DeepResearchEval: An Automated Framework for Deep Research Task Construction and Agentic Evaluation

Anonymous ACL submission

Abstract

Deep research systems are widely used for multi-step web research, analysis, and cross-source synthesis, yet their evaluation remains challenging. Existing benchmarks often require annotation-intensive task construction, rely on static evaluation dimensions, or fail to reliably verify facts when citations are missing. To bridge these gaps, we introduce **DeepResearchEval**, an automated framework for deep research task construction and agentic evaluation. For task construction, we propose a persona-driven pipeline generating realistic, complex research tasks anchored in diverse user profiles, applying a two-stage filter to retain only tasks requiring multi-source evidence integration and external retrieval. For evaluation, we propose an agentic pipeline with two components: an *Adaptive Point-wise Quality Evaluation* that dynamically derives task-specific evaluation dimensions, criteria, and weights conditioned on each generated task, and an *Active Fact-Checking* that autonomously extracts and verifies report statements via web search, even when citations are missing. We further conduct human verification, confirming the reliability of generated tasks, quality evaluation, and factual evaluation.

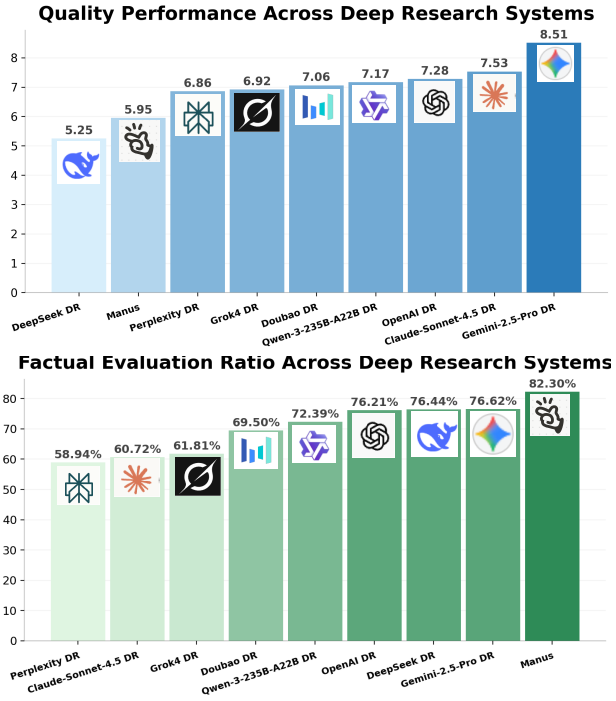

Figure 1: Overview of deep research systems' performance on our benchmark. The upper section reports quality evaluation results across deep research systems, with Gemini-2.5-Pro achieving the highest score (8.51/10). The bottom section reports factual correctness, where Manus achieves the highest ratio of correct statements (82.3%).

1 Introduction

The rapid advancement of Large Language Models (LLMs) has initiated a significant shift in AI capabilities, moving from passive text generation toward the development of agentic systems capable of tackling complex real-world tasks (Liu et al., 2025; Kimi et al., 2025; Zeng et al., 2025; Gemini, 2025; OpenAI, 2025a). Within this broader transition toward agentic intelligence, deep research systems have emerged as one of the representative paradigms (OpenAI, 2025b; Gemini, 2025; Team et al., 2025b; Perplexity, 2025). They autonomously conduct investigative processes that involve iterative web browsing, targeted information retrieval, cross-source verification, and multi-perspective synthesis. Through this structured workflow, deep research systems ultimately generate comprehensive, citation-grounded reports that traditionally require substantial human effort, e.g., "Please review AI compute investments in 2025 and looking

ahead to the trends of 2026".

With the transition of the technological paradigm, evaluating long reports generated by deep research systems poses a key challenge, as it differs substantially from conventional QA tasks. Several benchmarks have been proposed to assess long-form, research-style outputs, however most existing benchmarks suffer from three limitations: i) they rely on expert-driven task construction, which is annotation-intensive (Yao et al., 2025; Du et al., 2025; Abaskohi et al., 2025; Gou et al., 2025); ii) they employ static quality evaluation dimensions shared across tasks (Fan et al., 2025; Wang et al., 2025); and iii) they verify only citation-linked statements for factuality, leaving uncited factual claims unexamined (Du et al., 2025; Fan et al., 2025; Gou et al., 2025).

To bridge these gaps, we present **DeepResearchEval**, an automated framework for deep research task construction and agentic evaluation. To address the scarcity of high-quality deep research tasks and the high cost of expert annotation, we propose a persona-driven pipeline, which constructs tasks anchored in specific personas, ensuring realistic needs and high complexity. We then apply the Task Qualification Filter that assesses whether a task truly requires up-to-date evidence aggregation and multi-source investigation, and the Search Necessity Filter that discards simple questions solvable using an LLM's internal parametric knowledge. We retain 100 high-quality deep research tasks across multiple domains.

On the evaluation side, we develop an agentic evaluation pipeline with two key components. (i) *Adaptive Point-wise Quality Evaluation* preserves a fixed set of general evaluation dimensions shared across all tasks, while additionally generating task-specific dimensions, criteria, and corresponding weights for each task, allowing fine-grained and interpretable scoring tailored to individual deep research tasks. (ii) *Active Fact-Checking* performs active verification: it extracts verifiable statements (e.g., numbers, events, dates, entities), retrieves external evidence, and assigns structured labels (`Right`/`Wrong`/`Unknown`), thereby enabling factual verification of both cited and uncited claims. While factual correctness is an inherently core component of overall report quality, we treat it as a standalone evaluation module due to its distinct verification process, which requires external evidence retrieval, and its critical role in ensuring the reliability of deep research reports.

We apply our framework to evaluate leading deep research systems, spanning proprietary general-purpose systems (e.g., Gemini Deep Research (Gemini, 2025), OpenAI Deep Research (OpenAI, 2025b)), and agentic generalists with deep research capabilities (e.g., Manus (Manus, 2025)). We evaluate **900 deep research reports**, comprising 100 tasks per system across 9 deep research systems. As shown in Figure 1, our comprehensive evaluation reveals clear strengths and weaknesses across different dimensions of deep research capability. Gemini Deep Research achieve the strongest performance in report quality evaluation, reflecting superior coverage and insight. Manus and Gemini Deep Research attain the highest scores in factual evaluation, showing stronger robustness against hallucinations and citation errors during complex multi-source report synthesis. We also observe a consistent gap between general and task-specific evaluation: task-specific scores are systematically lower than those on fixed general dimensions. This gap indicates that current systems often fail to meet task-specific criteria, motivating task-adaptive evaluation beyond fixed general dimensions. We further conduct human verification of our framework: over 80% of generated tasks are retained by a majority of experts. The agreement between human rankings and our quality evaluation (Kendall's $\tau = 0.94$) confirms evaluation consistency, while factual evaluation achieves an F1 score of 82.5% against gold judgments.

2 Related Works

Deep research systems are a specialized class of agents designed for complex, multi-stage investigative tasks (OpenAI, 2025b; Gemini, 2025; xAI, 2025; Doubao, 2025; Perplexity, 2025; Manus, 2025; Anthropic, 2025). Unlike conventional QA systems, they autonomously plan long-horizon workflows, navigate heterogeneous web sources, and synthesize information into structured, citation-grounded reports. Existing systems broadly fall into proprietary solutions (OpenAI, 2025b; Gemini, 2025; Perplexity, 2025) with limited transparency, and open-source efforts (Li et al., 2025b; Zheng et al., 2025; Qiao et al.,

Table 1: Comparative Analysis of Deep Research Benchmarks.

Benchmark	Automated Task Generation	Output Format	Reference-free Evaluation	Adaptive Evaluation Dimensions	Active Fact Verification
Mind2Web 2 (Gou et al., 2025)	×	Report	×	×	×
DeepResearch Bench (Du et al., 2025)	×	Report	×	×	×
ResearcherBench (Xu et al., 2025)	×	Report	✓	×	×
BrowseComp-Plus (Chen et al., 2025b)	×	Short-Form Answer	×	×	×
WideSearch (Wong et al., 2025)	×	Table-Style Answer	×	×	×
ReportBench (Li et al., 2025a)	✓	Report	×	×	×
DeepResearch Arena (Wan et al., 2025)	✓	Report	✓	×	×
DRBench (Abaskohi et al., 2025)	×	Report	×	×	×
LiveResearchBench (Wang et al., 2025)	×	Report	✓	×	×
Finder (Zhang et al., 2025)	×	Report	✓	×	×
DeepResearchEval (Ours)	✓	Report	✓	✓	✓

2025) emphasizing modularity and reproducibility.

The emergence of deep research systems has motivated a broad range of benchmarks targeting different agentic capabilities (Gou et al., 2025; Mialon et al., 2024; Phan et al., 2025; Du et al., 2025; Wong et al., 2025; Wan et al., 2025; Abaskohi et al., 2025; Zhang et al., 2025; Wang et al., 2025; Li et al., 2025a; Chen et al., 2025b; Wei et al., 2025; Xu et al., 2025; Lei et al., 2025; Luo et al., 2025; Yao et al., 2025; Han et al., 2025). Early benchmarks such as GAIA (Mialon et al., 2024) and Humanity's Last Exam (HLE) (Phan et al., 2025) focus on general reasoning and tool use, while others emphasize persistent web navigation and retrieval, including WideSearch (Wong et al., 2025) and BrowseComp variants (Wei et al., 2025; Chen et al., 2025b). More recent benchmarks move toward report-level evaluation, including DeepResearch Bench (Du et al., 2025), LiveResearchBench (Wang et al., 2025), and DRBench (Abaskohi et al., 2025); however, they remain annotation-intensive, rely on fixed task-agnostic evaluation dimensions, and often restrict factual verification to cited statements, leaving uncited claims unchecked. In contrast, **DeepResearchEval** introduces an automated framework for task construction and agentic evaluation. In Table 1, it uniquely combines automatic task generation, reference-free evaluation, adaptive task-specific quality dimensions, and active fact verification over both cited and uncited statements for deep research systems.

3 Task Construction

Existing task collection relies heavily on expert annotators, suffering from three limitations: (1) high-quality annotation is costly and time-consuming; (2) tasks are constrained by annotators' individual backgrounds and domain knowledge; and (3) task collection is static and difficult to update over time.

To address these limitations, we introduce an automated persona-driven deep research task collection pipeline that mirrors real-world production workflows. As shown in Figure 2, we generate diverse personas conditioned on specific domains, which serve as seeds to produce expertise-aligned tasks, followed by multiple quality filtering stages to ensure a high-quality final set of deep research tasks. Implementation details and prompts are provided in Appendix H.

3.1 Construction Pipeline

Persona Synthesis. To ensure our evaluation covers a diverse spectrum of real-world information needs, we draw upon the domain taxonomy defined in (Wettig et al., 2025) and curate ten representative categories specifically suitable for deep research tasks, forming a domain set $\mathcal{D}$ that encompasses: *Transportation*, *Politics*, *Finance & Business*, *History*, *Software Development*, *Industrial*, *Sports & Fitness*, *Health*, *Science & Technology*, and *Education & Jobs*. For each domain $d \in \mathcal{D}$, we prompt an LLM to generate personas that are closely related to the domain while exhibiting diverse backgrounds. Each persona p is specified by attributes including *affiliation*, *role*, *background*, *name*, and *subdomain*. We generate five personas per domain, resulting in a total set P of 50 personas.

Task Construction. For each persona $p \in P$, we prompt an LLM to generate candidate deep research tasks conditioned on the persona's background. To ensure high task complexity, we enforce a generation

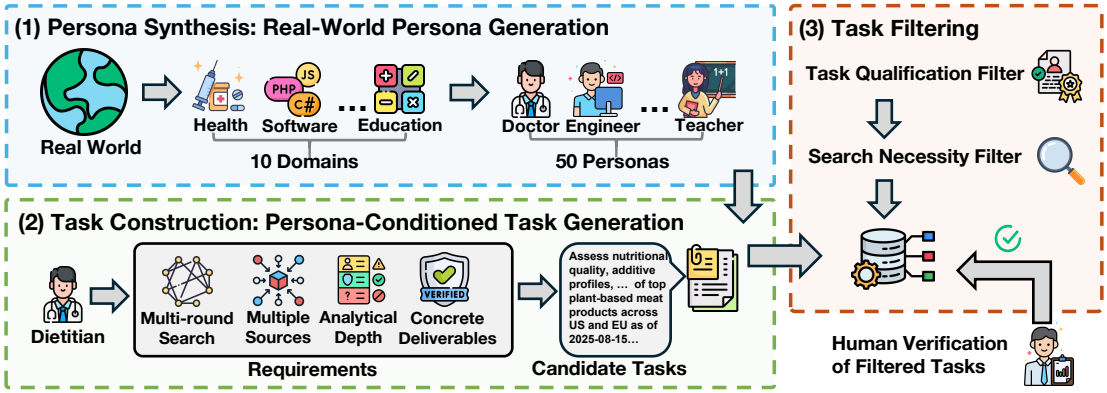

Figure 2: The proposed three-stage pipeline for constructing persona-driven deep research tasks.

Table 2: Distribution of expert approval counts for the 155 retained tasks.

Approval Count	0–1	2	3	4	5	6	7
Proportion	0%	4%	15%	15%	26%	30%	9%

schema requiring: (i) multi-round web searches; (ii) integration of evidence from diverse sources (e.g., papers, reports, and forums); (iii) sufficient analytical depth covering recent developments, data analysis, trend assessment, and comparative analysis; and (iv) concrete deliverables with explicit time constraints and 10–50 word descriptions. We generate four tasks per persona, yielding a total set of 200 candidates.

Task Filtering. To further ensure the quality of our benchmark, we employ a two-stage filtering pipeline for the candidate deep research tasks.

Task Qualification Filter: This distinguishes deep research tasks from simple tasks. An LLM-based evaluator assesses candidates on four criteria: requirement for up-to-date knowledge, multi-source evidence integration, multi-layered in-depth investigation, and persona's background and expertise alignment. Only tasks with a confidence score > 0.7 are retained.

Search Necessity Filter: To exclude tasks solvable by internal knowledge, an LLM attempts each retained task t using only parametric knowledge (no external tools). A separate evaluator assesses this non-search baseline across dimensions like accuracy, depth, and timeliness, professionalism, and structure. Tasks achieving high quality scores without search are filtered out, resulting in 155 retained tasks.

Human Verification of Retained Tasks. To validate the automated pipeline, we invited seven domain experts holding Ph.D. degrees to independently evaluate the 155 retained tasks against deep research criteria. The details are provided in Appendix **??**. Table 2 shows that 80% of tasks are deemed qualified by at least four experts. These results indicate the automated pipeline's effectiveness in reliably producing high-quality tasks.

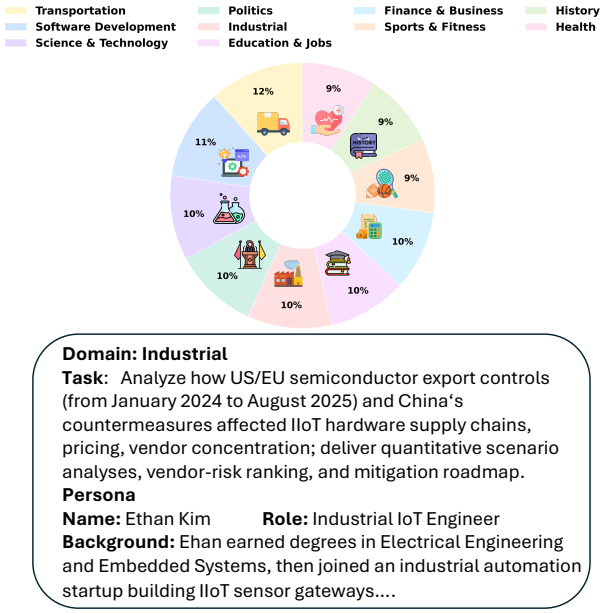

Figure 3: Domain Distribution and Example.

3.2 Benchmark Tasks

To mitigate evaluation costs, we curate 100 high-quality tasks based on human rankings (statistics and examples in Figure 3). This selection reflects practical constraints rather than pipeline deficiencies; as shown in Table 2, the majority of automatically gen-

erated tasks satisfy deep research criteria. Thus, our pipeline enables the continuous generation of fresh, high-quality tasks, allowing our framework to function as a dynamic "live" benchmark suitable for long-term monitoring (see Appendix D).

4 Agentic Evaluation

In this section, we present our agentic evaluation pipeline for assessing deep research reports. As illustrated in Figure 4, the pipeline comprises two components: (1) an adaptive point-wise quality evaluator that actively derives task-specific evaluation dimensions, criteria, and relative weights conditioned on the given research task, enabling fine-grained and task-aware scoring, and (2) an active fact checker that verifies both cited and uncited statements through external evidence retrieval.

4.1 Adaptive Point-wise Quality Evaluation

The deep research system produces long-form reports that vary substantially across tasks and domains, making it insufficient to evaluate all outputs using a fixed and uniform rubric. Prior work (Du et al., 2025; Fan et al., 2025) typically relies on a small set of pre-defined dimensions, which limits their ability to reflect task-specific evaluation aspects. Meanwhile, manually constructing customized rubrics for each task (Gou et al., 2025; Yao et al., 2025; Wang et al., 2025) is labor-intensive and does not scale.

To address these challenges, as illustrated in Figure 4, we propose an adaptive point-wise quality evaluation framework. For each task, the evaluator combines a fixed set of general dimensions with automatically generated task-specific dimensions, and assigns normalized weights to all dimensions to reflect their relative importance. Each dimension is further instantiated with weighted evaluation criteria, enabling fine-grained, criterion-level scoring. The final quality score is obtained by aggregating criterion scores within each dimension and then combining all dimensions according to their task-specific weights.

Agentic Quality Evaluation Framework. Formally, for a given task t, the evaluator first defines four general evaluation dimensions $\mathcal{D}_{\text{general}}$: *Coverage*, *Insight*, *Instruction-following*, and *Clarity*, capturing essential report qualities applicable across tasks. Dimension definitions are provided in Appendix G.1. Then the evaluator generates a set of task-specific dimensions $\mathcal{D}_{\text{task}}$, tailored to task t. For instance, *in a task that compares policies across different countries and requires the specification of quantitative indicators, the task-specific dimensions include Metric Utility and Comparative Synthesis, which are important evaluation metrics in political analysis but may not apply to more general tasks* (see Appendix J for descriptive statistics and more examples). The full dimension set is: $\mathcal{D} = \mathcal{D}_{\text{general}} \cup \mathcal{D}_{\text{task}}$, The evaluator assigns a normalized weight W_d to each dimension $d \in \mathcal{D}$ such that $\sum_{d\in\mathcal{D}} W_d = 1$, where higher weights indicate greater importance of a dimension for evaluating the task.

For each dimension d, a set of criteria $\{c\}$ are generated along with their corresponding weights $w_{d,c}$, where $\sum_c w_{d,c} = 1$. Given a report R, the evaluator scores each criterion on a scale of $[0, 10]$:

$$s_{d,c} = \text{LLM}_\theta(R, c)\,, \quad s_{d,c} \in [0, 10], \tag{1}$$

The final evaluation score for task t is computed as:

$$S_{\text{quality}} = \sum_{d\in\mathcal{D}} W_d \sum_c w_{d,c}\, s_{d,c}. \tag{2}$$

The evaluator generates a task-specific overall quality score, offering greater relevance to the evaluation context. Furthermore, it enables granular analysis of individual dimensions and criteria, ensuring a more comprehensive understanding of report quality. Details on prompts, LLM configurations, and full dimension set, dimension-level weights, criteria, and criterion-level weights can be found in Appendix G.1.

4.2 Active Fact-Checking

The adaptive point-wise quality evaluation provides fine-grained scoring over report quality dimensions. However, it does not explicitly evaluate factual correctness, which is particularly critical for deep research reports. Existing methods (Du et al., 2025; Fan et al., 2025) typically check if citations support the text, but this paradigm fails when: i) reports lack citations; ii) claims appear in uncited segments; and (iii) citation-based verification checks whether a cited source supports a claim, rather than its factual correctness. To address these issues, we propose an active

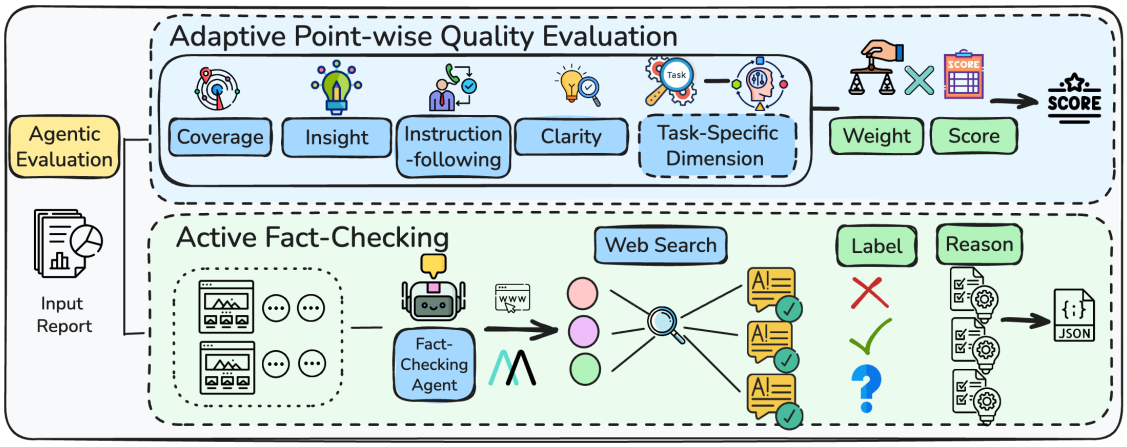

Figure 4: Overview of the proposed pipeline. (Top) Adaptive Point-wise Quality Evaluation augments $\mathcal{D}_{\text{general}}$ with task-specific $\mathcal{D}_{\text{task}}$. An LLM scores criteria $s_{d,c}$, aggregating them into S_{quality} via weights W_d and $w_{d,c}$. (Bottom) Active Fact-Checking extracts statements $\mathcal{S}_i$ from report segments $\{p_i\}$. An agent verifies claims using MCP-based retrieval, producing JSON labels (`Right`, `Wrong`, `Unknown`).

fact checking framework that actively retrieves and examines external evidence to assess the factual consistency of the entire report, as shown in Figure 4.

Agentic Fact-Checking Framework. Built upon *MiroFlow* (MiroMind, 2025), our agent iteratively invokes MCP tools to retrieve external evidence. Rather than relying solely on citations, it proactively identifies and verifies claims to ensure comprehensive statement-level checking.

Given a generated report that requires factual evaluation, our fact-checking agent follows a structured, multi-stage pipeline to support fine-grained, statement-level verification of long-form reports. To reduce the challenges associated with long-context processing in lengthy reports and to enable parallel verification across segments, the agent first segments R into a set of smaller parts $R \rightarrow \mathcal{P} = \{p_1, p_2, \ldots, p_N\}$, For each input part p_i, the agent extracts a set of statements $\mathcal{S}_i = \{s_{i1}, s_{i2}, \ldots\}$ involving verifiable entities such as *numbers, news, events, dates, locations, or people*.

For each statement $s \in \mathcal{S}_i$, the agent invokes a retrieval tool to search the web and collect relevant evidence $\mathcal{E}(s)$. Although verification is performed at the statement level, the agent holds the full segment context p_i as well as the associated deep research task, enabling context-aware and task-consistent judgments. Based on the consistency between s and the retrieved evidence, the agent assigns one of three labels: $y(s) \in \{\texttt{Right}, \texttt{Wrong}, \texttt{Unknown}\}$. `Right` denotes support, `Wrong` indicates contradiction, and `Unknown` marks insufficient evidence, explicitly distinguishing unverifiable claims from errors.

Results, including labels, evidence, and reasoning, are returned in JSON format. Implementation details and examples are in Appendix G.2 and K. Finally, the `Ratio` metric of factual evaluation is defined as the proportion of right statements over all statements: $\texttt{Ratio} = \frac{N_{\texttt{Right}}}{N_{\texttt{Statements}}}$.

5 Experiments

5.1 Experimental Setup

We evaluate 9 major commercial deep research systems, including OpenAI Deep Research (OpenAI, 2025b), Gemini-2.5-Pro Deep Research (Gemini, 2025), Grok4 Deep Research (xAI, 2025), Claude-Sonnet-4.5 Deep Research (Anthropic, 2025), Qwen3-235B-A22B Deep Research (Yang et al., 2025), DeepSeek Deep Research (Liu et al., 2024), Perplexity Deep Research (Perplexity, 2025), Doubao Deep Research (Doubao, 2025), and Manus (Manus, 2025). For **each DR system**, we collect **100 reports** by running the system on the deep research tasks constructed in our pipeline 3.1. Details about the collection of DR systems are provided in Appendix B.

Following Sec. 4, we utilize *Gemini-2.5-pro* (Gem-

Table 3: **Quality evaluation results across different deep research system.** Bold numbers indicate the best scores.

Model	Avg	Covera.	Insight	Instr.	Clarity	Task.
DeepSeek Deep Research	5.25	5.9	5.2	7.2	8.4	4.3
Manus	5.95	7.2	5.8	8.3	7.1	5.2
Perplexity Deep Research	6.86	8.2	6.6	9.3	8.6	5.9
Grok4 Deep Research	6.92	8.5	6.6	9.6	8.2	6.0
Doubao Deep Research	7.06	8.6	7.0	9.2	7.7	6.3
Qwen-3-235B-A22B Deep Research	7.17	8.0	7.9	8.7	8.3	6.6
OpenAI Deep Research	7.28	8.6	7.3	9.0	7.6	6.7
Claude-Sonnet-4.5 Deep Research	7.53	8.8	8.0	9.2	7.8	6.8
Gemini-2.5-Pro Deep Research	**8.51**	**9.2**	**9.0**	**9.7**	**9.1**	**8.0**

ini, 2025) for *Adaptive Point-Wise Quality Evaluation* to generate all adaptive components (dimensions, weights, criteria) and produce final scores. For active fact checking, we implement the agent on *MiroFlow* (MiroMind, 2025) using *GPT-5-mini* (OpenAI, 2025a) (default settings). The agent employs *Google Serper API* for retrieval, with a maximum of 30 agent turns. See Appendix G for details.

5.2 Main Results

Overall Quality Evaluation. Table 3 presents a point-wise quality evaluation across 9 representative deep research systems. We observe clear stratification: Gemini-2.5-Pro Deep Research achieves the highest average score (8.51) and leads across all dimensions, followed by Claude-Sonnet-4.5 Deep Research (7.53). This advantage is driven by *Coverage*, *Insight*, and *Instruction-following*, where these top systems exceed 8.5, indicating strong abilities in information gathering, synthesis, and complex instruction-following.

DeepSeek and Manus show moderate *Instruction-following* (7.2, 8.3) but lag in *Coverage* and *Insight*, resulting in lower overall scores (5.25, 5.95). Perplexity and Grok4 improve substantially in *Coverage* ($>$ 8.2) and *Instruction-following* (9.3, 9.6), reflecting stronger retrieval and planning. Doubao and Qwen-3-235B-A22B further enhance analytical depth, with Qwen achieving a high *Insight* score of 7.9. Finally, OpenAI and Claude-Sonnet-4.5 exhibit balanced performance across dimensions.

Notably, task-specific scores are consistently lower than general scores across all systems. This indicates that while systems excel at general synthesis, they often fail to optimize for task-specific criteria. This gap motivates our adaptive dimensions, which capture quality aspects missed by fixed rubrics. Ultimately, generating high-quality task-specific content remains a key challenge for current deep research systems.

Factual Evaluation. Table 4 presents factual evaluation results where our agent assesses statements per report. We report average checkable `Statements` and counts for `Right`, `Wrong`, and `Unknown` claims.

Top performers like Manus, Gemini-2.5-Pro, and DeepSeek achieve ratios $> 76\%$, indicating superior reliability. In contrast, Perplexity and Claude-Sonnet-4.5 exhibit lower ratios, implying more unverifiable or incorrect statements. Substantial variation exists in statement volume: Gemini-2.5-Pro and Doubao produce notably more claims (86.99, 80.75), yielding denser reports, whereas DeepSeek adopts a conservative strategy (25.08). These results suggest a potential trade-off between maintaining high factual accuracy and increasing the volume of reported statements.

Systems with higher `Ratio`, such as Manus and DeepSeek, exhibit consistently low `Wrong` counts (2.23, 1.81), indicating strong avoidance of false claims. Conversely, lower `Ratio` systems show higher `Unknown` values, implying claims are often unsupported rather than explicitly incorrect. Notably, `Wrong` statements are rare compared to `Unknown` across all systems, suggesting factual risks stem more from weakly grounded claims than outright errors.

5.3 Validation of Evaluation Methods

To validate the reliability of methods, we conduct extensive analyses.

Cross-judge Consistency of Quality Evaluation. To **mitigate self-preference bias** from Gemini-2.5-Pro, our primary judge, we additionally use GPT-5 (OpenAI, 2025a) as a secondary judge. Although

Table 4: **Factual evaluation results across different deep research system.** Bold numbers indicate the best values.

Model	Ratio	Statements	Right	Wrong	Unknown
Perplexity Deep Research	58.94%	61.34	36.16	9.08	16.10
Claude-Sonnet-4.5 Deep Research	60.72%	57.30	34.79	6.16	16.35
Grok4 Deep Research	61.81%	47.16	29.15	5.44	12.57
Doubao Deep Research	69.50%	80.75	56.12	7.43	17.20
Qwen-3-235B-A22B Deep Research	72.39%	37.45	27.11	3.36	6.34
OpenAI Deep Research	76.21%	45.98	35.04	2.72	8.22
DeepSeek Deep Research	76.44%	25.08	19.17	1.81	4.10
Gemini-2.5-Pro Deep Research	76.62%	**86.99**	66.65	4.16	16.18
Manus	**82.30%**	57.90	47.65	2.23	8.02

Table 5: Human preference study. Comparison between human **rankings** and DeepResearchEval **rankings**.

Model	Human	DeepResearchEval	ΔRank
DeepSeek Deep Research	8.50	9	=
Manus	7.40	8	=
Grok4 Deep Research	6.30	6	$\uparrow 1$
Perplexity Deep Research	6.20	7	$\downarrow 1$
Doubao Deep Research	5.50	5	=
Qwen-3-235B-A22B Deep Research	4.30	4	=
OpenAI Deep Research	3.70	3	=
Claude-Sonnet-4.5 Deep Research	1.60	2	=
Gemini-2.5-Pro Deep Research	1.50	1	=

Table 6: Quality evaluation across 3 independent runs.

Model	Score ($\mu \pm \sigma$)	Rank
DeepSeek Deep Research	5.24 ($\pm$0.02)	9.0
Manus	5.92 ($\pm$0.02)	8.0
Perplexity Deep Research	6.85 ($\pm$0.01)	7.0
Grok4 Deep Research	6.95 ($\pm$0.04)	6.0
Doubao Deep Research	7.08 ($\pm$0.02)	5.0
Qwen-3-235B-A22B Deep Research	7.21 ($\pm$0.06)	4.0
OpenAI Deep Research	7.30 ($\pm$0.08)	3.0
Claude-Sonnet-4.5 Deep Research	7.51 ($\pm$0.01)	2.0
Gemini-2.5-Pro Deep Research	8.52 ($\pm$0.03)	1.0

GPT-5 assigns stricter scores (Table 10), the resulting rankings remain highly consistent with Table 3: 7 of 9 models keep the same positions, with only a minor Doubao–Qwen swap ($|\Delta\text{Rank}| = 1$). GPT-5 also identifies the Task-Specific dimension as the lowest-scoring aspect, further highlighting its difficulty.

Human-Model Alignment of Quality Evaluation. To verify alignment with human preference, we conduct a human study where three experts independently rank all 9 deep-research systems on 10 randomly sampled benchmark tasks. The averaged human ranking strongly agrees with the Gemini ranking (Kendall's $\tau = 0.94$, Spearman's $\rho = 0.98$), confirming that our evaluation faithfully reflects human preferences.

Stochastic Stability of Quality Evaluation. We assess stochastic stability via three independent runs using Gemini-2.5-Pro. In Table 6, rankings remain unchanged with minimal score standard deviations, demonstrating the our stability against randomness.

Human-Model Alignment of Factual Evaluation. To validate our *active fact-checking*, four experts annotate 80 statements, treating both *Wrong* and *Unknown* as negative. The automated evaluation achieves 73% **agreement** with human annotations, approaching human-level performance. For the 20 inconsistent cases, re-annotation by a human expert assisted by GPT-5.2 shows that the automated judgment is correct in 70% of cases, mainly due to agentic verification. Correct and failed cases are provided in Appendices I.1 and I.2.

Gold Judgments of Factual Evaluation. We further establish gold judgments for all 80 statements to evaluate our factual evaluation in Table 11. The high precision indicates that statements labeled as correct are highly reliable. The overall F1 score **82.5%** demonstrates strong factual verification performance under the human-established gold standard.

6 Conclusions

We introduce an automated framework for DR task construction and agentic evaluation in report quality and factuality. Our persona-driven task construction enables generation of realistic, complex tasks without manual annotation. We propose an adaptive point-wise quality evaluation for quality assessment and an active fact-checking via external evidence retrieval. Experiments on 9 DR systems reveal substantial performance differences, demonstrating the effectiveness of our framework in evaluating generated long report.

Limitations

Despite its effectiveness, the proposed framework has several practical limitations. The current implementation is largely English-centric: although the persona-driven task construction and adaptive evaluation mechanisms are language-agnostic, the benchmark tasks, evidence sources, and reporting pipelines are grounded in English-speaking information ecosystems. As a result, performance in multilingual settings and the ability to synthesize evidence across diverse languages remain unexplored.

In addition, the agentic evaluation pipeline incurs substantial computational and financial costs. The framework relies on frequent interactions with frontier models, using Gemini-2.5-Pro for quality scoring and GPT-5-mini for factual verification, alongside extensive Google Serper API usage. While the fact checking agent's multi turn, tool intensive design enables high evaluation depth, it constrains scalability for large scale or real time deployment under limited resources.

References

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

Anthropic. 2025. Claude 4.5 model overview. Technical report, Anthropic.

Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, Chen Sun, Han Hou, Hui Yang, James Pan, Jianan Lou, Jiayi Mao, Jizheng Liu, Jinpeng Li, Kangyi Liu, and 14 others. 2025a. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations. *Preprint*, arXiv:2506.13651.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, and 1 others. 2025b. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. *arXiv preprint arXiv:2508.06600*.

Doubao. 2025. Doubao chat. https://www.doubao.com/chat/.

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

Tianyu Fan, Xinyao Niu, Yuxiang Zheng, Fengji Zhang, Chengen Huang, Bei Chen, Junyang Lin, and Chao Huang. 2025. Understanding deepresearch via reports. *arXiv preprint arXiv:2510.07861*.

Gemini. 2025. Gemini 3 pro model card.

Gemini. 2025. Gemini deep research. https://gemini.google/overview/deep-research/.

Google Gemini. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jiménez Gutiérrez, Yiheng Shu, and 1 others. 2025. Mind2web 2: Evaluating agentic search with agent-as-a-judge. *arXiv preprint arXiv:2506.21506*.

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

Kimi, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Fangyu Lei, Jinxiang Meng, Yiming Huang, Junjie Zhao, Yitong Zhang, Jianwen Luo, Xin Zou, Ruiyi Yang, Wenbo Shi, Yan Gao, Shizhu He, Zuo Wang, Qian Liu, Yang Wang, Ke Wang, Jun Zhao, and Kang Liu. 2025. Dacomp: Benchmarking data agents across the full data intelligence lifecycle. *Preprint*, arXiv:2512.04324.

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

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Haotian Luo, Huaisong Zhang, Xuelin Zhang, Haoyu Wang, Zeyu Qin, Wenjie Lu, Guozheng Ma, Haiying He, Yingsha Xie, Qiyang Zhou, Zixuan Hu, Hongze Mi, Yibo Wang, Naiqiang Tan, Hong Chen, Yi R. Fung, Chun Yuan, and Li Shen. 2025. Ultrahorizon: Benchmarking agent capabilities in ultra long-horizon scenarios. *Preprint*, arXiv:2509.21766.

Manus. 2025. Introducing manus: The general ai agent. https://manus.im/app.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2024. GAIA: a benchmark for general AI assistants. In *The Twelfth International Conference on Learning Representations*.

MiroMind. 2025. Miroflow: An open-source agentic framework for deep research. https://github.com/MiroMindAI/MiroFlow.

OpenAI. 2025a. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/.

OpenAI. 2025b. Introducing deep research. https://openai.com/index/introducing-deep-research/.

Perplexity. 2025. Introducing perplexity deep research. https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, and 1 others. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Zile Qiao, Guoxin Chen, Xuanzhong Chen, Donglei Yu, Wenbiao Yin, Xinyu Wang, Zhen Zhang, Baixuan Li, Huifeng Yin, Kuan Li, and 1 others. 2025. Webresearcher: Unleashing unbounded reasoning capability in long-horizon agents. *arXiv preprint arXiv:2509.13309*.

MiroMind Team, Song Bai, Lidong Bing, Carson Chen, Guanzheng Chen, Yuntao Chen, Zhe Chen, Ziyi Chen, Jifeng Dai, Xuan Dong, and 1 others. 2025a. Mirothinker: Pushing the performance boundaries of open-source research agents via model, context, and interactive scaling. *arXiv preprint arXiv:2511.11793*.

Tongyi DeepResearch Team, Baixuan Li, Bo Zhang, Dingchu Zhang, Fei Huang, Guangyu Li, Guoxin Chen, Huifeng Yin, Jialong Wu, Jingren Zhou, and 1 others. 2025b. Tongyi deepresearch technical report. *arXiv preprint arXiv:2510.24701*.

Haiyuan Wan, Chen Yang, Junchi Yu, Meiqi Tu, Jiaxuan Lu, Di Yu, Jianbao Cao, Ben Gao, Jiaqing Xie, Aoran Wang, and 1 others. 2025. Deepresearch arena: The first exam of llms' research abilities via seminar-grounded tasks. *arXiv preprint arXiv:2509.01396*.

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

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

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation. In *Forty-second International Conference on Machine Learning*.

Ryan Wong, Jiawei Wang, Junjie Zhao, Li Chen, Yan Gao, Long Zhang, Xuan Zhou, Zuo Wang, Kai Xiang, Ge Zhang, and 1 others. 2025. Widesearch: Benchmarking agentic broad info-seeking. *arXiv preprint arXiv:2508.07999*.

xAI. 2025. Grok deepsearch. https://x.ai/news/grok-3.

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang Yao, Yixu Wang, Yuxuan Zhang, Yi Lu, Tianle Gu, Lingyu Li, Dingyi Zhao, Keming Wu, Haozhe Wang, Ping Nie, and 1 others. 2025. A rigorous benchmark with multidimensional evaluation for deep research agents: From answers to reports. *arXiv preprint arXiv:2510.02190*.

Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

Dingling Zhang, He Zhu, Jincheng Ren, Kangqi Song, Xinran Zhou, Boyu Feng, Shudong Liu, Jiabin Luo,

Weihao Xie, Zhaohui Wang, and 1 others. 2025. How far are we from genuinely useful deep research agents? *arXiv preprint arXiv:2512.01948*.

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

A Usage of AI Assistant

We use ChatGPT solely for language refinement of the manuscripts text. All conceptual content, experimental design, analysis, and conclusions are developed entirely by the authors. We carefully review the AI-assisted edits to ensure that the meaning and technical accuracy of the original text are fully preserved.

B Deep Research Systems Details

Table 7 reports the main time windows during which reports from different deep research systems were collected. All dates correspond to the year 2025. Reports from the other 9 deep research systems were generated and downloaded from their official websites using automated tools.

Avg Length denotes the average length of valid Deep Research outputs produced by each deep research system across all evaluated tasks. Most Deep Research Agents produce responses exceeding ten thousand characters on average. In particular, Gemini-2.5-Pro, Doubao, and OpenAI Deep Research generate substantially longer outputs, with average lengths reaching several tens of thousands of characters. Qwen-3-235B-A22B Deep Research refers to the model selected within the official Qwen website when using the "Deep Research" function. At the time of our data collection, the strongest available model for this mode is Qwen-3-235B-A22B (Yang et al., 2025), and we report results based on reports generated under that configuration. DeepSeek DeepResearch refers to the configuration obtained through the official DeepSeek website by enabling both the "Deep Think" and "Search" options.

C Real-World Demand of Our Benchmark

We emphasize that our task generation is not unconstrained prompting, but a structured pipeline designed to approximate real-world deep research needs. The ten domains are derived from Organize the Web (Wettig et al., 2025), whose taxonomy is grounded in large-scale real-world web data and human-in-the-loop curation, ensuring empirically motivated and practically relevant domain coverage.

Our persona construction models concrete professional backgrounds to encourage realistic synthesis

Table 7: **Primary data collection periods of deep research systems (2025).**

Deep Research System	Data Collection Date Range	Avg Length
Claude-Sonnet-4.5 Deep Research	Aug 19 – Aug 28	26.3K
Doubao Deep Research	Aug 19 – Aug 26; Sep 1 – Sep 7	48.4K
Gemini-2.5-Pro Deep Research	Aug 19 – Aug 26; Sep 5 – Sep 6	51.8K
Perplexity Deep Research	Aug 22 – Aug 26	13.7K
OpenAI Deep Research	Aug 27 – Sep 8	41.3K
Grok4 Deep Research	Aug 28 – Sep 1	11.0K
Manus	Aug 28 – Sep 8	30.8K
Qwen3-235B-A22B Deep Research	Aug 29	29.8K
DeepSeek Deep Research	Nov 10	5.5K

tasks rather than generic prompts. Combined with the Task Qualification Filter, which removes shallow or non-expert-level requests, these design choices aim to ensure that the final tasks better reflect authentic deep research demand.

Our Human Verification can mitigate the concerns that the constructed tasks reflect real-world deep research demand. During actual evaluation, we require human experts to assess whether candidate tasks are qualified based on their own knowledge. This assessment considers multiple aspects, including quality, time constraints, and realism. Table 2 shows that 80% of tasks were deemed qualified by at least four experts, ensuring the realism of tasks.

Here we present two tasks from our benchmark, which cover real-world, timely demands such as AI-generated visuals and stablecoin regulatory debates: i) *Between January 2024 and August 2025, how have AI-generated visuals affected public trust in science communication across US, EU, and China? Provide case studies, policy responses, sentiment metrics, and recommendations.* ii) *How did stablecoin regulatory debates evolve across the US, EU, and major Asian jurisdictions from January 2024 to August 2025, and what are projected impacts on financial stability and crypto adoption through 2026? Deliver: timeline, risk analysis, policy recommendations.*

D Scalability of our Benchmark

The dataset size is limited to 100 tasks because running a single deep-research task requires substantial computational resources in terms of both time and cost. Moreover, several prior frontier benchmarks—such as XBench-DeepSearch (Chen et al., 2025a), Mind2Web2 (Gou et al., 2025), and DeepResearch Bench (Du et al., 2025) also contain roughly one hundred tasks, suggesting that this scale represents a practical trade-off between evaluation quality, cost, and stability.

To further demonstrate the scalability and "live" nature of our automated pipeline, we additionally generate a new 30-task dataset using the same fully automated framework. This set includes recently emerging topics, such as: *Deeply investigate updates to Claude Code, Cursor, Antigravity, and Codex in January 2026. Analyze the watershed characteristics of AI foundation companies pivoting from "coding assistance" to "autonomous agent" R&D paradigms*. We further evaluate several latest deep research systems (Team et al., 2025a) on this new task set. The results are shown in Table 8 and Table 9.

On the new 30 task "live" set, we observe consistent capability differences across systems. Gemini-3 DR achieves the highest quality score (8.05), while MiroThinker attains the strongest factual accuracy (72.83%), suggesting different trade-offs between synthesis and grounding. Lower performing systems (e.g., DeepSeek DR) remain clearly separated, indicating that the benchmark retains strong discriminative power on newly generated tasks.

More importantly, since both task construction and evaluation are fully automated, our framework can

Table 8: Quality Evaluation of New 30-task Dataset

Model	Avg	Coverage	Insight	Instr.	Clarity	Task
Gemini-3 DR	8.05	9.0	8.5	9.3	9.0	7.4
MiroThinker	7.71	8.3	7.8	9.2	9.1	7.1
OpenAI DR	7.13	7.8	7.2	8.6	8.2	6.5
Qwen-3-235B DR	7.05	7.2	7.8	8.7	8.8	6.3
Doubao DR	6.29	7.0	6.6	7.4	7.9	6.0
DeepSeek DR	2.56	2.3	2.1	2.9	7.0	2.2

Table 9: Factual Evaluation of New 30-task Dataset

Model	Statements	Right	Wrong	Unknown	Ratio
MiroThinker	1255	914	72	269	72.83%
OpenAI DR	974	698	56	220	71.66%
Qwen-3-235B DR	742	499	96	147	67.25%
DeepSeek DR	603	401	58	144	66.50%
Gemini-3 DR	2120	1325	246	549	62.50%
Doubao DR	1750	1039	192	519	59.37%

continuously generate new tasks and evaluate newly released models under a unified protocol. This demonstrates our scalability beyond the initial 100 task.

E Human Verification of Retained Tasks

Regarding the ranking criteria, since the seven experts each need to evaluate all 155 tasks, we only ask them to judge whether each task satisfies the deep research criteria based on their expertise. As shown in Table 2, the aggregation method and agreement are computed by simply counting how many experts consider each task qualified. The final result for each task is therefore the number of experts who approve it, referred to as the *approval count*.

F More Results

Table 10 presents the results of adaptive point-wise quality evaluation using GPT-5 judge.

Table 11 presents the performance of Active Fact-Checking under gold judgements.

G Evaluation Methods Details

G.1 Adaptive Point-wise Quality Evaluation

For the **adaptive point-wise quality evaluation**, we define four general evaluation dimensions. *Coverage*: Breadth, depth, and relevance of coverage. *Insight*: Depth, originality, logic, and value of analysis. *Instruction-following*: Accuracy in meeting all requirements and constraints. *Clarity*: Readability, fluency, structure, and ease of understanding. In addition, the framework automatically generates between one and three *task-specific* dimensions. For each dimension, we create between one and ten criteria, each scored on a $[0, 10]$ scale with two decimal places of precision.

For the base LLM, we employ **Gemini-2.5-Pro** to generate the dimensions, criteria, and weights, using a maximum of 8192 new tokens, a temperature of 0.1, and a random seed of 42. We additionally use **GPT-5** for scoring in Sec 5.3, with a maximum of 8192 new tokens and its default temperature setting.

Adaptive Point-wise Quality Evaluation Prompt. We present the prompts used for task-specific dimension generation, followed by the prompts for assigning weights to the four fixed dimensions and the addi-

Table 10: **Adaptive point-wise quality evaluation full results using GPT-5 judge.** The last column reports the absolute rank difference compared to Table 3.

Model	Avg	Covera.	Insight	Instr.	Clarity	Task.	|ΔRank|
DeepSeek Deep Research	2.72	3.5	3.5	5.0	4.0	1.8	0
Manus	3.60	5.1	4.3	6.4	5.0	2.3	0
Perplexity Deep Research	4.08	5.6	4.8	6.9	5.4	2.8	0
Grok4 Deep Research	4.18	5.9	4.8	7.5	5.3	2.8	0
Qwen-3-235B-A22B Deep Research	4.23	6.0	5.9	6.5	3.2	3.0	1
Doubao Deep Research	4.46	6.4	5.4	7.2	5.2	3.1	1
OpenAI Deep Research	4.63	6.6	5.9	7.3	5.0	3.2	0
Claude-Sonnet-4.5 Deep Research	4.73	6.6	6.0	7.0	4.8	3.4	0
Gemini-2.5-Pro Deep Research	**5.29**	**7.0**	**7.1**	**7.9**	**6.4**	**3.7**	0

Method	Precision	Recall	F1
Active Fact-Checking	97.1%	71.7%	82.5%

Table 11: Gold Judgements of the Active Fact-Checking.

tional task-specific dimensions, as well as the prompts for generating evaluation criteria and corresponding weights for each dimension. Finally, we provide the prompt used to perform scoring with an LLM-based judge. All prompts are designed to return outputs in a JSON format.

Point-wise Task-Specific Dimension Generation

```
<system_role>

You are an expert evaluator who designs **query-specific
meta-evaluation dimensions** for deep research reports. Your
 goal is to identify unique quality aspects that matter for
a given task, beyond the four standard meta-dimensions.
</system_role>

<user_prompt>
**Standard Meta-Dimensions** (already covered):
1. **Coverage**: Breadth, depth, and relevance of coverage
2. **Insight**: Depth, originality, logic, and value of
analysis
3. **Instruction Following**: Accuracy in meeting all
requirements
4. **Clarity**: Clarity, fluency, structure, and ease of
understanding

**Your Task**: For the research task below, generate **1-3
additional same-level meta-evaluation dimensions** that are:
- Highly specific to this query
- Distinct from the four standard meta-dimensions
- Crucial for assessing quality in this domain
- Actionable and measurable
- Serve as **upper-level meta-dimensions** that can be
further expanded into more detailed evaluation standards
- Do NOT include any factuality-related meta-dimensions,
since factual accuracy is handled by a separate evaluation
system

<research_task>
"{task_prompt}"
</research_task>

**Guidelines**:
1. Analyze the task carefully to understand its domain,
methodology, data needs, and unique challenges.
2. Identify domain-specific quality factors (e.g., for
finance: market timing; for science: experimental validity;
for policy: stakeholder impact).
3. Create meta-dimensions that are:
   - Unique to this query type (not generic)
   - Non-overlapping with the four standard meta-dimensions
   - Focused on specialized aspects relevant to this domain
4. For each meta-dimension, provide:
   - **Name** (1-3 words)
   - **Definition** (short explanation of what it measures
and why it matters)

**Output Format**:
Return only a JSON list of meta-dimensions, e.g.:

<json_output>
[
  {{
    "meta_dimension_name": "Xxx",
    "definition": "Clear, concise explanation"
  }},
  ...
]
</json_output>

</user_prompt>
```

Point-wise Weight Generation

```
<system_role>
You are a senior research evaluation expert. Your job is to
(1) consider both the four fixed meta-dimensions and the
provided query-specific meta-dimensions (each with a name
and definition), and (2) assign **dynamic, well-justified
weights** to all dimensions so that the total equals
**1.0**.
</system_role>

<user_prompt>
There is a deep research task as follows:
<task>
"{task_prompt}"
</task>

**Fixed Meta-Dimensions (always included):**
[
```

```
  {{
    "meta_dimension_name": "Coverage",
    "definition": "Breadth, depth, and relevance of coverage
."
  }},
  {{
    "meta_dimension_name": "Insight",
    "definition": "Depth, originality, logic, and value of
analysis."
  }},
  {{
    "meta_dimension_name": "Instruction Following",
    "definition": "Accuracy in meeting all requirements and
constraints."
  }},
  {{
    "meta_dimension_name": "Clarity",
    "definition": "Readability, fluency, structure, and ease
 of understanding."
  }}
]

**Provided Query-Specific Meta-Dimensions (each includes
name + definition)**
<additional_meta_dimensions_json>
{additional_dimensions_json}
</additional_meta_dimensions_json>

**Your Goals**
1. **Task-grounded analysis**: Carefully analyze the <task>
to identify goals, constraints, risks, and success criteria.
2. **Dynamic weighting**: Assign a weight (0-1) to **each**
dimension (fixed + provided).
   - The **sum across all dimensions must be exactly 1.0**.
   - Weights should reflect the **unique characteristics**
of the <task> (do not use fixed presets).
3. **Specific justification**: In <analysis>, explicitly
justify **each** weight by referencing the <task> and, for
the provided meta-dimensions, their **definitions**. Avoid
generic statements.

**Constraints & Notes**
- Do **not** introduce new dimensions. Only use the fixed
four plus the provided query-specific ones.
- Do **not** include any factuality-related dimensions (
factuality is evaluated elsewhere).
- Avoid overlap: if a provided dimension substantially
overlaps with a fixed one, explain how you differentiate it
in scope and why its weight is still necessary.
- If no additional dimensions are provided (empty list),
distribute weights among the four fixed dimensions only.

**Output Format (STRICT)**
First produce a concise yet concrete <analysis> that:
- Explains the task-grounded reasoning behind the overall
weighting strategy.
- Gives a 1-3 sentence justification **for each dimension**
tying the weight to the task and (for provided ones) their
definitions.

Then produce <json_output> containing only the final weights
, with keys exactly matching the dimension names:
- Fixed keys (always present): "comprehensiveness", "insight
", "instruction_following", "readability"
- For each provided meta-dimension, use its exact "
meta_dimension_name" as the key.

Example shape:
<analysis>
(Your reasoning here. One short paragraph about overall
trade-offs, then bullet points or short lines justifying
each dimension's weight.)
</analysis>

<json_output>
{{
  "coverage": 0.xx,
  "insight": 0.xx,
  "instruction_following": 0.xx,
  "clarity": 0.xx,
  "additional_dimension": 0.xx
}}
</json_output>

**Validation**
- Ensure all weights are in [0, 1].
- Ensure the sum across all keys equals **1.00** (allowing
up to +/-0.001 rounding; otherwise adjust and state the
adjustment briefly in <analysis>).
- Do not output anything other than <analysis> and <
json_output>.
</user_prompt>
```

Point-wise Criteria Generation

```
<system_role>
You are an expert evaluator of research reports. Your job is
 to break down an meta-evaluation dimension into clear,
specific, task-relevant criteria with explanations and
weights.
</system_role>

<user_prompt>
We evaluate a research report written for the task below
across {num_dimensions} meta evaluation dimensions:
{meta_dimensions}

<task>
"{task_prompt}"
</task>

<instruction>
Your goal: For the **{dimension_name}** dimension, generate
task-specific evaluation criteria.

Steps:
1. **Analyze Task**: Identify the essential areas and
coverage needed to satisfy "{dimension_name}".
2. **Formulate Criteria**: Write diverse, non-overlapping
criteria items.
3. **Explain Rationale**: Provide a short explanation (`
explanation`) for each criterion.
4. **Assign Weights**: Give each criterion a weight (`weight
`) so that the total = **1.0**. Adjust the last item if
needed.
5. **Focus**: Stay strictly within "{dimension_name}",
avoiding overlap with the other dimensions.

Output format:
1. First, provide `<analysis>` explaining your reasoning and
 weight allocation.
2. Then output `<json_output>` as a list of criteria in the
format:

<json_output>
[
  {{
    "criterion": "...",
    "explanation": "...",
    "weight": ...
  }},
  ...
]
</json_output>

Now begin for the task above and dimension = **{
dimension_name}**.
</user_prompt>
```

Point-wise Score Prompt

```
<system_role>
You are a strict, meticulous, and objective evaluator of
deep research reports.
You score the report on a **single evaluation dimension** at
 a time.
You must evaluate strictly according to the provided **
criteria under that dimension**.
Do **not** evaluate factual accuracy (handled by a separate
system).
</system_role>

<user_prompt>
**Task**
<task>
{task_prompt}
</task>
```

```
**Report to Evaluate**
<Report>
{report}
</Report>

**Evaluation Dimension and Criteria**
Below is the dimension to evaluate. It contains multiple
criteria, each with a short explanation.
<criteria_of_one_dimension_json>
{criteria_of_one_dimension_json}
</criteria_of_one_dimension_json>

**Scoring Rules**
- For **each criterion**, assign a **continuous score from 0
 to 10** (real number), and provide a concise justification
(`analysis`) grounded in the report content.
- **Revised Scale (more discriminative):**
  - **0-2 (Very Poor):** Severe deficiencies; almost
entirely fails the criterion.
  - **2-4 (Poor):** Major gaps or flaws; only marginally
addresses the criterion.
  - **4-6 (Fair):** Covers the basics but shallow,
inconsistent, or with notable weaknesses.
  - **6-7.5 (Good):** Solid attempt; generally clear and
adequate, but lacks depth or polish.
  - **7.5-9 (Very Good):** Strong and well-executed with
only minor issues; above average.
  - **9-10 (Excellent):** Outstanding; fully satisfies and
exceeds expectations in depth, clarity, and execution.
Reserved for exceptional cases.
- Scores should reflect only the current criterion, avoiding
 overlap with other dimensions.
- **Do not** judge factual correctness; only judge coverage,
 reasoning quality, clarity, structure, domain-
appropriateness, etc.
- Be conservative: most typical reports should fall in the
**4-8 range**. Scores above 9 should be rare.

**Output Format (STRICT)**
Output `<json_output>` as a list of criteria in a valid JSON
 object format:


<json_output>
[
    {{
        "criterion": "text of the criterion",
        "analysis": "your justification",
        "report_score_0_to_10": x.xx
    }},
    {{
        "criterion": "text of the criterion",
        "analysis": "your justification",
        "report_score_0_to_10": x.xx
    }},
    ...
]
</json_output>

**Validation**
- Use the exact criterion names as keys in the JSON.
- Each score must be a real number in [0,10], rounded to two
 decimals.
- Ensure the JSON is strictly valid and parseable (no
trailing commas, properly escaped characters).
</user_prompt>
```

G.2 Active Fact-checking

For the base LLM, we employ GPT-5-mini to segment the reports and perform fact-checking. The model is used with default temperature settings, a maximum of 128K new tokens, a 190K maximum context length, top-p = 1.0, and top-k = -1.

For retrieval, we use the Google Serper API to conduct searches and scrape website content. The number of returned results is set to 10, with a maximum of 5 retry attempts. For our agent workflow, we set the maximum number of turns to 30, with up to 10 tool calls allowed per turn. The agent interacts with the MCP server `tool-search`. The MCP tools available on the `tool-search` server mainly include `google_search()` for search, `scrape_website()` for retrieving webpage content, and `wiki_get_page_content()` for obtaining content from specific Wikipedia pages corresponding to target entities.

Active Fact-checking Prompt. We present the prompt used to segment a Deep Research report into multiple paragraphs, after which each paragraph is independently evaluated by a fact-checking agent. We further show the agent system prompt used in the MiroFlow framework, along with the task description prompt for fact-checking. Finally, a summary prompt aggregates the evaluations into a structured JSON-formatted result.

Factual Evaluation Report Segmentation

```
You are a text segmentation assistant. Your task is to
segment reports into logical parts:
- Avoid single-sentence parts.
- Do NOT alter the original content.
- Titles, headings, or section headers should be combined
with the following content as one part.
- If multiple sentences clearly belong to the same section,
keep them together in one part instead of splitting them.
- Output must be ONLY a valid JSON list, where each element
is one part.

Example output format:
[
  "This is the first part.",
  "This is the second part.",
  "This is the third part."
]

Please segment the following report into logical parts:

{Deep Research Report}
```

Agent System Prompt

```
In this environment you have access to a set of tools you
can use to answer the user's question.
Today is: {formatted_date}
Important Notes:
- Tool-use must be placed **at the end** of your response,
**top-level**, and not nested within other tags.
- Always adhere to this format for the tool use to ensure
proper parsing and execution.

String and scalar parameters should be specified as is,
while lists and objects should use JSON format. Note that
spaces for string values are not stripped. The output is not
 expected to be valid XML and is parsed with regular
expressions.

# General Objective

You accomplish a given task iteratively, breaking it down
```

```
into clear steps and working through them methodically.

## Task Strategy

1. Analyze the user's request and set clear, achievable sub-
goals. Prioritize these sub-goals in a logical order.
2. Start with a concise, numbered, step-by-step plan (e.g.,
1., 2., 3.) outlining how you will solve the task before
taking any action. Each sub-goal should correspond to a
distinct step in your task-solving process.
3. Work through these sub-goals sequentially. After each
step, the user may provide tool-use feedback, reflect on the
 results and revise your plan if needed. If you encounter
new information or challenges, adjust your approach
accordingly. Revisit previous steps to ensure earlier sub-
goals or clues have not been overlooked.
4. You have access to a wide range of powerful tools. Use
them strategically to accomplish each sub-goal.

## Tool-Use Guidelines

1. Each step must involve a single tool call, unless the
task is already solved.
2. Before each tool call:
- Briefly summarize and analyze what is currently known.
- Identify what is missing, uncertain, or unreliable.
- Be concise; do not repeat the same analysis across steps.
- Choose the most relevant tool for the current sub-goal,
and explain why this tool is necessary at this point.
- Verify whether all required parameters are either
explicitly provided or can be clearly and reasonably
inferred from context.
- Do not guess or use placeholder values for missing inputs.
- Skip optional parameters unless they are explicitly
specified.
3. All tool queries must include full, self-contained
context. Tools do not retain memory between calls. Include
all relevant information from earlier steps in each query.
4. Avoid broad, vague, or speculative queries. Every tool
call should aim to retrieve new, actionable information that
 clearly advances the task.
5. Even if a tool result does not directly answer the
question, extract and summarize any partial information,
patterns, constraints, or keywords that can help guide
future steps.

## Tool-Use Communication Rules

1. Do not include tool results in your response — the user
will provide them.
2. Do not present the final answer until the entire task is
complete.
3. Do not mention tool names.
4. Do not engage in unnecessary back-and-forth or end with
vague offers of help. Do not end your responses with
questions or generic prompts.
5. Do not use tools that do not exist.
6. Unless otherwise requested, respond in the same language
as the user's message.
7. If the task does not require tool use, answer the user
directly.

# Agent Specific Objective

You are a task-solving agent that uses tools step-by-step to
 answer the user's question. Your goal is to provide
complete, accurate and well-reasoned answers using
additional tools.

Before verifying all the statements, **always** use the tool
 to search for information of the statement.
```

Factual Evaluation

```
Your task is to perform factual verification of a [part]
from an LLM-generated report.
The report is produced in response to a [user query], and
both the [user query] and the [part] are provided above.

Task requirements:
1. Use retrieval tools to fact-check key statements in the [
part], collecting evidence from reliable sources whenever
possible.
2. You should focus on validating key statements, especially
 those involving numbers, news, events, dates, locations, or
 people that can be fact-checked with external sources.
3. Do not verify every sentence; focus on the core, publicly
 verifiable claims.
4. For each checked statement, clearly record:
   - verification: one of [Right, Wrong, Unknown] (choose
only one).
     * "Right": the statement matches reliable public
information.
     * "Wrong": the statement is contradicted by reliable
sources or a statement mixes accurate and inaccurate content
.
     * "Unknown": the statement lacks sufficient relevant
information, or contains some correct details but other
parts can't be verified.
   - evidence: a list of objects, each containing:
     * source: the URL of the supporting or contradicting
evidence
     * excerpt: the quoted text or summary from the source
that directly relates to the statement
   - reasoning: a clear explanation of the verification
reasoning and process, including how the evidence supports
the assigned verification result.
```

Factual Evaluation Summary

```
This is a direct instruction to you (the assistant), not the
 result of a tool call.

We are now ending this session, and your conversation
history will be deleted.
You must NOT initiate any further tool use. This is your
final opportunity to report
*all* of the information gathered during the session.

Summarize the above conversation, and output strictly in the
 following JSON format, with no additional text,
explanations, or casual remarks:

{
  "core_state": [
    {
      "statement": "The statement to be verified",
      "verification": "Right / Wrong / Unknown",
      "evidence": [
        {
          "source": "Evidence source URL or reference",
          "excerpt": "Quoted text or summary from the
evidence"
        }
      ],
      "reasoning": "Explanation of the verification
reasoning and process"
    },
    ...
  ]
}

Requirements:
1. Do not add any output other than the JSON object.
2. Each `evidence` entry must always include both `source`
and `excerpt`; do not omit either field.
3. The `source` and `excerpt` must always be included if
content is provided.
4. If multiple pieces of evidence exist, include all of them
 in the `evidence` array.
```

H Task Construction Details

Our automated deep research task construction pipeline begins with a predefined set of empirical domains. For each domain, we generate multiple personas with clearly defined roles and well-specified backgrounds. Given each persona, we then construct corresponding deep research tasks tailored to that per-

sona. These tasks are subsequently filtered using an LLM-based judge according to predefined criteria.

Next, we generate a no-search baseline for each task and evaluate its performance, discarding tasks that can already be answered well by the baseline alone. Throughout the pipeline, GPT-5-mini is used as the base LLM with default API settings. In total, we define 10 fixed domains, generate 5 personas per domain, and create 4 related deep research tasks for each persona.

Persona Generation Prompt

```
You are a persona generator. For the domain "{{domain}}",
generate 5 distinct personas.

Requirements:
1. Personas can be ordinary individuals (e.g., students,
hobbyists, employees, community members) or domain experts.
2. Each persona must include a clear background: education,
career path, interests, or life experience (100-150 words).
3. Each persona should have specific interests or
responsibilities related to the domain.

Example subdomains:
Domain: Software Development
Subdomain: Artificial Intelligence/Machine Learning
Development & Tools

Domain: Science & Technology
Subdomain: Large Language Models
Subdomain: Artificial Intelligence & Machine Learning

Domain: Sports & Fitness
Subdomain: Professional Sports (Football, Basketball, etc.)

Strictly output the following JSON only, with no extra text:
{
  "domain": "{{domain}}",
  "personas": [
    {
      "name": "First persona name",
      "role": "Role or identity (e.g., university student,
software engineer, policy advisor)",
      "affiliation": "Affiliation/Organization/Community (if
 applicable)",
      "background": "Detailed background, including
education, work/interest history, and relevant experience
(100-150 words)",
      "subdomain": "subdomain within the domain",
    },
    {
      "name": "Second persona name",
      "role": "Role or identity",
      "affiliation": "Affiliation/Organization/Community (if
 applicable)",
      "background": "Detailed background (100-150 words)",
      "subdomain": "subdomain within the domain",
    },
    ...
  ]
```

Deep Research Generation Prompt

```
You are a deep research query designer. Based on the persona
 profile below, generate 4 realistic deep research queries
that the persona would like to ask, which MUST rely on deep
web search, reasoning, and information synthesis to complete
.

Persona information:
- Name: {{persona_name}}
- Role: {{persona_role}}
- Affiliation: {{persona_affiliation}}
- Background: {{persona_background}}
- Subdomain: {{subdomain}}

Query requirements:
1. Each query must require multiple rounds of web search (at
 least 2 rounds from different perspectives).
2. Each query must integrate information from multiple
credible sources (e.g., academic papers, industry reports,
news articles, policy documents, statistics, online forums).
 ]
3. Cover latest developments, data analysis, trend
assessment, comparisons, or case studies — aligned with the
persona's role and domain interest.
4. Complexity should be proportional to the persona's
profile:
   - For experts: advanced analytical or strategic-level
queries.
   - For non-experts: practical, decision-making, or problem
-solving queries with substantial information gathering.
5. 70% of queries should include explicit time constraints (
e.g., "from January 2024 to August 2025", "before 2025", "as
 of 2025-08-05").
6. Each query must be concrete, specific, and have clear
deliverables.
7. Queries must reflect the persona's realistic needs and
challenges.
8. Length: 10-50 words per query.
9. It would be better if query is **tied to a current or
emerging hot topic** (e.g., stablecoin regulations in May
2025, AI safety debates, semiconductor export bans) that
requires synthesis of conflicting perspectives and time-
sensitive evidence.
10. Encourage queries that involve **controversial
viewpoints, cross-country comparisons, or forward-looking
predictions** (e.g., market impacts, regulatory outcomes,
adoption scenarios).

Example queries:
1. My son is about to start his university applications in
2025 for postgraduates but he's still uncertain about both
his major and which universities to apply to. Could you help
 me find the top five universities in each of the five broad
 subjects from the QS World University Rankings by Subject
2025, and also check their standings in the QS World
University Rankings 2025 and the Times Higher Education
World University Rankings 2025? And I need the home page of
the university's official website, standard application
deadline for regular decision as well as the application fee
 without the fee waiver.
2. Please provide a detailed explanation of the differences
and connections between Google's recently released A2A
protocol and the MCP protocol. Furthermore, elaborate on the
 innovative aspects of the A2A protocol and the specific
problems it is designed to address.
3. What are the investment philosophies of Duan Yongping,
Warren Buffett, and Charlie Munger?
4. How did discussions on stablecoin regulations evolve
across US, EU, and Asia in May 2025, and what are the
projected implications for financial stability and crypto
adoption by 2026?

Strictly output the following JSON only, with no extra text:
{
  "persona_name": "persona_name",
  "tasks": [
    {
      "task_id": "task_1",
      "deep_research_query": "Deep research query (10-50
words)",
      "key_challenges": "Main challenges and why deep search
 is required",
      "expected_search_rounds": 4,
      "time_sensitivity": true,
      "time_constraint": "Explicit time constraint
description"
    },
    {
      "task_id": "task_2",
      "deep_research_query": "Deep research query (10-50
words)",
      "key_challenges": "Main challenges",
      "expected_search_rounds": 3,
      "time_sensitivity": false,
      "time_constraint": null
    },
    ...
```

```
  ]
}
```

Task Qualification Filter Prompt

```
You are a deep research query analysis expert. Evaluate
whether the following task truly requires deep web search
and information synthesis to complete.

Query: {{deep_research_query}}
Persona background: {{persona_background}}
Query challenges: {{key_challenges}}

Evaluation criteria:
1. Requires up-to-date information that cannot be accurately
 answered using pre-2023 knowledge alone.
2. Requires cross-verification and integration of multiple
credible sources (e.g., academic papers, news, industry
reports, policy documents, statistics).
3. Requires multi-angle, multi-layered deep investigation.
4. Complexity matches the persona's role and capabilities:
   - For experts: advanced analytical or strategic-level
complexity.
   - For non-experts: realistic but still substantial
research complexity beyond casual searching.


Strictly output the following JSON only, with no extra text:
{
  "needs_deep_research": (true/false),
  "confidence_score": (0~1),
  "reasoning": "Detailed rationale explaining why deep
search is needed and what information collection and
synthesis are required (100-150 words)",
  "search_complexity": ("High"/"Medium"/"Low"),
  "information_sources_needed": ["academic papers", "news",
"technical reports", "market data", "policy documents"],
  "latest_info_required": (true/false),
  "cross_domain_integration": (true/false)
}
```

Search Necessity Baseline Prompt

```
Based solely on your existing knowledge, without using any
external search tools, answer the following query. Provide
the best answer you can.

Persona background: {{persona_background}}

Query: {{deep_research_query}}

Requirements:
1. Provide as detailed and complete an answer as you can,
considering the persona's role, knowledge scope, and domain
interest.
2. If some information is uncertain or may be outdated,
clearly state this.
3. Do not fabricate specific data, dates, or facts you are
not confident about.
4. Ensure logical structure and clear organization.
5. Demonstrate appropriate depth of understanding for the
persona's role:
   - For experts: in-depth analysis with technical or
strategic insights.
   - For non-experts: practical, clear, and well-explained
reasoning.
6. If the query involves recent developments, acknowledge
any knowledge cutoff limitations.

Please provide your best possible answer:
```

Search Necessity Assessment Prompt

```
You are an answer quality evaluator for deep research
queries. Assess the quality and completeness of the
following answer for the given query.

Original query: {{deep_research_query}}
Query challenges: {{key_challenges}}
Persona background: {{persona_background}}
No-search answer: {{baseline_answer}}

Evaluation dimensions:
1. Accuracy: Are statements reliable and free of obvious
errors?
2. Completeness: Does it cover the core requirements and key
 aspects of the query?
3. Depth: Is the analysis sufficiently deep and appropriate
for the persona's role?
4. Timeliness: Could information be outdated? Does it
acknowledge time limitations?
5. Professionalism: Does it meet the expected standards for
the persona's role and demonstrate relevant insight?
6. Structure: Is the answer well-organized and logical?

Special focus:
- If the task involves time requirements such as "latest", "
recent", or "current", assess whether the answer meets them.
- If the task requires multi-source synthesis, assess
whether a single knowledge source is sufficient.
- Assess whether additional search is truly necessary to
provide a better answer, based on the persona's role and
context.

Strictly output the following JSON only, with no extra text:
{
  "overall_quality": "low",
  "quality_score": 0.3,
  "completeness_score": 0.4,
  "accuracy_score": 0.7,
  "depth_score": 0.2,
  "timeliness_score": 0.1,
  "accuracy_concerns": "Specific accuracy concerns or issues
",
  "missing_aspects": "Important aspects and information that
 are missing",
  "outdated_info": "Potentially outdated content",
  "requires_search": true,
  "search_necessity_reasoning": "Detailed reasoning for why
search is needed and what key information is missing
(150-200 words)"
}
```

I Human Study

I.1 Correct Examples

Case 1

Statement

DeepSeek V3 was trained for approximately $5.6 million using 2,788M H800 GPU hours.

Model Response

Verification: Wrong.

Reasoning: The arXiv paper reports 2.788M H800 GPU hours (i.e., approximately 2.788 million hours). The statement incorrectly uses "2,788M," which implies 2,788 million (i.e., 2.788 billion) GPU hours, introducing a three-orders-of-magnitude unit error. In addition, the $5.6M figure is a community-derived estimate based on assumed GPU hourly costs, rather than a training cost explicitly reported in the paper. Therefore, the claim as written is incorrect.

Evidence:

- *arXiv*: "Despite its excellent performance, DeepSeek-V3 requires only 2.788M H800 GPU hours for its full training."
- *Medium*: "2.788M H800 GPU hours (âĽ$5.6M at $2/h) ..."

Human Annotation
Right

Final Judge
The model judgment is **correct**. Although the discrepancy appears minor, the original statement mistakenly interprets "2.788M" as "2,788M," resulting in a thousand-fold overestimation of training compute. This unit-level error materially changes the factual meaning of the claim.

Case 2

Statement
The FIBA Europe Coaching Certificate (FECC), a two-year, mentorship-intensive program run by elite European coaches, is specifically designed to elevate the knowledge base of coaches working with players aged 15–20.

Model Response
Verification: Unknown.
Reasoning: The FIBA Europe Coaching Certificate (FECC) is a real and highly regarded coaching program focused on the development of youth coaches in Europe. Publicly available sources confirm its prestige and its emphasis on developing coaches working with young players. However, descriptions of the program vary regarding its duration, with some sources referring to a three-year structure, and publicly available materials do not consistently substantiate the precise phrasing "two-year, mentorship-intensive" nor explicitly specify a target player age range of 15–20. Due to these inconsistencies and insufficient public detail, the exact claim cannot be fully verified.
Evidence:

- *FIBA Europe*: "The FIBA Europe Coaching Certificate (FECC) program has established its position as one of the most prestigious coaching programs in European sport."
- *FIBA Europe Coaching*: "The graduates from the FECC will receive an extremely rare and valuable qualification and they are FIBA Europe approved coaches for developing young players."
- *African Sports Monthly*: "The exclusive three-year coaching program, which focuses on the development of young players, has a limited intake of 60 up-and-coming European coaches..."

Human Annotation
Right

Final Judge
The model judgment is **correct**. A review of publicly available information on the FIBA Europe Coaching Certificate (FECC) confirms that most aspects of the statement are broadly accurate. However, official descriptions do not explicitly specify a player age range of 15–20, and the program duration is described inconsistently across sources. As a result, the claim cannot be fully verified and is appropriately classified as Unknown.

Correct Case 3

Statement
European Union: Dutch and EU-wide measures broadened controls on advanced lithography and semiconductor manufacturing equipment (e.g., immersion DUV tools, EUV pellicles).

Model Response
Verification: Unknown.
Reasoning: Public documentation from the Dutch government and ASML confirms that the Netherlands expanded export license requirements on advanced semiconductor manufacturing equipment, notably including immer-

sion DUV lithography tools, in September 2024. However, no publicly available primary EU regulatory documents were found that explicitly demonstrate an EU-wide measure adding EUV pellicles to the EU dual-use export control list by the relevant time frame. Because the statement jointly attributes such controls to both Dutch national measures and EU-wide actions, including EUV pellicles, the combined claim cannot be fully verified.
Evidence:

- *ASML Press Release*: "The Dutch government today published an updated license requirement regarding the export of immersion DUV semiconductor equipment."
- *Government of the Netherlands*: "As of 7 September 2024, the national export control measure applicable to advanced semiconductor manufacturing equipment will be expanded."

Human Annotation
Right

Final Judge
The model judgment is **correct**. A review of relevant public sources confirms that export controls were expanded at the Dutch national level, while no clear evidence supports an EU-wide regulatory action specifically extending controls to EUV pellicles. The original statement implies coordinated Dutch and EU-wide expansion of such controls, which is not fully supported by available evidence; therefore, the claim is appropriately classified as Unknown.

I.2 Incorrect Examples

We present two examples in which overly strict verification leads to incorrect judgments; however, it should be noted that such cases occur only occasionally.

Incorrect Case 1

Statement
AI-based virtual triage is effective at identifying individuals with potentially life-threatening conditions who initially did not intend to seek professional care.

Model Response
Verification: Unknown.
Reasoning: There are individual examples of digital triage flagging urgent conditions, but the overall performance of symptom checkers and AI-based triage tools varies across systems and clinical settings. The broad and unconditional claim that AI virtual triage is effective in this role lacks a single, universal supporting source. Therefore, the claim is classified as Unknown.
Evidence:

- *BMJ*: "Evaluations of symptom checkers show widely variable performance; some identify urgent conditions, but overall sensitivity and safety vary by tool and condition."
- *PMC*: "AI-based triage systems show promise for improving emergency department efficiency but require rigorous multicenter validation and standardized outcome reporting."

Human Annotation
Right

Final Judge
The model judgment is **incorrect**. Although the model identified concrete examples supporting the claim, it applied an overly strict verification standard by requiring a universal supporting source. The evidence it retrieved is already sufficient to substantiate the statement, and therefore the claim should be considered correct.

Incorrect Case 2

Statement
Technology procurement for basic AI triage systems ranges from $50,000 to $200,000.

Model Response
Verification: Unknown.
Reasoning: Industry and vendor analyses report wide and overlapping ranges for AI implementation costs in healthcare, spanning from tens of thousands to well over one million dollars depending on scale and complexity. Although the 50k to 200k range is plausible for certain deployments, it is not established as a general or authoritative standard in peer-reviewed or governmental sources. Therefore, the specific figure cannot be conclusively verified and is classified as Unknown.
Evidence:

- *Biz4Group*: "AI in healthcare applications ranges from $50KâĂŞ$1M+ depending on scale and complexity."
- *ITRex Group*: "The costs of implementing AI in healthcare range from $40,000 for simple AI functionality to $100,000 and much more for a comprehensive solution."
- *Aalpha*: "The cost of implementing AI in healthcare can range from $50,000 to $300,000 for small clinics."
- *Provided Part*: "Procurement âĂę can range from $50,000 to $200,000 for basic AI triage systems depending on customization needs."

Human Annotation
Right

Final Judge
The model judgment is **incorrect**. The original source explicitly reports this numerical range, but the model dismissed the claim by treating the specific figure as inappropriate and applying an overly strict verification criterion. In fact, the statement is supported by the source and should be judged as correct.

Table 12: Descriptive statistics of task-specific dimensions.

Statistic	Value
Total task-specific dimensions	282
Total tasks	100
Average number of task-specific dimensions per task	2.82
Minimum number of dimensions per task	2
Maximum number of dimensions per task	3

J Task-Specific Dimension

We provide descriptive statistics of task-specific dimensions, three examples of task-specific dimensions here, and also present an example illustrating how they function in practice.

Task-Specific Dimension Example 1


```
{"topic": "Transportation", "language": "en", "prompt": "
Compare eàĂŠscooter regulatory frameworks, safety outcomes,
and injury trends in US, EU, and China from January 2023 to
August 2025; deliver harmonized policy recommendations, five
 measurable safety metrics, and data sources."}

  "dimensions": [
    {
      "meta_dimension_name": "Policy Pragmatism",
      "definition": "Assesses the real-world viability and
implementability of the harmonized policy recommendations,
evaluating their adaptability to the distinct political,
economic, and cultural contexts of the US, EU, and China."
    },
    {
      "meta_dimension_name": "Comparative Synthesis",
      "definition": "Evaluates the quality of the cross-
jurisdictional analysis, measuring how effectively the
report integrates disparate regulatory frameworks and safety
 data into a coherent, unified narrative, rather than merely
 juxtaposing separate regional summaries."
    },
    {
      "meta_dimension_name": "Metric Utility",
      "definition": "Measures the quality of the proposed
safety metrics, evaluating if they are genuinely comparable,
 measurable, and actionable across the diverse data
collection ecosystems of the US, EU, and China."
    }
  ]
```


Task-Specific Dimension Example 2


```
{"topic": "Health", "language": "en", "prompt": "Assess
nutritional quality, additive profiles, sodium levels, and '
ultra-processed' classification of top plant-based meat
products across US and EU as of 2025-08-15; deliver
comparative tables and public-health recommendations."}

  "dimensions": [
    {
      "meta_dimension_name": "Classification Rigor",
      "definition": "Assesses the clarity, consistency, and
scientific validity of applying the 'ultra-processed'
classification framework (e.g., NOVA) to the products. This
is crucial because the classification is a central, non-
trivial analytical task of the query, not just a simple data
 point."
    },
```

```
    {
      "meta_dimension_name": "Cross-Regional Synthesis",
      "definition": "Evaluates the depth of the comparative
analysis between US and EU markets, assessing not just the
presentation of data from both regions, but the synthesis of
 differences in product formulation, nutritional standards,
and regulatory landscapes that explain the observed data."
    },
    {
      "meta_dimension_name": "Recommendation Actionability",
      "definition": "Measures how specific, practical, and
evidence-based the public-health recommendations are, and
how directly they are derived from the report's comparative
findings on nutrition, additives, and processing, rather
than being generic health advice."
    }
  ]
```


Task-Specific Dimension Example 3


```
{"topic": "Sports & Fitness", "language": "en", "prompt": "
Compare community youth basketball coaching curricula (US
AAU/USAB, Spain club youth pathway, China school/club models
): skill retention, coach training, and transition rates;
deliver a scalable municipal implementation toolkit."}

  "dimensions": [
    {
      "meta_dimension_name": "Implementation Viability",
      "definition": "Assesses the practicality, scalability,
 and adaptability of the proposed municipal toolkit. This
evaluates whether the deliverable includes actionable
templates, realistic budget considerations, and clear
guidance for tailoring the program to diverse local contexts
 (e.g., city size, funding levels, existing infrastructure)
."
    },
    {
      "meta_dimension_name": "Comparative Synthesis",
      "definition": "Evaluates the ability to analyze each
country's basketball development system within its unique
socio-cultural and structural context, and to synthesize
these disparate findings into transferable principles. This
goes beyond a side-by-side description to assess the quality
 of drawing actionable lessons from one system for
application in another."
    }
  ]
```


Consider the task: *Compare community youth basketball coaching curricula (US AAU/USAB, Spain club youth pathway, China school/club models): skill retention, coach training, and transition rates; deliver a scalable municipal implementation toolkit.*

For this task, Doubao achieves the following scores on the four fixed general dimensions: Comprehensiveness (9.5), Insight (7.8), Instruction Following (9.9), and Readability (7.9). Manual inspection confirms that the report performs well on these general aspects. However, under the task-specific dimension *Implementation Viability*, the score drops to 6.02, as the report only outlines high-level components of the toolkit without providing sufficiently detailed or actionable implementation plans.

Importantly, the evaluation rationale generated by our quality assessment module aligns with this judgment: *The report outlines the components of a toolkit but does not provide the actual resources.* This case demonstrates that task-specific dimensions impose more targeted and stringent evaluation requirements, and the inclusion of such dimensions helps improve consistency with user preferences.

K Factual Evaluaion Examples

Factual Evaluation Example 1


```
{
  "statement": "China produces two-thirds of the world's
consumer electronics products.",
  "verification": "Wrong",
  "evidence": [
    {
      "source": "https://www.ey.com/en_cn/consumer-products/
ey-market-trends-report-2024-en.pdf",
      "excerpt": "China's dominant position in consumer
electronics exports is undeniable, with its global share
reaching an impressive 42% in 2022."
    },
    {
      "source": "https://rhg.com/research/china-and-the-
future-of-global-supply-chains/",
      "excerpt": "We review China's role in consumer
electronics..China is the world's largest electronics
manufacturing hub."
    }
  ],
  "reasoning": "Authoritative industry analyses indicate
that China is indeed the world's largest manufacturing hub
for consumer electronics, but recent estimates typically
place its share at around 40-45%, rather than approximately
66%. The available sources do not support the claim that
China produces 'two-thirds' of global consumer electronics."
}
```


Factual Evaluation Example 2


```
{
  "statement": "China / SMIC developed 7nm chips without EUV
 equipment and has further advanced to mass production at 5
nm.",
  "verification": "Wrong",
  "evidence": [
    {
      "source": "https://www.bloomberg.com/news/features
/2023-09-04/look-inside-huawei-mate-60-pro-phone-powered-by-
made-in-china-chip",
      "excerpt": "Huawei's Mate 60 Pro is powered by the new
 Kirin 9000S chip...manufactured in China by SMIC...the
processor is the first product made using SMIC's most
advanced 7-nanometer process."
    },
    {
      "source": "https://americanaffairsjournal.org/2024/02/
a-new-era-for-the-chinese-semiconductor-industry-beijing-
responds-to-export-controls/",
      "excerpt": "Assuming Beijing's new semiconductor
industrial policies achieve significant results...by 2025
SMIC could significantly expand capacity at the 5/6/7 nm
nodes...in the most optimistic scenario...SMIC may obtain a
prototype EUV system for testing in 2025."
    }
  ],
  "reasoning": "The evidence supports the factual claim that
 SMIC has produced 7nm-class chips using advanced DUV
techniques without EUV lithography (e.g., the Kirin 9000S
used in the Mate 60). However, during the period from 2023
to mid-2024, no reliable sources confirm that SMIC has
achieved stable mass production at the 5nm node; discussions
```

```
 of 5nm progress remain speculative or forward-looking.
Because the statement combines a verified fact (7nm without
EUV) with an unverified forward-looking claim (5nm mass
production), and under the verification rule that mixed
accuracy results in an overall negative judgment, the
statement is rated as Wrong."
}
```

Factual Evaluation Example 3


```
{
  "statement": "Surveys show that the public holds a low
level of optimism about the social impact of artificial
intelligence, with only 27% of Americans believing that AI
has a positive effect on science and medicine.",
  "verification": "Wrong",
  "evidence": [
    {
      "source": "https://www.pewresearch.org/internet
/2025/04/03/how-the-us-public-and-ai-experts-view-artificial
-intelligence/",
      "excerpt": "There are also large differences in views
of AI's impact on the economy, health care, education and
the arts....Health care is the area where the public is most
 optimistic about AI's impact. Experts are much more likely
than the public to say it will have a positive effect, with
a gap of as much as 40 percentage points (84% vs. 44%)."
    },
    {
      "source": "https://www.pewresearch.org/wp-content/
uploads/sites/20/2025/04/pi_2025.04.03_us-public-and-ai-
experts_report.pdf",
      "excerpt": "These surveys reveal deep divides and some
 common ground on AI issues. Experts are far more positive
than the public about AI's potential...The public is more
likely to say AI will harm them (43%) than benefit them
(24%)....Experts are more likely than the public to say AI
will have a very or somewhat positive effect on the United
States over the next 20 years (56% vs. 17%)."
    }
  ],
  "reasoning": "The Pew Research Center reports do not show
that '27% of Americans believe AI has a positive effect on
science and medicine.' Instead, the reports indicate that
44% of U.S. adults believe AI will have a positive impact on
 health care, and 17% believe it will have a positive impact
 on the United States overall. Therefore, the '27%' figure
cited in the statement is not supported by the Pew findings
and is incorrect."
}
```