.98$).

Validity of Each Component. Stability is not accuracy, so each layer is checked against human labels (Appendix L). Three annotators adjudicated 270 stratified claims blind to the verifier: precision is 93.3% on `RIGHT` and 82.2% on `WRONG`, corpus-weighted accuracy is 84.7%, and correcting system scores for the measured error rates leaves the factuality ranking essentially unchanged ($\rho = 0.988$). A pairwise study of 150 same-query pairs (30 per dimension, three annotators blind to scores) finds 75.7–85.9% agreement between the judges and human majorities on clearly separated pairs; a fully static rubric lowers this agreement from 85.4% to 67.3%. Because the judges and one evaluated system are GPT-family, we re-scored every system with a Claude-family judge and found no consistent excess decline for the GPT-family system, whereas the Claude-family system gains relative to the field under its own judge (Appendix L.6).

Benchmark Reliability. The 70 Text-Only tasks include both user-derived and auto-generated queries (§2). System rankings remain consistent across the two subsets despite the user-derived queries being measurably harder, indicating that the benchmark does not advantage either source. Per-source results are provided in Appendix H.1.

5 Related Work

Deep Research Systems. Deep research systems autonomously plan multi-step web investigations and produce structured reports [31, 13, 3, 36, 2]. Several benchmarks evaluate these capabilities from a search or QA perspective, covering general agentic reasoning [24], persistent web navigation [41], expert-level knowledge [33], search breadth [43], and grounded page interaction [6]. These benchmarks assess retrieval accuracy or short-answer correctness rather than output quality.

Report-Level Evaluation. Beyond short-answer settings, report-level benchmarks evaluate the quality of synthesized long-form outputs. Existing benchmarks evaluate synthesis quality via fixed human-annotated rubrics [9, 1] and address specific facets such as temporal grounding [38], factual grounding of cited statements [22], and multi-step workflow benchmarking [46]. Other text-only efforts cover live evaluation [32], expert-level assessment [15], personalization [26], and interactive research [11]. All of these focus on the final report and apply uniform criteria across tasks.

Multimodal Deep Research Evaluation. Multimodal extensions remain less mature. MM-BrowseComp [23] adds multimodal retrieval but stays in short-form QA, and MMDeepResearch-Bench [16] evaluates multimodal reports under fixed evaluation dimensions. Deep-Reporter [47] introduces M2LongBench for grounded multimodal long-form generation, but like prior work does not assess the underlying multimodal research process. Other related benchmarks [48, 19, 40, 17] cover joint visual-textual search and long-form evaluation.

Position of ApodexEval. Together, existing benchmarks share four limitations: rubrics are fixed and task-agnostic, factual verification is restricted to a narrow scope of cited claims, the underlying research process is rarely assessed, and multimodal evaluation seldom goes beyond short-form QA over attachments. ApodexEval addresses all four with its three multimodal evaluation layers.

6 Conclusion

We introduced **ApodexEval**, a three-layer evaluation framework for deep research systems accompanied by a refreshable 100-task benchmark. Across 11 leading systems, our results support a diagnostic reading in which deep-research capability is not a single quantity: synthesis and factuality rank systems almost independently, the research process reveals a traceability gap that the report hides, and multimodal tasks degrade synthesis and process while leaving factuality nearly unchanged. Each component is validated against human judgment. Limitations and future work are discussed in Appendix A.

Acknowledgments and Disclosure of Funding

This work was funded entirely by Apodex; it received no third-party funding or other support.

References

[1] Amirhossein Abaskohi, Tianyi Chen, Miguel Muñoz-Mármol, Curtis Fox, Amrutha Varshini Ramesh, Étienne Marcotte, Xing Han Lù, Nicolas Chapados, Spandana Gella, Christopher Pal, et al. Drbench: A realistic benchmark for enterprise deep research. *arXiv preprint arXiv:2510.00172*, 2025.

[2] Manus AI. Introducing wide research. `https://manus.im/blog/introducing-wide-research`, 2025.

[3] Anthropic. Introducing claude opus 4.6, 2026. URL `https://www.anthropic.com/news/claude-opus-4-6`.

[4] S. Bai, L. Bing, L. Lei, R. Li, X. Li, X. Lin, E. Min, L. Su, B. Wang, L. Wang, L. Wang, S. Wang, X. Wang, Y. Zhang, Z. Zhang, G. Chen, L. Chen, Z. Cheng, Y. Deng, Z. Huang, D. Ng, J. Ni, Q. Ren, X. Tang, B. L. Wang, H. Wang, N. Wang, C. Wei, Q. Wu, J. Xia, Y. Xiao, H. Xu, X. Xu, C. Xue, Z. Yang, Z. Yang, F. Ye, H. Ye, J. Yu, C. Zhang, W. Zhang, H. Zhao, and P. Zhu. Mirothinker-1.7 & h1: Towards heavy-duty research agents via verification, 2026. URL `https://arxiv.org/abs/2603.15726`.

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

[6] 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.

[7] Kuicai Dong, Shurui Huang, Fangda Ye, Wei Han, Zhi Zhang, Dexun Li, Wenjun Li, Qu Yang, Gang Wang, Yichao Wang, et al. Doc-researcher: A unified system for multimodal document parsing and deep research. *arXiv preprint arXiv:2510.21603*, 2025.

[8] Doubao. Doubao chat, 2026. URL `https://www.doubao.com/chat/`.

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

[10] Yuwen Du, Rui Ye, Shuo Tang, Xinyu Zhu, Yijun Lu, Yuzhu Cai, and Siheng Chen. Openseeker: Democratizing frontier search agents by fully open-sourcing training data, 2026. URL `https://arxiv.org/abs/2603.15594`.

[11] Yingchaojie Feng, Qiang Huang, Xiaoya Xie, Zhaorui Yang, Jun Yu, Wei Chen, and Anthony KH Tung. Idrbench: Interactive deep research benchmark. *arXiv preprint arXiv:2601.06676*, 2026.

[12] Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. Wikimxqa: a multimodal benchmark for question answering over tables and charts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24941–24958, 2025.

[13] Gemini. Gemini deep research. `https://gemini.google/overview/deep-research/`, 2025.

[14] Jeremy Hadfield, Barry Zhang, Kenneth Lien, Florian Scholz, Jeremy Fox, and Daniel Ford. How we built our multi-agent research system. `https://www.anthropic.com/engineering/multi-agent-research-system`, 2025.

[15] Janghoon Han, Heegyu Kim, Changho Lee, Dahm Lee, Min Hyung Park, Hosung Song, Stanley Jungkyu Choi, Moontae Lee, and Honglak Lee. Deer: A benchmark for evaluating deep research agents on expert report generation. *arXiv preprint arXiv:2512.17776*, 2025.

[16] Peizhou Huang, Zixuan Zhong, Zhongwei Wan, Donghao Zhou, Samiul Alam, Xin Wang, Zexin Li, Zhihao Dou, Li Zhu, Jing Xiong, et al. Mmdeepresearch-bench: A benchmark for multimodal deep research agents. *arXiv preprint arXiv:2601.12346*, 2026.

[17] Yukun Huang, Leonardo FR Ribeiro, Momchil Hardalov, Bhuwan Dhingra, Markus Dreyer, and Venkatesh Saligrama. Deepfact: Co-evolving benchmarks and agents for deep research factuality. *arXiv preprint arXiv:2603.05912*, 2026.

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

[19] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, et al. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*, 2024.

[20] Nathan Kuissi, Suraj Subrahmanyan, Nandan Thakur, and Jimmy Lin. Still fresh? evaluating temporal drift in retrieval benchmarks. *arXiv preprint arXiv:2603.04532*, 2026.

[21] Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.

[22] Minghao Li, Ying Zeng, Zhihao Cheng, Cong Ma, and Kai Jia. Reportbench: Evaluating deep research agents via academic survey tasks. *arXiv preprint arXiv:2508.15804*, 2025.

[23] Shilong Li, Xingyuan Bu, Wenjie Wang, Jiaheng Liu, Jun Dong, Haoyang He, Hao Lu, Haozhe Zhang, Chenchen Jing, Zhen Li, et al. Mm-browsecomp: A comprehensive benchmark for multimodal browsing agents. *arXiv preprint arXiv:2508.13186*, 2025.

[24] Xiaochuan Li, Ryan Ming, Pranav Setlur, Abhijay Paladugu, Andy Tang, Hao Kang, Shuai Shao, Rong Jin, and Chenyan Xiong. Benchmark test-time scaling of general llm agents. *arXiv preprint arXiv:2602.18998*, 2026.

[25] Yibo Li, Zijie Lin, Ailin Deng, Xuan Zhang, Yufei He, Shuo Ji, Tri Cao, and Bryan Hooi. Just-in-time reinforcement learning: Continual learning in llm agents without gradient updates. *arXiv preprint arXiv:2601.18510*, 2026.

[26] Yuan Liang, Jiaxian Li, Yuqing Wang, Piaohong Wang, Motong Tian, Pai Liu, Shuofei Qiao, Runnan Fang, He Zhu, Ge Zhang, et al. Towards personalized deep research: Benchmarks and evaluations. *arXiv preprint arXiv:2509.25106*, 2025.

[27] Hongzhan Lin, Yang Deng, Yuxuan Gu, Wenxuan Zhang, Jing Ma, See Kiong Ng, and Tat-Seng Chua. Fact-audit: An adaptive multi-agent framework for dynamic fact-checking evaluation of large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 360–381, 2025.

[28] Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. Verifact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17919–17936, 2025.

[29] MiniMax. Minimax-m2.5, 2026. URL `https://www.minimax.io/news/minimax-m25`.

[30] Xuan-Phi Nguyen, Shrey Pandit, Revanth Gangi Reddy, Austin Xu, Silvio Savarese, Caiming Xiong, and Shafiq Joty. Sfr-deepresearch: Towards effective reinforcement learning for autonomously reasoning single agents, 2025. URL `https://arxiv.org/abs/2509.06283`.

[31] OpenAI. Introducing deep research. `https://openai.com/index/introducing-deep-research/`, 2025.

[32] 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.

[33] Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

[34] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[35] Shiqian Su, Sen Xing, Xuan Dong, Muyan Zhong, Bin Wang, Xizhou Zhu, Yuntao Chen, Wenhai Wang, Yue Deng, Pengxiang Zhu, et al. Miroflow: Towards high-performance and robust open-source agent framework for general deep research tasks. *arXiv preprint arXiv:2602.22808*, 2026.

[36] Kimi Team. Kimi-researcher: End-to-end rl training for emerging agentic capabilities. `https://moonshotai.github.io/Kimi-Researcher/`, 2025.

[37] Nandan Thakur, Jimmy Lin, Sam Havens, Michael Carbin, Omar Khattab, and Andrew Drozdov. Freshstack: Building realistic benchmarks for evaluating retrieval on technical documents. *arXiv preprint arXiv:2504.13128*, 2025.

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

[39] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024.

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

[41] 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.

[42] Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, et al. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827, 2024.

[43] 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.

[44] xAI. Grok deepsearch. `https://x.ai/news/grok-3`, 2025.

[45] Zhifei Xie, Zongzheng Hu, Fangda Ye, Xin Zhang, Haobo Chai, Zihang Liu, Pengcheng Wu, Guibin Zhang, Yue Liao, Xiaobin Hu, et al. Pask: Toward intent-aware proactive agents with long-term memory. *arXiv preprint arXiv:2604.08000*, 2026.

[46] Tianze Xu, Pengrui Lu, Lyumanshan Ye, Xiangkun Hu, and Pengfei Liu. Researcherbench: Evaluating deep ai research systems on the frontiers of scientific inquiry. *arXiv preprint arXiv:2507.16280*, 2025.

[47] Fangda Ye, Zhifei Xie, Yuxin Hu, Yihang Yin, Shurui Huang, Shikai Dong, Jianzhu Bao, and Shuicheng Yan. Deep-reporter: Deep research for grounded multimodal long-form generation. *arXiv preprint arXiv:2604.10741*, 2026.

[48] Yu Zeng, Wenxuan Huang, Zhen Fang, Shuang Chen, Yufan Shen, Yishuo Cai, Xiaoman Wang, Zhenfei Yin, Lin Chen, Zehui Chen, et al. Vision-deepresearch benchmark: Rethinking visual and textual search for multimodal large language models. *arXiv preprint arXiv:2602.02185*, 2026.

[49] Wenlin Zhang, Xiaopeng Li, Yingyi Zhang, Pengyue Jia, Yichao Wang, Huifeng Guo, Yong Liu, and Xiangyu Zhao. Deep research: A survey of autonomous research agents. *arXiv preprint arXiv:2508.12752*, 2025.

[50] Zhipu AI. Chatglm, 2026. URL `https://chatglm.cn`.

[51] Yutao Zhu, Xingshuo Zhang, Maosen Zhang, Jiajie Jin, Liancheng Zhang, Xiaoshuai Song, Kangzhi Zhao, Wencong Zeng, Ruiming Tang, Han Li, et al. Gisa: A benchmark for general information-seeking assistant. *arXiv preprint arXiv:2602.08543*, 2026.

Appendix Contents

A Limitations and Future Work **16**

B Data Collection and Report Statistics **17**

C User-Derived Query Curation Pipeline **17**
C.1 Privacy-Preserving Processing . . . 17
C.2 Feature-Driven Classification and Routing . . . 18
C.3 Difficulty-Stratified Rewriting . . . 19
C.4 Worked Example: A Hard-Tier Rewrite under Strategy E . . . 19

D Automated Query Generation Pipeline **20**
D.1 Trend-Grounded Generation . . . 20
D.2 Three-Stage Filtering . . . 21
D.3 Canonical Domain Labels for Auto-Generation Routing Only . . . 22
D.4 Worked Example: A High-Quality Auto-Generated Query . . . 22

E Benchmark Composition, Verification, and Release **23**
E.1 Benchmark Overview . . . 23
E.2 Assembly, Human Verification, and Release . . . 24

F Annotation Protocols and Source-Data Provenance **27**
F.1 Annotator Protocols . . . 27
F.2 Source-Data Provenance . . . 28

G Construction Prompts **29**

H Further Analysis of Evaluation Results **50**
H.1 User-Derived vs. Auto-Generated Queries . . . 50
H.2 Text-Only vs. MultiModal Comparison . . . 50

I Detailed Outcome-Level Analysis **51**

J Detailed Process-Level Analysis **52**

K Evaluation Robustness and Human Study **53**
K.1 Intra-Judge Stability . . . 53
K.2 Cross-Judge Consistency and Prompt Sensitivity . . . 54
K.3 Human Study . . . 54

L Validation of the Evaluation Framework **56**
L.1 Factuality Verifier Accuracy and Scoring Rules . . . 56
L.2 Per-Dimension Pairwise Human Study . . . 57
L.3 Static-Rubric Ablation . . . 58
L.4 Process Ablation: Stripping Marked Reasoning . . . 58
L.5 Trace-Exposure Control . . . 59
L.6 Judge-Family Analysis . . . 59
L.7 Uncertainty . . . 61

M Case Study **62**
M.1 Synthesis Evaluation . . . 62
M.2 Factuality Evaluation . . . 65

A Limitations and Future Work

Scale and statistical scope. The benchmark contains 100 tasks and our conclusions rest on 11 systems. The task grid of Figure 5 is a coverage map rather than a fully crossed design, and many domain-by-task cells hold one or two items, so we draw no cell-level conclusions. System-level results carry task-bootstrap confidence intervals (Appendix L.7); intervals on the 30-task MultiModal split are notably wider than on Text-Only, and the process–outcome correlation is a limited-sample association over 11 systems.

Aggregate and process–outcome association. The three dimensions capture complementary aspects of capability, so the per-dimension diagnosis is the result and the Overall score of Table 1, an equal-weight mean that includes Process, is a summary of it. The association between process and outcome (Spearman $\rho = 0.88$, 95% CI $[0.73, 0.94]$) is a limited-sample rank correlation over 11 systems; we report it as an association and do not claim that process quality predicts outcome quality.

Absolute scores. Substituting the judge model shifts absolute scores by 13 to 19 points while largely preserving rankings (Appendix K.2). We therefore state our claims as orderings rather than point magnitudes, and absolute gaps between systems should not be read as calibrated quantities.

Factuality scoring. The right ratio penalizes `UNKNOWN` verdicts in the same way as `WRONG` ones, yet nearly half of sampled `UNKNOWN` verdicts are claims the verifier did not reach rather than claims that cannot be verified (Appendix L.1). We report the `UNKNOWN`-excluded ratio and the separate verdict rates alongside the right ratio, and the two should be read together. The `CONFLICT` label flags disagreements between sources without resolving which source is correct, and verdicts based on live web retrieval inherit the reliability of the sources retrieved; the verifier does not independently audit source bias. Resolving conflicts and weighting sources by reliability are directions for future work.

Dependence on observable traces. Process evaluation can only assess what a system exposes, which limits applicability to fully closed systems and makes cross-system comparisons of process *levels*, notably Report→Process traceability, partly reflect how much a system reveals (Appendix L.5). The Process→Report versus Report→Process asymmetry is robust to exposure, but its absolute level is not.

Adaptive rubric. The adaptive rubric improves agreement with human judgments over a static rubric (Appendix L.3), but its per-query criteria are themselves model-generated, and the human validation covers a sample of 30 pairs per dimension.

Judge-family overlap. All three primary judges are GPT-family models. A Claude-family re-scoring finds no consistent GPT-family score inflation (Appendix L.6), but with one candidate per judge family among eleven systems this bounds a large effect rather than excluding a small one.

Provenance and representativeness. The user-derived queries come from internal testing of a single product, so their distribution reflects that product's users. This path, which supplies all 30 multimodal tasks, refreshes only as new internal-testing data accrues and cannot be regenerated independently by others; the auto-generated path can be re-run on demand. The consent basis for the user-derived queries rests on a contemporaneous notice and retrospective written confirmation from every contributor rather than study-specific consent at collection time (Appendix F.2).

Benchmark verification. Difficulty tiers were checked in a flag-and-correct pass rather than labeled independently by the three verifiers, so inter-annotator agreement is not reported for them.

Future work. We plan to leverage the refreshable dual-path construction pipeline to periodically update the benchmark with new queries reflecting evolving real-world deep-research demands and the latest web trends. The auto-generated path is benchmark-agnostic in structure (trend grounding, a taxonomy-constrained generator, and a cascaded filter), but porting it to another benchmark is not automatic: the filter's necessity criterion must be replaced by that benchmark's defining property, for example requiring multi-hop tool use, and the taxonomy must be re-specified. The process-evaluation layer does not depend on our benchmark: it takes a system's trajectory and the task, so it can also

wrap accuracy-scored benchmarks such as GAIA or BrowseComp and report how an answer was reached alongside whether it is right.

B Data Collection and Report Statistics

All evaluated systems are accessed through their official online service frontends. For each system, we develop automated pipelines that interact with the service interface to submit queries (including multimodal attachments where applicable), select the deep research mode, and collect the resulting outputs upon completion. The collected outputs include both the final report and, where exposed by the system, the intermediate research process such as search queries, retrieved sources, and reasoning steps. For systems that do not natively display structured process traces, we reconstruct available process information from any visible intermediate outputs provided during the research session. All system outputs were collected within a controlled time window in March 2026 to ensure temporal consistency across systems.

Table 4 summarizes the report length statistics. A consistent pattern is that text-only reports are generally longer than their multimodal counterparts. Several systems, including Kimi-K2.5 Deep Research and Doubao Deep Research, do not support multimodal deep research; for these systems, only text-only statistics are reported.

Table 4: **Average report length across different systems.**

System	Text-Only	Multimodal	Overall
Grok Deep Research	7,585	4,977	6,803
Manus-1.6-Max Wide Research	10,263	5,585	8,860
OpenAI Deep Research	17,669	12,751	16,194
MiroThinker-1.7	21,138	11,293	18,185
Qwen-3.5-Plus Deep Research	24,299	9,081	19,734
ChatGLM Agent	24,386	10,313	20,164
MiniMax-M2.5 Research	26,747	9,593	21,601
Claude-Opus-4.6 Research	23,624	20,129	22,576
Doubao Deep Research	43,160	–	–
Gemini-3.1-Pro Deep Research	49,343	32,568	44,311
Kimi-K2.5 Deep Research	61,739	–	–

C User-Derived Query Curation Pipeline

This appendix details the user-derived query curation pipeline summarized in §2, including the privacy filter (with a schematic illustration of entity replacement), evaluation feature taxonomy (Table 6), rewrite strategy router (Table 7), and a worked example illustrating a Hard-tier rewrite under Strategy E.

C.1 Privacy-Preserving Processing

Throughout the entire pipeline, all data handling follows strict confidentiality protocols: raw user queries are processed only on access-controlled internal infrastructure, and every language-model call that touches raw or partially-identified user content (privacy filtering, entity replacement, and the user-derived rewriting strategies) is routed to internally-deployed LLM endpoints that do not transmit raw data to external services. Subsequent steps that operate on fully anonymized synthetic content (trend-grounded generation and three-stage filtering for the auto-generated path) may use external LLM APIs, as documented in Appendix D. At the entry point, an automated filter removes all queries containing privacy-sensitive content (personal communications, confidential documents, and private financial records) through a combination of rule-based and model-assisted filtering. For the queries that pass this filter, all named entities (institutions, organizations, and individuals) are replaced with realistic same-type substitutes through an automated anonymization pass, ensuring that identifiable references are systematically removed before the rewrite stages.

Schematic of Entity Replacement. Table 5 specifies which tokens are substituted from same-tier pools and which are preserved verbatim, so that anonymized queries remain plausible same-class instances while no longer pointing to the original entities.

Table 5: Entity replacement rules. Identifying tokens are substituted with same-tier pool draws; surrounding intent-bearing tokens are preserved verbatim.

Token type	Action
Named institution (university, company, lab)	Substitute drawn from a pool of same-tier institutions, optionally cross-country, of comparable size and prestige
Named individual (PI, advisor, colleague)	Substitute drawn from a pool of researchers in the same subfield and at a similar career stage
Named group / project	Substitute drawn from comparable groups at the substitute institution / organization
Named dataset / benchmark / tool / product (when identifying)	Substitute drawn from publicly available alternatives in the same domain; if no plausible substitute exists, the reference is generalized (e.g., a proprietary dataset name → "an internal dataset")
Role descriptor (postdoc, PhD candidate, MD, founder, . . .)	Preserved verbatim
Motivation, analytical scope, query intent	Preserved verbatim
Non-identifying technical content (subfield terms, methodology, data references)	Preserved verbatim

A minimal synthetic pair illustrates the rules in action (placeholders `[University-A]`, `[University-B]`, `[PI-X]`, `[PI-Y]` stand for tokens that the LLM substitutes from the respective pools):

> *Original*: "As a postdoc at `[University-A]`, I work with Prof. `[PI-X]`'s group on protein-structure prediction; could you compare our pipeline against recent AlphaFold updates and summarize the methodological differences?"
>
> *After replacement*: "As a postdoc at `[University-B]`, I work with Prof. `[PI-Y]`'s group on protein-structure prediction; could you compare our pipeline against recent AlphaFold updates and summarize the methodological differences?"

Identifying tokens (`[University-A]`→`[University-B]`, `[PI-X]`→`[PI-Y]`) are drawn from same-tier pools; the role descriptor (*postdoc*), the technical scope (*protein-structure prediction*, *AlphaFold*), and the analytical question (*compare and summarize methodological differences*) are passed through verbatim, so the downstream rewriter inherits the same query intent.

Privacy filtering, entity replacement, and the user-derived rewriting strategies (Appendix C.2) are performed on internally-deployed LLM endpoints; raw and partially-identified user queries do not leave internal infrastructure at any stage of the user-derived pipeline.

C.2 Feature-Driven Classification and Routing

Pre-Classification Normalization. Two prompts normalize each anonymized query before classification. Prompt P4 (Appendix G) detects and extracts long inline content (e.g., pasted articles, code blocks) from the query body, so that classification and downstream rewriting operate on the analytical request rather than the embedded text. Prompt P5 merges multi-turn dialogue cases into a single self-contained query.

An LLM classifies each anonymized query along seven dimensions: attachment type, information density, domain, complexity, attachment role, rewrite potential, and a target evaluation-feature vector drawn from the taxonomy of eight capabilities defined in Table 6. These eight features capture the core competencies a deep research system must exercise to answer a complex query end-to-end.

Table 6: Definitions of the 8 evaluation features.

Feature	Definition
Goal adherence	Whether the system maintains focus on all specified goals and constraints throughout a multi-step task without deviating from original objectives or silently dropping sub-tasks.
Repetition avoidance	Whether the system avoids repeating the same information or analysis across different sections of its output when the query contains multiple similar but distinct sub-tasks.
Planning	Whether the system can decompose a complex query into a coherent, logically ordered sequence of execution steps with clear dependencies between stages.
Search	Whether the system can formulate effective search queries and retrieve relevant external information, rather than relying solely on parametric knowledge or the provided attachments.
Report generation	Whether the system can organize retrieval results into a well-structured, logically coherent report (comparison tables, analytical summaries, or recommendation lists) that synthesizes information from multiple sources.
Factuality	Whether factual claims in the system's output are accurate and verifiable against authoritative sources, with proper citation where appropriate.
Error correction	Whether the system can detect errors, contradictions, or problematic premises in the query or attachments and proactively correct them, rather than blindly following flawed instructions.
Multimodal understanding	Whether the system can correctly parse, interpret, and utilize non-textual information in attachments (charts, tables, images, diagrams, structured data), extracting accurate values and understanding spatial/visual relationships. This feature is only assigned to queries with attachments.

Based on the classification, the router selects one of six rewrite strategies (A through F; Table 7) according to four factors: (i) *hard constraints* that exclude strategies incompatible with the query's attachment type; (ii) *feature matching* that scores how well each strategy's target capabilities align with those of the query; (iii) *quota bonuses* that up-weight strategies covering underrepresented evaluation features; and (iv) *usage decay* that penalizes recently selected strategies to maintain diversity across the benchmark.

C.3 Difficulty-Stratified Rewriting

Each routed query is rewritten into a benchmark-ready instance by an LLM following the selected strategy. Easy queries require basic retrieval together with attachment comprehension; Medium queries involve multi-step reasoning across heterogeneous sources; Hard queries demand contradiction identification or erroneous-premise detection. Difficulty labels are first assigned by the LLM judge during routing and additionally verified by the three expert annotators of Appendix E.2 during the validity and non-triviality annotation pass.

Post-Rewrite Quality Verification. Prompt P3 (Appendix G) performs a quality-verification pass over each rewritten query, flagging residual factual errors, contradictions, or outdated information for author revision before the validity and non-triviality annotation.

The resulting set of 65 queries covers all eight evaluation features with balanced representation across difficulty tiers.

C.4 Worked Example: A Hard-Tier Rewrite under Strategy E

This worked example illustrates how the difficulty-stratified rewriter handles a Hard-tier query routed to Strategy E (Text-Only Research Synthesis; Table 7) for a representative methodologically dense user query in the medical domain. The router selected Strategy E based on the query's match against the feature set {*planning*, *search*, *report generation*, *goal adherence*, *repetition avoidance*} combined with the absence of attachments. The original query (excerpted from our internal-testing logs)

Table 7: The 6 rewrite strategies used for user-derived query curation. "Requires attachments" indicates the strategy is excluded for text-only queries during routing.

ID	Difficulty	Target Features	Description
A	Easy	search, multimodal understanding	Extract 1–2 key points from the attachment, perform one round of retrieval for supplementary context, and generate a concise response. Requires attachments.
B	Medium	planning, search, report generation, factuality, repetition avoidance	Compare attachment data against at least 2 external public sources and produce a structured comparative analysis report. Requires attachments.
C	Hard	factuality, error correction, multimodal understanding	Embed contradictions between the query text and attachment content (e.g., numerical discrepancies, date misalignment). The system must discover the inconsistency through reading the attachment and/or retrieval. Requires high-density attachments.
D	Hard	error correction, goal adherence	Embed false premises or ambiguous expressions in the query. The system should identify the erroneous premise, correct it, and still complete the core task.
E	Medium / Hard	planning, search, report generation, goal adherence, repetition avoidance	Multi-step research query with no attachment dependency. Answers must be synthesized from multiple public sources through iterative retrieval.
F	Easy / Medium	multimodal understanding, report generation	Primary focus on attachment processing: structured extraction, summarization, format conversion, or cross-page synthesis. Retrieval is auxiliary only. Requires attachments.

declares an explicit expert role and employs distinctive specialized jargon (RoB 2, GRADE Certainty Domains, CONSORT 2025, ROBINS-I, the "Total Justified Confidence" framework). The rewrite (i) abstracts the requester persona to a practitioner role ("strength coach working with elite athletes"), (ii) replaces insider methodological terminology with domain-accessible language while preserving the analytical scope (RCT-grade evidence comparison, biomechanical reasoning about joint angles, length-tension relationships across exercise variations), and (iii) translates the original structured output specification into a free-form research request that still demands the same evidence-synthesis depth. This worked example focuses on the persona-and-jargon component of the rewrite pipeline; entity replacement on queries containing named institutions or individuals operates as a separate mechanism, illustrated schematically in §C.

D Automated Query Generation Pipeline

This appendix details the trend-grounded automated query generation pipeline summarized in §2, including the topic taxonomy shown in Table 9, the three-stage filtering protocol with quantitative retention statistics, the inverse quality formula, the canonical domain labels shown in Table 10, and a worked example.

D.1 Trend-Grounded Generation

We organize generation around 12 topics, each with three subtopics as shown in Table 9. For each topic, we issue per-subtopic web searches via the Serper API to collect recent headlines and snippets as trend context for LLM-based query generation.

For each topic, an LLM generates 15 candidate queries conditioned on the topic description, the retrieved trend context, and a small set of anonymized seed exemplars sampled from a broader pool

Table 8: Strategy E worked example: persona and jargon abstraction on a Hard-tier user-derived query (medical / sports-science domain). The rewrite preserves analytical scope while abstracting identity markers and specialized terminology.

Original (excerpt; identity markers in italics)	**Rewritten (excerpt)**
ROLE: Senior Research Scientist in Biomechanics, Meta-Epidemiology, and Human Kinetics. TASK: Construct an exhaustive full-body training protocol by identifying the "Optimal Tier" exercise for every distinct muscle group, using the *Total Justified Confidence* framework. Apply RoB 2 and ROBINS-I to current literature, evaluate evidence under GRADE Certainty Domains, and align with CONSORT 2025, PRISMA 2020, and STROBE reporting rigor. Output structure: muscle group, optimal exercise, biomechanical rationale, evidence synthesis, programming variables.	I'm a strength coach working with elite athletes and I need to design evidence-based training programs that can withstand scrutiny from sports scientists. Recent research seems to disagree on optimal exercise selection across muscle groups—some studies favor compound movements while others argue isolation exercises are superior for hypertrophy. Can you identify the current scientific consensus by reviewing recent meta-analyses and systematic reviews, compare evidence quality across exercise variations (e.g., high-bar vs. low-bar squats, different rowing angles), and explain what the latest biomechanical research says about optimal joint angles and muscle length–tension relationships? I want this backed by RCT-grade evidence so I can defend my programming choices to the performance team.

Table 9: Topic taxonomy for automated query generation: 12 topics, each with 3 subtopics.

#	**Topic**	**Subtopics**
1	AI Policy & Regulation	EU AI Act implementation; US state AI laws; AI safety frameworks
2	Cybersecurity	Zero-day exploits; Agentic SOC; AI-powered social engineering
3	Finance & Macro	Central bank policy; Sovereign debt; Infrastructure investment
4	Crypto & Digital Assets	Stablecoin regulation; DeFi compliance; CBDC adoption
5	Healthcare & Pharma	Gene therapy trials; GLP-1 market dynamics; FDA regulatory shifts
6	International Trade	Global supply chain restructuring; Free trade agreements impact; Cross-border regulatory harmonization
7	AI Engineering	LLM benchmarking; Agentic coding tools; Model deployment architecture
8	Climate & Energy	Data center sustainability; Carbon pricing; Grid constraints
9	Education & Workforce	AI in K-12 policy; Workforce reskilling; Immigration & talent
10	Legal & Compliance	AI privilege doctrine; GDPR enforcement; Algorithmic discrimination
11	Biotech & Science	Computational biology; Quantum computing; Open access publishing
12	Supply Chain & Industrial	Nearshoring trends; Autonomous logistics; Semiconductor supply

of real user queries. Each candidate query is required to demand investigation from multiple distinct angles, draw on diverse source types, and adopt a specific professional persona such as analyst, engineer, journalist, investor, or graduate student. This produces an initial pool of 180 candidates. Generation uses the OpenRouter alias `openai/gpt-5.2`, routed to provider snapshots active in March 2026, with sampling temperature $T{=}0.8$ and a maximum of 8192 output tokens; the full generation prompt appears as P6 in Appendix G.

D.2 Three-Stage Filtering

We apply three filters in cascade to remove unsuitable candidates.

Stage 1: Search Validation. Each candidate query is submitted to a live web search; we require at least three results returned from at least two distinct source domains. This stage retains 152 of the 180 candidates (84.4%). A representative rejected candidate from a development run reads: *"As a supply-chain analyst, I'm mapping 'regional reset' strategies in 2025–2026. Compare how firms are restructuring supplier networks across North America (USMCA), Europe, and Southeast Asia in response to tariff volatility, shifting shipping routes, and logistics cost instability . . . "* The query was rejected because the live search returned only a single distinct source domain, falling below the

two-domain threshold: "regional reset" is an ad-hoc analyst phrasing without sufficient corroborating coverage across independent outlets.

Stage 2: Deep-Research Necessity. Each surviving query is evaluated by an LLM along four criteria: (i) it requires up-to-date information not answerable from pre-2023 parametric knowledge alone; (ii) it requires cross-verification across multiple credible sources; (iii) it requires multi-angle, multi-layered investigation; and (iv) it is not a textbook question. We retain queries with a necessity confidence score ≥ 0.7, yielding 96 queries (63.2% of Stage 1 survivors). A representative rejected candidate reads: *"With ongoing debates about AI alignment intensifying, summarize the philosophical foundations of value alignment by synthesizing Asimov's Three Laws of Robotics, Bostrom's Superintelligence (2014), and Russell's Human Compatible (2019). Compare their treatments of corrigibility, instrumental convergence, and the orthogonality thesis. Discuss how these foundational ideas shape 2024–2025 alignment research priorities at frontier labs."* Despite the contemporary framing, the substantive task is exposition of pre-2023 canonical works on alignment foundations; the named technical concepts (corrigibility, instrumental convergence, the orthogonality thesis) are extensively covered in pre-2024 literature, and an LLM can answer this query from parametric knowledge without time-sensitive retrieval. The filter assigned a necessity confidence of approximately 0.35, well below the retention threshold. This filter uses `openai/gpt-5.2` at $T{=}0.3$; the prompt appears as P7 in Appendix G.

Stage 3: Inverse Quality Assessment. The most discriminative filter targets a key principle: effective benchmark queries should expose the limitations of parametric knowledge. We first elicit a baseline answer without search access with $T{=}0.3$ and prompt P8 in Appendix G using only parametric knowledge, then assess this baseline in a separate call using prompt P9 in Appendix G that produces three signals: a continuous quality score $\sigma \in [0, 1]$, a categorical label $\ell \in \{\texttt{low}, \texttt{medium}, \texttt{high}\}$, and a binary `requires_search` flag. We retain only queries where the baseline is demonstrably inadequate:

$$\mathcal{Q}_{\text{gen}} = \{q \mid \sigma(q) \leq 0.75 \ \wedge \ \ell(q) \neq \texttt{high} \ \wedge \ \texttt{requires_search}(q)\}\,. \tag{3}$$

The joint condition on all three signals provides robustness against boundary cases where any single indicator may be unreliable. This stage yields 50 queries; a final manual selection pass curates the 35 queries used in the benchmark. A representative rejected candidate (from development logs) reads: *"Chinese OEMs are aggressively establishing local assembly footprints in Mexico to bypass US tariffs, fundamentally altering North American competitive dynamics . . . "* The no-retrieval baseline received σ just above 0.75 with $\ell{=}$`medium` and a high timeliness sub-score, indicating that an LLM could already produce a medium-quality answer from parametric knowledge; the query missed the strict $\sigma \leq 0.75$ retention threshold, illustrating the filter's discrimination near boundary cases.

The cumulative retention rate across all three stages is 19.4%, selecting a total of 35 queries from 180 candidates.

D.3 Canonical Domain Labels for Auto-Generation Routing Only

The auto-generation prompt constrains each generated query to declare exactly one of 11 canonical domain labels as shown in Table 10; this constraint is enforced through prompt-level rules (see P6 in Appendix G) and verified by post-hoc matching of the generated declaration against the table. A simple rule-based normalization function, i.e. substring matching with keyword fallback, defaulting to `tech`, maps free-form domain strings to these canonical labels.

Relationship to the benchmark's user-facing domains. These 11 canonical labels are used as a generation-time constraint within the auto generation pipeline; they are *not* the same as the 12 user-facing domains reported in the final benchmark composition (Appendix E.1, Figure 5). The benchmark's 12 user-facing domains emerge from post-hoc analysis across all 100 queries including user-derived queries that are not bound by the canonical constraint and include broader categories such as *Medical*, *Humanities*, and *Lifestyle* that fall outside the canonical taxonomy.

D.4 Worked Example: A High-Quality Auto-Generated Query

The following query from `crypto` domain in our auto-generated pool illustrates the design intent of trend-grounded generation. The query anchors on two named regulatory developments: the

Table 10: The 11 canonical domain labels used as a generation constraint in the auto-gen pipeline.

#	Label	Scope
1	`finance`	Financial markets, investment analysis, banking, macroeconomics, corporate earnings, and economic data.
2	`policy`	Government policy, regulation, governance, and institutional rule-making at local, national, and international levels.
3	`tech`	Technology, software engineering, AI/ML, hardware, and internet products. Also serves as the default fallback label.
4	`cybersecurity`	Digital security threats, defenses, vulnerability research, threat intelligence, and security operations.
5	`health`	Healthcare, medicine, pharmaceuticals, clinical research, medical devices, and public health.
6	`science`	Natural sciences, academic research, engineering, mathematics, and research infrastructure.
7	`education`	Education systems, learning, workforce training, talent development, and professional reskilling.
8	`legal`	Law, legal practice, compliance, data protection enforcement, and algorithmic accountability.
9	`energy`	Energy systems, climate policy, carbon markets, sustainability, and data center environmental impact.
10	`trade`	International trade, supply chains, logistics, manufacturing, and cross-border commerce.
11	`crypto`	Cryptocurrencies, digital assets, decentralized finance, blockchain technology, and CBDCs.

MAS Stablecoin Regulatory Framework of Singapore and the US GENIUS Act. Both of them postdate the LLM training cutoff and thus require real-time retrieval to answer accurately. The query naturally decomposes into three orthogonal investigative tracks: (i) regulatory comparison between two jurisdictions, (ii) quantitative compliance-cost modeling, and (iii) a technical architecture question about cross-border atomic settlement. This combination of named real-world triggers, decomposable structure, and authentic professional voice exemplifies the targets of the auto generation pipeline.

> *"We are currently evaluating whether to domicile and launch our B2B stablecoin cross-border settlement product under the new MAS Stablecoin Regulatory Framework in Singapore or under the emerging provisions of the US GENIUS Act. Compare the severe operational burden of the MAS requirement for 100% segregated reserve assets and the S$1 million base capital mandate against the projected compliance costs in the US. Crucially, how do the strict prohibitions on direct central bank payment system access in both jurisdictions impact our technical ability to execute true atomic cross-border settlements?"*

The filter signals on this query yield a necessity confidence of 0.9, search complexity `High`, and inverse-quality $\sigma{=}0.7$. These results indicate that the query is valid and is thus retained.

E Benchmark Composition, Verification, and Release

This appendix reports the final benchmark composition (domains, task types, joint distribution), construction statistics, the human verification protocol, and the release, privacy, and intended-use disclosures for the dataset introduced in §2.

E.1 Benchmark Overview

The final benchmark comprises 100 queries: 70 text-only and 30 multimodal (Figure 5).

Domain Coverage. Queries span 12 domains reflecting the breadth of real-world deep research needs. Tech (20) and Finance (17) are the most represented, followed by Science (13). Nine mid-frequency domains (Engineering, Medical, Business, Policy, Legal, Humanities, Cybersecurity, Lifestyle, and Education) each contribute 2 to 8 queries, ensuring that evaluation is not dominated

Table 11: Crosswalk between the three taxonomies, mapped by scope definition. User-facing domains marked † arise only from the user-derived path. The canonical label of an individual auto-generated query is assigned from its generated content, so it can differ from the row of its seed topic.

Generation topic (Table 9)	**Canonical label (Table 10)**	**User-facing domain (Fig. 5)**
AI Policy & Regulation	`policy`	Policy
Cybersecurity	`cybersecurity`	Cybersecurity
Finance & Macro	`finance`	Finance
Crypto & Digital Assets	`crypto`	Finance
Healthcare & Pharma	`health`	Medical
International Trade	`trade`	Business
Supply Chain & Industrial	`trade`	Business / Engineering
AI Engineering	`tech`	Tech
Climate & Energy	`energy`	Engineering / Science
Education & Workforce	`education`	Education
Legal & Compliance	`legal`	Legal
Biotech & Science	`science`	Science
—	—	Humanities†, Lifestyle†

by a narrow set of topics. These 12 user-facing domains are derived from post-hoc analysis of the final benchmark; they are distinct from the 11 canonical labels in Table 10 (Appendix D), which are used solely as a generation-time constraint within the auto-generated path. The user-derived path produces broader categories such as *Medical*, *Humanities*, and *Lifestyle* that fall outside the canonical generation taxonomy. Table 11 reconciles the three taxonomies used in the paper: the 12 generation topics (Table 9), which seed trend retrieval; the 11 canonical labels (Table 10), which constrain generation; and the 12 user-facing domains of Figure 5, which describe the final benchmark.

Task Distribution. Each query is annotated with one of 10 task types that characterize the reasoning pattern required. The three most common types are Decision & Recommendation (17), Comparative Analysis (16), and Fact Enumeration & Verification (15), which together account for nearly half the benchmark. Policy & Regulation Analysis (12), Causal Explanation (11), and Survey & Synthesis (11) form a second tier. The remaining four types (Trend & Forecast, Data Analysis & Computation, Code Generation, and Document Editing) cover specialized research patterns at lower frequency.

Cross-Distribution. Figure 5(a) shows the joint distribution of domains and task types. It is a coverage map rather than a fully crossed design: many domain-by-task cells hold one or two queries, so it shows that task types are spread across domains (Comparative Analysis appears in Finance, Science, Engineering, and Policy; Decision & Recommendation spans Tech, Medical, Business, and Legal), but it does not support conclusions about individual cells.

E.2 Assembly, Human Verification, and Release

Assembly. The final benchmark combines 65 user-derived and 35 auto-generated queries for a total of 100 (Table 12). Each query is annotated with its source, a domain label drawn from the 12 user-facing domains (Appendix E.1), a task type, and source-specific metadata: a feature vector and difficulty tier for user-derived queries; topic, necessity confidence, and baseline quality score for auto-generated queries (the auto-generated path additionally records the canonical domain label assigned during generation, for routing-constraint validation).

Human Verification. We validate the assembled benchmark using three annotators with graduate-level research experience. All three annotators are internal research staff at the authors' institution who participated as part of their salaried routine work; no separate compensation was issued. Each annotator independently assesses every query on two binary criteria: (i) *validity*, i.e., whether the query constitutes a legitimate deep-research task, and (ii) *non-triviality*, i.e., whether it requires web search to be answered adequately. During the same pass, annotators also verify the difficulty tier assigned by the LLM judge for user-derived queries (§C), flagging any obviously mislabeled tier for correction. Because the tier check was flag-and-correct rather than independent labeling,

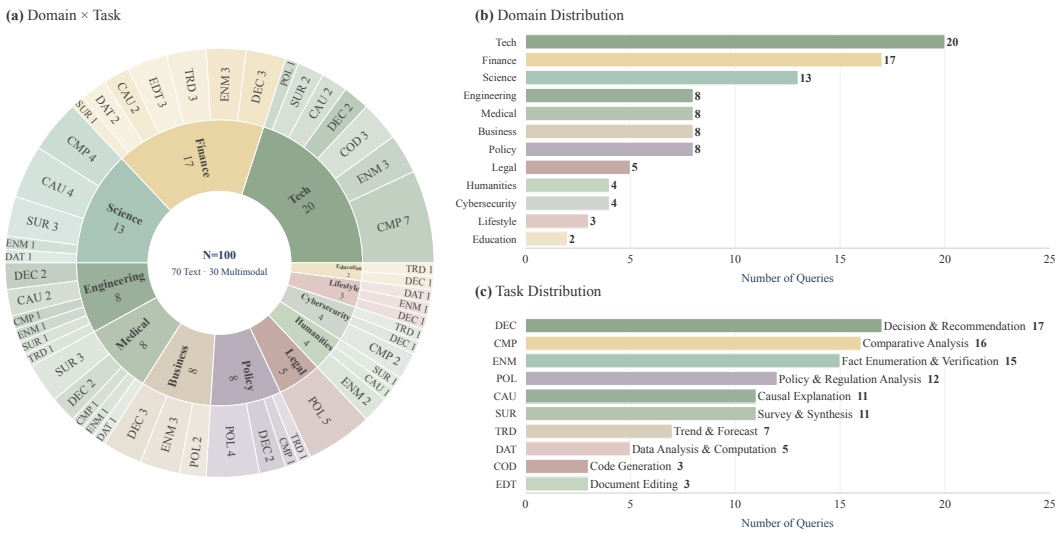


Figure 5: Query distribution across 12 domains and 10 task types (N=100; 70 text-only + 30 multimodal). **(a) Joint distribution**, a coverage map: task types spread across domains, but many cells hold only one or two queries and support no cell-level conclusions. **(b) Domain distribution.** Led by Tech (20), Finance (17), and Science (13); nine mid-frequency domains (Engineering, Medical, Business, Policy, Legal, Humanities, Cybersecurity, Lifestyle, Education) contribute 2–8 queries each. **(c) Task type distribution.** Decision & Recommendation, Comparative Analysis, and Fact Enumeration & Verification together account for ∼half of queries.

Table 12: Benchmark construction statistics. User-derived queries are fully rewritten from patterns observed during internal testing; auto-generated queries are produced by trend-grounded generation with three-stage filtering.

Stage	Count	Retention
User-Derived Path		
Internal testing patterns → rewritten queries	**65**	—
Auto-Generated Path		
Trend-grounded generation	180	—
+ Search validation	152	84.4%
+ Deep-research necessity	96	63.2%
+ Inverse quality assessment	50	52.1%
+ Manual selection	**35**	—
Cumulative from generation		*19.4%*
Final benchmark	**100**	—

inter-annotator agreement is reported for validity and non-triviality only. As shown in Table 13, both sources achieve substantial inter-annotator agreement ($\kappa \geq 0.74$) and acceptance rates of at least 90%. The acceptance rate is the fraction of queries the annotator majority judges valid and non-trivial; it measures agreement with the annotators rather than accuracy against an external gold standard. Weighted by source size (65 user-derived, 35 auto-generated), the overall acceptance rate is 92.6%.

Pipeline Operability and Refreshability. Both construction paths are designed for periodic re-execution, but their refresh mechanisms are not equivalent. The auto-generated path can be refreshed at any time, by anyone, by re-fetching trend context via the Serper API. The user-derived path, which supplies all 30 multimodal tasks, can incorporate new internal-testing rounds without re-engineering its prompts or classifiers, but it refreshes only as new internal-testing data accrues, under the same privacy and provenance constraints, and cannot be regenerated independently by others. The end-to-end pipeline (privacy filter, classification, rewriting and trend-grounded generation, and three-stage filtering) is implemented as parameterized modules so that the underlying language-model backend can be replaced without disturbing the data contract. This design directly addresses the staleness

Table 13: Human verification results. Three annotators assess validity and non-triviality. Fleiss' κ is computed within each source; the overall acceptance rate is weighted by source size.

Metric	User-Derived	Auto-Generated
Fleiss' κ (validity)	0.83	0.79
Fleiss' κ (non-triviality)	0.78	0.74
Acceptance rate (majority)	94.0%	90.0%
Unanimous agreement	86.0%	82.0%
Overall acceptance rate (size-weighted)	92.6%	

risk of static deep-research benchmarks: time-sensitive queries (e.g., recent regulatory changes) would otherwise become answerable from parametric knowledge as world events recede into training corpora.

Release and Maintenance. The evaluation framework code is released under the Apache-2.0 license. Author-created dataset components (the task schema, rewritten queries, curation metadata, and annotations) are released under the MIT license. Third-party attachments are not relicensed by us and remain subject to their original rights; the dataset card states this scope and provides a notice-and-takedown route for rights holders. The code is hosted at `https://github.com/MiroMindAI/MiroEval` and the dataset, with a Croissant file including Responsible AI fields, at `https://huggingface.co/datasets/miromind-ai/MiroEval-data`. ApodexEval was previously released as MiroEval (arXiv:2603.28407). The benchmark is maintained by the authors; bug reports, refresh requests, and pull requests are accepted via the project's issue tracker. We plan annual refresh releases using the dual-path pipeline described above, with version tags `vYYYY.MM` and a changelog documenting added, revised, and deprecated queries. Refresh cadence may be adjusted based on community demand.

Privacy. No raw user query, attachment, or session metadata appears in the released benchmark. All user-derived queries pass an automated privacy filter (rule-based plus model-assisted) that excludes content with personal communications, confidential documents, or private financial records, followed by entity replacement that substitutes all named institutions, organizations, and individuals with realistic same-type alternatives (§C). Internal testing participants were informed, through the notice that opened the testing channel, that de-identified test content may be used in benchmarks and released research; every contributor whose queries are included has given retrospective written confirmation, and an audit of all 65 user-derived queries found no identifiable or re-linkable content (Appendix F.2). The released artifact contains no identifiers linkable to individual users, and we do not redistribute the original session traces.

Intended Use. ApodexEval is intended for evaluating deep research agents on three dimensions (synthesis quality, factuality, and process integrity) on tasks that demand multi-step retrieval, cross-source synthesis, and long-form report generation. It is *not* intended for: (i) direct use as training data, due to evaluation-leakage risk; (ii) short-form question-answering benchmarking, since queries are deliberately constructed to require multi-round investigation; or (iii) safety, jailbreak, or red-teaming evaluation, which is orthogonal to the benchmark's design objective. Users planning derivative releases should respect the license terms above (Apache-2.0 for code, MIT for author-created data, original rights for third-party attachments) and clearly indicate any modifications.

Computational Cost. No GPU or TPU training is performed in this work. Benchmark construction (one-time) splits across two compute tiers. Privacy filtering, entity replacement, and user-derived rewriting are performed on internally-deployed LLM endpoints running on existing institutional infrastructure (no external API cost; raw queries do not leave internal infrastructure). Trend-grounded generation and three-stage filtering use `openai/gpt-5.2` via OpenRouter, consuming approximately 20M tokens at an approximate cost of $80; the auto-generation pipeline also issued approximately 250 Serper API calls (36 trend-retrieval calls across 12 topics $\times$ 3 subtopics, plus $\sim$180 Stage-1 search-validation calls), reported as call counts so that reproducers can size their API quota. Evaluating all 11 deep research systems on the 100-query benchmark across the three evaluation dimensions (synthesis quality, factuality, and process integrity, with GPT-5.1, GPT-5.2, and GPT-5-mini as judge models; §4)

consumed approximately 370M tokens at a total cost of $1,760. A single-system evaluation runs the three dimensions in parallel and completes in 3–6 hours of wall-clock time, consuming approximately 34M tokens (≈$160). The robustness experiments in Appendix K consist of (i) intra-judge re-runs and a prompt-sensitivity study, which re-invoke the same GPT-5.x judges on selected MultiModal-30 subsets and add approximately $80 to the GPT-5.x budget above, and (ii) a cross-judge re-evaluation of 11 systems × Text-Only 70 tasks × three dimensions using Gemini-2.5-Pro (synthesis and process) and Gemini-3-Flash (factuality), consuming approximately 400M tokens at an approximate cost of $600. Human verification (Appendix E.2) and the human ranking study (Appendix K.3) consumed approximately 80 person-hours in total, across three expert verifiers (validity, non-triviality, and difficulty review) and ten annotators, all of whom are internal research staff at the authors' institution who participated as part of their salaried routine work. Preliminary and failed experiments did not materially add to these budgets. External API calls (OpenRouter, Serper, evaluation judges) are issued from standard CPU machines. The internally-deployed LLM endpoints used for the user-derived path run on existing institutional GPU infrastructure shared with other ongoing research; no additional hardware was procured for this work. Token and dollar figures are approximate, based on internal usage logs and self-reported annotator estimates.

F Annotation Protocols and Source-Data Provenance

Two distinct groups of people contributed to this work, and they are documented separately. *Annotators* verified the benchmark and validated the evaluation framework (§F.1). *Internal staff* produced, as part of pre-release testing, the work-product from which the user-derived queries were rewritten (§F.2); they were not recruited as research participants.

F.1 Annotator Protocols

All annotators are salaried research staff at the authors' institution, paid above the statutory minimum wage of the jurisdiction, with no piece-rate payment and no compensation contingent on individual annotations or on which items were included. Before each task, annotators received a briefing on its purpose, the planned academic-benchmark use and public release of aggregate annotations, the absence of personal information in the materials, and their right to withdraw. The studies are:

- **Benchmark verification** (three annotators, all 100 queries; Appendix E.2): binary validity and non-triviality judgments per query, with a flag-and-correct check of the assigned difficulty tier.
- **Holistic ranking study** (ten annotators, 20 sampled multimodal queries; Appendix K.3): a single end-to-end ranking of systems from the report and the research process.
- **Verifier adjudication** (three annotators, 270 claims; Appendix L.1): a verdict for each claim given the verifier's retrieved evidence, blind to the verifier's own verdict, with written adjudication of disagreements.
- **Pairwise per-dimension study** (three annotators per pair, 150 pairs; Appendix L.2): which of two same-query outputs is better on one named dimension, blind to all scores, with written adjudication of disagreements.

Annotation materials. In the verifier-adjudication and pairwise studies, each annotator received a spreadsheet with one row per item, the blinded materials, a static HTML page for browsing them, and a written guide. System identities, judge scores, and the verifier's verdicts were withheld in a separate key held by the coordinator. The benchmark-verification and holistic-ranking tasks are specified in Appendices E.2 and K.3. The task instructions for the two component-validation studies follow.

Verifier adjudication: annotator guide

You receive a CSV of claims extracted from research reports. For each row, verify the claim with web search and fill the `human_label` column with exactly one of: `RIGHT`, `WRONG`, `CONFLICT`, `UNKNOWN`.
Label definitions. `RIGHT`: the claim is supported by evidence from reliable sources. `WRONG`: the claim is contradicted by evidence from reliable sources. `CONFLICT`: credible sources genuinely disagree about the claim (document both sides in `notes` if convenient). `UNKNOWN`: the claim cannot be verified either way after a reasonable search (no usable evidence found).
Procedure. (1) Read the `statement`. The `evidence_excerpt` column (may be empty) shows material previously associated with the claim; treat it as a starting point only and verify independently. (2) Search the web for primary or authoritative sources. Budget roughly 2 minutes per claim; if you are still unsure after that, label `UNKNOWN` rather than guessing. (3) Fill `human_label`. Use `notes` for anything noteworthy (key URL, ambiguity, why sources conflict). Do not edit any other column.
Independence rules. Work alone: no discussion with the other annotators until all CSVs are returned. No LLM/AI assistance of any kind (chatbots, AI summaries, AI search answers); plain web search engines are allowed. Do not attempt to guess what an automatic system said about the claim; the file intentionally hides that information. Label every row; partially filled files cannot be scored.

Pairwise per-dimension study: annotator instructions

Each row is one comparison: two anonymized outputs, A and B, for the same task, and one named dimension. Judge only which output is better on that dimension.
For each row. (1) Read the definition of the dimension in the `dimension` column. (2) Open the two files named in `output_A_ref` and `output_B_ref`. (3) Read only for this dimension. You do not need to read the full text; locate the parts relevant to the dimension. (4) Fill `preference` (`A`, `B`, or `TIE`; use `TIE` only if you genuinely cannot rank the two), `magnitude` (`negligible`, `moderate`, or `large`; `negligible` for a tie), and `notes`: one sentence that points to specific content in the outputs, not a generic remark such as "A is more comprehensive".
Dimensions. *Synthesis*: overall writing quality, meaning coverage, depth of analysis, fulfillment of the task's requirements, and clarity; set aside whether the facts are true (report only). *Task-Specific criteria*: whether the output meets the professional requirements particular to this task, rather than generic specificity (report only). *Intrinsic process quality*: the quality of the research process itself, meaning search breadth, progressive deepening, critical thinking, and absence of redundancy (research trace only). *Process→Report alignment*: whether the key findings in the trace are taken up in the report (report and trace). *Report→Process alignment*: whether the report's claims can be traced back to investigation actually carried out in the trace (report and trace).
Rules. Do not open the answer key or try to guess which system produced A or B. If an output names its own identity, ignore it and write `self-identified` in `notes`. Do not discuss any comparison with other annotators. Judge only the named dimension; the same pair may go the other way on a different dimension. No LLM/AI assistance of any kind. Take a break about every 20 comparisons.

F.2 Source-Data Provenance

Data subjects and released content. The user-derived queries originate from internal pre-release testing by full-time salaried staff of the authors' institution; no external users and no recruited participants are involved. Only rewritten query text is released: no original queries, sessions, contributor identities, or behavioral data.

Notice at collection time. The following notice was posted on 1 December 2025 in the internal testing workgroup, as the announcement that opened the dedicated testing channel, through which every user-derived query in the benchmark was collected. Product identifiers, links, and the contact are redacted.

> "Our [product] [version] is now open for internal testing through a dedicated channel: [link]. Everyone is invited to test actively and report issues. Queries, attachments, model outputs, and feedback submitted during testing may be recorded and used for product improvement, research analysis, and evaluation. After de-identification and necessary processing, some test content may be used in subsequent research, evaluation benchmarks, papers, or externally released materials. Please do not submit personal sensitive information, customer data, unpublished business information, or other confidential content. If you have questions about data use, or do not wish your content to be used in externally released research, please contact [contact]. Continued participation in testing indicates that you have read and acknowledged the above."

Consent basis. The notice is a general data-use notice rather than study-specific consent, and institutional authorization to use employee work-product is distinct from individual research consent. The basis for use is that the material is work-product created under ordinary employment and internal-testing terms, that its research use and release were institutionally authorized, and that only transformed, de-identified derivatives are released. In addition, we obtained retrospective written confirmation from every contributor whose queries are included. There is no consumer terms-of-service relationship: the queries were collected during internal testing by staff, governed by the employment and internal-testing terms and the notice above.

IRB review. We did not seek IRB review because the study did not recruit or interact with the internal staff as research participants, and it does not study their attributes, behavior, or employment outcomes. The source queries were used as task content rather than as observations about the individuals who produced them, and the released benchmark contains no contributor identities, sessions, or contributor-level records.

Anonymization audit. Beyond the privacy filter and entity replacement of Appendix C, we audited all 65 user-derived queries for identifiability and re-linkability and found no identifiable or re-linkable content. Three annotators independently screened every user-derived query and its attachments for (i) identifiable individuals or organizations, (ii) quasi-identifiers and distinctive role, project, or domain combinations, (iii) near-verbatim phrasing of the source query, and (iv) plausible re-linkability via public search.

G Construction Prompts

This appendix lists all prompts used by the construction pipelines (Appendix C, Appendix D). Prompts are reproduced verbatim from the production code; placeholders such as `{topic}` and `{query}` are bound at run time.

Table 14: Index of construction prompts.

ID	Pipeline	Stage	Purpose
P1	Rewrite	Classification	7-dimension classification and 8-feature tagging of anonymized queries
P2A	Rewrite	Strategy A (Easy)	Basic retrieval over an attachment plus one round of search
P2B	Rewrite	Strategy B (Medium)	Multi-source comparative analysis grounded in an attachment
P2C	Rewrite	Strategy C (Hard)	Contradiction injection between query and attachment
P2D	Rewrite	Strategy D (Hard)	Erroneous-premise injection requiring correction
P2E	Rewrite	Strategy E (Mid–Hard)	Multi-step text-only research synthesis
P2F	Rewrite	Strategy F (Easy–Mid)	Attachment-centric structured extraction
P3	Rewrite	Quality verification	Fact-check rewrite for factual errors, contradictions, and outdated info
P4	Rewrite	Inline extraction	Detect and extract long inline content from the query body
P5	Rewrite	Multi-turn merge	Merge multi-turn dialogue cases into a single query
P6	Auto-Gen	Generation	Trend-grounded query generation
P7	Auto-Gen	Stage 2 filter	Deep-research necessity assessment
P8	Auto-Gen	Stage 3 baseline	Baseline answer without retrieval
P9	Auto-Gen	Stage 3 quality	Inverse quality assessment of baseline answer

G.1 P1 — Classification and Feature Tagging

```
# Role
You are a professional benchmark data annotation assistant. Your job is to perform
     multi-dimensional structured analysis of a user query and any accompanying
     attachment, and emit a tag JSON to be consumed by a downstream automated
     rewriting pipeline.

# Task
You will receive:
- A user query (text)
- Optionally, an attachment description or attachment content (image / PDF / table,
     etc.); the attachment may also be absent

You must analyze the input along the dimensions below and produce a strict JSON
     object of tags.

# Tag Dimensions

## Dimension 1: `attachment_type`
The format or modality of the attachment.
Allowed values:
- `chart`: data visualizations such as line charts, bar charts, pie charts, scatter
     plots
- `table`: structured tables (HTML / Markdown / Excel / table screenshot)
- `document`: text-dominant content such as papers, reports, web-page screenshots,
     PDF pages
- `code`: code screenshots, terminal output, log files
- `image_other`: non-chart images (product photos, maps, natural images, screenshots
     without structured information)
- `mixed`: an attachment that combines multiple types above
- `none`: no attachment
- `uncertain`: cannot be determined; explain in the `analysis` field

## Dimension 2: `info_density`
How much extractable, citable, and verifiable information the attachment contains.
Allowed values: `high`, `medium`, `low`, `none`, `uncertain`.

## Dimension 3: `domain`
The primary subject area of the query.
Allowed values: `finance`, `tech`, `science`, `business`, `policy`, `health`,
     `education`, `lifestyle`, `other`, `uncertain`.

## Dimension 4: `complexity`
The intrinsic complexity of the original query.
Allowed values:
- `simple`: a single goal, no constraints, clearly stated
- `moderate`: 2-3 constraints or sub-tasks; some reasoning required
- `complex`: many constraints, multi-step, or cross-domain reasoning
- `uncertain`

## Dimension 5: `attachment_role`
The role the attachment plays in answering the query.
Allowed values:
- `essential`: the attachment must be read to answer
- `supplementary`: the attachment provides supporting context
- `irrelevant`: the attachment is weakly related or unrelated
- `none`: no attachment
- `uncertain`

## Dimension 6: `rewrite_potential`
A holistic estimate of the potential to rewrite this (query + attachment) pair into a
     high-quality benchmark item.
Allowed values: `high` / `medium` / `low`. Provide a one-sentence justification in
     `analysis`.
```

(continues on next page)

(continued from previous page)

```
## Dimension 7: `target_features`
Capabilities of the system under test that this case is suited to evaluate. Pick all
    that apply (typically 2-4):
- `goal_adherence`: query has multiple sub-goals that may be dropped
- `repetition`: similar sub-tasks easy to repeat
- `atomic_plan`: query must be decomposed into multi-step plan
- `atomic_search`: requires actively retrieving external information
- `atomic_report`: demands a structured output
- `factuality`: answer involves verifiable objective facts
- `error_correction`: query or attachment has room for embedded errors
- `multimodal_understanding`: attachment contains non-text information requiring
    parsing

# Output format
Output strict JSON only. Do not output anything besides the JSON.
{
  "attachment_type": "<value>",
  "info_density": "<value>",
  "domain": "<value>",
  "complexity": "<value>",
  "attachment_role": "<value>",
  "rewrite_potential": "<value>",
  "target_features": ["<feature_key_1>", ...],
  "analysis": "<short paragraph explaining each dimension>"
}

# Additional rules
1. If there is no attachment: set `attachment_type`, `info_density`,
    `attachment_role` to `none`, and do NOT include `multimodal_understanding`.
2. If a dimension genuinely cannot be determined, set it to `uncertain` and explain
    in `analysis`; do not force a guess.
3. Pick 2-4 `target_features` as a rule; prefer features reflecting the core
    characteristics, not edge cases.
4. Scoring logic for `rewrite_potential`:
   - `high`: attachment info density `high`/`medium` AND `attachment_role`
     `essential`/`supplementary` AND query has rewrite headroom
   - `medium`: partially meets the above
   - `low`: no attachment and information-poor query, or attachment unrelated to query

# Example

Input:
- Query: "What is this chart showing?"
- Attachment: A line chart of monthly sales for new-energy vehicle brands from 2019
    to 2023, covering 12 brands; Y-axis is sales (units), X-axis is months; legend
    complete and data points dense.

Output:
{
  "attachment_type": "chart",
  "info_density": "high",
  "domain": "business",
  "complexity": "simple",
  "attachment_role": "essential",
  "rewrite_potential": "high",
  "target_features": ["atomic_search", "atomic_report", "factuality",
    "multimodal_understanding"],
  "analysis": "The attachment is a sales line chart for multiple EV brands, with
    dense data points (12 brands x 60 months) and high information density. The
    original query is broad, leaving abundant rewrite headroom. High info density
    supports factuality assessment, the chart format requires
    multimodal_understanding, and rewriting it into a comparison task naturally
    covers atomic_search and atomic_report."
}
```

(continues on next page)

(continued from previous page)

```
# Now analyze the following input and output the JSON:

Query: {query}
Attachment description / content: {attachment}
```

G.2 P2A — Strategy A: Basic Retrieval (Easy)

```
You are a benchmark query rewriting assistant. Your task is to transform an original
     user query and its accompanying attachment into an "easy"-tier evaluation query.

## Goals of This Prompt

The query produced by this prompt should primarily evaluate the following
     capabilities of the system under test:
- Retrieval (atomic_search): the rewritten query must require the system to actively
     retrieve information beyond what is in the attachment.
- Multimodal understanding (multimodal_understanding): the rewritten query must
     require the system to extract specific information from the attachment.

When rewriting, ensure that:
1. Answering the query necessarily exercises the capabilities above and cannot bypass
     them.
2. If the rewrite cannot cover one of these capabilities, explain why in the output.

## Rewriting Objectives

- The rewritten query should require the model to: extract 1-2 key pieces of
     information from the attachment, perform one round of simple retrieval to add
     context, and then produce a concise, complete answer.
- Difficulty calibration: 1-2 constraints; the answer is readily available from
     mainstream public sources; the attachment is a supporting input rather than the
     sole necessary one.
- The query must read like something a real user would actually ask, not like an exam
     item.

## Strict Constraints

1. The rewritten query must not depend on paywalled databases, internal systems,
     login-gated pages, or any source that cannot be publicly crawled.
2. **[Language constraint]** The language of the rewritten query must exactly match
     the "Query language" field. This is a hard requirement.
3. **[Information preservation constraint]** All important supplementary information
     in the original query (job descriptions, technical specifications, policy
     clauses, lists of specific requirements, data parameters, evaluation criteria)
     must be preserved in full. You may adjust wording for naturalness, but you must
     not omit information or incorrectly relegate it to "see attachment".
4. The rewritten query must read like a natural, real-user question. Specifically:
   - Do not use enumerators such as 1)/2)/3); weave the requests together with
     natural prose.
   - Add a plausible user persona or motivation where appropriate.
   - Do not begin every query with "I've uploaded..."; vary the lead-in.
   - Anchor with specific times, scenarios, or entities to make it feel realistic.
   - Colloquial phrasing is allowed; avoid being overly polished.

## Output Format (strict JSON)
{
  "rewritten_query": "the full rewritten query text",
  "target_difficulty": "easy",
  "covered_features": ["atomic_search", "multimodal_understanding"],
  "feature_coverage_notes": "explain how this item covers the target features",
  "evaluation_anchors": [
    "Anchor 1: the model should correctly identify and reference the key information
     from the attachment",
```

(continues on next page)

(continued from previous page)

```
    "Anchor 2: the model should retrieve at least one public source to supplement or
     verify the attachment information",
    "Anchor 3: the model should not ignore the attachment and rely solely on
     retrieved results"
  ],
  "expected_search_sources": ["list 2-4 plausible public source types or specific
     websites"],
  "notes": "issues for human reviewers to watch for"
}

## Few-shot Example

**Original query**: How do this company's financials look in this image?
**Attachment description**: A table containing the quarterly revenue and net profit
     of a listed company for Q1-Q4 of 2023.
**Rewrite**:
{
  "rewritten_query": "I'm a finance writer working on an analysis piece about listed
     companies in the consumer sector. The attachment is a table with one company's
     quarterly revenue and net profit for 2023 -- could you take a look at how the
     full-year performance trended, and also check what the average profit margin in
     the consumer industry was in 2023 so I can do a quick comparison?",
  "target_difficulty": "easy",
  "covered_features": ["atomic_search", "multimodal_understanding"],
  "feature_coverage_notes": "The query requires extracting revenue and profit figures
     from the table (multimodal_understanding) and retrieving the industry average
     profit margin for comparison (atomic_search).",
  "evaluation_anchors": [
    "Anchor 1: the model should correctly read each quarter's revenue and net profit
     from the table",
    "Anchor 2: the model should retrieve the industry average profit margin and
     briefly compare it against the attachment data",
    "Anchor 3: the model should not fabricate any financial figures not present in
     the attachment"
  ],
  "expected_search_sources": ["company annual reports (official site)", "public pages
     of major financial data sites", "industry association public reports"],
  "notes": "Verify that the industry data the model cites comes from a publicly
     accessible source."
}

Now produce a rewrite based on the following inputs:

**Query language**: {{query_language}}
**Original query**: {{original_query}}
**Attachment description**: {{attachment_description}}
**Labels**: {{labels}}
```

G.3 P2B — Strategy B: Multi-Source Synthesis (Medium)

```
You are a benchmark query rewriting assistant. Your task is to rewrite an original
     user query and its associated attachment into a "medium"-difficulty evaluation
     query.

## Evaluation Goals of This Prompt

The query produced by this rewrite should primarily probe the system-under-test on
     the following capabilities:
- Planning (atomic_plan): the rewritten query should contain 2-3 logically dependent
     sub-tasks that require the system to plan a sensible execution order.
- Retrieval (atomic_search): the rewritten query should require the system to
     retrieve information from multiple distinct sources.
```

(continues on next page)

(continued from previous page)

```
- Report generation (atomic_report): the rewritten query should require the system to
    produce a structured comparative analysis.
- Factuality (factuality): the answer to the rewritten query should rest on
    verifiable, objective facts.
- Anti-repetition (repetition): the rewritten query should force multiple distinct
    comparison dimensions so the model cannot answer by repeating one point.
- Multimodal understanding (multimodal_understanding): the rewritten query must
    require the system to read and interpret data in the attachment.

When rewriting, ensure that:
1. Answering the rewritten query inevitably exercises the capabilities above; they
    cannot be bypassed.
2. If the rewrite cannot cover one of these capabilities, explain the reason in the
    output.

## Rewrite Objective
- The rewritten query should require the model to: compare the data or information in
    the attachment against at least 2 external public sources and produce a
    structured, synthesized report.
- Difficulty calibration: include 2-3 logically dependent sub-tasks (e.g.,
    extraction, comparison, evaluation) that demand a complete plan -> search ->
    report capability chain, with the answer drawing on multiple sources.
- It must read like a question a real user would actually ask.

## Strict Constraints
1. The rewritten query must not depend on paid databases, internal systems,
    login-gated pages, or any source that cannot be publicly crawled.
2. **[Language constraint]** The language of the rewritten query must exactly match
    the "Query language" field. This is a hard requirement.
3. Sub-tasks should be logically connected, not a simple enumeration.
4. **[Information preservation]** All important supplementary information in the
    original query must be preserved in full in the rewrite.
5. The rewritten query must read like a real user's natural question:
   - Do not use markers like (1)(2)(3); weave the requirements together in natural
    prose.
   - Add appropriate user identity or motivational background where suitable.
   - Vary the lead-in instead of always starting with "I've uploaded...".
   - Add concrete time, scenario, or entity anchors where appropriate.
   - Colloquial phrasing is allowed but avoid sounding overly polished.

## Output Format (strict JSON)
{
  "rewritten_query": "the full text of the rewritten query",
  "target_difficulty": "medium",
  "covered_features": ["atomic_plan", "atomic_search", "atomic_report", "factuality",
    "repetition", "multimodal_understanding"],
  "feature_coverage_notes": "explain how this item covers the target features",
  "evaluation_anchors": [
    "Anchor 1: the model should formulate a reasonable analysis plan that covers all
    sub-tasks",
    "Anchor 2: the model should obtain comparative data from at least 2 distinct
    public external sources",
    "Anchor 3: the model should clearly distinguish attachment data from externally
    retrieved data and cite sources",
    "Anchor 4: the model should not merely restate the attachment while ignoring the
    external comparison dimensions"
  ],
  "expected_search_sources": ["list 3-5 plausible public source types or specific
    websites"],
  "notes": "items that human reviewers should pay attention to"
}

## Few-shot Example

**Original query**: Take a look at our company's electricity usage data
```

(continues on next page)

(continued from previous page)

```
**Attachment description**: An Excel spreadsheet containing monthly electricity
     consumption and electricity expenditure for a manufacturing enterprise from 2022
     to 2023
**Rewrite**:
{
  "rewritten_query": "We're a mid-sized manufacturing company, and the attachment is
     our factory's monthly electricity consumption and cost breakdown for 2022
     through 2023. My boss recently asked me to look into whether our electricity
     spend is on the high side -- I'd like to understand where we stand relative to
     others in the industry and whether there have been any recent industrial
     electricity pricing policy changes we could take advantage of. If there's room
     for energy-saving and cost-reduction measures, some directional suggestions
     would be appreciated as well.",
  "target_difficulty": "medium",
  "covered_features": ["atomic_plan", "atomic_search", "atomic_report", "factuality",
     "repetition", "multimodal_understanding"],
  "feature_coverage_notes": "The query contains three logically dependent sub-tasks
     (atomic_plan); requires reading and interpreting the monthly electricity data in
     the attachment (multimodal_understanding); retrieves industry benchmarks and
     policy data from public sources (atomic_search); produces a structured
     comparative evaluation report (atomic_report); grounds conclusions in verifiable
     industry data and policy information (factuality); and forces multiple distinct
     comparison dimensions so the answer cannot collapse into a single point
     (repetition).",
  "evaluation_anchors": [
    "Anchor 1: the model should produce a plan that covers trend analysis, external
     benchmarking, and policy lookup",
    "Anchor 2: the model should correctly read the monthly data in the attachment and
     identify seasonality or anomalies",
    "Anchor 3: the model should retrieve manufacturing-sector electricity benchmarks
     and pricing policies from publicly accessible sources",
    "Anchor 4: the model should not directly conclude 'above/below the industry
     average' without external data support"
  ],
  "expected_search_sources": ["national statistics bureau energy bulletins",
     "government pricing-authority electricity policy announcements", "public reports
     from major electric power industry associations", "energy coverage from
     mainstream financial media"],
  "notes": "Watch whether the model proactively recognizes the data unit in the
     attachment"
}

Now please perform the rewrite based on the following input:

**Query language**: {{query_language}}
**Original query**: {{original_query}}
**Attachment description**: {{attachment_description}}
**Labels**: {{labels}}
```

G.4 P2C — Strategy C: Contradiction Injection (Hard)

```
You are a benchmark query rewriting assistant. Your task is to rewrite an original
     query into a "hard"-tier evaluation query by embedding, in the query text,
     factual claims that contradict the attachment content or publicly verifiable
     facts.

## Evaluation Goals of This Prompt

The query produced by this rewrite should specifically test the following
     capabilities:
- **Factuality**: the rewritten query should reference concrete, verifiable data so
     that the system must cross-check facts.
```

(continues on next page)

(continued from previous page)

```
- **Error correction**: the query text contains incorrect references that contradict
    either the attachment content or publicly known facts; the system must detect
    the inconsistency.
- **Multimodal understanding**: the system must first correctly parse the attachment
    in order to discover the false reference in the query.

When rewriting, please ensure:
1. Answering the query necessarily exercises the capabilities above and they cannot
    be bypassed.
2. If the rewritten query fails to cover any one of these capabilities, explain why
    in the output.

## Rewrite Objective

- The rewritten query should **actively cite information from the attachment but
    state it incorrectly on purpose**, creating a contradiction between the query
    and the attachment.
- Alternatively, a public fact stated in the query may contradict the attachment's
    actual data.
- The system under evaluation must understand the attachment **and** consult public
    sources in order to detect the contradiction.
- On the surface, the query should look like a normal information request, with no
    hint that any error may be present.

## Contradiction-Embedding Strategies

- **Contradiction with the attachment**: cite data from the attachment but state it
    incorrectly (e.g., a 15-30% numerical deviation, a misaligned date, a swapped
    name).
- **Contradiction with public facts**: background information in the query
    contradicts a publicly verifiable fact.
- The contradiction must "look plausible": correct order of magnitude, correct
    format, consistent with domain conventions, but different from the true value.
- It must not be obviously fake at a glance.
- Prefer contradictions that can be directly checked against the attachment or
    verified through public data sources.

## Strict Constraints

1. The rewritten query must not depend on data sources that cannot be publicly
    crawled.
2. **[Language constraint, mandatory]** The language of the rewritten query must
    exactly match the "query language" field. This is a hard requirement.
3. The rewritten query must not contain explicit cue words such as "verify,"
    "double-check," or "check whether there are errors."
4. The embedded contradictory information must be specific and directly comparable.
5. **[Information-preservation constraint]** All important supplementary information
    in the original query must be fully retained.
6. The rewritten query must read like a real user's natural question (no enumerator
    markers; varied lead-ins; concrete anchors; colloquial phrasing acceptable).

## Output Format (Strict JSON)

{
  "rewritten_query": "The full rewritten query text (with the embedded contradictory
    information)",
  "target_difficulty": "hard",
  "covered_features": ["factuality", "error_correction", "multimodal_understanding"],
  "feature_coverage_notes": "Explain how this item covers the target features",
  "embedded_contradictions": [
    {
      "error_type": "contradiction with the attachment / contradiction with public
    facts",
      "claimed_in_query": "The (incorrect) value claimed in the query",
      "actual_value": "The true value from the attachment or a public source",
      "error_description": "A specific description of the contradiction",
```

(continues on next page)

(continued from previous page)

```
      "detectable_via": "How the contradiction can be detected"
    }
  ],
  "evaluation_anchors": [
    "Anchor 1: the model should read the attachment and find that the data referenced
     in the query is inconsistent with the attachment's actual content",
    "Anchor 2: the model should retrieve public sources to cross-check the key data",
    "Anchor 3: the model should not directly accept the incorrect reference in the
     query when forming its conclusion",
    "Anchor 4: the model should clearly identify the contradiction and defer to the
     attachment / public data"
  ],
  "expected_search_sources": ["List the public sources that can be used to verify the
     contradiction"],
  "notes": "During human review, confirm that the embedded contradiction actually
     exists and that it is not too obvious"
}

## Few-shot Example

**Original query**: Help me analyze this company's annual report data.
**Attachment description**: A summary of a listed company's 2022 annual report (the
     attachment's actual net profit is RMB 1.23 billion).
**Rewrite result**:
{
  "rewritten_query": "I've been looking at investment opportunities in the consumer
     sector recently, and a friend sent me a 2022 annual report summary for a listed
     company (see attachment). I skimmed it quickly -- looks like net profit last
     year was around RMB 1.46 billion? Could you give me a systematic breakdown of
     its core financial performance? I'm mainly interested in revenue growth, profit
     levels, and the other key indicators, and I'd also like a read on the company's
     investment value in the context of the overall industry.",
  "target_difficulty": "hard",
  "covered_features": ["factuality", "error_correction", "multimodal_understanding"],
  "feature_coverage_notes": "The item requires reading financial data from the annual
     report (multimodal_understanding); the 1.46 billion claim contradicts the
     attachment's actual 1.23 billion (error_correction); it concerns verifiable
     financials (factuality).",
  "embedded_contradictions": [
    {
      "error_type": "contradiction with the attachment",
      "claimed_in_query": "Net profit of RMB 1.46 billion",
      "actual_value": "Attachment actually states RMB 1.23 billion",
      "error_description": "The query claims the attachment shows net profit of about
     1.46 billion, but the attachment actually records 1.23 billion -- roughly an 18%
     deviation.",
      "detectable_via": "Reading the net profit field in the attached annual report
     summary directly reveals the inconsistency."
    }
  ],
  "evaluation_anchors": [
    "Anchor 1: the model should read the attachment and find that the actual net
     profit is 1.23 billion.",
    "Anchor 2: the model should retrieve the company's officially disclosed 2022
     annual report to cross-check.",
    "Anchor 3: the model should not base its investment judgment on the incorrect
     1.46 billion figure.",
    "Anchor 4: the model should explicitly state that it defers to the attachment /
     official disclosure."
  ],
  "expected_search_sources": ["stock exchange disclosure systems", "major financial
     data platforms", "the company's official investor relations page"],
  "notes": "Human reviewers must confirm that the company's actual 2022 net profit in
     the attachment is indeed 1.23 billion."
}
```

(continues on next page)

(continued from previous page)

```
Now, given the following inputs, perform the rewrite by embedding into the query text
     information that contradicts the attachment content:

**Query language**: {{query_language}}
**Original query**: {{original_query}}
**Attachment description**: {{attachment_description}}
**Labels**: {{labels}}
```

G.5 P2D — Strategy D: Erroneous-Premise Injection (Hard)

```
# Benchmark Query Rewriting Assistant -- Strategy D: Hard-Tier Erroneous-Premise
     Injection

You are a benchmark query rewriting assistant. Your task is to transform an original
     user query into a **hard**-tier evaluation query that contains a false premise
     or an ambiguous expression.

## Evaluation Targets of This Prompt

The query produced by this rewrite should specifically probe:
- **Error correction (`error_correction`)**: the query text must embed a false
     premise or a common misconception, and the system under test should correct it
     rather than execute blindly.
- **Goal adherence (`goal_adherence`)**: even after detecting the false premise, the
     system under test must still complete the core task of the query -- it should
     not stop at correction alone.

When rewriting, ensure that:
1. Answering the query necessarily exercises the capabilities above; they cannot be
     bypassed.
2. If the rewritten query fails to cover any of the target capabilities, explain why
     in the output.

## Rewriting Objectives
- The rewritten query embeds one or more false premises, fabricated assumptions, or
     deliberately ambiguous expressions.
- The correct behavior of the system under test is: do **not** execute the query
     directly; first identify and flag the false premise or ambiguity, then provide a
     properly directed answer.
- If the system under test follows the false premise verbatim and produces an answer
     based on it, that is treated as a capability failure.
- **Strategy boundary with C**: The false premise must live in the query's background
     framing or world-knowledge claims, NOT in citations of the attachment. If an
     attachment is present, treat it only as task context; do NOT embed
     contradictions against the attachment -- that is Strategy C's role.

## Types of False Premises (select 1-2 to embed)
- **Factual error**: world-knowledge claims (data, dates, people, or institutions)
     that contradict publicly verifiable facts unrelated to any attachment content.
- **Logical contradiction**: premises that contradict each other.
- **Category confusion**: conflates concepts from two different domains.
- **Outdated information**: treats a superseded or rescinded policy/datum as a
     current fact.

## Strict Constraints
1. The embedded false premise must be verifiable against publicly available
     information.
2. The error must not be too obvious; it should sit within a reasonable boundary of
     plausible knowledge.
3. **[Mandatory language constraint]** The language of the rewritten query must match
     the `query_language` field exactly. This is a hard requirement.
4. The rewritten query must not depend on data sources that cannot be obtained from
     publicly crawlable channels.
```

(continues on next page)

(continued from previous page)

```
5. The output must explicitly state what error was embedded and what the correct
     answer is.
6. **[Information-preservation constraint]** All important supplementary information
     in the original query must be preserved in full.
7. The rewritten query must read like a real user's natural question (no enumerator
     markers; varied lead-ins; concrete anchors; colloquial phrasing acceptable).

## Output Format (strict JSON)
{
  "rewritten_query": "Full rewritten query text (with the false premise embedded but
     not flagged)",
  "target_difficulty": "hard",
  "covered_features": ["error_correction", "goal_adherence"],
  "feature_coverage_notes": "Explain how this item covers the target features",
  "embedded_errors": [
    {
      "error_type": "factual_error / logical_contradiction / category_confusion /
     outdated_information",
      "error_description": "What error was concretely embedded",
      "correct_fact": "What the correct fact is",
      "detectable_via": "Which public source can be used to verify it"
    }
  ],
  "evaluation_anchors": [
    "Anchor 1: the model should identify and explicitly flag the false premise in the
     query",
    "Anchor 2: the model should provide the correct factual information",
    "Anchor 3: the model should not execute the task on top of the false premise and
     produce conclusions based on incorrect information",
    "Anchor 4: if the premise is ambiguous, the model should ask the user for
     clarification rather than making its own assumptions"
  ],
  "expected_search_sources": ["Public sources that can be used to verify the false
     premise"],
  "notes": "Human reviewers should confirm: (i) the false premise is genuinely
     incorrect; (ii) the error is verifiable via public sources; (iii) the error is
     appropriately subtle"
}

## Few-Shot Example

**Original query**: Where are Tesla's factories in China and how much capacity do
     they have?
**False-premise design**: cite a stale capacity figure as if it were the "current"
     fact.
**Rewritten output**:
{
  "rewritten_query": "I've been looking into the capacity landscape of the new-energy
     vehicle industry, and I've come across materials saying that Tesla's Shanghai
     Gigafactory has an annual capacity of around 250,000 vehicles. Where would that
     capacity level rank among the world's major EV plants? And does Tesla have
     further expansion plans in China going forward?",
  "target_difficulty": "hard",
  "covered_features": ["error_correction", "goal_adherence"],
  "feature_coverage_notes": "The query embeds an outdated capacity figure
     (error_correction); after correcting it, the system under test must still
     complete the global ranking and expansion-plan analysis (goal_adherence) rather
     than stopping at the correction.",
  "embedded_errors": [
    {
      "error_type": "outdated_information",
      "error_description": "Shanghai Gigafactory annual capacity is described as
     250,000 vehicles, which is an early-2021 figure; actual capacity has since
     increased substantially.",
      "correct_fact": "As of 2023-2024, the Shanghai Gigafactory's annual capacity
     exceeds 950,000 vehicles (defer to the latest official figure).",
```

(continues on next page)

(continued from previous page)

```
      "detectable_via": "Tesla official investor reports; coverage in Bloomberg /
     Reuters and other major outlets."
    }
  ],
  "evaluation_anchors": [
    "Anchor 1: the model should point out that the 250,000 figure is outdated and
     provide a more current capacity number.",
    "Anchor 2: the model should cite the source and time point for its data.",
    "Anchor 3: the model should not perform a global ranking comparison on top of the
     250,000 false premise."
  ],
  "expected_search_sources": ["Tesla official site / IR page", "Bloomberg",
     "Reuters"],
  "notes": "When constructing the benchmark, verify the most current capacity figure
     to avoid the 'correct answer' itself becoming outdated."
}

Now produce a rewrite based on the following inputs:

**Query language**: {{query_language}}
**Original query**: {{original_query}}
**Attachment description**: {{attachment_description}} (used only as task context; do
     NOT inject errors that contradict the attachment -- that is Strategy C's role)
**Labels**: {{labels}}
```

G.6 P2E — Strategy E: Text-Only Research Synthesis (Mid–Hard)

```
You are a benchmark query rewriting assistant. Your task is to rewrite an original
     user query into a deep-research query that does not depend on any attachment and
     instead requires multi-step retrieval and synthesis of information across
     multiple sources, with a target difficulty of "medium" or "hard".

## Evaluation Goals of This Prompt

The query produced by this rewrite should primarily probe:
- **Planning (atomic_plan)**: the rewritten query must require multi-step execution
     with dependencies between steps.
- **Retrieval (atomic_search)**: the answer must require synthesis from multiple
     sources.
- **Report generation (atomic_report)**: multi-source information must be organized
     into a coherent analytical report.
- **Goal adherence (goal_adherence)**: the query should contain multiple constraints,
     and the system under test must keep all of them in view throughout.
- **Anti-repetition (repetition)**: the rewritten query should span multiple related
     but distinct analytical dimensions.

When rewriting, please ensure that:
1. Answering the query inevitably exercises the capabilities above and cannot bypass
     any of them.
2. If the rewritten query fails to cover one of the capabilities above, explain why
     in the output.

## Rewriting Objectives
- The rewritten query must not depend on any attachment and must be answerable purely
     through retrieval of public information.
- Adjust the number of constraints and the accessibility of the answer based on the
     target difficulty:
  - **Medium**: 2-3 constraints, with the answer distributed across 2-4 mainstream
     public sources.
  - **Hard**: 4+ constraints or constraints with dependencies, with the answer
     scattered across specialized/long-tail sources, requiring multiple rounds of
     retrieval and deep synthesis.
```

(continues on next page)

(continued from previous page)

```
## Strict Constraints
1. The rewritten query must not rely on data sources that cannot be publicly crawled.
2. **[Language constraint]** The language of the rewritten query must exactly match
     the "Query language" field. This is a hard requirement.
3. The rewritten query must exercise the full plan -> search -> report capability
     chain.
4. Hard-difficulty queries must include at least one sub-task that requires
     cross-domain retrieval or temporal analysis.
5. **[Information preservation constraint]** All important supplementary information
     in the original query must be fully retained.
6. The rewritten query must read like a real user's natural request:
   - Do not use enumerator markers; weave the requirements together in natural prose.
   - Where appropriate, add a user persona or motivation/background.
   - Add concrete time, scenario, and entity anchors to make the question feel
     authentic.
   - Colloquial phrasing is allowed, but avoid sounding artificially polished.

## Output Format (strict JSON)
{
  "rewritten_query": "the full text of the rewritten query",
  "target_difficulty": "medium or hard",
  "covered_features": ["atomic_plan", "atomic_search", "atomic_report",
     "goal_adherence", "repetition"],
  "feature_coverage_notes": "explain how this item covers the target features",
  "constraint_count": "number of constraints and a brief description",
  "answer_accessibility": "description of answer accessibility (mainstream sources /
     requires multi-source synthesis / scattered across long-tail sources)",
  "evaluation_anchors": [
    "Anchor 1: the model should formulate a retrieval plan that covers all sub-tasks",
    "Anchor 2: the model should retrieve information from at least N distinct sources
     and cite them",
    "Anchor 3: the model should not give a definitive conclusion in the absence of
     key sources",
    "Anchor 4: the model should identify and handle conflicts between sources"
  ],
  "expected_search_sources": ["list 3-6 plausible public sources"],
  "notes": "items that human reviewers should pay attention to"
}

## Few-shot Example (Medium)

**Original query**: The current state of large AI models
**Rewrite result**:
{
  "rewritten_query": "I'm an editor at a tech media outlet and I'm putting together a
     feature on large AI models. Could you help me map out how today's mainstream
     LLMs differ in their public benchmark performance, what their commercial pricing
     models look like, and what the main regulatory differences are between the US
     and China when it comes to large models? I'm looking for a fairly comprehensive
     side-by-side comparison that I can cite in my piece.",
  "target_difficulty": "medium",
  "covered_features": ["atomic_plan", "atomic_search", "atomic_report",
     "goal_adherence", "repetition"],
  "feature_coverage_notes": "The query requires planning retrieval along three
     dimensions (atomic_plan), pulling information from sources across different
     domains (atomic_search), generating a structured multi-dimensional comparison
     report (atomic_report), and keeping all three dimensional constraints in view
     throughout (goal_adherence).",
  "constraint_count": "3 parallel constraints: capability benchmarks,
     commercialization, regulatory policy",
  "answer_accessibility": "Requires synthesis across multiple sources",
  "evaluation_anchors": [
    "Anchor 1: the model should plan a search covering all three dimensions",
    "Anchor 2: benchmark data should come from credible public evaluation platforms",
    "Anchor 3: policy information should come from official government sources and
     not be conflated",
```

(continues on next page)

(continued from previous page)

```
    "Anchor 4: the model should not state information from after its training cutoff
      as established fact"
  ],
  "expected_search_sources": ["public LLM evaluation leaderboards", "official model
      vendor blogs", "national regulatory agency portals", "academic and industry
      reports"],
  "notes": "Watch whether the model distinguishes between vendor 'official reports'
      and 'third-party benchmarks'"
}

Now rewrite the query based on the following input:

**Query language**: {{query_language}}
**Original query**: {{original_query}}
**Attachment description**: {{attachment_description}} (if any; for background
    reference only)
**Labels**: {{labels}}
**Target difficulty**: {{target_difficulty}}
```

G.7 P2F — Strategy F: Attachment-Centric Task (Easy–Mid)

```
You are a benchmark query rewriting assistant. Your task is to rewrite an original
    user query into an evaluation query whose core task is centered on attachment
    processing, with a target difficulty of "easy" or "medium".

## Evaluation Goals of This Prompt

The query produced by this rewrite should primarily probe:
- Multimodal understanding (`multimodal_understanding`): the core task of the query
    depends on correctly parsing the attachment content.
- Report generation (`atomic_report`): the system under test must organize the
    information extracted from the attachment into a structured output.

When rewriting, please ensure:
1. Answering the query necessarily exercises the capabilities above; they cannot be
    bypassed.
2. If the rewrite cannot cover one of the capabilities above, explain why in the
    output.

## Rewriting Goals
- The main task of the rewritten query is to process the attachment (e.g., structured
    extraction, summary generation, format conversion, key-information localization,
    cross-page synthesis), with retrieval playing only a supporting role.
- Adjust the attachment-processing complexity according to the target difficulty:
  - **Easy**: a single attachment, a single processing task; with no or minimal
    retrieval needs.
  - **Medium**: the attachment contains multiple parts or has a relatively complex
    structure; the processing task involves 2-3 steps and requires one supporting
    retrieval.

## Strict Constraints
1. The core value of the rewritten query lies in processing the attachment content;
    retrieval must not become the main task.
2. If the rewritten query involves retrieval, the retrieved content must come from
    publicly crawlable sources.
3. **[Mandatory Language Constraint]** The language of the rewritten query must
    exactly match the "Query Language" field. This is a hard requirement.
4. The rewritten query should clearly state the expected output format (e.g., table,
    bullet list, structured summary, Markdown).
5. **[Information-Preservation Constraint]** All important supplementary information
    in the original query must be preserved in full.
6. The rewritten query must read like a real user's natural request (no enumerator
    markers; varied lead-ins; concrete anchors; colloquial phrasing acceptable).
```

(continues on next page)

(continued from previous page)

```
## Output Format (Strict JSON)

{
  "rewritten_query": "the full rewritten query text",
  "target_difficulty": "easy or medium",
  "covered_features": ["multimodal_understanding", "atomic_report"],
  "feature_coverage_notes": "explain how this item covers the target features",
  "primary_task": "core attachment-processing task",
  "secondary_task": "supporting retrieval task description (fill in null if none)",
  "evaluation_anchors": [
    "Anchor 1: the model should accurately parse the attachment structure and
     complete the specified processing task",
    "Anchor 2: the model's output should conform to the format requirements specified
     in the query",
    "Anchor 3: the model should not omit key information fields from the attachment",
    "Anchor 4: (if retrieval is involved) the supporting retrieval content should be
     genuinely verifiable, not fabricated"
  ],
  "expected_search_sources": ["if there is supporting retrieval, list possible
     sources; otherwise null"],
  "notes": "items to pay attention to during human review"
}

## Few-shot Example (Easy)

**Original query**: Help me organize this contract
**Attachment description**: An 8-page commercial contract PDF containing party A and
     party B information, contract amount, delivery terms, liability for breach, etc.
**Rewritten result**:
{
  "rewritten_query": "Our legal team asked me to pull together the key clauses of a
     commercial contract we recently signed so I can walk our manager through it. The
     contract PDF is attached -- could you extract the names of party A and party B,
     the total contract amount, the delivery milestones, the acceptance criteria, and
     the liquidated-damages clause, and put it all in a clean table? No need to
     translate the full text verbatim.",
  "target_difficulty": "easy",
  "covered_features": ["multimodal_understanding", "atomic_report"],
  "feature_coverage_notes": "The task requires parsing and extracting multiple
     structured fields from the contract PDF (multimodal_understanding) and
     organizing them into a clear tabular output (atomic_report).",
  "primary_task": "Structured information extraction with tabular output",
  "secondary_task": null,
  "evaluation_anchors": [
    "Anchor 1: the model should accurately identify and extract all specified fields
     from the contract without omissions",
    "Anchor 2: the output should be a clean table where field labels correspond
     correctly to their content",
    "Anchor 3: the model should not subjectively interpret contract terms or provide
     legal advice",
    "Anchor 4: if a field does not exist in the contract, the model should explicitly
     mark it 'not mentioned in the contract'"
  ],
  "expected_search_sources": null,
  "notes": "Check whether the model correctly preserves details such as currency
     units and date formats"
}

Now please rewrite based on the following inputs:

**Query language**: {{query_language}}
**Original query**: {{original_query}}
**Attachment description**: {{attachment_description}}
**Labels**: {{labels}}
**Target difficulty**: {{target_difficulty}}
```

G.8 P3 — Rewrite Quality Verification

```
You are a fact-checking assistant. Your task is to independently review a query text
     and determine whether it contains factual errors, outdated information, logical
     contradictions, or inconsistencies with attached materials.

## Task Description

You will be given a query text (possibly accompanied by attachments). Read the query
     carefully and identify all potential factual errors.

The scope of inspection includes:
1. **Contradictions with attached content**: The query cites data from the
     attachment, but values, dates, names, etc., do not match what the attachment
     actually contains.
2. **Contradictions with publicly known facts**: Facts stated in the query (e.g.,
     company data, policy provisions, historical events, scientific common knowledge)
     conflict with publicly verifiable information.
3. **Outdated information**: The query cites old, superseded data as if it were the
     current fact.
4. **Logical contradictions**: The premises within the query contradict one another.
5. **Category confusion**: Concepts from different domains are conflated.

## Important Notes
- Only report errors you have high confidence in; do not guess or over-infer.
- For information you are unsure about, do not flag it as an error.
- If an attachment is provided, carefully compare the data cited in the query against
     the actual data in the attachment.
- If no attachment is provided, use your own knowledge to judge whether the factual
     statements in the query are correct.

## Output Format (Strict JSON)
{
  "errors_found": [
    {
      "error_description": "Description of the error identified",
      "evidence": "Why you believe this is an error"
    }
  ],
  "has_errors": true/false
}

If no errors are found, output:
{
  "errors_found": [],
  "has_errors": false
}

Please review the following query:

**Query text**: {{rewritten_query}}
```

G.9 P4 — Inline Content Extraction

```
# Role
You are a text analysis assistant responsible for detecting whether a user query has
     large blocks of user-authored content embedded within it (e.g., personal
     statement drafts, reports, letters, articles), and for separating and
     anonymizing that content.

# Task
```

(continues on next page)

(continued from previous page)

```
You will receive a relatively long user query. You must perform the following three
    steps:

1. **Detect**: Determine whether the query contains a large embedded block of
    user-authored content.
2. **Separate**: If such content is present, split the query into a "core
    instruction" part and an "inline content" part.
3. **Anonymize**: Anonymize the inline content to protect personal privacy.

# Detection Criteria

**Counts as inline content**:
- The query embeds a complete user-authored document or a long passage of text (e.g.,
    personal statement, application letter, report draft, article draft).
- This content is the material the user wants the AI to process / revise / polish,
    rather than the instructional part of the query itself.
- Such content typically has independent structure and substantial length (several
    hundred words or more; rough trigger threshold ~2000+ characters of pasted text).

**Does not count as inline content**:
- The query itself is simply a long, detailed question or instruction.
- Short example snippets cited within the query.
- Background notes or contextual information included in the query.

# Separation Rules

Split the query into:
- **core_instruction**: The part that expresses what the user actually wants the AI
    to do. Replace concrete references to the embedded material with phrases like
    "the [content type] in the attachment".
- **anonymized_content**: The embedded large block of user content, after
    anonymization.

# Anonymization Rules

**Must be replaced**:
- Real personal names -> generic placeholder names
- Specific school / institution names -> a generic description of the same tier
- Specific program / major names -> keep the disciplinary direction but generalize
- Addresses that could identify the individual -> "a certain city" / "a certain
    region"
- Phone numbers / emails / ID numbers -> [redacted]
- Specific company names (non-public) -> "a certain company" / "a certain studio"

**Must be preserved**:
- The original structure (paragraph breaks, heading hierarchy)
- The writing style and tone
- Approximate length and content density
- General facts and viewpoints
- Disciplinary direction (after generalization)
- Time expressions: years may be kept, but blur specific dates

# Output Format

## If inline content is detected:
{
  "has_inline_content": true,
  "content_type": "<personal_statement|report|letter|article|essay|other>",
  "core_instruction": "the trimmed core instruction",
  "anonymized_content": "the full anonymized content",
  "anonymization_notes": ["original info A -> replacement A", "original info B ->
    replacement B"]
}

## If no inline content is detected:
{
```

(continues on next page)

(continued from previous page)

```
  "has_inline_content": false,
  "content_type": null,
  "core_instruction": null,
  "anonymized_content": null,
  "anonymization_notes": null
}

# Output Requirements
- Output strict JSON, with nothing outside the JSON.
- Do not wrap the output in Markdown code blocks.
- All fields must be present; fields that do not apply should be set to null.
- Newlines inside `anonymized_content` must be represented as `\n`.
- `core_instruction` should be concise but complete, preserving all of the user's key
    requirements.

Now please analyze the following query:
```

G.10 P5 — Multi-Turn Dialogue Merge

```
You are a multi-turn dialogue merging assistant. Your task is to consolidate a text
    containing multiple turns of dialogue into a single complete, self-contained
    query.

## Input Format

The input is a text segment containing markers `[Turn 1]`, `[Turn 2]`, ..., `[Turn
    N]`. Each turn may consist of:
- A user question or instruction
- A user follow-up, addition, or correction in response to an AI answer
- AI response content
- User constraints on output format, language, or scope

## Task

Synthesize the user's intent across all turns and generate a complete, self-contained
    single query. This query should enable a fresh AI assistant -- with no prior
    context -- to fully understand and execute the user's complete request.

## Merge Rules

1. **Preserve all of the user's core requirements and constraints**: do not omit any
    requirement raised in any turn.
2. **Incorporate the user's format/language directives**: if a turn instructs "answer
    in English", "output as a table", or similar, fold it into the query as a
    constraint.
3. **Discard AI replies**: the AI's own responses should not appear in the merged
    query, but the user's feedback or follow-ups regarding those responses must be
    retained.
4. **Merge progressive refinements**: combine progressive user requests into a single
    complete task description.
5. **Handle corrections and overrides**: if a later turn revises an earlier
    requirement (e.g., "don't use Python, use Java instead"), the final version
    takes precedence.
6. **Preserve the original language**: the merged query should use the same primary
    language as the original conversation.

## Edge Cases

- **Single-turn input**: If the input contains only one turn or has no `[Turn k]`
    markers, return the input verbatim with `turn_count: 1` and `merge_notes:
    "single turn, no merge required"`.
```

(continues on next page)

(continued from previous page)

```
- **All AI replies**: If every turn is an AI reply with no user instruction or
    feedback, return `merged_query: ""` with `merge_notes` explaining the degenerate
    input.

## Output Format

Output strictly in JSON format:

{
  "merged_query": "the complete merged query text",
  "turn_count": <integer>,
  "merge_notes": "a brief explanation of the merge logic"
}

## Few-shot Example

**Input**:
[Turn 1] User: I want a Python script that scrapes news headlines from a few major
    outlets.
[Turn 2] AI: Sure, here's a sample using BeautifulSoup that hits cnn.com and
    bbc.com...
[Turn 3] User: Actually use Java instead, and I want it to also extract publication
    dates and author names.

**Output**:
{
  "merged_query": "I want a Java program that scrapes news headlines from a few major
    outlets and also extracts publication dates and author names for each headline.",
  "turn_count": 3,
  "merge_notes": "Turn 3 overrides Turn 1's language choice (Python -> Java) per rule
    5; Turn 3 also adds two new requirements (publication dates, author names) per
    rule 4. The AI reply in Turn 2 is discarded per rule 3."
}
```

G.11 P6 — Trend-Grounded Query Generation

```
You are a deep research query designer. Generate {n} realistic, expert-level
queries that REQUIRE deep web search and multi-source synthesis to answer well.

## Topic Area
{topic} -- subtopics: {subtopics}

## Recent Real-World Context (use as background -- DO NOT invent events)
{trends}

## Example Queries from Real Users (match this specificity and natural language)
{seeds}

## Requirements
1. Frame queries around structural trends, ongoing debates, or comparative
   analyses that remain meaningful over 3-6 months. Do NOT pin to a single
   fleeting news event.
2. Each query must need 2-3 rounds of search from different angles.
3. Queries must require synthesis across multiple credible sources
   (reports, papers, news, official documents).
4. Mix personas: expert professionals (analyst, engineer, physician) and
   informed non-experts (grad student, startup founder, journalist, investor).
5. About 40% may include broad time context (e.g., "in 2025-2026") but
   avoid exact dates/weeks.
6. Length: 40-120 words. Specific but not verbose.
7. Include cross-country comparisons, conflicting expert views,
   forward-looking predictions, or quantitative analysis.
8. NO generic or textbook questions -- each must need CURRENT information.
```

(continues on next page)

(continued from previous page)

```
9. AVOID politically sensitive topics: specific leaders, partisan conflicts,
   military operations, territorial disputes, regime criticism. Focus on
   policy mechanisms, industry, technology, economics.
10. Domain label must be one of: finance, policy, tech, cybersecurity,
    health, science, education, legal, energy, trade, crypto.

Output ONLY a JSON array:
[
  {
    "query": "...",
    "domain": "...",
    "persona": "...",
    "anchored_event": "...",
    "time_sensitive": true/false,
    "expected_difficulty": "high/very_high"
  }
]
```

G.12 P7 — Deep-Research Necessity Assessment

```
Evaluate whether this query truly requires deep web search and multi-source
synthesis.

Query: {query}

Criteria:
1. Needs information that may have changed or emerged after the model's
   training cutoff (not answerable from a static LLM knowledge base alone)
2. Needs cross-verification from multiple credible sources
3. Needs multi-angle, multi-layered investigation
4. Is NOT a textbook question LLMs can already answer well

Decision rule:
Output `needs_deep_research=true` only if criterion 1 AND
(criterion 2 OR criterion 3) AND criterion 4 hold; otherwise false.
`needs_deep_research` is the retention signal -- queries with `true` are
kept; the other fields below are diagnostic-only and not used for filtering.

Output JSON only:
{
  "needs_deep_research": true/false,
  "confidence_score": 0.0-1.0,
  "reasoning": "Brief rationale (<=50 words) citing which criteria are satisfied",
  "search_complexity": "High"/"Medium"/"Low"
}
```

G.13 P8 — Baseline Answer Generation (No Retrieval)

```
Based solely on your existing knowledge, without any search tools, answer
this query concisely but thoroughly.

Query: {query}

If information is uncertain or possibly outdated, say so. Do not fabricate
facts.
```

G.14 P9 — Inverse Quality Assessment

```
Assess this no-search baseline answer for the given query.

Query: {query}
No-search answer: {baseline}

Decision rule:
Set `requires_search=true` if the baseline omits time-sensitive facts
(post-cutoff developments), lacks multi-source corroboration, or admits
uncertainty on load-bearing claims. `requires_search` is the retention
signal -- queries with `true` are kept; otherwise the query is dropped
as not needing deep research. Score `timeliness_score` low (<0.4) when
the query depends on developments after the model's knowledge cutoff.

Output JSON only:
{
  "overall_quality": "low"/"medium"/"high",
  "quality_score": 0.0-1.0,
  "timeliness_score": 0.0-1.0,
  "requires_search": true/false,
  "missing_aspects": "key missing info (<=30 words)"
}
```

Table 15: Performance comparison of models on user derived query and auto generation query.

Model	User-Derived				Auto-Generation				Overall
	Synthesis	Factuality	Process	Overall	Synthesis	Factuality	Process	Overall	
Grok Deep Research	59.8	65.6	63.4	62.9	57.8	62.0	53.4	57.7	60.3
Doubao Deep Research	63.2	60.8	47.9	57.3	65.2	68.8	58.1	64.0	60.7
Manus-1.6-Max Wide Research	57.1	67.9	64.6	63.2	53.7	77.2	63.9	64.9	64.1
Qwen-3.5-Plus Deep Research	57.9	70.3	58.7	62.3	62.1	75.8	63.5	67.1	64.7
ChatGLM Agent	61.1	62.9	64.5	62.9	65.3	74.0	66.7	68.7	65.8
MiniMax-M2.5 Research	63.5	67.5	65.4	65.5	63.1	76.0	68.9	69.3	67.4
Claude-Opus-4.6 Research	65.9	70.1	66.3	67.4	68.7	69.6	65.7	68.0	67.7
Kimi-K2.5 Deep Research	**74.9**	63.5	64.1	67.5	**76.5**	67.5	64.3	69.5	68.5
Gemini-3.1-Pro Deep Research	70.1	69.5	65.8	68.5	72.3	73.0	68.4	71.2	69.9
MiroThinker-1.7	73.6	78.5	**71.2**	**74.4**	75.0	80.5	74.3	76.6	75.5
OpenAI Deep Research	71.4	**80.3**	71.0	74.2	76.3	**86.4**	**75.1**	**79.3**	**76.7**

H Further Analysis of Evaluation Results

This appendix examines benchmark behavior across two complementary axes: query source and modality setting.

H.1 User-Derived vs. Auto-Generated Queries

The 70 Text-Only tasks comprise two equally sized subsets: 35 user-derived queries curated from real-world usage patterns through privacy-preserving rewriting, and 35 auto-generated queries produced by a trend-grounded pipeline. Table 15 compares system performance across these two sources.

Auto-generated queries are consistently easier. Nearly all systems with complete data score higher on the auto-generated subset, with overall improvements ranging from 0.6 points for Claude-Opus-4.6 to 6.7 points for Doubao Deep Research. This gap likely reflects the greater complexity and ambiguity inherent in queries grounded in real-world usage patterns, which often involve underspecified goals, domain-specific jargon, and multi-faceted information requirements that are difficult to replicate through automated generation. Despite the difficulty gap, the relative ranking of systems remains largely stable across the two subsets. The top positions are occupied by the same systems in both cases, and the lower tier (Doubao, Qwen-3.5-Plus) is also consistent. Notably, the top systems all improve by a few points on auto-generated queries (OpenAI 5.1, MiroThinker-1.7 2.2, and Gemini-3.1-Pro 2.7 points), indicating that the user-derived subset does not systematically advantage any particular system.

Factuality also shows a systematic source effect. The average Factuality score across systems is approximately 4 to 5 points higher on auto-generated queries, suggesting that trend-grounded queries, which are anchored in recent and well-documented web events, are easier to verify than the more niche topics arising from real usage.

These results carry two implications for benchmark design. First, the ranking stability validates that auto-generated queries provide a reasonable proxy for real-world difficulty, supporting the scalability of automated benchmark construction. Second, the consistent difficulty gap highlights that user-derived queries capture a dimension of complexity that automated generation does not fully reproduce, arguing for the inclusion of both sources in a comprehensive benchmark.

H.2 Text-Only vs. MultiModal Comparison

In §4.2 we observed that multimodal tasks degrade most systems. Here we quantify this effect more systematically. Across the nine systems with both Text-Only and MultiModal overall scores, the average overall score drops by 3.3 points. However, the degradation is highly uneven across systems and dimensions.

By dimension, Synthesis quality suffers the largest average decline at about 5.5 points, with Qwen-3.5-Plus experiencing an extreme drop of 15.4 points from 60.0 to 44.6 and MiniMax-M2.5 declining by 6.6 points. Process scores decrease by an average of about 4.6 points, with ChatGLM Agent showing

the sharpest decline of 7.9 points from 65.6 to 57.7. In contrast, Factuality Ratios remain remarkably stable, with the multimodal–text-only difference within ± 0.2 points on average, suggesting that multimodal tasks do not systematically degrade factual precision.

This pattern reinforces a finding from §4.3. The multimodal bottleneck lies in report generation, particularly task-specific quality and coverage of visual content, and in research process quality, particularly analytical depth, but not in factual verification. Systems that already struggle with these capabilities in the Text-Only setting experience disproportionate degradation when visual understanding is required. The relative ranking between systems nevertheless remains broadly consistent across both settings.

I Detailed Outcome-Level Analysis

This appendix provides the full sub-metric breakdowns and per-system analyses summarized in §4.3. We focus primarily on the Text-Only setting (70 tasks) due to its broader system coverage. Table 2 reports the full breakdown.

Synthesis Sub-Metrics: Task-Specific Quality is the Bottleneck, Insight is the Differentiator. Among the five Synthesis sub-metrics, the Task-Specific (TS) dimension, the aggregate of the per-query dynamic dimensions, emerges as the universal bottleneck: it is the lowest-scoring sub-metric for nearly every system in both settings (for example, OpenAI Deep Research scores 78.2 on Coverage and 69.1 on TS). Because absolute levels depend on the judge (Appendix K.2), we rely on this ordering rather than on the size of the gaps. This consistent shortfall indicates that current systems can identify relevant topics with reasonable breadth but struggle to provide the granular, evidence-grounded details that distinguish thorough research from surface-level summaries. Instruction-following, by contrast, is uniformly high among top systems and is no longer a meaningful differentiator.

While Task-Specific quality marks the shared weakness, Insight is what most separates systems from one another: its spread across systems, from Manus-1.6-Max (lowest) to Kimi-K2.5 (highest), is substantially wider than that of Coverage or Instruction-following. This variance reveals that the ability to synthesize non-obvious analytical observations, rather than merely aggregating retrieved information, is the most discriminative report-writing capability. Several systems with moderate overall performance, such as Gemini-3.1-Pro at 76.6 and Claude-Opus-4.6 at 72.0, score relatively well on Insight, suggesting analytical strengths offset by weaknesses in other dimensions.

Factual Claims: A Precision–Volume Trade-off. The Factuality sub-metrics reveal a fundamental tension between how many claims a system generates and how often those claims are correct as shown in Figure 4. At one extreme, ChatGLM Agent and Gemini-3.1-Pro produce over 4,000 correct claims each, but this high volume comes with 580 and 526 wrong claims respectively, plus over 900 unverifiable ones, pulling their Factuality Ratios down to around 70. At the other extreme, OpenAI Deep Research generates fewer correct claims at 3,335 but keeps wrong claims to just 170 and unverifiable claims to 496, achieving the highest per-task right Ratio of 83.3. These profiles reflect fundamentally different generation strategies. The spread of profiles indicates that the precision–volume trade-off is a real design axis along which deep research systems can be tuned. Because large pools of unverifiable claims lower the right ratio as much as wrong ones do, these profiles should be read together with the `UNKNOWN`-excluded ratio and the separate verdict rates (Table 20).

Connecting the Two Dimensions: What Drives the Synthesis–Factuality Misalignment? The sub-metric breakdowns above help explain the synthesis–factuality misalignment observed in §4.2. Kimi-K2.5, which achieves the highest Synthesis Avg yet one of the lowest Factuality Ratios, combines the leading Insight score of 79.8 with a high wrong-claim count of 595 and the second-largest pool of unverifiable claims at 1,256. Kimi-K2.5's reports are analytically rich but insufficiently grounded. It generates insightful interpretations that are not always backed by verifiable evidence. Manus-1.6-Max Wide Research presents the mirror image. Its Insight score of 54.8 is the lowest among all systems, dragging its Synthesis Avg down to 55.4, yet it produces only 191 wrong claims across nearly 2,000 correct ones, yielding a competitive Factuality Ratio of 72.6. Manus-1.6-Max appears to prioritize factual caution over analytical depth, a defensible strategy for high-stakes tasks but one that limits report usability. These contrasting profiles suggest that the synthesis–factuality

gap is not random and is systematically driven by how systems balance analytical ambition against factual verification.

Takeaways. The outcome-level analysis yields two actionable insights. First, improving *task-specific* quality is the most impactful path to better synthesis quality, as coverage and instruction-following are approaching saturation among top systems. Second, the precision–volume trade-off in factual claims is a salient design axis. Systems can be characterized by where they sit along it, and improving research and verification strategies offers a path to relax the trade-off without sacrificing either dimension.

J Detailed Process-Level Analysis

This appendix provides the full per-dimension analyses summarized in §4.4. Table 3 reports both Intrinsic metrics (Search Breadth, Analytical Depth, Progressive Refinement, Critical Thinking, and Efficiency) and Alignment metrics (Process→Report coverage, Report→Process traceability, and Contradiction detection).

Intrinsic Quality: Systems Search Wide but Fail to Go Deep. The Intrinsic sub-metrics reveal a consistent structural imbalance. Most systems achieve reasonable Search Breadth but substantially lower Analytical Depth. In the Text-Only setting, Breadth scores cluster between 71.9 and 79.1 for eight of the eleven systems, whereas Depth scores spread far more widely from 41.6 for Doubao to 67.3 for OpenAI Deep Research. This makes Depth the single most discriminative Intrinsic metric, echoing the role that Insight plays among Report sub-metrics (Appendix I). Claude-Opus-4.6 offers a particularly instructive case. Its Breadth of 79.1 is the highest among all systems, but its Depth of 58.8 trails the best Depth score (67.3) by about 8 points, suggesting a search strategy that retrieves broadly but rarely follows up with targeted, iterative investigation.

Beyond Depth, Efficiency is a widespread weakness: it is the lowest Intrinsic score for five of the eleven systems, and every system except Manus-1.6-Max (68.8) falls in the 53 to 65 range. This indicates that current research processes contain substantial redundancy, including repeated queries, circular exploration paths, and retrieved information that is never utilized, pointing to a clear avenue for future optimization.

Alignment: Findings Reach the Report, but Reports Outrun the Process. The Alignment metrics expose a revealing asymmetry between two directions of process–report consistency. Process→Report (P→R) scores are generally high. Top systems exceed 83, and even mid-tier systems such as MiniMax-M2.5 and ChatGLM Agent remain above 70. This means that information uncovered during the research process is, for the most part, successfully incorporated into the final report.

Report→Process (R→P) tells a different story. Scores are dramatically lower across the board. Even the best system reaches only 59.4, and most others fall below 55, with Doubao at 36.8 and Qwen-3.5-Plus at 39.7. R→P falls well below P→R for every system, and the asymmetry persists within the most-exposed tier (Appendix L.5), revealing that a substantial portion of report content cannot be traced back to the research process. The R→P level itself partly reflects how much of its process a system exposes, so we compare systems on the asymmetry rather than on absolute R→P levels. Systems routinely introduce claims, interpretations, or synthesized content that do not originate from their documented search and analysis steps. Whether this reflects implicit reasoning, hallucination, or unlogged intermediate steps, the practical implication is the same. Current deep research systems exhibit a significant traceability gap that undermines the auditability of their outputs.

Contradiction Detection (Contr) further differentiates systems on a complementary axis. Top systems (above 79) score well above mid-tier systems clustered in the low 70s, while Doubao and Qwen-3.5-Plus fall below 57, suggesting limited capacity to handle conflicting sources. This spread of nearly 30 points highlights contradiction resolution as a critical and highly variable capability for complex research tasks where authoritative sources frequently disagree.

Process and Outcome Quality. In §4.2 we noted that process quality is broadly aligned with outcome quality. Here we deepen this observation by examining how Process relates to Synthesis and Factuality individually versus jointly. When correlated with Synthesis alone, the relationship

Table 16: Intra-judge stability on MultiModal (30 tasks). Three independent runs with the same GPT judge configuration. Each dimension reports scores from Run 1 / Run 2 / Run 3, followed by the mean and standard deviation.

Model	Synthesis			Factuality			Process			Overall			Avg	Std
	R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3		
OpenAI Deep Research	66.7	66.5	66.7	77.0	75.0	76.9	66.8	65.0	66.2	70.2	68.8	69.9	69.6	0.6
Gemini-3.1-Pro	66.4	66.3	66.9	73.7	70.0	72.7	64.1	63.0	62.8	68.1	66.4	67.5	67.3	0.7
MiroThinker-1.7	69.0	68.7	68.6	78.4	80.9	79.6	67.4	67.7	67.4	71.6	72.4	71.9	72.0	0.3

is moderate. Doubao achieves a Synthesis of 64.2 despite a Process score of only 53.1, and Qwen-3.5-Plus attains a Process score of 61.1 that substantially outranks its Synthesis of 60.0 relative to peers. The correlation with Factuality alone is similarly imperfect. Kimi-K2.5's Process score of 64.2 would not predict its unusually low Factuality Ratio of 65.4. However, when Synthesis and Factuality are combined into the overall outcome score, these individual irregularities partially cancel out and the alignment with Process becomes stronger. A strong research process benefits both dimensions simultaneously, while idiosyncratic strategies that inflate one dimension at the expense of the other are averaged away. Empirically, the Spearman correlation between Process and the combined outcome score is $\rho = 0.88$ (95% CI $[0.73, 0.94]$; leave-one-system-out range $[0.84, 0.93]$; Pearson $r = 0.83$), compared with 0.62 for Synthesis and 0.69 for Factuality alone. With 11 systems this is a limited-sample association; we report it as evidence that process and outcome quality are aligned, not as a predictive model.

Takeaways. The process-level analysis identifies two systemic weaknesses shared by current deep research systems. First, Analytical Depth and Efficiency are the primary Intrinsic bottlenecks. Systems retrieve broadly but rarely investigate deeply, and much of the retrieval effort is wasted. Second, the P→R versus R→P asymmetry reveals a fundamental traceability gap. Reports consistently contain more than what the research process can account for. Process quality is also strongly aligned with overall outcome across systems, and its diagnostic value lies in revealing weaknesses, such as the traceability gap, that output-level metrics alone cannot show.

K Evaluation Robustness and Human Study

A key concern for any LLM-based evaluation framework is whether the results are sensitive to random variation, the choice of judge model, or minor prompt differences. We address this through three controlled robustness experiments (§K.1–§K.2), followed by a human study that validates consistency with expert judgments (§K.3). These analyses establish stability and aggregate agreement; the validity of each individual component is assessed separately in Appendix L.

K.1 Intra-Judge Stability

LLM-based evaluation can exhibit non-trivial variance across runs due to sampling randomness. To quantify this, we re-run the primary judge configuration (GPT series) two additional times on the MultiModal setting (30 tasks) for three systems: OpenAI Deep Research, Gemini-3.1-Pro, and MiroThinker-1.7. Together with the original run, this yields three independent evaluations per system. Table 16 reports the mean and standard deviation across runs. This is a targeted diagnostic of judge stochasticity on three systems rather than a variance estimate for every system; system-level uncertainty for all systems is reported through task-bootstrap intervals (Appendix L.7).

The standard deviations on Overall are remarkably low, ranging from 0.3 to 0.7 across systems, and the system ranking is identical across all three runs. At the sub-dimension level, Synthesis scores are the most stable with variations under 1 point, while Factuality shows slightly larger fluctuations of up to about 4 points for individual systems (e.g., Gemini-3.1-Pro: 73.7 / 70.0 / 72.7), likely due to the stochastic nature of web search during claim verification. Despite these per-dimension fluctuations, the Overall ranking remains perfectly preserved, confirming that the evaluation results are stable under repeated execution.

K.2 Cross-Judge Consistency and Prompt Sensitivity

Beyond run-level variance, we further examine whether system rankings are robust to the choice of judge model and the formulation of judge prompts. Table 17 summarizes both experiments. Each cell reports scores in the format *original / alternative / Δ*.

Cross-Judge Consistency. We re-evaluate the Text-Only setting with 70 tasks using Gemini as an alternative judge for all three dimensions. Specifically, we use Gemini-2.5-Pro for synthesis and process evaluation, and Gemini-3-Flash for factuality evaluation. The Gemini judge produces substantially higher absolute scores across the board, with Overall deltas ranging from $+13.2$ of OpenAI Deep Research to $+18.7$ of Grok Deep Research. This systematic shift is most pronounced on Process with deltas of $+17.1$ to $+24.0$ and least on Factuality from $+6.0$ to $+14.2$, suggesting that Gemini applies more lenient criteria for process evaluation than for factual verification, and that lower-scoring systems receive larger increases, which compresses inter-system gaps. Despite these large absolute shifts, per-dimension rankings are preserved closely (Kendall's $\tau = 0.93$ for Synthesis, 0.90 for Factuality, and 0.79 for Process), with Process the least stable dimension, and the Overall ranking changes only by three swaps of adjacent systems (Gemini-3.1-Pro and Kimi-K2.5, Qwen-3.5-Plus and Manus-1.6-Max, Doubao and Grok; $\tau = 0.89$). Cross-judge differences are thus largely systematic, which supports comparative conclusions stated as orderings; because absolute levels move by up to 24 points on a single dimension, we do not interpret absolute scores or gaps as calibrated quantities. Preserved ordering alone does not rule out a judge favoring its own model family in absolute terms; Appendix L.6 tests this directly.

Prompt Sensitivity. We re-evaluate all systems on the MultiModal setting with 30 tasks using the same GPT judge but with a modified prompt that rephrases the scoring criteria in a more concise format and adjusts the ordering of evaluation dimensions. In contrast to the cross-judge experiment, the prompt modification produces only modest score changes: Overall deltas fall within -2.2 to -0.7 points, with Synthesis shifts below 1 point for all systems. The dimension showing the largest variation is Factuality, up to -3.3 for Grok Deep Research, consistent with the higher sensitivity of claim-level verification to prompt phrasing. The Overall ranking is fully preserved, confirming that the evaluation outcomes are robust to reasonable prompt reformulations.

K.3 Human Study

To validate that our automated evaluation aligns with expert human judgment, we conduct a human study with 10 qualified annotators. All ten annotators are internal research staff at the authors' institution who participated as part of their salaried routine work; no per-task or hourly piece-rate compensation was issued, and routine salaries materially exceed the statutory minimum wage in the jurisdiction where the work was performed. We randomly sample 20 queries from the benchmark and collect reports from all deep research systems that support multimodal attachments. For each case, annotators are provided with both the final report and the associated research process.

Annotators are asked to rank the systems based on overall quality, jointly considering the effectiveness of the research process and the quality of the resulting report. We deliberately adopt a single end-to-end ranking rather than requesting separate per-dimension judgments along Synthesis, Factuality, and Process. When evaluating a long-form deep-research output, human assessors form an integrated impression that is difficult to cleanly decompose: judgments about analytical depth interact with perceptions of factual reliability, and the readability of a report colors the assessor's view of its underlying evidence. Forcing annotators to disentangle these dimensions tends to introduce additional subjective noise rather than yield cleaner per-dimension signals. We therefore treat humans as holistic end-to-end evaluators and use this overall human judgment as the reference for validating ApodexEval's aggregate ranking. The goal of this study is not to verify that ApodexEval matches human assessment on each individual dimension, but to verify that the aggregate ranking produced by ApodexEval is consistent with how human experts holistically judge system quality. Per-dimension validity is assessed with a separate pairwise design that asks about one dimension at a time (Appendix L.2).

Table 18 reports the average human ranking alongside the ApodexEval ranking for each system. The two rankings exhibit strong agreement. The top six systems (OpenAI Deep Research, MiroThinker-1.7, Gemini-3.1-Pro, Claude-Opus-4.6, MiniMax-M2.5, and ChatGLM Agent) match exactly in both

Table 17: Robustness to judge model choice and prompt variation. Each cell reports *original / alternative / Δ*. Overall is the equal-weight mean of the three dimensions; ΔRank: change in Overall rank under the alternative (positive = higher). **Upper**: Cross-judge consistency on Text-Only (70 tasks), GPT vs. Gemini. **Lower**: Prompt sensitivity on MultiModal (30 tasks), original vs. modified prompt with the same GPT judge.

Model	Synthesis	Factuality	Process	Overall	ΔRank
Cross-Judge: Text-Only (70 Tasks) — GPT Series (orig.) / Gemini Series (alt.) / Δ					
OpenAI Deep Research	73.8 / 90.2 / +16.4	83.3 / 89.3 / +6.0	73.1 / 90.2 / +17.1	76.7 / 89.9 / +13.2	0
MiroThinker-1.7	74.3 / 90.8 / +16.5	79.4 / 87.6 / +8.2	72.7 / 91.0 / +18.3	75.5 / 89.8 / +14.3	0
Gemini-3.1-Pro Deep Research	71.2 / 89.5 / +18.3	71.3 / 81.8 / +10.5	67.1 / 88.0 / +20.9	69.9 / 86.4 / +16.5	-1
Kimi-K2.5 Deep Research	75.7 / 93.3 / +17.6	65.4 / 79.6 / +14.2	64.2 / 87.9 / +23.7	68.4 / 86.9 / +18.5	+1
Claude-Opus-4.6 Research	67.3 / 87.3 / +20.0	69.8 / 81.8 / +12.0	66.0 / 87.9 / +21.9	67.7 / 85.7 / +18.0	0
MiniMax-M2.5 Research	63.3 / 82.4 / +19.1	71.8 / 81.8 / +10.0	67.1 / 86.5 / +19.4	67.4 / 83.6 / +16.2	0
ChatGLM Agent	63.2 / 82.2 / +19.0	68.6 / 80.3 / +11.7	65.6 / 85.7 / +20.1	65.8 / 82.7 / +16.9	0
Qwen-3.5-Plus Deep Research	60.0 / 79.4 / +19.4	73.1 / 83.3 / +10.2	61.1 / 82.7 / +21.6	64.7 / 81.8 / +17.1	-1
Manus-1.6-Max Wide Research	55.4 / 76.4 / +21.0	72.6 / 83.5 / +10.9	64.2 / 86.0 / +21.8	64.0 / 82.0 / +18.0	+1
Doubao Deep Research	64.2 / 80.8 / +16.6	64.9 / 76.3 / +11.4	53.1 / 77.1 / +24.0	60.7 / 78.1 / +17.4	-1
Grok Deep Research	58.7 / 78.0 / +19.3	63.7 / 77.2 / +13.5	58.3 / 81.4 / +23.1	60.2 / 78.9 / +18.7	+1
Prompt Sensitivity: MultiModal (30 Tasks) — Original / Modified / Δ					
MiroThinker-1.7	69.0 / 68.8 / -0.2	78.4 / 76.3 / -2.1	67.4 / 67.6 / +0.2	71.6 / 70.9 / -0.7	0
OpenAI Deep Research	66.7 / 66.3 / -0.4	77.0 / 74.6 / -2.4	66.8 / 65.0 / -1.8	70.2 / 68.6 / -1.6	0
Gemini-3.1-Pro Deep Research	66.4 / 66.3 / -0.1	73.7 / 71.0 / -2.7	64.1 / 63.1 / -1.0	68.1 / 66.8 / -1.3	0
Claude-Opus-4.6 Research	62.5 / 62.2 / -0.3	70.7 / 67.8 / -2.9	65.9 / 64.4 / -1.5	66.4 / 64.8 / -1.6	0
ChatGLM Agent	61.6 / 60.9 / -0.7	71.6 / 68.8 / -2.8	57.7 / 55.5 / -2.2	63.6 / 61.7 / -1.9	0
MiniMax-M2.5 Research	56.7 / 55.9 / -0.8	71.0 / 67.9 / -3.1	62.2 / 59.5 / -2.7	63.3 / 61.1 / -2.2	0
Manus-1.6-Max Wide Research	54.3 / 53.8 / -0.5	70.0 / 67.2 / -2.8	61.8 / 59.4 / -2.4	62.0 / 60.1 / -1.9	0
Grok Deep Research	56.3 / 56.2 / -0.1	71.5 / 68.2 / -3.3	53.9 / 51.0 / -2.9	60.5 / 58.5 / -2.0	0
Qwen-3.5-Plus Deep Research	44.6 / 44.6 / +0.0	69.9 / 66.9 / -3.0	53.8 / 51.3 / -2.5	56.1 / 54.3 / -1.8	0

Table 18: Comparison between human rankings and ApodexEval rankings. Human rankings are averaged across 10 annotators on a holistic end-to-end basis. ΔRank: positive values indicate higher human ranking than ApodexEval.

System	Human	ApodexEval	ΔRank
OpenAI Deep Research	1.7	1	=
MiroThinker-1.7	2.8	2	=
Gemini-3.1-Pro Deep Research	2.9	3	=
Claude-Opus-4.6 Research	3.1	4	=
MiniMax-M2.5 Research	5.4	5	=
ChatGLM Agent	5.5	6	=
Qwen-3.5-Plus Deep Research	7.3	8	↑ 1
Manus-1.6-Max Wide Research	8.0	7	↓ 1
Grok Deep Research	8.3	9	=

rankings, and the only mismatch occurs as a local swap of one position between Qwen-3.5-Plus and Manus-1.6-Max in the lower tier. No system shifts by more than one rank position. This high agreement (Spearman's $\rho = 0.98$) indicates that, despite operating along structured per-dimension criteria, ApodexEval's aggregate scoring closely tracks the holistic judgments that expert humans naturally produce when evaluating deep-research outputs end-to-end.

Summary. Across all four analyses, the relative ranking of systems is remarkably stable. Repeated runs produce Overall standard deviations of 0.3 to 0.7; switching from GPT to Gemini as judge shifts absolute Overall scores by 13 to 19 points but preserves per-dimension rankings closely and changes the Overall ranking only by adjacent swaps; modifying the judge prompt shifts Overall scores within 2.2 points with no rank changes; and expert human annotators converge on the same

ranking as ApodexEval with a maximum rank shift of only 1 position (Spearman's $\rho = 0.98$). These results indicate that the comparative conclusions drawn in the main text are robust to the evaluation configuration, and that absolute score differences between judge models are largely systematic offsets rather than disagreements about relative system quality; Appendix L.6 tests whether such offsets hide a self-family preference.

L Validation of the Evaluation Framework

The holistic ranking study of Appendix K.3 validates ApodexEval's aggregate ranking, and the robustness experiments establish stability. This appendix validates the components individually: the factuality verifier against adjudicated human labels (§L.1), the synthesis and process judges against per-dimension human preferences (§L.2), the adaptive rubric against a static one (§L.3), the process layer's sensitivity to documented reasoning (§L.4) and to trace exposure (§L.5), and the judges against a different model family (§L.6). §L.7 reports full confidence intervals.

L.1 Factuality Verifier Accuracy and Scoring Rules

Adjudication protocol. We sampled 270 claims stratified across the three verdict classes that carry the scoring (120 `RIGHT`, 90 `WRONG`, 60 `UNKNOWN`), oversampling the rarer classes so that per-class recall is estimable, and spread across evaluated systems. Three annotators labeled each claim independently, blind to the verifier's verdict, with the verifier's retrieved evidence available. Disagreements were resolved by written adjudication rather than majority vote, because a human majority is not a clean gold standard here: Appendix M includes a case in which the human annotation was overturned in the verifier's favor.

Table 19: Factuality verifier accuracy against adjudicated human labels (270 claims).

Verdict class	n	Precision	Recall
`RIGHT`	120	93.3%	84.3%
`WRONG`	90	82.2%	78.7%
`UNKNOWN`	60	51.7%	75.4%

Weighted to the corpus verdict distribution, in which these three classes cover 99.3% of all claims, verifier accuracy is 84.7%. The two classes that carry the headline ratio are precise, while `UNKNOWN` is mixed: 48.3% of sampled `UNKNOWN` verdicts could be resolved to a definite verdict by an annotator holding the same evidence, so nearly half of `UNKNOWN` reflects claims the verifier did not reach rather than claims that cannot be verified. Correcting each system's score for the error rates measured in the adjudication leaves the factuality ranking essentially unchanged (Spearman's $\rho = 0.988$), so verification error does not drive the cross-system comparisons.

Scoring rules. Table 20 reports, for every system and setting, the right ratio used in the main results, the ratio with `UNKNOWN` excluded, and the pooled `WRONG`, `CONFLICT`, and `UNKNOWN` rates. `CONFLICT` is zero on Text-Only, where there are no attachments for web evidence to conflict with. The two ratios rank systems similarly (Spearman's $\rho = 0.80$ on the task-weighted overall): the middle of the ordering moves, but the leading and trailing groups do not, and the comparisons drawn in §4.3 hold under both.

Abstention and the spread of the right ratio. The pooled right ratio (all of a system's claims pooled) equals 100 minus the `WRONG`, `UNKNOWN`, and `CONFLICT` rates, so its cross-system variance decomposes exactly into their contributions. On Text-Only, where `CONFLICT` is identically zero, differences in `WRONG` rates alone account for 23% of the variance, differences in `UNKNOWN` rates alone for 35%, and their positive covariance for the remaining 42%: systems that make more wrong claims also tend to leave more claims unverifiable. Consistently, excluding `UNKNOWN` narrows the cross-system range from 63.7–83.3 to 82.1–94.6. The MultiModal setting differs: its ratios span only 69.9–78.4 across nine systems, and `UNKNOWN` there is both rarer (14.5% vs. 18.8% on average) and far more uniform across systems, plausibly because the attachments give the verifier an additional evidence source that is always at hand. Abstention-driven spread is therefore chiefly a Text-Only phenomenon.

Table 20: Factuality under both scoring rules, with separate verdict rates (%). Ratio is the per-task average right ratio of Table 2; Ratio$_{\neg U}$ excludes UNKNOWN; rates are pooled over claims.

System	Ratio	Ratio$_{\neg U}$	WRONG	CONFLICT	UNKNOWN
Text-Only (70 tasks)					
Doubao Deep Research	64.9	83.1	12.9	–	23.0
Kimi-K2.5 Deep Research	65.4	85.2	10.7	–	22.6
Grok Deep Research	63.7	82.1	12.3	–	23.4
Qwen-3.5-Plus Deep Research	73.1	87.5	10.5	–	16.3
Manus-1.6-Max Wide Research	72.6	88.5	7.3	–	17.5
ChatGLM Agent	68.6	85.0	10.3	–	17.3
MiniMax-M2.5 Research	71.8	87.4	9.2	–	17.4
Claude-Opus-4.6 Research	69.8	90.2	8.3	–	22.3
Gemini-3.1-Pro Deep Research	71.3	88.1	9.3	–	19.0
MiroThinker-1.7	79.4	94.6	4.3	–	16.0
OpenAI Deep Research	83.3	94.5	4.2	–	12.4
MultiModal (30 tasks)					
Grok Deep Research	71.5	83.1	10.0	3.6	15.7
Qwen-3.5-Plus Deep Research	69.9	79.9	12.5	2.4	12.7
Manus-1.6-Max Wide Research	70.0	81.6	8.7	3.4	14.4
ChatGLM Agent	71.6	83.0	10.0	3.2	14.9
MiniMax-M2.5 Research	71.0	82.4	10.5	3.4	14.5
Claude-Opus-4.6 Research	70.7	87.0	6.3	3.3	18.2
Gemini-3.1-Pro Deep Research	73.7	86.3	7.7	4.6	14.7
MiroThinker-1.7	78.4	89.8	6.0	3.7	13.8
OpenAI Deep Research	77.0	89.2	7.4	2.7	11.6

L.2 Per-Dimension Pairwise Human Study

Protocol. The holistic study cannot validate individual dimensions, so we ran a pairwise study that asks about one dimension at a time. We constructed 150 same-query pairs of outputs from different systems, 30 for each of five dimensions: overall synthesis, Task-Specific criteria, intrinsic process quality, Process→Report alignment, and Report→Process alignment. Pairs were stratified by the judge's score gap. Annotators saw the two outputs side by side, blind to all scores and with anonymized, per-pair labels, and chose the better output on the named dimension; for synthesis, they were told to set factual correctness aside, since factuality is validated claim by claim (§L.1) and "which report is more factual" is not a readable pairwise judgment. Three annotators judged each pair, with written adjudication of disagreements.

Results. On pairs the judge scores as near-ties, humans themselves split almost evenly (54.9% direction agreement on synthesis), so the informative stratum is the pairs humans can consistently tell apart. Table 21 reports, per dimension, inter-annotator agreement and the rate at which the judge's ordering matches the human majority on that stratum. All five dimensions fall in the same band, including both alignment directions on which the traceability finding of §4.4 rests.

Table 21: Per-dimension agreement between the judges and human majorities (30 pairs per dimension).

Judge	Dimension	Fleiss' κ	Agreement, clear pairs
Synthesis	Overall synthesis	0.811	85.4%
Synthesis	Task-Specific criteria	0.741	83.7%
Process	Intrinsic process quality	0.709	85.9%
Process	Alignment, Process→Report	0.628	80.0%
Process	Alignment, Report→Process	0.710	75.7%

L.3 Static-Rubric Ablation

Setup. To isolate the effect of the adaptation, we re-scored the same Text-Only reports under a fully static rubric: the 20 generic criteria of Table 22, five per fixed dimension, shared by every query and weighted equally, with no Task-Specific dimensions. Each criterion was scored with the synthesis judge of the main results (GPT-5.1, temperature 0.1, the standard scoring prompt, full reports), and a report's score is the mean of its 20 criterion scores. Scores for the adaptive configuration are taken from the main run, so the two arms were scored at different times with the same judge and protocol. As a same-batch control, a third arm with 20 per-query generated criteria and no fixed backbone, scored together with the static arm, ranks the 11 systems almost exactly as the adaptive configuration does (Spearman's $\rho = 0.99$). The Text-Only split keeps the rubric as the only variable; on the MultiModal split, attachment visibility would be a second one.

Table 22: The 20 generic criteria of the static rubric, shared by every query and weighted equally.

Dimension	Criteria
Coverage	(1) Covers all sub-questions and angles required by the task; (2) Treats the central issues in depth rather than at surface level; (3) Supports claims with concrete evidence, data, or examples; (4) Stays on topic without irrelevant filler; (5) Draws on a diverse set of sources rather than a single origin
Insight	(6) Provides insight that goes beyond aggregating retrieved information; (7) Explains causes and mechanisms, not just phenomena; (8) Cross-references sources and resolves or flags conflicts; (9) Reasons soundly with no internal contradictions or unsupported leaps; (10) Offers non-obvious judgments of practical value
Instruction-following	(11) Satisfies every explicit requirement stated in the task; (12) Delivers the requested output form; (13) Respects the scope and boundary constraints of the task; (14) Complies with language, length, and level-of-detail requirements; (15) Directly answers the user's core question
Clarity	(16) Organizes content in a clear, navigable structure; (17) Maintains coherent flow between sections and paragraphs; (18) Uses fluent, precise language with appropriate terminology; (19) Uses tables, lists, or other formatting devices effectively; (20) Is concise and free of redundancy

Table 23: Static versus adaptive rubric on the same reports and judge, 11 systems, Text-Only (judge's 0–10 scale).

	Static rubric	ApodexEval (adaptive)
Mean synthesis score	7.66	6.61
Mean absolute score change vs. adaptive	1.05	—
Score spread across systems (max − min)	1.04	2.03
Criterion scores at 9 or above	11.4%	—
Agreement with human majorities, clear pairs	67.3%	85.4%

Results. Turning the adaptation off changes the measurement materially. Scores shift by about a point on average, and individual systems land in visibly different places: the static rubric ranks Claude-Opus-4.6 2nd where the adaptive rubric ranks it 5th, and Kimi-K2.5 3rd where the adaptive rubric ranks it 1st. The systems are also squeezed together: the spread across systems nearly halves, and 11.4% of the static rubric's criterion scores sit at 9 or above, a generic checklist scoring most competent reports alike. Against human judgment, the adaptive rubric agrees with the human majority on 85.4% of clearly separated synthesis pairs, the static one on 67.3%: where humans can reliably separate two reports, the adaptive rubric agrees with them and the static one substantially less so.

L.4 Process Ablation: Stripping Marked Reasoning

The three evaluation layers are parallel measurements of the same rollout rather than pipeline stages, so removing a layer discards a measurement rather than degrading the others; component ablations are therefore within layers. For the process layer, a clean ablation is harder, because a vendor's trace may omit steps that cannot be observed. The one component that can be removed operationally is reasoning text that the trace format explicitly marks as such (think blocks or equivalent structured fields); no

semantic classifier or manual judgment decides what counts as reasoning, and the experiment makes no claim about reasoning a system does not expose. We strip those segments and re-score, with the judge unchanged.

Table 24: Change in process score after stripping explicitly marked reasoning segments.

Traces	Δ Text-Only	Δ MultiModal
No marked reasoning to strip (OpenAI DR, Doubao, Manus, Qwen)	-0.02 to $+0.06$	-0.08 to -0.02
Marked reasoning stripped (Claude, MiroThinker-1.7, Kimi)	-0.40 to -0.94	-1.41 to -1.58

For systems whose traces contain no marked segments the input is unchanged, so their near-zero differences provide an empirical reference for rerun variation; it is therefore unlikely that the substantially larger decreases for reasoning-dense traces are explained by judge stochasticity alone. The ablation supports a narrow conclusion: the process evaluator assigns measurable weight to documented reasoning. It also identifies a scope limitation: cross-system process comparisons depend partly on what each system makes observable. The alignment metrics are less sensitive to this specific removal, because they compare report claims with extracted actions and findings rather than with reasoning text, but they are not invariant to overall trace exposure (§L.5). Doubao and Kimi-K2.5 have no MultiModal results, so the MultiModal ranges exclude them.

L.5 Trace-Exposure Control

Vendors expose very different amounts of their research process, so the low Report→Process traceability of §4.4 could partly measure how much a system reveals. As a descriptive diagnostic rather than a formal statistical control, we grouped systems into three tiers by observable trace exposure (trace length and the concreteness of recorded steps) and checked whether the finding depends on the tier (Table 25).

Table 25: Process–report alignment by trace-exposure tier (task-weighted over settings).

Tier	Systems	n	P→R	R→P	Gap
Full	OpenAI DR, Gemini-3.1-Pro, MiroThinker-1.7	3	79.4	55.8	23.6
Partial	Manus-1.6-Max, Kimi-K2.5, Claude-Opus-4.6, ChatGLM	4	75.5	49.2	26.3
Minimal	MiniMax-M2.5, Qwen-3.5-Plus, Doubao, Grok	4	69.3	42.8	26.5

The asymmetry does not depend on exposure. The P→R minus R→P gap stays within a three-point band across tiers, and within the Full tier alone it averages 23.6 points with a 95% CI of $[22.6, 24.6]$ (tasks resampled as clusters shared across the three systems, stratified by setting; 10,000 resamples), and the gap is positive in 99.7% of the individual system–task units. What does track exposure are the levels: from the most- to the least-exposed tier, P→R falls from 79.4 to 69.3 and R→P from 55.8 to 42.8; cross-system comparisons of R→P levels therefore carry an exposure caveat (Appendix A).

L.6 Judge-Family Analysis

All three primary judges are GPT-family models, and one evaluated system, OpenAI Deep Research, is GPT-family. The cross-judge check of Appendix K.2 shows largely preserved ordering, which is consistent with a judge inflating one family's absolute scores, so we test absolute scores directly.

Diagnostic on Table 17. Let $d(s)$ be a system's score change when the judge is swapped from GPT to Gemini. The GPT-family candidate has the smallest d, which on its face fits self-family inflation, but it is also the top-scoring system under the GPT judge, and the Gemini judge compresses all systems into a narrow high band, so a small rise is what compression alone predicts. Conditioning on the baseline, we fit d against the GPT-judge score on the nine systems belonging to neither judge family and read the family candidates' out-of-sample residuals (Table 26). Most of the raw gap is explained by compression: the remaining shortfalls are 0.8 to 1.4 times the neutral systems' residual spread on the three dimensions and stable under leave-one-out, and 1.8 times on Overall; the Gemini-family candidate's residuals are of similar size with mixed signs. With one candidate per family we treat this as a supporting diagnostic, not proof.

Table 26: How much less the GPT-family candidate rises than the nine neutral systems when the judge is swapped from GPT to Gemini (points).

	Overall	Synthesis	Factuality	Process
Raw gap	4.0	2.3	5.3	4.4
Conditioned on baseline score (out-of-sample)	1.8	0.9	1.0	1.7

Direct test with a third judge family. We re-scored every system with Claude-Opus-4.8 as the judge for all three layers on both settings, keeping the prompts and applying the same aggregation to both judges (two departures are noted below and in the caption of Table 27); Table 27 gives every system's scores under both judges. If the GPT judge had inflated its candidate's absolute scores by some margin, removing the GPT judge should make that candidate fall by roughly that margin more than the field. Table 28 reports, per panel, the candidate's score change minus the field's average change. It is positive in three of six panels; where it is negative it is small, at most 0.25 points on the 0–10 scales and 0.8 percentage points on factuality. The Claude judge assigns lower absolute scores and largely preserves the ordering of report and process (Spearman's ρ of 0.84 to 0.92), and less so for factuality (0.75 on Text-Only and 0.43 on the nine-system MultiModal panel). As a sensitivity check, the same quantity computed for the Claude-family candidate under the Claude judge is the largest of any system in five of six panels ($+0.56$ to $+1.09$ on the 0–10 scales and $+4.4$ points on Text-Only factuality), so the procedure surfaces family-level effects when they exist; the GPT-family candidate shows no effect of comparable size. We find no consistent evidence of GPT-family score inflation, although this does not establish that the judges are unbiased; we therefore keep the GPT judges as primary. On the MultiModal report panel, rubrics were regenerated with the standard prompt rather than the cached key-facts variant, a small departure from the main-results protocol.

Table 27: Every system under the Claude-family judge (C) and the primary judges (P), per layer and setting. Report and process are on the judge's 0–10 scale and factuality is the per-task average right ratio (%). Process here is the mean of the eight process sub-scores, the aggregation used for both judges in this comparison, so it differs from the Process column of Table 1. Tasks that could not be scored under the Claude judge are excluded, so some systems have fewer than 70 or 30 scored tasks.

	Report, TO		Report, MM		Process, TO		Process, MM		Factuality, TO		Factuality, MM	
System	C	P	C	P	C	P	C	P	C	P	C	P
Doubao Deep Research	5.60	6.42	–	–	3.35	5.33	–	–	58.9	64.9	–	–
Kimi-K2.5 Deep Research	6.66	7.57	–	–	6.43	6.46	–	–	65.2	65.4	–	–
Grok Deep Research	4.84	5.87	5.52	5.63	5.03	5.77	4.84	5.24	59.0	63.7	63.7	71.5
Qwen-3.5-Plus Deep Research	4.62	6.00	4.04	4.46	5.88	6.34	4.78	5.42	64.8	73.1	60.5	69.9
Manus-1.6-Max Wide Research	5.12	5.54	5.23	5.43	5.22	6.34	5.73	6.03	67.8	72.6	70.7	70.0
ChatGLM Agent	5.69	6.32	5.37	6.16	6.04	6.52	4.97	5.63	63.9	68.6	64.3	71.6
MiniMax-M2.5 Research	5.00	6.33	5.21	5.67	5.95	6.69	5.73	6.07	68.7	71.8	65.0	71.0
Claude-Opus-4.6 Research	6.47	6.73	6.51	6.25	7.21	6.57	7.17	6.57	70.1	69.8	67.6	70.7
Gemini-3.1-Pro Deep Research	6.42	7.12	6.39	6.64	6.72	6.74	6.06	6.36	69.3	71.3	62.8	73.7
MiroThinker-1.7	6.42	7.43	6.54	6.90	7.00	7.20	7.05	6.62	72.1	79.4	75.3	78.4
OpenAI Deep Research	6.26	7.38	6.30	6.67	7.49	7.28	6.89	6.64	78.4	83.3	74.0	77.0

Table 28: Claude-family judge versus the primary judges. Positive differences mean the GPT-family candidate fell less than the field average (report and process on the 0–10 scale, factuality in percentage points).

Panel	**Ranking ρ**	**GPT-family Δ minus field Δ**
Report, Text-Only	0.836	-0.25
Report, MultiModal	0.850	-0.08
Process, Text-Only	0.891	$+0.66$
Process, MultiModal	0.917	$+0.40$
Factuality, Text-Only	0.745	-0.76
Factuality, MultiModal	0.433	$+2.62$

L.7 Uncertainty

All intervals are 95% task-level bootstrap intervals with 10,000 resamples, in which a task is resampled as a cluster shared by all systems and all three layers, stratified by setting; with no resampling, the procedure reproduces every cell of Table 1. Half-widths are 1.0 to 3.6 points on Text-Only cells and on the task-weighted Overall, but 2.3 to 8.7 points on the 30-task MultiModal cells. The process–outcome correlation of §4.4 is Spearman's $\rho = 0.88$ with CI $[0.73, 0.94]$ and leave-one-system-out range $[0.84, 0.93]$; the Pearson coefficient is $r = 0.83$ with CI $[0.73, 0.89]$ and range $[0.78, 0.91]$. The two axes differ: the bootstrap redraws tasks with all eleven systems present, and leave-one-out removes a system. Table 29 repeats Table 1 with the half-width of every interval.

Table 29: Table 1 with 95% task-bootstrap confidence intervals. Grey $\pm$ values give the half-width of each interval; adjacent systems within these margins are not reliably separated.

Model	**Text-Only**				**MultiModal**				**Overall**
	Synthesis	Factuality	Process	Overall	Synthesis	Factuality	Process	Overall	
Doubao Deep Research	64.2$\pm$1.7	64.9$\pm$3.3	53.1$\pm$2.8	60.7$\pm$1.7	–	–	–	–	–
Kimi-K2.5 Deep Research	**75.7**$\pm$1.7	65.4$\pm$3.0	64.2$\pm$1.1	68.4$\pm$1.2	–	–	–	–	–
Grok Deep Research	58.7$\pm$2.1	63.7$\pm$3.2	58.3$\pm$2.3	60.2$\pm$1.8	56.3$\pm$3.9	71.5$\pm$5.6	53.9$\pm$4.9	60.5$\pm$3.3	60.3$\pm$1.6
Qwen-3.5-Plus Deep Research	60.0$\pm$1.9	73.1$\pm$2.5	61.1$\pm$1.7	64.7$\pm$1.6	44.6$\pm$7.4	69.9$\pm$7.1	53.8$\pm$3.9	56.1$\pm$4.0	62.1$\pm$1.6
Manus-1.6-Max Wide Research	55.4$\pm$2.7	72.6$\pm$3.5	64.2$\pm$2.9	64.0$\pm$2.0	54.3$\pm$5.2	70.0$\pm$8.7	61.8$\pm$4.2	62.0$\pm$4.6	63.4$\pm$2.0
ChatGLM Agent	63.2$\pm$2.4	68.6$\pm$3.6	65.6$\pm$1.4	65.8$\pm$2.0	61.6$\pm$4.3	71.6$\pm$5.8	57.7$\pm$4.0	63.6$\pm$3.1	65.1$\pm$1.7
MiniMax-M2.5 Research	63.3$\pm$1.8	71.8$\pm$2.7	67.1$\pm$1.3	67.4$\pm$1.5	56.7$\pm$4.5	71.0$\pm$6.0	62.2$\pm$3.2	63.3$\pm$3.4	66.2$\pm$1.5
Claude-Opus-4.6 Research	67.3$\pm$2.8	69.8$\pm$2.7	66.0$\pm$2.0	67.7$\pm$1.6	62.5$\pm$5.2	70.7$\pm$6.4	65.9$\pm$2.8	66.4$\pm$3.0	67.3$\pm$1.5
Gemini-3.1-Pro Deep Research	71.2$\pm$1.7	71.3$\pm$2.0	67.1$\pm$1.3	69.9$\pm$1.2	66.4$\pm$3.6	73.7$\pm$3.9	64.1$\pm$2.5	68.1$\pm$2.4	69.3$\pm$1.1
MiroThinker-1.7	74.3$\pm$1.3	79.4$\pm$1.9	72.7$\pm$1.1	75.5$\pm$1.0	**69.0**$\pm$2.9	**78.4**$\pm$4.6	**67.4**$\pm$2.3	**71.6**$\pm$2.3	74.3$\pm$1.0
OpenAI Deep Research	73.8$\pm$1.7	**83.3**$\pm$2.4	**73.1**$\pm$1.3	**76.7**$\pm$1.5	66.7$\pm$5.3	77.0$\pm$3.8	66.8$\pm$4.2	70.2$\pm$3.8	**74.8**$\pm$1.6

M Case Study

M.1 Synthesis Evaluation

We present two representative case studies to illustrate how the adaptive synthesis quality evaluation operates on concrete tasks, including the generated dimensions, criteria, and scoring. Both cases are multimodal tasks with attachment-augmented queries.

Case 1: Analyzing the 50 Fastest-Growing Software Vendors

Task. The user provides a screenshot of the BREX BENCHMARK "50 fastest-growing software vendors of 2025" ranking (an image listing vendor names and ranks from #01 Cursor to #50 Wispr) and asks the system to extract all company names, research each company's product, business model, culture, and competitive advantages, and produce an integrated report.

Extracted Key Facts. The framework extracts the following factual anchors from the attachment:

- Top-10 roster: 01 Cursor, 02 OpenRouter, 03 Kling AI, 04 Retell AI, 05 Perplexity, 06 Windsurf, 07 FireCrawl, 08 Clay, 09 Replit, 10 Exa.
- The graphic contains *only* vendor names and ranks; no growth rates, categories, business models, or financial metrics are shown.
- Multiple vendors include ".ai" in their names, but this alone does not establish AI-native status.

Generated Evaluation Dimensions and Weights. Beyond the four fixed dimensions, three task-specific Grounding & Expertise dimensions are generated. The assigned weights are:

Dimension	Weight
Coverage	0.27
Insight	0.30
Instruction-following	0.14
Clarity	0.09
Attachment-Grounded Scope & Methodology Integrity	0.08
Evidence-Linked Business Model & AI-Native Typology	0.07
Investment-Grade Profitability & Go-to-Market Evaluation	0.05

The first grounding dimension checks whether the report uses the exact top-20 roster as its sampling frame and avoids inferring attributes from the graphic. The second checks whether business-model labels are backed by external evidence rather than assumed. The third checks whether profitability claims are sourced from credible references.

Cross-System Scoring. We compare MiroThinker-1.7 (overall: 8.5) against ChatGLM Agent (overall: 4.5).

Coverage — "Complete per-company classifications for the top-20."
MiroThinker-1.7 provides a per-company table classifying all 20 vendors by business model, customer segment, and technology route. The evaluator notes: "*The report provides a table that lists all top-20 companies with columns for business model, core customer group, and technology route, which it uses to indicate AI-native status.*" Score: **9.4**.
ChatGLM Agent does not produce a per-company classification table. The evaluator notes: "*The report does not provide a per-company matrix or table that classifies each of the top-20 by business model, primary customer segment, and AI-native vs. traditional. It relies on anecdotes and placeholders.*" Score: **1.0**.

Grounding: Evidence-Linked Business Model & Typology — "Correct top-20 sampling frame and coverage."
MiroThinker-1.7 explicitly lists all 20 companies with correct names and ranks matching the attachment. The evaluator notes: "*The report clearly works off the correct Brex 'top 20' roster. It explicitly lists all 20 companies in Table 1 with correct names and ranks matching the key facts.*" Score: **9.8**.
ChatGLM Agent introduces companies outside the top-20 and conflates them into its analysis. The evaluator notes: "The report repeatedly goes beyond the top-20 scope, using ElevenLabs—

not in the roster—as a core example for business model and profitability conclusions." Score: **2.5**.

Grounding: Scope & Methodology Integrity — "Acknowledges graphic content limitations and avoids inference."
ChatGLM Agent fabricates growth rates from the graphic (e.g., "Cursor 1000%, Kling AI 1900%"), when the attachment contains no growth data at all. The evaluator notes: "*The attachment shows only names and ranks. The report neither acknowledges this limitation nor avoids inference; it asserts growth rates, profitability, and ARR figures not present in the graphic.*" Score: **0.5**.

Takeaway. The dynamically generated grounding dimension catches a critical failure mode that a fixed rubric would not detect: fabricating growth rates and financials from a graphic that contains only names and ranks. The key-facts extraction step transforms the abstract instruction into precise checkpoints (e.g., "does the report use the exact top-20 roster?"), enabling fine-grained discrimination.

Case 2: Veterinary Nutrition Planning for a Senior Cat

Task. A user provides medical records and 10 product photographs (canned wet foods from brands including Schesir After Dark, RAWZ, AIXIA, Unicharm, Zealandia, and Miaw Miaw) for a 12-year-old cat with chronic pancreatitis and mild CKD. The system must compare the nutritional profiles and produce a prioritized feeding recommendation.

Extracted Key Facts. The framework extracts product-level facts from each photograph. Crucially, most product images do *not* display nutritional analysis panels:
- AIXIA Kenko-can 11+: phosphorus 0.08%, sodium 0.10% (visible on label); **no fat content shown**.
- Schesir After Dark Chicken 80 g: labeled "Complete," "Grain Free"; **no fat or phosphorus visible**.
- RAWZ Shredded Chicken & Pumpkin ∼3 oz: **no nutrient panel visible**.
- Unicharm Silver Spoon 13+ (multiple tuna variants, 70 g each): **no nutrient values visible**.

Generated Evaluation Dimensions and Weights. Three task-specific dimensions are generated:

Dimension	Weight
Coverage	0.28
Insight	0.26
Instruction-following	0.16
Clarity	0.10
SKU-Verified Nutrient Provenance & Regional Fidelity	0.07
Clinically-Normalized Nutrition Modeling & Suitability Ranking	0.09
Uncertainty-Aware Data Gap Governance	0.04

The first grounding dimension checks whether nutrient values are matched to the exact SKU and regional variant (e.g., Japan-market vs. US-market formulations). The second checks whether nutrients are converted to comparable clinical metrics for pancreatitis and CKD. The third checks whether the report explicitly acknowledges missing data.

Cross-System Scoring. We compare Gemini-3.1-Pro (overall: 7.6) against MiniMax-M2.5 (overall: 4.0).

Insight — "Trade-off reasoning and decision logic."
Gemini-3.1-Pro constructs a two-dimensional fat–phosphorus framework and proposes mixed-feeding strategies. The evaluator notes: "*Demonstrates nuanced balancing with a 2D fat–phosphorus framework and proposes tactical mixed feeding to reconcile low-P vs low-fat trade-offs. It classifies products by which axis they satisfy.*" Score: **8.9**.
MiniMax-M2.5 assigns subjective star ratings without explicit trade-off analysis. The evaluator notes: "*Trade-offs are only superficially addressed. The report assigns star ratings but does not explicitly weigh key compromises.*" Score: **4.6**.

Grounding: Uncertainty-Aware Data Gap Governance — "Explicitly acknowledges missing fat and phosphorus for most SKUs and avoids inventing values."
Gemini-3.1-Pro partially acknowledges missing panels and marks some values as unknown, but still uses retail aggregator data without confirming regional SKU matches. Score: **3.2**.
MiniMax-M2.5 asserts specific nutrient values (e.g., "Silver Spoon fat: 0.3%," "RAWZ fat: 1.5%") for products whose labels show *no nutritional data at all*, without citing any source. The evaluator notes: "*Attachments show no fat/phosphorus panels for Schesir After Dark, RAWZ, Unicharm Silver Spoon 13+, SEEDS Golden Cat, Zealandia Wallaby. The report nonetheless provides specific values without acknowledging data gaps or citing external sources.*" Score: **0.8**.

Grounding: SKU-Verified Nutrient Provenance — "Authoritative, region-matched nutrient sourcing."
Both systems struggle with regional fidelity. Gemini-3.1-Pro sources from retail aggregators rather than manufacturer datasheets, and sometimes conflates Japan-market and US-market variants. Score: **3.4**. MiniMax-M2.5 provides no manufacturer URLs or datasheets for any value. Score: **0.5**.

Takeaway. The *Uncertainty-Aware Data Gap Governance* dimension, generated because the key-facts extraction identified that most product images lack nutrient panels, catches a failure mode where systems fabricate precise-looking numbers from non-existent label data. A fixed rubric evaluating only "nutritional accuracy" would miss this distinction, as it cannot know *a priori* which data points are available in the attachments.

Overall observation. The two cases illustrate how adaptive evaluation captures failure modes invisible to fixed rubrics. In Case 1, the key-facts extraction reveals that the attachment contains only names and ranks, prompting a grounding dimension that penalizes fabricated growth rates. In Case 2, the extraction identifies pervasive missing nutrient panels, prompting an uncertainty-governance dimension that penalizes invented values. In both cases, the dynamically generated criteria provide task-specific discrimination, while the shared fixed dimensions maintain cross-task comparability.

M.2 Factuality Evaluation

We present several representative case studies to illustrate typical examples of verifying a statement using the proposed agentic evaluation framework. We also include two cases where the judgments are incorrect; Appendix L.1 quantifies how often such errors occur.

Correct Case 1

Extracted Statement
In the second half of 2014, global oversupply driven by U.S. shale production growth and slowing demand caused oil prices to plummet from approximately $110 to below $40 per barrel.

Agentic Verification: Wrong.
Reasoning: While the qualitative explanation (U.S. shale-driven oversupply combined with weakening demand) is well supported by EIA and World Bank analyses, the specific numeric claim that prices fell "below $40" during H2 2014 is incorrect. EIA monthly data show Brent crude averaging approximately $106.77 in July 2014 and $62.34 in December 2014. WTI ended December 2014 at approximately $59.29. Prices did not fall below $40 per barrel until early 2016, more than a year after the period referenced in the statement.

Evidence

- *EIA (Brent)*: "EIA monthly table for 2014: Brent July 2014 $\approx$ $106.77; Brent December 2014 $\approx$ $62.34."
- *EIA (WTI)*: "WTI December 2014 $\approx$ $59.29 (not below $40 at year-end 2014)."

Human Annotation
Right

Final Judge
The model judgment is correct. Although the qualitative explanation is accurate, the claim that prices fell "below $40" in H2 2014 is false. EIA data show end-2014 prices near $62 (Brent) and $59 (WTI), with sub-$40 levels only reached in early 2016. This case illustrates a common hallucination pattern: correct directional reasoning but incorrect magnitude or timing.

Correct Case 2

Statement
In 2019, Huawei reported revenue of CNY 858,833M, operating profit of CNY 77,835M, and an operating margin of 9.1%.

Agentic Verification: Right.
Reasoning: The system extracted all three figures from the five-year financial highlights table in the user-uploaded screenshot, and then independently retrieved the same table from Huawei's 2019 Annual Report. All three values (revenue, operating profit, and the stated margin) are consistent across both sources. An internal arithmetic cross-check further confirms the margin: $77{,}835 \div 858{,}833 \approx 9.07\%$, which rounds to the reported 9.1%.

Evidence

- *Uploaded screenshot (image)*: "Revenue row: ... 858,833 ...; Operating profit row: ... 77,835 ...; Operating margin row: ... 9.1% ..."
- *Huawei 2019 Annual Report (PDF)*: "Five-Year Financial Highlights (2019 column): Revenue (CNY Million) 858,833; Operating profit 77,835; Operating margin 9.1%."

Human Annotation
Right

Final Judge
Both the model and the human annotator correctly judge the statement as accurate. The three figures are directly verifiable from Huawei's publicly available 2019 Annual Report, and the margin is arithmetically consistent with the revenue and profit values: $77,835 \div 858,833 =$

$9.07\% \approx 9.1\%$. This case illustrates a distinctive multimodal verification workflow: the system first extracts numeric data from an embedded image (a screenshot of the financial highlights table), then independently confirms the same figures via an authoritative external document, and finally validates internal arithmetic consistency, a three-layer cross-check that is not possible in text-only settings.

Correct Case 3

Statement
The LiuGong 856HE electric wheel loader is equipped with a CATL LFP battery rated at 350 kWh / 580 V, with a charging time of under 75 minutes.

Agentic Verification: Conflict.
Reasoning: The attached technical review explicitly reports a battery capacity of **350 kWh** and a rated voltage of **580 V**, with a charging time under **75 minutes**. However, LiuGong's official product page lists the 856HE with an LFP battery rated at **432 kWh**, resulting in an 82 kWh (23%) discrepancy. Since both sources are credible yet inconsistent, and the difference cannot be explained by rounding, the claim cannot be definitively verified without further clarification (e.g., model variants or specification updates).

Evidence

- *Attachment — technical review PDF (*Electric Wheel Loader Overview.pdf*, p. 8)*: "the LiuGong 856HE EWL … has a rated energy capacity of **350 kWh** and a rated voltage of 580 V, with a charging time of under 75 min."
- *Web — LiuGong North America product page (856HE)*: "Battery Type: LFP. Rated Storage Energy: **432 kWh**. Operating Weight: 46,297 lb (21,000 kg)."

Human Annotation
Conflict

Final Judge
The model judgment is correct. The discrepancy between 350 kWh and 432 kWh indicates conflicting specifications across credible sources, likely due to different model variants or updates. This case highlights a common multimodal conflict scenario: consistent evidence within each source but irreconcilable differences across sources.

Incorrect Case 1

Statement
Star Raiders was released in 1979.

Agentic Verification: Wrong.
Reasoning: The system retrieved the Wikipedia article for *Star Raiders* and identified the phrase "published in 1980 by Atari, Inc.", concluding that the correct release year is 1980 and that 1979 likely refers only to development or internal demonstrations.

Evidence

- *Wikipedia — Star Raiders*: "Star Raiders is a space combat simulator video game created by Doug Neubauer and published in 1980 by Atari, Inc. Originally released for the Atari 400/800 computers…" [excerpt truncated]

Human Annotation
Right

Final Judge
The agentic verdict is incorrect. While the retrieved evidence emphasizes the 1980 publication date, *Star Raiders* was already available in 1979 as a pack-in title for the Atari 400/800 launch. Therefore, the statement is factually valid under a broader definition of release. This case highlights a temporal ambiguity: different sources refer to different release events, and the agent fails to reconcile these interpretations.

Incorrect Case 2

Statement
Revenue from products built on generative AI models grew over 200% year-over-year.

Agentic Verification
Result: Right.
Reasoning: The system retrieved the CEO's remarks from both the uploaded earnings document and the company's official blog, which report the figure as "nearly 400% year-over-year." Since 400% exceeds 200%, the system interpreted the statement "over 200%" as a conservative but numerically valid description and returned a Right verdict.

Evidence

- *Attachment — Alphabet Q4 2025 Earnings Release (PDF, p. 1)*: "In Q4, revenue from products built on our generative AI models grew **nearly 400%** year-over-year."
- *Web — Alphabet CEO Blog*: "nearly 400% year-over-year." (`https://blog.google/company-news/inside-google/message-ceo/alphabet-earnings-q4-2025/`)

Human Annotation
Wrong

Final Judge
The model judgment is incorrect. Although "nearly 400%" logically satisfies the lower-bound claim "over 200%," the statement materially understates the reported magnitude and thus misrepresents the underlying fact. This case highlights a common failure mode: treating logical entailment as sufficient for factual correctness, particularly for vague lower-bound expressions (e.g., "over X%"), leading to false positives.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The contributions stated in the Abstract and §1, namely a three-layer evaluation framework (adaptive synthesis, agentic factuality, process-centric), a refreshable 100-task dual-path benchmark, and an evaluation across 11 deep research systems, are supported in §3 (methodology), §2 (construction), and §4 (experiments), with per-component validation against human judgment in Appendix L. Findings are stated as orderings with uncertainty rather than as point magnitudes; the process–outcome relation is reported as a limited-sample rank correlation with its confidence interval, not as a predictive claim. Aspirational goals such as resolving `CONFLICT` cases are flagged as future work in Appendix A.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Appendix A discusses benchmark scale and the 11-system scope (with wider intervals on the 30-task MultiModal split), judge-dependence of absolute scores, the conflation of unverifiable and wrong claims in the right ratio and the reliability of live web sources, the dependence of process evaluation on observable traces, the scope of the adaptive-rubric validation, judge-family overlap, single-product provenance and path-specific refreshability of the user-derived queries, and the absence of a difficulty-tier agreement statistic. The *Intended Use* paragraph of Appendix E delimits the benchmark's applicable scope.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This work is a dataset and evaluation paper containing no formal theorems or proofs; the scoring formulas in §3.1 and the process-score equation in §3.3 are evaluation metric definitions rather than theoretical propositions to be proved.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The benchmark construction pipeline is described in §2, with detailed parameters (privacy filter, entity replacement, six rewrite strategies, three-stage filtering, sampling temperatures, and token budgets) in Appendix C–D. The evaluation protocol is in §4 and Appendix B: 11 evaluated systems accessed through their official online interfaces during a March 2026 collection window; GPT-5.1 (synthesis), GPT-5.2 (process), and GPT-5-mini (factuality) as judge models, pinned to the OpenAI provider snapshots active in March 2026; Gemini-2.5-Pro and Gemini-3-Flash (used for the cross-judge robustness experiment in Appendix K.2) similarly pinned to the Google provider snapshots active in March 2026; Claude-Opus-4.8 serves as the judge for the judge-family check (Appendix L.6). The full construction prompt set (P1, P2A–P2F, P3–P9) is reproduced verbatim in Appendix G, and the judge prompts for synthesis, factuality, and process evaluation are released in the repository. The benchmark and evaluation framework code are released at the URLs listed in Appendix E (*Release and Maintenance*). *Caveat*: factuality evaluation issues live web searches via the Serper API, and judge LLMs and the evaluated deep-research systems themselves are stochastic; bit-exact numerical reproduction is therefore not achievable. The robustness experiments in Appendix K characterize the resulting variance and show that the comparative ranking is largely preserved across re-runs, judge changes, and prompt changes.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The evaluation framework code is released under Apache-2.0 at `https://github.com/MiroMindAI/MiroEval`; author-created dataset components (schema, rewritten queries, curation metadata, annotations) are released under MIT, and third-party attachments retain their original rights (Appendix E, *Release and Maintenance*). The full construction prompt set is provided verbatim in Appendix G; reproduction commands, environment specifications, and the scripts that reproduce every table and interval in this paper are in the repository. The benchmark is designed for periodic refresh, with version tags `vYYYY.MM` and a changelog.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: §4 reports the experimental setup: the list of 11 evaluated deep research systems with their version identifiers, the per-dimension judge assignment (GPT-5.1 for synthesis, GPT-5.2 for process, GPT-5-mini for factuality), and a unified March 2026 data-collection window via official online interfaces. Appendix B reports report-length and collection statistics. Construction-stage sampling parameters (generation T=0.8, filtering T=0.3, 8192-token output cap) are listed in Appendix D. The work involves no model training, so no optimizer or learning-rate hyperparameters apply.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Every score in Table 1 has a 95% task-bootstrap confidence interval, reported in Table 29 (10,000 resamples, tasks resampled as clusters shared by all systems and layers, stratified by setting; Appendix L.7). The process–outcome association is reported as Spearman's $\rho = 0.88$ with CI $[0.73, 0.94]$ and leave-one-system-out range $[0.84, 0.93]$, and the trace-exposure control reports a CI for the within-tier asymmetry (Appendix L.5). Additional variability is characterized by intra-judge re-runs (a targeted three-system diagnostic), cross-judge and prompt-sensitivity experiments, a third-judge-family re-scoring, and human validation studies (Appendices K and L). Factors of variability, the resampling unit, and the interval construction are stated in the corresponding paragraphs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).

- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This work involves no local GPU training. Benchmark construction (internally-deployed LLM endpoints for privacy filtering, entity replacement, and user-derived rewriting; OpenRouter routing to `openai/gpt-5.2` for trend-grounded generation and three-stage filtering; Serper API retrieval) and system evaluation (11 deep research systems via official online interfaces; GPT-5.1 / GPT-5.2 / GPT-5-mini judges; Gemini-2.5-Pro and Gemini-3-Flash for the robustness experiments in Appendix K; Claude-Opus-4.8 for the judge-family check in Appendix L.6) are all API-based. Token usage, dollar costs, wall-clock time, Serper API call counts, robustness-experiment budgets, and annotator-hours are documented in the *Computational Cost* paragraph of Appendix E.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. The user-derived construction path follows strict privacy protocols: raw queries are processed only on access-controlled internal infrastructure; all LLM calls go through internally-deployed instances that do not transmit data externally; entries pass an automated rule-plus-model privacy filter followed by named-entity replacement (Appendix C, *Privacy-Preserving Processing*). The released benchmark contains no original queries or person-linkable metadata (Appendix E, *Privacy* paragraph).

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix E (*Intended Use* paragraph) discusses both directions of societal impact: positively, the benchmark supports more reliable AI use in high-stakes domains such as finance, healthcare, and legal analysis; on the negative side, three categories of misuse are explicitly excluded (use as direct training data to avoid evaluation leakage, short-form QA benchmarking, and safety / red-teaming evaluation). Appendix A further notes that process evaluation may not apply to fully closed-source systems, preventing the benchmark's conclusions from being misread as a universal verdict across all deployments.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: The released benchmark exposes no original user queries, attachments, or session metadata. The user-derived path applies two complementary safeguards: an automated privacy filter (rule-plus-model, removing personal communications, confidential documents, and private financial records) and named-entity replacement that substitutes institutions, organizations, and individuals with realistic same-type alternatives (Appendix C, *Privacy-Preserving Processing*; Appendix E, *Privacy* paragraph). Beyond entity-level anonymization, the rewrite stage also abstracts content-level identity markers: distinctive professional roles are generalized to broader practitioner descriptors, idiosyncratic methodological vocabulary is rephrased into domain-accessible language, and structured output specifications are converted into free-form research requests, while preserving the analytical scope (illustrated by the Strategy E worked example in Appendix C). ApodexEval's MIT (data) and Apache-2.0 (code) licenses permit broad reuse subject to notice preservation; the *Intended Use* paragraph (Appendix E) frames discouraged uses (training-data inclusion, short-form QA benchmarking, safety/red-teaming) as community norms enforced through attribution and contamination audits rather than contract.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: External services used during construction are credited and used under their published terms: Serper API for trend retrieval (Appendix D); OpenRouter as a routing layer to `openai/gpt-5.2` for trend-grounded generation and three-stage filtering (Appendix D), pinned to provider snapshots active in March 2026. Privacy filtering, entity replacement, and user-derived rewriting are performed on internally-deployed LLM endpoints (Appendix C) and do not transmit raw user data to external services. Judge models for evaluation (GPT-5.1 / GPT-5.2 / GPT-5-mini) and the alternative judges used in the robustness and judge-family experiments (Gemini-2.5-Pro, Gemini-3-Flash, Claude-Opus-4.8) are accessed in accordance with the OpenAI Terms of Use, Google AI Terms of Service, Anthropic Commercial Terms of Service, OpenRouter Terms of Service, and Serper Terms of Service (as of March 2026). The benchmark does not redistribute any external dataset; related benchmarks (e.g., BrowseComp, HLE, MMSearch) are cited only for comparison. We use three classes of third-party software assets in the released artifact: (i) the MiroFlow agent framework [35] (Apache-2.0), integrated into the factuality evaluator; (ii) the OpenAI, Anthropic, and Google AI vendor SDKs (each released under Apache-2.0 or MIT), used to access judge models under their respective Terms of Service; and (iii) the Serper API, used as a web-search service for benchmark construction and for evidence retrieval in factuality evaluation, and not redistributed (governed by Serper Terms of Service).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: ApodexEval: evaluation framework code under Apache-2.0 at `https://github.com/MiroMindAI/MiroEval`; author-created dataset components under MIT on Hugging Face (`https://huggingface.co/datasets/miromind-ai/MiroEval-data`), with a Croissant file including Responsible AI fields; third-party attachments are not relicensed and retain their original rights, with a notice-and-takedown route on the dataset card. The release includes a README, separate LICENSE files, and a maintenance plan of annual refresh releases, with version tags `vYYYY.MM` and a changelog; bug reports, refresh requests, and pull requests are accepted via the issue tracker (Appendix E). Documentation covers source labels and per-query metadata, the joint distribution over 12 user-facing domains and 10 task types with a crosswalk to the generation taxonomies (Table 11), the human verification protocol (Appendix E.2), the annotation protocols and source-data provenance (Appendix F), and the full construction prompt set (Appendix G).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Two kinds of contributors are involved, documented separately in Appendix F. Annotators, all salaried research staff at the authors' institution paid above the statutory minimum wage with no piece-rate or inclusion-contingent compensation, took part in four studies: benchmark verification (three annotators, all 100 queries; size-weighted acceptance rate 92.6%), a holistic ranking study (ten annotators; Spearman's $\rho = 0.98$ vs. ApodexEval), verifier adjudication (270 claims), and a per-dimension pairwise study (150 pairs); their briefing and instructions are given in Appendix F.1. Internal staff whose testing work-product informed the user-derived path were not recruited as research participants; their provenance, the notice in force at collection time, and their retrospective written confirmation are documented in Appendix F.2.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: We did not seek IRB review because the study did not recruit or interact with internal staff as research participants and does not study their attributes, behavior, or employment outcomes: the source queries were used as task content, and the released benchmark contains only transformed queries with no contributor identities, sessions, or contributor-level records (Appendix F.2). Annotators acted as research personnel in the studies of Appendix F.1, with minimal risk (only anonymized text is processed). For the user-derived queries, contributors were notified at collection time and every contributor has given retrospective written confirmation; an audit of all 65 released queries found no identifiable or re-linkable content.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are core components of both the benchmark construction and the evaluation framework, and are explicitly declared in the paper. (i) Construction: privacy filtering, entity replacement, and user-derived rewriting are performed on internally-deployed LLM endpoints (raw user queries do not leave internal infrastructure); `openai/gpt-5.2` via OpenRouter is used for trend-grounded generation ($T{=}0.8$) and three-stage filtering ($T{=}0.3$), pinned to March 2026 active provider snapshots, with full prompts in Appendix G. (ii) Evaluation: GPT-5.1, GPT-5.2, and GPT-5-mini serve as judge models for synthesis, process, and factuality respectively (§4); Gemini-2.5-Pro and Gemini-3-Flash are used for cross-judge robustness verification (Appendix K), and Claude-Opus-4.8 for the judge-family check (Appendix L.6). This declaration covers the LLM components of the core method. LLMs were additionally used during paper preparation for routine writing assistance such as language polishing, sentence-level rewriting, and LaTeX formatting; per the NeurIPS LLM policy this writing-assistance use is not declarable as core methodology, but we disclose it here for transparency.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.