SUPERRESEARCH BENCH: STRESS-TESTING LONG-HORIZON RESEARCH WITH EVIDENCE GRAPHS AND RUBRIC-GUIDED AUDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

While Large Language Models (LLMs) have demonstrated proficiency in Deep Research or Wide Search, their capacity to solve highly complex questions requiring long-horizon planning, massive evidence gathering, and synthesis across heterogeneous sources remains largely unexplored. We introduce **SuperResearch Bench**, a benchmark for *Super Research*: complex autonomous research tasks that integrate (i) structured decomposition, (ii) **super wide** retrieval for diverse perspectives, and (iii) **super deep** investigation through iterative queries. To evaluate this capability, we curated 300 expert-written questions across diverse domains, each requiring up to 100+ retrieval steps and 1,000+ web pages to reconcile conflicting evidence. SuperResearch Bench provides source-linked evidence graphs, reference reports with fine-grained citations, and intermediate artifacts to support traceable reasoning. The benchmark artifacts undergo automated validation and expert review. Our graph-anchored, rubric-guided auditing protocol evaluates generated reports along four primary dimensions: Coverage, Logical Consistency, Report Utility, and Objectivity. Citation Health is reported separately. While such questions may be infrequent in standard applications, SuperResearch Bench serves as a ceiling evaluation and stress test for long-horizon research. A comparison of fourteen systems and protocols reveals substantial variation in evidence recovery, reasoning consistency, and report usefulness, enabling fine-grained diagnosis of current research systems.

1 INTRODUCTION

The evolution of Large Language Models (LLMs) has marked a paradigm shift in artificial intelligence, moving from simple statistical text completion to sophisticated reasoning and agentic problem-solving. Early iterations focused on scaling parameters to capture world knowledge within static weights. However, the recent shift toward Retrieval-Augmented Generation (RAG) Lewis et al. (2020) and agentic workflows Yao et al. (2022); Wu et al. (2024) has empowered LLMs to interact with the external web, allowing them to overcome the limitations of their training cutoff and hallucination tendencies. Today, LLMs are no longer just repositories of information but are increasingly treated as "reasoning engines" capable of orchestrating complex sequences of tool use and information retrieval to answer non-trivial queries.

In tandem with these advancements, the concept of Deep Research Du et al. (2025); Google (2025) has emerged as a specialized capability for LLMs. This approach prioritizes vertical exploration, where an agent iteratively follows a chain of evidence to resolve specific, nuanced questions. By employing "Chain-of-Thought" prompting Wei et al. (2022) and recursive search loops, Deep Research agents can dive into technical documentation, legal archives, or scientific papers to extract precise answers that are hidden beneath layers of text. The focus here is on the depth of inquiry, which ensures that a specific thread of reasoning is followed to its logical conclusion through rigorous verification and iterative refinement.

Conversely, Wide Search Wong et al. (2025) addresses horizontal information coverage. This paradigm prioritizes large-scale data acquisition, striving to capture an extensive array of information nodes and web sources. In contrast, we define Wide Research as a more sophisticated evolution that

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

focuses on thematic width and perspective synthesis. Rather than mere data accumulation, Wide Research systems are designed to "survey the landscape", ensuring that no critical angle—be it economic, social, or technical—is overlooked. This is achieved through multi-query expansion and aspect-based retrieval, allowing the model to identify the multifaceted dimensions of a problem and curate a representative sample of evidence from across the digital ecosystem.

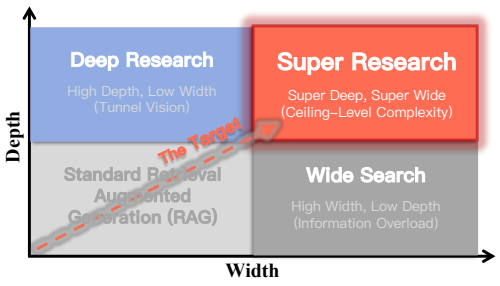

Figure 1: Research breadth and depth across RAG, Deep Research, Wide Search, and Super Research.

In this paper, we present **SuperResearch Bench**, a benchmark designed to evaluate research systems on *Super Research* tasks. These tasks include super-complex questions such as *"optimizing immunopharmacological mechanisms where T cell activation must be precisely balanced against tumor microenvironment immune escape pathways to minimize autoimmune risks"*, which require long-horizon planning, broad evidence gathering, and multi-step synthesis. Such tasks are designed to push beyond the typical operating scope of current Deep Research or Wide Search paradigms.

Super Research explicitly couples three core pillars: (i) Structured Decomposition, which breaks a monolithic query into a multi-layered research plan; (ii) Super Wide Retrieval, which explores the search space horizontally to ensure total coverage of diverse perspectives; and (iii) Super Deep Investigation, which utilizes iterative follow-up queries to resolve uncertainties and verify the reliability of individual data points. To rigorously test this framework, we introduce **SuperResearch Bench**, a benchmark of 300 expert-refined questions that require extensive evidence gathering and synthesis, each paired with a source-linked evidence graph (Appendix I).

Compared to existing Deep Research or Wide Search that typically operate within a range of 10–20 retrieval iterations and approximately 100 web pages, Super Research is designed for a higher tier of complexity. Those truly "super complex" questions are found in professional intelligence, scientific discovery, or strategic planning, and usually demand long-horizon planning that can manage 100+ retrieval steps and 1,000+ web pages to synthesize massive evidence while maintaining both extreme width and extreme depth.

However, evaluating Super Research is itself extremely challenging. Existing evaluation paradigms Mialon et al. (2023); Du et al. (2025); Wang et al. (2025); Hu et al. (2025) are ill-suited for the scale and nuance of complex inquiry. ① **Limitations of LLM-as-a-Judge:** Standard report-to-report comparisons Zheng et al. (2023) are prone to alignment errors and often fail to capture deep reasoning. ② **Limited Assessment of Evidence Synthesis:** Atomic factuality metrics Min et al. (2023); Wei et al. (2024) assess support for individual claims but do not directly evaluate the higher-level evidence dependencies involved in complex synthesis. ③ **Costly Human Evaluation:** Expert reviews and Elo-based systems Wan et al. (2025) are rigorous but prohibitively slow, expensive, and difficult to scale. ④ **Neglected Ambiguity:** Most metrics fail to reward "uncertainty expression", penalizing models that correctly identify lack of consensus while rewarding false confidence.

To bridge this gap, **SuperResearch Bench** combines evidence graphs with rubric-guided audits to evaluate generated reports along four primary dimensions: Coverage and Comprehension, Logical Consistency, Report Utility, and Objectivity. Citation Health provides a separate source-use diagnostic. Furthermore, we introduce an automated, graph-anchored auditing tool (Appendix G.1). By projecting generated reports onto an expert-curated evidence graph and applying explicit evaluation rubrics, this tool enables: ① **Structured Claim Verification:** Matching report claims to expert-reviewed reference evidence. ② **Interpretable Error Tracing:** Identifying missing evidence links in the generated report. ③ **Reproducible Scoring:** Applying explicit evaluation criteria at scale.

Although highly complex, multi-faceted queries may be infrequent in routine consumer applications, SuperResearch Bench serves a vital role as a frontier stress test and a high-ceiling evaluation framework for LLMs. As current benchmarks begin to suffer from performance saturation, there is a growing need for "ceiling protocols" that push models to their operational limits. By demanding long-horizon planning and the orchestration of 100+ retrieval steps and 1,000+ web pages, SuperResearch

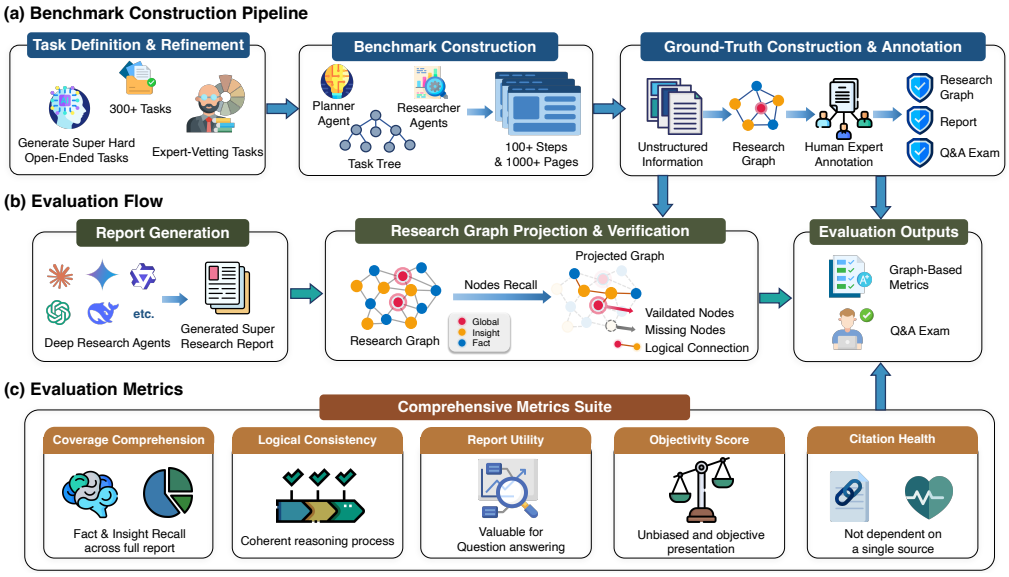

Figure 2: SuperResearch Bench pipeline: task refinement, reference-artifact construction, report generation, and graph-based evaluation.

Bench exposes latent weaknesses in reasoning consistency and context management that simpler tasks may fail to trigger. Consequently, a model's proficiency on this benchmark provides insight into its research competence; success in such a demanding setting suggests robustness and "agentic" stability in handling complex, long-horizon research tasks.

2 Related Work

Autonomous Research Agent Frameworks. Autonomous research agents have evolved from basic reasoning techniques to sophisticated multi-agent systems. Foundational approaches such as Chain-of-Thought Wei et al. (2022) and ReAct Yao et al. (2022) established core mechanisms for task decomposition and tool integration, operationalized in early systems like WebGPT Nakano et al. (2021) and WebCPM Qin et al. (2023). Framework libraries, including LangChain LangChain AI (2022) and LangGraph LangChain (2024) formalized complex, stateful reasoning workflows. Multi-agent paradigms, exemplified by AutoGen Wu et al. (2024) and MetaGPT Hong et al. (2023), introduced collaborative role-based architectures for intricate research processes. Specialized systems like STORM Shao et al. (2024) for open-domain synthesis and ChemCrow Bran et al. (2023) for scientific discovery advanced long-form research automation. However, a fundamental limitation persists: existing agents typically optimize for either breadth of information gathering or depth of verification (via mechanisms like Reflexion Shinn et al. (2023)), but rarely achieve both simultaneously. Recent efforts such as DeepWideSearch Lan et al. (2025) and Step-DeepResearch Hu et al. (2025) address this trade-off but remain constrained by limited planning horizons and insufficient robustness for high-entropy, conflicting information landscapes.

Knowledge Synthesis and Retrieval Methods. Information retrieval has evolved from simple vector-based similarity to structured knowledge integration paradigms. Early Retrieval-Augmented Generation (RAG) approaches Lewis et al. (2020); Guu et al. (2020) mitigated hallucination but struggled with contextual fragmentation across disparate sources. Recent graph-based methods, including GraphRAG Edge et al. (2024), HiPPoRAG Jimenez Gutierrez et al. (2024), and LightRAG Jiang et al. (2024), leverage structured representations for multi-hop reasoning and heterogeneous information aggregation. By incorporating schema-based grounding Pan et al. (2024); Wu et al. (2023), these approaches model semantic relationships between sources, enabling synthesis of conflicting narratives. Unlike traditional RAG systems treating documents as isolated chunks, graph-anchored methods capture evidence topology, essential for resolving contradictory information streams.

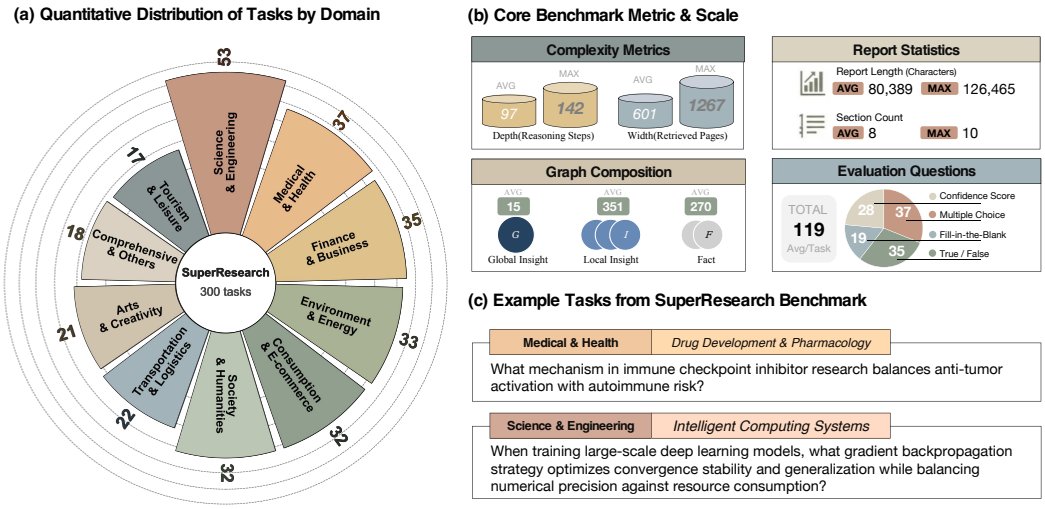

Figure 3: Benchmark profile. (a) Distribution of the 300 tasks across ten domains; (b) research and evaluation scale; (c) representative questions.

Evaluation Frameworks and Quality Metrics. Assessing open-ended research outputs requires metrics beyond surface-level plausibility. General-purpose benchmarks like GAIA Mialon et al. (2023) and AgentBench Liu et al. (2023) provide broad evaluations but lack precision for detailed research quality assessment. Specialized frameworks include DeepResearch Bench Du et al. (2025), LiveResearchBench Wang et al. (2025), DeepResearch Arena Wan et al. (2025), and DeepResearch-Gym Coelho et al. (2025). DeepResearch Bench II Li et al. (2026) derives fine-grained binary rubrics from expert-written articles to assess information recall, analysis, and presentation. SuperResearch Bench instead centers on long-horizon questions with source-backed research graphs, task-linked QA, and stance calibration to diagnose how reports assemble and use evidence.

3 SUPERRESEARCH BENCH

In this section, we detail the methodology used to construct SUPERRESEARCH BENCH. We employ an expert-involved, axiomatic construction process aiming to establish a "Gold Standard" for deep research tasks characterized by high cognitive complexity and rigorous logical structures. The core design philosophy of SUPERRESEARCH BENCH lies in achieving a dual balance: the breadth of information coverage (*SuperWide*) and the depth of synthetic reasoning (*SuperDeep*). Unlike existing datasets that focus on fact retrieval, our benchmark is engineered to test the asymptotic limits of an agent's ability to navigate unstructured and high-entropy information landscapes.

3.1 TASK DEFINITION

To stress-test the upper limits of Deep Research Agents, we synthesized criteria from frontier models and domain experts to establish a rigorous design protocol. Our construction follows three core principles: ① *Frontier-Level Specificity* targeting unresolved challenges intelligible to domain experts and excluding general-knowledge queries to ensure the benchmark measures deep expertise rather than surface-level reasoning. ② *Search-Intensive Complexity* demands broad, multi-hop synthesis across heterogeneous sources. Beyond simple fact retrieval, tasks require the analysis of conflicting data and long causal chains to test information integration. ③ *Professional Precision* employs unambiguous, industry-standard terminology to define the research scope, ensuring performance gaps reflect genuine capability limitations rather than interpretation errors.

3.2 EXPERT REFINEMENT

We employ a top-down, expert-driven task refinement process within the construction workflow shown in Figure 2. Task refinement integrates the generative breadth of LLMs with domain-expert review and consists of two steps:

To generate high-complexity programs, we utilized advanced LLMs (GPT-4o, Gemini-deep-research-agent) under our *Cognitive-Rank Constrained Expert Simulation* framework. By injecting strict negative constraints to enforce a senior investigator persona, models were instructed to: ① traverse sub-domain topologies to identify structural tensions; ② prioritize multi-layered analysis over factual lookups; and ③ scale task volume by theoretical field density. Subsequently, domain experts (senior practitioners and Ph.D. candidates) audited the raw candidates to eliminate hallucinations. The screening retained only tasks meeting three criteria: ① theoretical solvability via public retrieval; ② genuine complexity resistant to simple keyword search; and ③ unambiguous formulation, ensuring alignment with real-world research demands.

3.3 HUMAN AI COLLABORATIVE COLLECTION

Our four-stage dataset construction pipeline is designed to mimic the cognitive workflow of a domain expert. This process integrates autonomous agent collaboration with human-in-the-loop verification to ensure both the breadth of information retrieval and the logical depth of synthesis.

Phase 1: Hierarchical Task Decomposition and Retrieval. This phase addresses the challenge of information breadth and logical dependency. Given a complex research query, a *Planner* agent B.3.1 decomposes the root topic into a structured, Directed Acyclic Graph (DAG) of research tasks, organized into chronological phases and specific investigation chapters. To achieve progressive generation, the *Researcher* B.3.2 and *Summarizer* B.3.3 operate in a collaborative loop: as the Researcher executes sub-tasks in a dependency-aware sequence, the Summarizer continuously synthesizes the results into a *Dynamic Memory*.

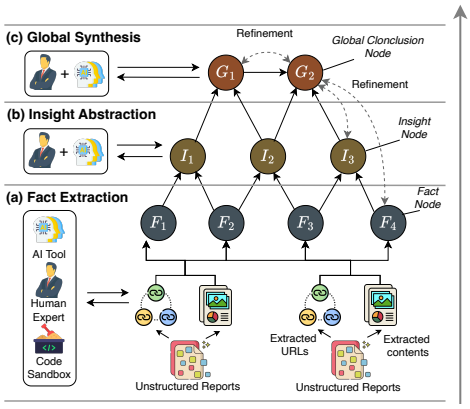

Figure 4: From source-linked facts to local insights and global conclusions.

This memory is iteratively injected into subsequent steps, ensuring that the global context evolves alongside the execution.

Phase 2: Research Graph Construction. We transform unstructured sub-reports into a structured research graph, which serves as an evidence graph linking Facts, Insights, and Conclusions, through a collaborative Human-AI process (Figure 4). This phase integrates subgraph extraction, expert-guided merging, and global insight generation. First, the system extracts atomic "Facts" (anchored to specific URLs) and derived "Insights" from sub-reports. Subsequently, adopting a human-in-the-loop workflow, experts refine the graph structure, constructing a bottom-up topology where higher-order insights are rigorously connected to supporting factual nodes. This verification is augmented by LLM-driven code sandboxes and search tools for automated calculation and fact-checking B.3.4. Finally, both experts and the LLM traverse this graph to identify evidence clusters, iteratively generating global insights that bridge disparate sub-reports.

Phase 3: Report Synthesis and Expert Editing. This phase focuses on producing definitive canonical reports. Leveraging the structured research graph and granular sub-reports, a *Writer* agent B.3.5 iteratively constructs the manuscript section by section. Crucially, this process integrates a real-time human-in-the-loop workflow: domain experts interact via a dashboard to validate logical soundness and intervene with on-the-fly corrections during text generation.

Phase 4: Automated Evaluation Metric Construction. Leveraging both the research graph and the final report, we derive a comprehensive suite of Question-Answer (QA) pairs B.3.6. Beyond standard fact retrieval, this suite specifically includes "Bias Calibration" queries designed to assess the neutrality and objectivity of the report's narrative, ensuring the description remains impartial

even when synthesizing subjective or conflicting sources. The generated QA pairs and calibrated bias scores serve as the ground truth for benchmarking the *SuperDeep* and *SuperWide* capabilities of the models evaluated in this study.

As part of human-in-the-loop verification, automated checks and expert review assess Fact support, graph dependencies, and evaluation-question quality. Identified issues are corrected, and reports affected by reference revisions are rescored. The main results use the audited benchmark artifacts; Appendix B.1 summarizes the checks and corrections.

Rubric-guided construction and scoring. The pipeline uses explicit criteria at several decision points. Task refinement checks public answerability, multi-source dependence, and clarity. Graph prompts distinguish verifiable facts, analytical insights, and global conclusions, requiring supporting node identifiers for higher-level claims. Question generation records the research step and assigns a depth label from 0 (local recall) to 5 (cross-step inference), alongside the question type and answer key. For competing claims, the stance audit scores support for each side on a 0–10 anchored scale. These operational rubrics make the assessment criteria explicit without reducing open-ended reports to text overlap.

4 Evaluation Framework and Metrics

Evaluating frontier-level, open-ended research poses unique challenges. Unlike factoid retrieval, open-ended queries allow models to explore divergent research trajectories, rendering traditional string-matching metrics ineffective. Furthermore, frontier questions often lack static ground truths, necessitating a deeper assessment of the reasoning process and derivation depth rather than mere surface-level overlap.

To address these challenges, we introduce a comprehensive evaluation suite anchored in rigorously constructed Research Graphs. The primary score combines four dimensions: ① Coverage and Comprehension (graph-based information recall), ② Logical Consistency (coherence of reasoning chains), ③ Report Utility (QA-based functional assessment), and ④ Objectivity Score (stance calibration). Each dimension receives an equal weight of 25%. For system m, $\text{Overall}_m = \frac{1}{4}\sum_{i=1}^{4} \bar{s}_{mi}$, where $\bar{s}_{mi}$ is its mean score on dimension i. Citation Health provides separate source-use diagnostics and is not included in Overall.

4.1 Coverage and Comprehension

To comprehensively assess both the breadth and depth of the generated reports, we introduce a metric based on research graph projection (as shown in Figure 2(b)). The core principle involves projecting the ground-truth graph structure onto the generated text to verify the presence of essential information. We first calculate the recall rate for individual graph nodes, categorizing them into three hierarchical levels: Atomic Facts Level ($\mathcal{R}_1$) for specific data points, Key Insights Level ($\mathcal{R}_2$) for intermediate reasoning connections, and Global Insights Level ($\mathcal{R}_3$) for high-level strategic conclusions. To derive a single, holistic indicator, we compute the Depth-Weighted Recall (denoted as $\mathcal{R}_{\text{weighted}}$). This metric assigns progressive weights to nodes based on their hierarchical depth, ensuring that the evaluation prioritizes the capture of complex synthesis and core theses over the mere enumeration of trivial facts. The detailed mathematical formulation is provided in Appendix C.1.

4.2 Logical Consistency

Research fundamentally demands that conclusions be derived from evidence, not merely stated. To verify the integrity of these deduction paths, we propose Logical Consistency (denoted as $\mathcal{C}_{\text{logic}}$). This metric assesses whether the generated Global Conclusions "are algorithmically grounded in Atomic Facts" via complete, unbroken citation chains. Formally, $\mathcal{C}_{\text{logic}}$ is defined as the product of the valid inference ratio (the proportion of conclusions fully supported by valid evidence edges) and the overall graph coverage. This multiplicative formulation serves as a rigorous filter: it penalizes reports that hallucinate correct high-level answers without the necessary supporting evidence, ensuring that high scores reflect both accurate outcomes and valid reasoning processes. The mathematical derivation is detailed in Appendix C.2.

4.3 Report Utility

Beyond structural completeness, a research report should help a reader answer task-specific questions. We introduce Report Utility (denoted as $\mathcal{U}_{\text{qa}}$), which evaluates whether the generated content encodes usable knowledge. We construct a set of exam questions from the reference artifacts. An independent exam-taker receives the generated report as context and answers these questions; its responses are checked against the reference answers. The resulting score measures report-conditioned question-answering performance. The detailed evaluation protocol is provided in Appendix C.3.

4.4 Objectivity Score

For complex, open-ended research topics, a high-quality report must navigate conflicting viewpoints without succumbing to one-sided bias. To measure this, we utilize the **Objectivity Score** (denoted as $\mathcal{O}_{\text{bias}}$), which quantifies the model's ability to maintain **multi-perspective balance**.

Instead of verifying binary truth, this metric conducts a double-blind audit to assess how the report weighs competing arguments (e.g., Thesis vs. Antithesis). The core objective is **Calibration of Stance**: the evaluation ensures that the degree of certainty and the distribution of evidence in the generated report accurately mirror the inherent ambiguity found in the ground-truth Research Graph. A high $\mathcal{O}_{\text{bias}}$ indicates that the model successfully represents the diversity of perspectives, avoiding the pitfall of presenting a highly debated topic as a settled fact. The detailed calculation methodology is provided in Appendix C.4.

4.5 Citation Health

While not part of the primary score, Citation Health serves as a diagnostic measure of **information synthesis quality**. Its primary objective is to distinguish between genuine multi-source integration and mere single-document summarization, flagging potential risks of Single-Source Dependency. We monitor two diagnostic indicators (where lower is better): ① *Source Dominance* ($\mathcal{D}_{\text{src}}$) checks the total volume, measuring if a single reference constitutes a disproportionately high percentage of the total citations. ② *Narrative Monopolization* ($\mathcal{M}_{\text{mono}}$) evaluates the citation layout across the full text. Unlike localized clusters (where using one source for a specific paragraph is acceptable), this metric penalizes cases where a single source drives the narrative flow from beginning to end, implying the report is merely a summarization of one document rather than a multi-source synthesis. Formal definitions are in Appendix C.5.

5 Experiments

5.1 Experimental Setup

Evaluation Methods. To comprehensively benchmark the capabilities of current systems in super research tasks, we evaluate a diverse set of methods. The evaluated methods are grouped into three primary tiers:

(1) *Deep Research Systems*. We assess leading commercial agents explicitly optimized for autonomous, long-horizon research tasks. These systems are designed to handle complex planning, multi-step evidence gathering, and information synthesis natively. Our evaluation includes Sonar Deep Research Perplexity (2025), Gemini Deep Research Google (2025), Tongyi Deep Research Qwen et al. (2025) and OpenAI's latest reasoning-enhanced models, o3 Deep Research and o4-mini Deep Research OpenAI (2025).

(2) *Native Search-Integrated Agents*. To examine systems with integrated real-time web access, we evaluate Perplexity Pro Search, Kimi K2 Turbo Team et al. (2025), and Grok-4-1-fast xAI (2025). Their tested API identifiers are listed in Table 1; Perplexity Pro Search is distinct from Sonar Deep Research.

(3) *Search-Augmented Baselines*. Five foundation models use a LangGraph workflow LangChain (2024) with Tavily Search Tavily (2024): DeepSeek R1 Guo et al. (2025), MiniMax M2.1 MiniMax (2025), Claude 3.5 Sonnet Anthropic (2024), Qwen 2.5 72B Qwen et al. (2025), and Llama 3.3 70B Meta (2024). We additionally test a constructed Seed 2.1 Turbo agent with OpenRouter web-

Table 1: Results for 14 tested systems on the 300 tasks. Overall combines Coverage (Cov), Consistency (Con), Utility (Util), and Objectivity (Obj) with equal 25% weights; bold and underlined Overall values denote the highest and second-highest scores. Dom (source dominance) and Mono (narrative monopolization) are lower-is-better citation diagnostics. Exact model or API identifiers are shown beneath each system name.

System (tested model/API ID)	Overall ↑	Cov ↑	Con ↑	Util ↑	Obj ↑	Dom ↓	Mono ↓
Seed 2.1 Turbo bytedance-seed/seed-2-1-turbo	**63.92**	65.31	51.92	76.00	62.45	–	–
Perplexity Deep Research perplexity/sonar-deep-research	<u>62.24</u>	64.55	51.92	71.00	61.48	8.31	81.46
Perplexity Pro Search perplexity/sonar-pro-search	61.80	62.37	44.46	76.50	63.87	–	–
Gemini Deep Research deep-research-pro-preview-12-2025	57.66	60.29	48.62	65.00	56.70	5.74	69.92
MiniMax M2.1 minimax/minimax-m2.1	50.63	53.96	41.58	53.50	53.48	7.56	72.60
Kimi K2 Turbo (web search) moonshot/kimi-k2-turbo-preview-web-search	45.86	45.94	31.79	55.00	50.69	–	–
Grok 4.1 Fast (web search) grok-4-1-fast-web-search	44.83	47.77	36.58	47.00	47.98	12.42	65.98
Tongyi DeepResearch alibaba/tongyi-deepresearch-30b-a3b	44.20	46.77	32.67	49.00	48.36	–	–
DeepSeek R1 deepseek/deepseek-r1	42.86	45.54	33.29	46.00	46.60	7.46	64.20
OpenAI o3 Deep Research openai/o3-deep-research	40.66	43.47	33.54	43.00	42.61	4.36	43.87
OpenAI o4-mini Deep Research openai/o4-mini-deep-research	39.61	43.07	30.38	43.00	42.01	6.01	58.30
Claude 3.5 Sonnet anthropic/claude-3.5-sonnet	28.30	33.56	21.00	24.50	34.13	4.25	13.29
Qwen 2.5 72B Instruct qwen/qwen-2.5-72b-instruct	26.05	33.03	21.42	18.50	31.24	5.31	19.58
Llama 3.3 70B Instruct meta-llama/llama-3.3-70b-instruct	17.83	24.60	13.17	9.50	24.05	1.29	2.46

search and web-fetch tools. The Seed result measures this search-equipped protocol, not a native Seed Deep Research product. The table combines archived January 2026 reports for twelve systems with September 2026 reports for Seed and Perplexity Pro Search.

5.2 Main Results

5.2.1 Overall Performance

Table 1 summarizes the fourteen tested systems under the specified 300-task scenario. Seed 2.1 Turbo achieves the highest Overall score of 63.92, followed by Perplexity Deep Research (62.24), Perplexity Pro Search (61.80), and Gemini Deep Research (57.66). These scores characterize the tested systems and protocols; the small differences among the leading systems do not by themselves establish statistically significant performance gaps.

The remaining Deep Research systems score 44.20 for Tongyi, 40.66 for o3, and 39.61 for o4-mini. MiniMax M2.1 scores 50.63, Kimi K2 Turbo 45.86, Grok 4.1 Fast 44.83, and DeepSeek R1 42.86. The lowest Overall score is 17.83 for Llama 3.3 70B. Because report dates and search protocols differ, these values describe the tested systems rather than an isolated effect of model architecture.

5.2.2 Coverage and Consistency

Coverage exceeds Consistency for every system in Table 1; the table alone does not establish a causal relationship between them.

① **Coverage–Consistency Gap.** Seed has the highest Coverage (65.31), while Seed and Perplexity Deep Research tie for the highest Consistency (51.92). Gemini scores 60.29 on Coverage and 48.62 on Consistency. At the lower end, Llama scores 24.60 and 13.17, respectively. The gap between

recovering reference content and connecting it consistently persists across the tested systems (see Appendix D.1).

② **Differences Among Search-Integrated Systems.** Kimi K2 Turbo has Coverage 45.94 and Consistency 31.79, while Grok 4.1 Fast has Coverage 47.77 and Consistency 36.58. Perplexity Pro Search scores higher on both dimensions (62.37 and 44.46). These differences reflect the complete tested protocols, including their search behavior.

5.2.3 Utility and Objectivity

The revised scores show substantial variation in Report Utility ($\mathcal{U}_{qa}$) and Objectivity ($\mathcal{O}_{bias}$).

① **High Utility Among the Leaders.** Perplexity Pro Search has the highest Utility (76.50) and Objectivity (63.87). Seed follows closely at 76.00 and 62.45, while Perplexity Deep Research scores 71.00 and 61.48. These component scores refer to the tested reports.

② **Lower-Scoring Reports.** The o3 and o4-mini Deep Research reports both score 43.00 on Utility, with Objectivity 42.61 and 42.01. Qwen scores 18.50 on Utility and 31.24 on Objectivity, and Llama scores 9.50 and 24.05. The table does not support a general Utility–Objectivity trade-off: the three highest-Utility systems also have the highest Objectivity scores.

5.2.4 Citation Health Diagnostics

Citation diagnostics are available for ten of the fourteen systems; the number of reports with recorded citation diagnostics varies across systems. Lower Source Dominance ($\mathcal{D}_{src}$) and Narrative Monopolization ($\mathcal{M}_{mono}$) values indicate less concentration within the scored reports.

① **Variation in Citation Concentration.** Perplexity Deep Research has Dominance 8.31 and Monopolization 81.46 over 300 scored reports; Gemini has 5.74 and 69.92 over 300. Among the OpenAI systems, o3 has 4.36 and 43.87, while o4-mini has 6.01 and 58.30. These diagnostics are separate from Overall and the four quality components.

② **Availability Limits.** Seed, Perplexity Pro Search, Kimi, and Tongyi have no recorded citation diagnostics in this result set. Llama has low recorded values (1.29 and 2.46) on 85 reports, but those values alone do not establish citation quality or source diversity. We therefore avoid ranking systems with missing or uneven citation records on these diagnostics.

Evaluator sensitivity. In the controlled 120-report perturbation study, the graph-based evaluator responds in the expected direction in 57.4–79.6% of cases across the tested dimensions and interventions, compared with 14.8–22.2% for the LLM-judge baseline. The full perturbation protocol, figure, and evaluator-consistency comparison are given in Appendix E.1.

6 Conclusion

We introduce **SuperResearch Bench**, a benchmark that couples *Super Wide Retrieval* and *Super Deep Investigation* across 300 expert-refined tasks. Evidence graphs and rubric-guided audits measure how research systems recover evidence, connect it to conclusions, answer task-linked questions, and calibrate competing claims. Automated checks and expert review support benchmark quality, with affected reports rescored after correction. The four-dimensional comparison of fourteen systems is complemented by Citation Health diagnostics and evaluator ablations. Together, these analyses expose different aspects of long-horizon research performance and provide a structured testbed for developing more capable research agents.

AI Use Statement

Generative AI tools assisted task candidate generation, reference-artifact construction, evaluation, and manuscript drafting. The authors are responsible for the final content, analyses, and artifacts.

REFERENCES

Anthropic. Claude 3.5 sonnet system card. https://www.anthropic.com/research/claude-3-5-sonnet-research, 2024. Accessed: 2026-01-26.

Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Chemcrow: Augmenting large-language models with chemistry tools. *arXiv preprint arXiv:2304.05376*, 2023.

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

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

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

Google. Gemini deep research agent: Autonomous information synthesis. https://ai.google.dev/gemini-api/docs/deep-research, 2025. Accessed: 2026-01-26.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645 (8081):633–638, September 2025. ISSN 1476-4687. doi: 10.1038/s41586-025-09422-z. URL http://dx.doi.org/10.1038/s41586-025-09422-z.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*, 2023.

Chen Hu, Haikuo Du, Heng Wang, Lin Lin, Mingrui Chen, Peng Liu, Ruihang Miao, Tianchi Yue, Wang You, Wei Ji, et al. Step-deepresearch technical report. *arXiv preprint arXiv:2512.20491*, 2025.

Zirui Jiang et al. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*, 2024.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569, 2024.

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

LangChain. Langgraph: Building stateful agents with langchain. `https://langchain-ai.github.io/langgraph/`, 2024. Accessed: 2026-01-26.

LangChain AI. Langchain. `https://github.com/langchain-ai/langchain`, 2022. Accessed: 2026-01-26.

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

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

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

Meta. Llama 3.3 model card. `https://github.com/meta-llama/llama-models/blob/main/models/llama3_3/MODEL_CARD.md`, 2024. Accessed: 2026-01-26.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

MiniMax. Minimax-m2.1: Scaling foundation models for agentic capabilities. `https://www.minimaxi.com/research/m2-1`, 2025. Accessed: 2026-01-26.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

OpenAI. Introducing deep research. `https://openai.com/index/introducing-deep-research/`, 2025. Accessed: 2026-01-26.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024.

Perplexity. Sonar deep research: The new standard for ai search. `https://www.perplexity.ai/hub/blog/sonar-deep-research`, 2025. Accessed: 2026-01-26.

Yujia Qin, Zihan Cai, Dian Jin, Lan Yan, Shihao Liang, Kunlun Zhu, Yankai Lin, Xu Han, Ning Ding, Huadong Wang, et al. Webcpm: Interactive web search for chinese long-form question answering. *arXiv preprint arXiv:2305.06849*, 2023.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

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

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

Tavily. Tavily api: Search engine optimized for ai agents. https://tavily.com/, 2024. Accessed: 2026-01-26.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. URL https://arxiv.org/abs/2507.20534.

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

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

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

Lingfei Wu, Yu Chen, Kai Shen, Xiaojie Guo, Hanning Gao, Shucheng Li, Jian Pei, and Bo Long. Graph neural networks for natural language processing: A survey. *Foundations and Trends in Machine Learning*, 16(2):119–328, 2023.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

xAI. Grok-4 technical report: Advancing speed and reasoning. `https://x.ai/blog/grok-4`, 2025. Accessed: 2026-01-26.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

A BENCHMARK COMPARISON

Table 2 compares SuperResearch Bench with representative agentic benchmarks in terms of task scope, reasoning depth, retrieval breadth, and evaluation criteria. The comparison highlights our emphasis on long-horizon research supported by evidence graphs and rubric-guided assessment.

Table 2: Comparison of SuperResearch Bench with representative agentic benchmarks.

Dimensions	Metric	GAIA (Level 3) Mialon et al. (2023)	WideSearch Wong et al. (2025)	DeepWideSearch Lan et al. (2025)	DeepResearch Bench Du et al. (2025)	Super Research (Ours)
Task Scope	Core Paradigm	General Assistant	Broad Info-Seeking	Deep & Wide Info-Seeking	Deep Research Agent	Super Research
	Task Goal	Precise Problem Solving	Structured Aggregation	Complex Table Synthesis	Analyst-Grade Report Gen.	**Strategic Planning & Discovery**
Depth	Steps / Iterations	10–40 (Variable)	$\sim$1.2 / Entity	$\sim$4.2 / Entity	10–20 (Iterative)	**$\sim$100 (Max 140+)**
	Methodology	Tool Orchestration	Parallel Execution	Recursive / Hierarchical	Hierarchical Synthesis	Structured Decomposition
Width	Web Pages	$<$ 20 (Targeted Search)	$\sim$44 (High Width)	$\sim$414 Units (High Vol.)	$\sim$110 (Hundreds)	**$\sim$600 (Max 1,200+)**
	Search Scope	Specific Factoid	Multi-Entity Coverage	Hidden Target Entities	PhD-Level Research Topic	Thematic Perspective Synthesis
Evaluation	Primary Method	Quasi-Exact Match	Hybrid (F1 & SR)	Multi-Granularity F1	Adaptive Ref-Based Judge	**Graph-Anchored Auditing**
	Focus Dimensions	Result Correctness	Completeness & Fidelity	Depth & Width Integration	Quality & Citation Accuracy	**Consistency, Bias & Coverage**

B MORE DETAILS ABOUT SUPERRESEARCH BENCHMARK.

B.1 DATA CURATION AND QUALITY AUDIT

Benchmark construction includes automated validation and expert review of reference Facts, graph dependencies, and evaluation questions. Reviewers check factual claims against their original sources, including the study subjects, conditions, comparisons, outcomes, and temporal scope. Graph review checks whether higher-level Insights and Conclusions are supported by their cited predecessor nodes, and identifies duplicated claims, unsupported generalizations, and conclusions that require qualification. Agreement across generated reports serves as a cue for further source checking, rather than independent verification.

During this process, 54 Facts across 35 tasks were revised into statements directly supported by their sources. Model-assisted entailment screening covered 629 revised Insight and Conclusion nodes, followed by corrections to substantive connection errors in nine tasks. Structural validation of the exported graphs found no directed cycles or illegal cross-layer edges. We also revised the answer options of 140 multiple-choice questions to reduce answer-length cues. Among these 140 questions, the number for which the correct answer was the longest option decreased from 95 to 16.

These checks and revisions form an integral part of benchmark construction. The resulting reference graphs and evaluation questions serve as the basis for rubric-guided report assessment, and the main results are computed using these audited artifacts.

B.2 PRINCIPLES FOR TASK GENERATION

Prompt: Cognitive-Rank Constrained Task Generation

Role: Senior Principal Investigator (NSF/ERC Level)
Objective: Propose high-complexity research programs for the domain: {domain}.
Cognitive-Rank Constraints (Negative Constraints):

- **NO "How-to" Guides:** Reject procedural or tutorial-style topics.
- **NO Fact Lookups:** Reject questions answerable by a single retrieval.
- **NO Settled Science:** Avoid topics with established consensus unless proposing a paradigm shift.

Generation Instructions:

1. **Traverse Topology:** Identify structural tensions and unresolved mechanisms.
2. **Multi-Layered Analysis:** Proposals must require synthesizing economic, technical, and ethical layers.
3. **Theoretical Density:** Generate more tasks for rapidly evolving fields.

B.3 System Prompts

B.3.1 Planner Agent

System Prompt: Planner Agent (Hierarchical Decomposition)

Role: Chief Research Architect
Task: According to the topic and expert advice. Create a MECE (Mutually Exclusive, Collectively Exhaustive) hierarchical research plan.
Root Question: {root_topic}
Structure Requirements:

1. **Phases:** Chronological or Logical progression (e.g., Background - Mechanism - Implication).
2. **Chapters:** Specific, actionable investigation areas.
3. **Queries:** 3-5 specific high-value search queries per chapter.

Output JSON Schema:

```
{
    "root_task": "{root_topic}",
    "phases": [
        {
            "phase_title": "Phase 1: ...",
            "chapters": [
                {
                    "chapter_title": "...",
                    "search_queries": ["query1", "query2"]
                }
            ]
        }
    ]
}
```

B.3.2 RESEARCHER AGENT

System Prompt: Researcher Agent (Sub-task Execution)

Role: Senior Research Analyst (Deep Research Agent)
Mission: Conduct a rigorous investigation on the specific chapter below.
Target:

- **Root Topic:** {root_topic}
- **Current Phase:** {task_obj.phase_title}
- **Current Chapter:** {task_obj.chapter_title}
- **Search Strategy:** {task_obj.question}

Context Awareness (DAG Memory): The following findings have already been established in previous chapters. **CRITICAL: Do NOT repeat general facts found below. Build upon them. Focus on "What's missing" or "Next order effects".**
{previous_context if previous_context else "[No previous context]"}

Output Requirements (Professional Standard):

1. **Format:** Markdown.
2. **Visuals, Formatting & References:**
   - **Markdown Tables & Figures:** Must have numbered captions. Cell content must be CONCISE.
   - **LaTeX Formulas:** Use standard LaTeX syntax. Append tags/IDs to important equations.
   - **Mermaid Diagrams:** Strictly use `graph LR` or `gantt`.

B.3.3 SUMMARIZER AGENT

System Prompt: Summarizer Agent (Context Transition)

Task: Summarize Section Context
Input Text: {input_text}

Goal: Write a concise summary (2-3 sentences) of the key findings and arguments in this section. This summary will be used by the writer of the NEXT section to ensure a smooth logical transition.

B.3.4 GRAPH CONSTRUCTION METHODOLOGY

The following prompts illustrate the pipeline for transforming unstructured text into a structured Research Graph.

Step 1: Construct Insight (Atomic Extraction)

Task: Identify Analytical Insights
Input Sentences: {batch_json}
Expert Advice: {}

Instruction: Following the Expert Advice, Review the sentences above from a research report. **Select ONLY** the sentences that represent **Analysis, Synthesis, Deductions, or Conclusions**.

- **REJECT:** Pure descriptions of data, procedural steps, or obvious definitions.
- **KEEP:** Sentences that connect dots, explain "why", or summarize trends/implications.

Output JSON Schema:

```
{ "selected": ["Exact text of insight 1", "Exact text of insight 2"] }
```

Step 2: Extract Residual Context (Missing Info)

Task: Extract Residual Useful Information from Research Report
Original Report Snippet: {task['report_text']}
Context: We have already extracted {facts_count} atomic facts/insights from this report into a database.

Goal: Capture the **Missing Information Context**. What important context, statistics, table, figure, latex or descriptive details are lost but is useful if we only keep the atomic facts/insights?

Instruction: Write a dense summary (1000 words) of the "Residual Useful Information". Focus on:

1. Figures or Tables or latex formulas which keep dense information.
2. Implicit tensions or qualitative descriptions not easily captured as "facts".
3. The evolution of the situation.

Output JSON: { `"dense_summary": "..."` }

Step 3: Construct Links (Logical Entailment)

Task: Build Logical Entailment Links
Scenario: We are analyzing the thought process at **Step {child_id}**. This step logically follows/relies on **Step(s) {parent_ids}**.

Nodes in Current Step (Findings): {c_json}
Nodes in Previous Step (Evidence): {p_json}

Instructions: Identify logical connections (`source - target`). **Allowed Link Types:**

1. **Internal Link (Self-Consistency):** Source and Target are BOTH in Current Step.
2. **Cross-Layer Link (Dependency):** Source is from Previous Step, Target is in Current Step.

Constraint: Arrows generally flow forward in logic (Evidence - Conclusion) or forward in time.

Output JSON Schema:

```
{
    "links": [
        { "source": "insight_1",
          "target": "insight_2", "relation": "inference" },
        { "source": "fact_1",
          "target": "insight_5", "relation": "supports" }
    ]
}
```

Step 4: Construct Global Insight (Multi-Source Synthesis)

Role: Chief Research Scientist (Meta-Synthesis Architecture)
Task: Multi-Source Pattern Recognition & Iterative Insight Generation
Input Data: Compilation of knowledge nodes (atomic facts OR high-level insights).

Instructions:

1. **Cluster & Converge:** Look for **clusters** of evidence where 3 or more disparate nodes point to a single, larger truth.
2. **Iterative Synthesis:**
   - Generate **5-15 Global Insights**.
   - Combine existing Insights to form a "Higher-Order Meta-Insight".
   - Each Global Insight must be a novel synthesis, not regurgitation.
3. **Reasoning Trace:** Articulate *how* these scattered nodes converge.
4. **Strict Citation:** Every Global Insight must be supported by a list of specific `supporting_node_ids`.

Output JSON Schema:

```
{
    "global_insights": [
        {
            "synthesis_logic":
                "We observe a convergence across [Node A]...",
            "content": "The finalized high-level conclusion...",
            "supporting_node_ids": [
                "chap_1_fact_5", "global_insight_2"]
        }
    ]
}
```

B.3.5 WRITER AGENT

System Prompt: Writer Agent (Report Generation)

Role: Senior Strategic Analyst (Think Tank Level)
Task: Write Chapter **{title}**
Input Data: Verified Quiz Facts, Knowledge Graph Nodes, and Narrative Context.

Writing Guidelines:

1. **Flow:** {flow_context}
2. **Deep & Dense:** Avoid "AI-style" lists. Write comprehensive paragraphs.
3. **Header Formatting:** Identify Section Number. Use `###` (H3) for subsections.
4. **Visuals:**
   - Use Python Plotting Agent (SciencePlots/Seaborn) for statistical data.
   - Use LaTeX for formulas with continuous numbering tags.

Negative Constraints:

- NO CITATIONS IN DRAFT (Tags are added later).
- NO CHINESE.
- NO BROKEN TABLES.

B.3.6 AUTOMATED EVALUATION METRIC CONSTRUCTION

Prompt: Exam QA Generation (Quiz)

Task: Quiz Generation
Requirements:

1. **Quantity & Format:** Generate exactly one question per step (MCQ, True/False, Fill-in-Blank).
2. **Self-Contained Context (CRITICAL):** The testing LLM has NO access to the original report. Must embed all context (dates, names) in the question stem.
3. **Depth Scoring (0-5):**
   - 0-1: Surface-level fact.
   - 4-5: Requires deep inference or cross-referencing.

Output JSON Schema:

```
{
    "quiz": [
        {
            "type": "multiple_choice",
            "question": "...",
            "options": ["A...", "B..."],
            "answer": "...",
            "depth_metric": 4
        }
    ]
}
```

Prompt: Dialectical Confidence Audit (Bias Calibration)

Task: Dialectical Confidence Audit Generation
Goal: Identify topics characterized by **interpretative ambiguity** or **balanced conflict** ("Gray Zones").

Rules for Formulation:

1. **Blind Audit Format:** Question must NOT reference report structure.
2. **Dialectical Framing:** Use "Thesis vs. Antithesis" logic.
3. **Hidden Consensus:** Test if respondent can identify inconclusive data.

Output Format:

```
{
    "question": "Regarding the net impact of 'Green Initiative'...",
    "options": [
        "Thesis: It is a success due to 30% reduction...",
        "Antithesis: Impact negated by 10% rise in total emissions..."
    ],
    "answer": [
        "Reason: Report confirms efficiency gains. Score: 6.",
        "Reason: Report acknowledges rebound effect. Score: 5."
    ],
    "type": "confidence_score"
}
```

B.4 TASK SELECTION

The SUPERRESEARCH benchmark is meticulously constructed to challenge the limits of autonomous research agents. We selected 300 professional research tasks across 10 high-stakes domains, ensuring a balanced distribution of theoretical complexity and empirical depth.

As illustrated in Figure 5, our dataset covers a wide spectrum of professional fields ranging from *Science & Engineering* and *Medical & Health* to specialized sectors like *Environment & Energy* and

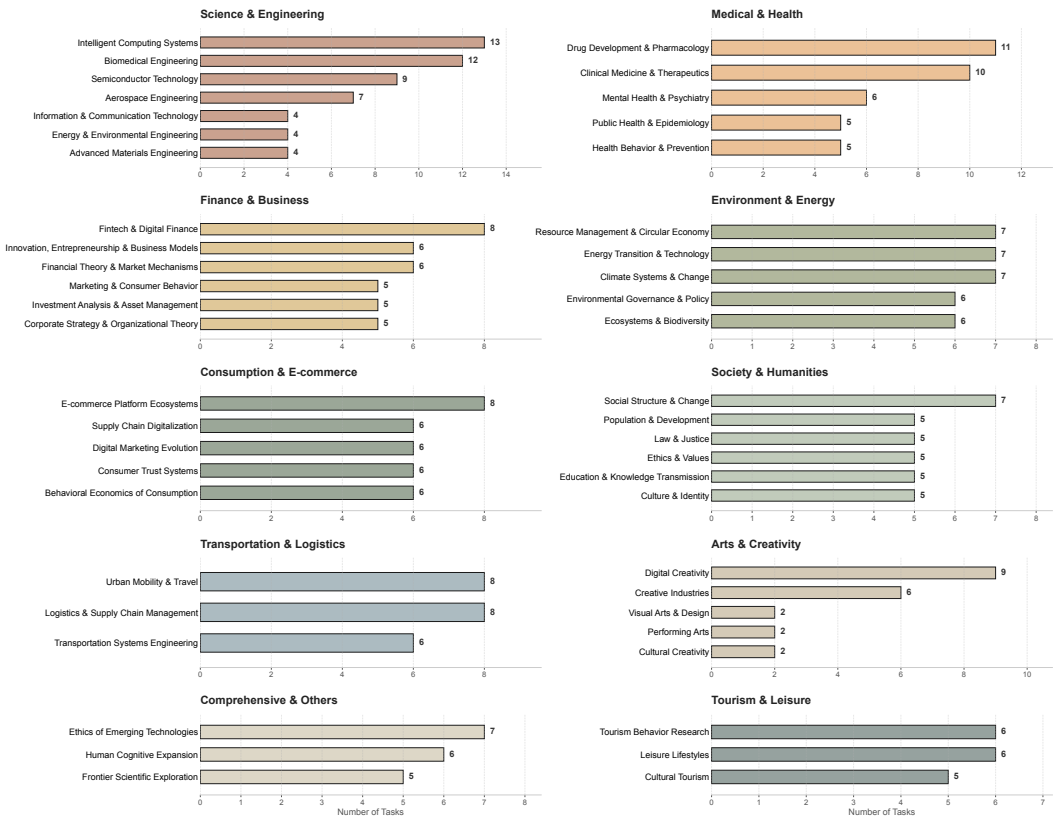

Figure 5: Detailed Task Distribution Across Subdomains. Supplementing the main text, this figure visualizes the granular distribution of tasks within each specific subdomain. This breakdown highlights the diversity of the SUPERRESEARCH benchmark, confirming its coverage across a wide spectrum of specialized professional fields.

Finance & Business. The task selection process, spearheaded by a panel of Ph.D. candidates and industry experts, follows a "ceiling-level" philosophy:

- **Cognitive Nuance:** We explicitly prioritized tasks containing conflicting evidence or unresolved scientific tensions, requiring agents to perform dialectical reasoning.
- **Structural Depth:** Unlike traditional QA benchmarks, each task in SUPERRESEARCH is designed with long-horizon dependencies, where finding a definitive conclusion necessitates traversing multiple levels of intermediate evidence.
- **High Entropy Landscapes:** The information environment for these tasks is intentionally unstructured, forcing agents to demonstrate superior information filtering and synthesis capabilities.

C MORE DETAILS ABOUT EVALUATION METRICS.

C.1 COVERAGE AND COMPREHENSION ($\mathcal{R}_{\text{WEIGHTED}}$)

Unlike flat recall metrics, Coverage and Comprehension emphasize the retrieval of high-value structural information. We assign a topological weight $w(v)$ to each node v in the Research Graph based on its hierarchical level (Atomic Fact, Key Insight, or Global Insight). Let $Children(v)$ be the set of nodes supporting v. The weight is defined recursively to reflect derivation depth:

$$w(v) = \begin{cases} 1 & \text{if } v \text{ is a leaf (Atomic Fact level)} \\ 1 + \max_{u \in Children(v)} w(u) & \text{otherwise (Insight levels)} \end{cases} \tag{1}$$

The Depth-Weighted Recall is calculated as:

$$\mathcal{R}_{\text{weighted}} = \frac{\sum_{v \in \mathcal{V}_{hit}} w(v)}{\sum_{v \in \mathcal{V}_{total}} w(v)} \times 100 \tag{2}$$

where $\mathcal{V}_{total}$ is the set of all graph nodes, and $\mathcal{V}_{hit}$ is the subset of nodes semantically verified as present in the generated report by an LLM judge.

C.2 LOGICAL CONSISTENCY ($\mathcal{C}_{\text{LOGIC}}$)

Logical Consistency evaluates the integrity of the reasoning chains. Let $\mathcal{G}$ be the set of Global Conclusion nodes (roots). A global node $g \in \mathcal{G}$ is considered *supported* if and only if there exists a valid, unbroken citation path in the recovered subgraph from g down to a recovered Atomic Fact node. We formally define $\mathcal{C}_{\text{logic}}$ as the product of the valid inference ratio and overall graph coverage:

$$\mathcal{C}_{\text{logic}} = \left(\frac{|\mathcal{G}_{supported}|}{|\mathcal{G}_{hit}| + \epsilon} \times \frac{|\mathcal{G}_{hit}|}{|\mathcal{G}_{total}|} \right) \times 100 \tag{3}$$

where ϵ is a small constant to prevent division by zero. This multiplicative formulation penalizes reports that hallucinate correct high-level answers without the necessary supporting evidence.

C.3 REPORT UTILITY ($\mathcal{U}_{\text{QA}}$)

We quantify report-conditioned question-answering performance using a set of N exam questions $Q = \{q_1, ..., q_N\}$ derived from the reference artifacts. Let a_i be the exam-taker's answer to question q_i after receiving the generated report as context, and $a_i^\star$ be the reference answer.

$$\mathcal{U}_{\text{qa}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(a_i \approx a_i^\star) \times 100 \tag{4}$$

where $\mathbb{I}(\cdot)$ is the semantic equivalence indicator function evaluated by the LLM judge.

C.4 OBJECTIVITY SCORE ($\mathcal{O}_{\text{BIAS}}$)

To measure the Calibration of Stance, we define a Thesis A and Antithesis B for controversial topics. Let $S_A, S_B \in [0, 10]$ be the evidentiary support scores extracted from the generated report by a double-blind auditor, and GT_A, GT_B be the ground-truth support levels derived from the graph. The calibration error E is:

$$E = |S_A - GT_A| + |S_B - GT_B| \tag{5}$$

The final Objectivity Score is derived as:

$$\mathcal{O}_{\text{bias}} = \max(0, 100 - E \times \lambda) \tag{6}$$

The factor is fixed by the score range: two 0–10 stance ratings have maximum total absolute error 20, so $\lambda = 100/20 = 5$ maps that error to a 100-point deduction.

C.5 CITATION HEALTH

Citation Health serves as a diagnostic suite to detect Single-Source Dependency. We compute two specific indicators based on the citation distribution C. Let n be the total number of unique sources cited, and p_i be the proportion of citations attributed to source i.

Source Dominance ($\mathcal{D}_{\text{src}}$). This metric checks for volume imbalance. We utilize normalized Shannon Entropy (H_{norm}) to measure the "peakedness" of the citation distribution:

$$H(C) = -\sum_{i=1}^{n} p_i \log_2 p_i, \quad \mathcal{D}_{\text{src}} = \left(1 - \frac{H(C)}{\log_2 n}\right) \times 100 \tag{7}$$

A high $\mathcal{D}_{\text{src}}$ indicates that the report relies disproportionately on a single source for its information volume.

Narrative Monopolization ($\mathcal{M}_{\text{mono}}$). This metric evaluates the spatial layout of citations. Let K be the total number of distinct sections in the report, and k_i be the number of sections where source i appears. We measure the maximum narrative coverage:

$$\mathcal{M}_{\text{mono}} = \max_i \left(\frac{k_i}{K}\right) \times 100 \tag{8}$$

A high $\mathcal{M}_{\text{mono}}$ suggests that a single source drives the narrative flow across the entire document, rather than being synthesized with others.

Table 3: Coverage and Comprehension Breakdown. Detailed performance analysis of **Coverage & Comprehension** across different knowledge granularities. Evaluation covers three architectural paradigms: Deep Research System, Native Search-Integrated Agent, and Search-Augmented Baseline. **Coverage** ($\mathcal{R}_{\text{weighted}}$) measures overall information coverage using graph-depth weights across Atomic Facts, Key Insights, and Global Conclusions. **Bold** and underline denote the best and second-best performance, respectively; ↑ indicates that higher values are preferred.

Method	Knowledge Recall Breakdown			Overall
	Atomic Facts ↑	Key Insights ↑	Global Conclusions ↑	Coverage ($\mathcal{R}_{\text{weighted}}$) ↑
Deep Research System				
Gemini Deep Research	30.85	71.88	69.96	60.29
Sonar Deep Research	**35.79**	74.12	<u>75.21</u>	<u>64.55</u>
Tongyi Deep Research	17.96	56.56	57.25	46.77
o3 Deep Research	19.40	53.99	50.62	43.47
o4 Mini Deep Research	19.46	53.25	50.21	43.07
Native Search-Integrated Agent				
Perplexity Pro Search	26.86	<u>75.03</u>	74.88	62.37
Kimi-k2	18.49	56.62	55.04	45.94
Grok-4-1-fast	22.16	57.54	56.37	47.77
Search-Augmented Baseline				
Seed 2.1 Turbo + web tools	<u>34.72</u>	**75.68**	**76.58**	**65.31**
Deepseek-r1	19.72	54.82	54.71	45.54
Claude-3.5-sonnet	10.96	41.83	41.42	33.56
Minimax-m2.1	27.05	62.94	63.92	53.96
Llama-3.3-70B	8.63	32.38	28.79	24.60
Qwen-2.5-72B	14.04	41.27	38.79	33.03

D More Experiment Results.

D.1 Coverage and Comprehension Details

The breakdown below uses the 300 questions scored with the reference graphs for all fourteen systems ($n = 4200$ reports). Each knowledge-granularity column is the mean node-coverage percentage within a task, averaged across the 300 tasks; Overall Coverage is computed separately using the graph-depth weights.

D.2 Sensitivity Analysis Details

D.2.1 Responsiveness Rate Formulation

To quantify the sensitivity of an evaluator, we define the **Responsiveness Rate (RR)** as the probability that the evaluator's score shifts in the correct direction following a controlled perturbation.

Let $\mathcal{D} = \{r_1, r_2, ..., r_N\}$ be the set of original reports derived from N experiment instances (where $N = 120$ in our setup, comprising 12 models $\times$ 10 tasks). For each report r_i, we generate two perturbed variants: a degraded version r_i^- (with fact removal) and an improved version r_i^+ (with fact injection).

Let $S(r)$ denote the score assigned by an evaluator to report r. We define the responsiveness indicator function $\mathbb{I}(\cdot)$ for the two scenarios:

Degradation Responsiveness (RR_{deg}). In the degradation scenario, a responsive evaluator should penalize the quality drop, i.e., $S(r_i) > S(r_i^-)$. The rate is calculated as:

$$RR_{deg} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(S(r_i) - S(r_i^-) > \epsilon\right) \times 100\% \tag{9}$$

Improvement Responsiveness (RR_{imp}). In the improvement scenario, a responsive evaluator should reward the quality gain, i.e., $S(r_i^+) > S(r_i)$. The rate is calculated as:

$$RR_{imp} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left(S(r_i^+) - S(r_i) > \epsilon\right) \times 100\% \tag{10}$$

where ϵ is a minimal threshold (set to 0 in our experiments) to strictly enforce directional correctness.

D.2.2 Score Normalization and Comparison

To ensure a fair comparison between the Graph Metric (scale 0-100) and the LLM Judge (typically scale 0-10), we normalized the scalar outputs during the delta calculation. Specifically, for the LLM Judge, we applied a scaling factor of $\alpha = 10$ before computing the differentials: $\Delta_{LLM} = (S_{LLM}^{org} - S_{LLM}^{pert}) \times 10$.

This normalization ensures that the magnitude of score changes is comparable in the sensitivity distribution analysis, although the Responsiveness Rate itself is scale-invariant as it only measures the *directionality* of the change.

D.2.3 Perturbation Methodology

The controlled perturbation was conducted via a semi-automated pipeline to ensure ground-truth validity:

1. **Degradation (Fact Removal):** We explicitly removed 1$\sim$3 Atomic Facts (nodes in the research graph) or Logical Links from the report's reasoning section. The expected outcome is a decrease in the *Breadth/Recall* or *Depth/CoT* metric.
2. **Improvement (Fact Injection):** We retrieved 1$\sim$3 high-relevance facts from the ground-truth Knowledge Graph that were originally missing from the report and injected them into the appropriate context. The expected outcome is an increase in the corresponding metric.

D.3 Cross-Domain Performance Analysis

To examine variation in research-system performance across professional contexts, we compare fourteen systems across 10 distinct domains, ranging from technical fields like *Science & Engineering* and *Medical & Health* to social and creative dimensions such as *Society & Humanities* and *Arts & Creativity*.

Figure 7 shows that relative system performance varies across domains. Perplexity Deep Research (Sonar) achieves the highest displayed scores in *Medical & Health* and *Finance & Business*, while Seed 2.1 Turbo leads in *Transportation & Logistics*. Perplexity Pro Search leads in *Consumption & E-commerce* and *Tourism & Leisure*. These differences highlight the value of domain-level analysis alongside aggregate scores.

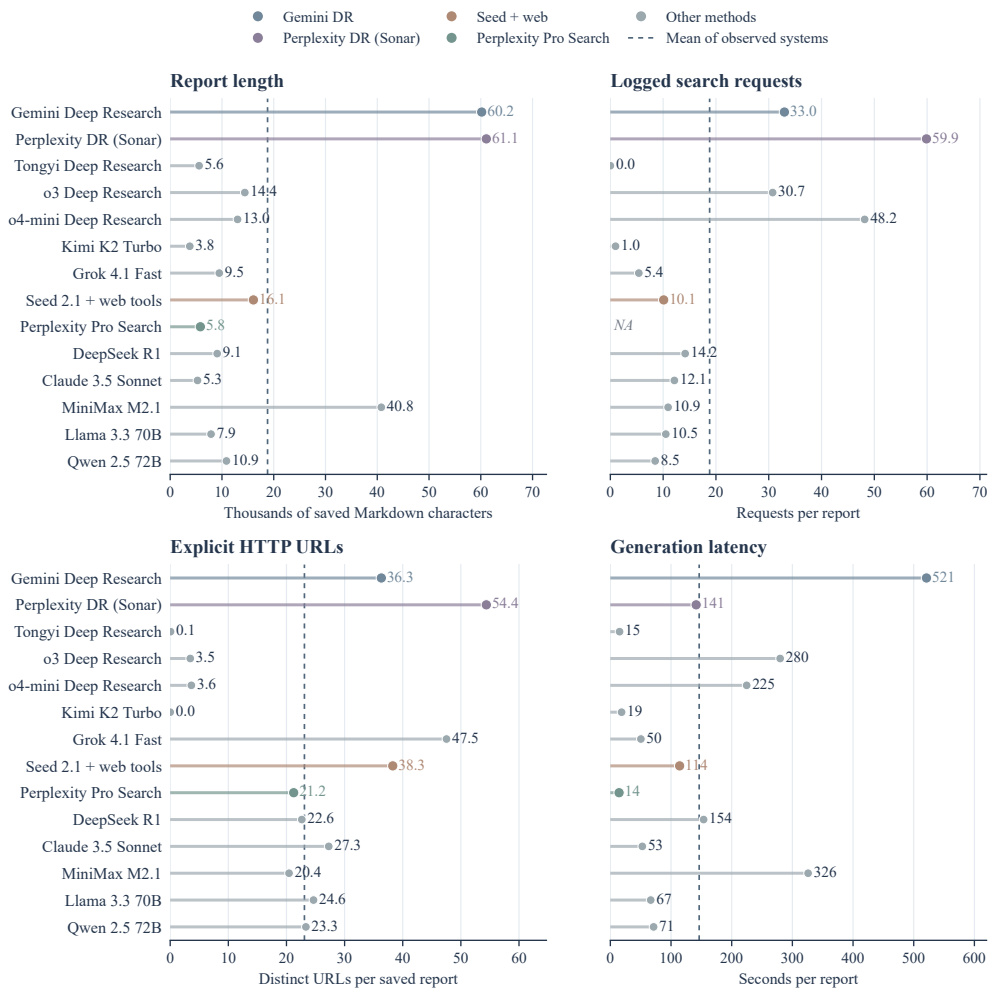

Figure 6: Operational Characteristics of Fourteen Research Systems. Panels show mean report length (thousands of saved Markdown characters), logged search requests per report, distinct explicit HTTP URLs per saved report, and generation latency (seconds per report). Vertical dashed lines indicate the mean across systems with recorded values for each metric. NA denotes an unrecorded metric; a recorded zero does not establish that the system lacks the corresponding capability.

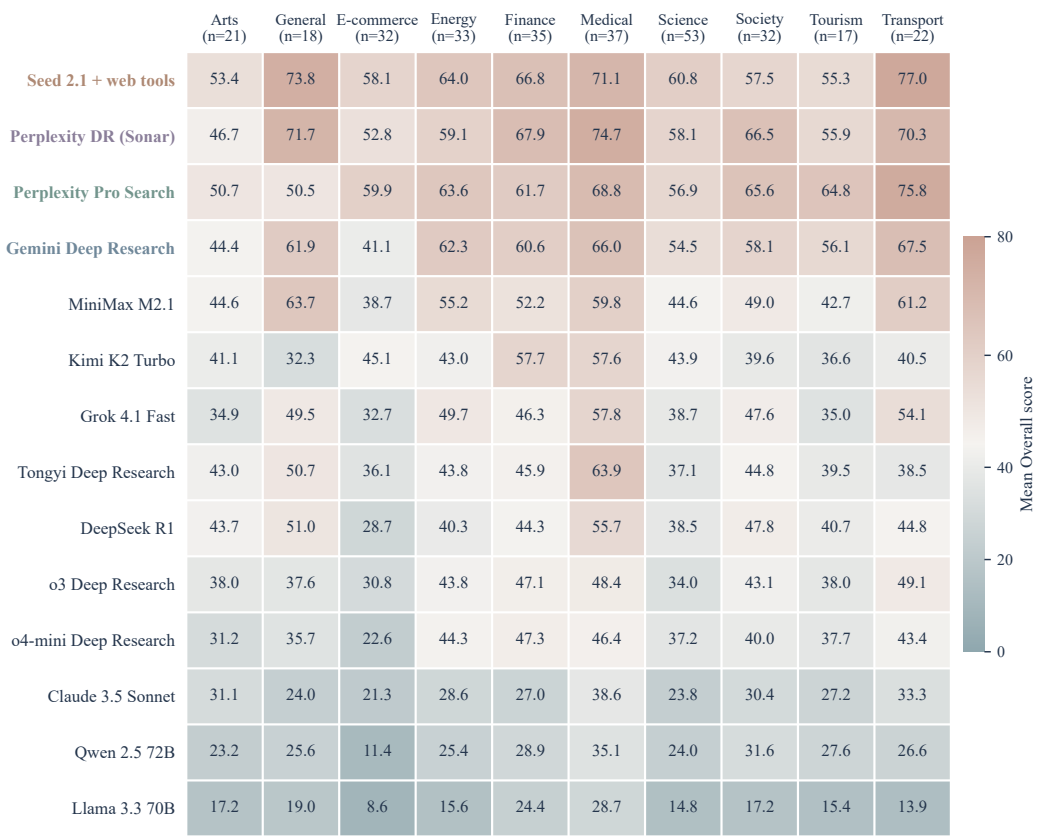

Figure 7: Cross-Domain Performance of Fourteen Research Systems. The heatmap shows mean Overall scores across 10 domains, with systems arranged by row and domains by column. Darker shading indicates higher scores.

D.4 Ablation Study Prompts

To validate the responsiveness of our metrics, we employed controlled perturbation experiments. The following prompts were used to artificially degrade or improve report quality, as well as to conduct the baseline "LLM Judge" evaluation.

Prompt: Report Degradation (Malicious Editor)

System Prompt: You are a malicious editor. Your goal is to degrade the report quality.
User Prompt: Original Report: {report_text}

Task: REWRITE the report to specifically **REMOVE 1∼3 facts of evidence**.

- The report must remain readable and coherent, but these specific points must be gone.
- For sections not directly related to Facts, please copy them in full, and only make surgical adjustments to the relevant parts.

Output: ONLY the rewritten report text.

Prompt: Report Improvement (Helpful Editor)

System Prompt: You are a helpful editor. Your goal is to improve the report.
User Prompt: Original Report: {report_text}

Task: REWRITE the report to **INSERT the following missing facts**. Integrate them naturally into relevant sections.

Facts to INSERT: {facts_list}

- For sections not directly related to Facts, please copy them in full, and only make surgical adjustments to the relevant parts.

Output: ONLY the rewritten report text.

Prompt: LLM Direct Judge (Baseline Evaluator)

Role: Expert Research Evaluator (Senior Research Analyst)
User Question: {root_question}
Report Content: {report_text}...

Task: Rate this report on two dimensions (0-10 scale):

1. **Breadth (Recall):** How comprehensively does it cover the necessary key facts?
2. **Depth (CoT/Logic):** How well does it explain the logical connections and underlying mechanisms?

Output JSON:

```
{
    "breadth_score": float,
    "depth_score": float,
    "reason": "string"
}
```

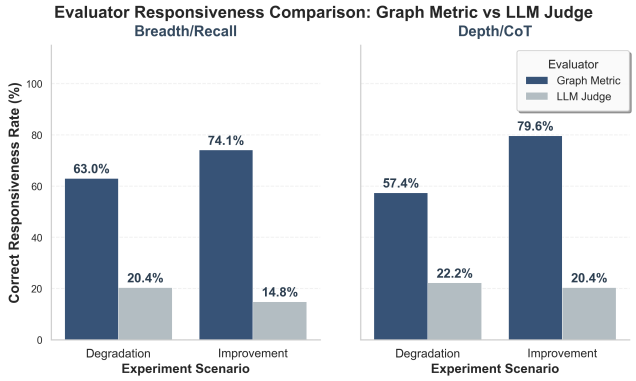

Figure 8: Evaluation Sensitivity Analysis. Compared to the LLM Judge (grey), our Graph Metric (dark blue) shows superior responsiveness to quality fluctuations. It aligns with quality trends—significantly increasing for optimized reports and decreasing for degraded ones—whereas the LLM Judge exhibits higher bias and lower sensitivity.

Prompt: Comprehensive Quality Evaluation (5-Metric Baseline)

Role: Expert Research Evaluator
Task: Evaluate the quality of a research report generated in response to a specific user query.

Input Report: {report}

Evaluation Metrics (Scale 0.0 - 10.0):

1. **Depth:** Does it provide deep technical insights and causal analysis?
2. **Breadth:** Does it cover all necessary aspects and sub-topics?
3. **Utility:** Is it practically useful and directly answers the intent?
4. **Citation Quality:** Are citations frequent, relevant, and well-distributed?
5. **Objective Score:** Does it provide multiple perspectives on controversial issues?

Output JSON:

```
{
    "depth_score": <float>,
    "breadth_score": <float>,
    "utility_score": <float>,
    "citation_quality_score": <float>,
    "objective_score": <float>
}
```

E Evaluator Ablations

E.1 Ablation Study

E.1.1 Sensitivity Analysis: Graph-Based Metrics vs. LLM-as-a-Judge

To validate the reliability, we conducted a comparative study against the standard "LLM Judge" baseline using 120 reports randomly generated from different tasks and models. We designed a controlled perturbation experiment to measure evaluator responsiveness in two scenarios: ① *Degradation:* Removing 1∼3 key facts or insights from the original report; ② *Improvement:* Injecting 1∼3 relevant facts into appropriate sections (Details in D.4). We computed the *Responsiveness Rate* (formal definition in Appendix D.2), defined as the percentage of instances where the metric's score correctly shifts in alignment with the perturbation (i.e., decreases for degradation and increases for improvement).

Table 4: Evaluator Consistency Analysis. Cross-evaluator score variation on the 120-report sample from the earlier experiment. The four-score mean uses equal weights; σ_{norm} expresses the standard deviation as a percentage of the score range (lower is better).

Metric	Evaluator	Norm STD (σ_{norm}) ↓
Coverage ($\mathcal{R}_{\text{weighted}}$)	Ours Graph	4.42%
	LLM Judge	**3.88%**
Consistency ($\mathcal{C}_{\text{logic}}$)	Ours Graph	**0.83%**
	LLM Judge	4.69%
Utility ($\mathcal{U}_{\text{qa}}$)	Ours Q&A	**2.05%**
	LLM Judge	4.37%
Objectivity ($\mathcal{O}_{\text{bias}}$)	Ours Q&A	**1.07%**
	LLM Judge	7.48%
Four-score mean	Ours Method	**1.67%**
	LLM Judge	3.02%

The experimental results, as illustrated in Figure 8, highlight a significant and observable performance gap between the methods. The **Graph Metric** demonstrated exceptional sensitivity to qualitative changes, consistently achieving responsiveness rates ranging from 57.4% to 79.6% across both Breadth and Depth dimensions. In contrast, the **LLM Judge** struggled to distinguish these fine-grained perturbations, as evidenced by its responsiveness rates fluctuating widely between 14.8% and 22.2%. This demonstrates the robustness of the graph metric.

E.1.2 Evaluator Consistency

To assess consistency across evaluator models, we retain an earlier evaluation on a random sample of 120 reports. Three models (Gemini-2.5-Flash, GPT-4o-mini, and Qwen2.5-72B-Instruct) served as direct LLM judges (Appendix D.4) and as parsers for our graph-based evaluation. This experiment examines how scores vary with the choice of evaluator model. We employ range-normalized standard deviation (σ_{norm}) to compare consistency across metrics with different scales (e.g., 0-300 for Graph metrics vs. 1-10 for LLM judges). As shown in Table 4, our evaluation procedure exhibits lower normalized score variation across most metrics on this report sample.

Across evaluator models, the graph-linked procedure has lower normalized score variation for Consistency (0.83% versus 4.69%), Utility (2.05% versus 4.37%), Objectivity (1.07% versus 7.48%), and the four-score mean (1.67% versus 3.02%). Coverage is the exception, with lower variation for the LLM Judge (3.88% versus 4.42%). These results characterize cross-evaluator consistency on this report sample and complement the controlled perturbation analysis in Figure 8.

F CASE STUDY

In this section, we provide detailed qualitative analyses of representative failure modes identified in our main results. These cases illustrate the specific behavioral patterns underpinning the quantitative metrics.

F.1 CASE STUDY: THE COMPLEXITY TRAP AND LOGIC BOTTLENECK

Figure 9 presents a case study illustrating how models struggle with the complexity of such queries, leading to significant logic bottlenecks.
Query: "How are geometry and mechanical balance explored through spatial dynamics in ballet choreography?"

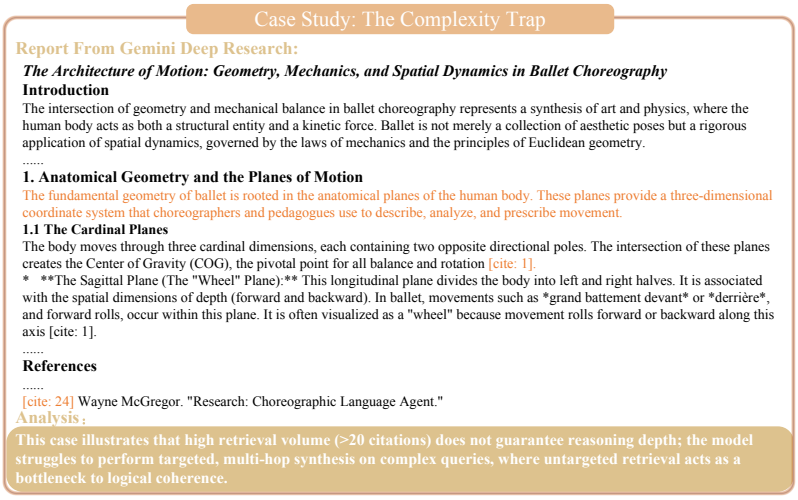

Figure 9: Case Study: The Complexity Trap

F.2 CASE STUDY: SINGLE-SOURCE REGRESSION

Figure 10 illustrates a typical instance of single-source regression, where the model's narrative is monopolized by a limited set of references despite the query's complexity. **Query:** "What represents the theoretical limit for blockchain Byzantine Fault Tolerance mechanisms when balancing system throughput and security against network latency and message complexity?"

F.3 CASE STUDY: THE ILLUSION OF DIVERSITY VIA SPATIAL OVERLOAD

Figure 11 depicts a critical utilization gap, where the model lists numerous sources to imply comprehensive coverage but fails to integrate them into the narrative, effectively creating an illusion of diversity. **Query:** "How do digital health interventions for type 2 diabetes balance clinical HbA1c monitoring with remote guidance to maintain adherence to lifestyle changes?"

F.4 CASE STUDY: RELIABILITY PARADOX VIA DEFENSIVE SUMMARIZATION

Figure 12 highlights the defensive summarization paradox, where the model employs a high density of citations to validate vague generalizations, resulting in technically 'safe' but shallow content that lacks specific, actionable detail. **Query:** "Is it possible for MEMS sensors to maintain long-term resonant frequency stability and detection accuracy despite the combined effects of process deviations and environmental vibration?"

Case Study: Single-Source Regression

Report From Llama-3.3-70B:

Theoretical Limit for Blockchain Byzantine Fault Tolerance Mechanisms: Balancing System Throughput and Security against Network Latency and Message Complexity

.....

Introduction

Blockchain systems rely on consensus algorithms to ensure the integrity and security of transactions. BFT mechanisms, such as Practical Byzantine Fault Tolerance (PBFT) [1], are designed to tolerate Byzantine faults, which occur when a node in the network behaves arbitrarily or maliciously. However, BFT mechanisms face significant challenges in scaling to meet the demands of large-scale blockchain systems.

.....

Analysis

Our analysis reveals that the theoretical limit for BFT mechanisms is bounded by the following constraints:

* **Scalability**: The number of nodes in the system, n, is limited by the message complexity, $O(n^2)$, which makes it difficult to scale to large-scale systems.

* **Security**: The security of the system, s, is limited by the number of Byzantine faults, f, which can be at most $n/3$ for PBFT [1].

.....

References

[1] M. Castro and B. Liskov, "Practical Byzantine Fault Tolerance," Technical Report, 1999.

[2] L. Lamport, R. Shostak, and M. Pease, "The Byzantine Generals' Problem," ACM Transactions on Programming Languages and Systems, vol. 4, no. 3, pp. 382-401, 1982.

Analysis :

This case exhibits a severe Over-reliance failure, relying on only two references while allowing a single source to monopolize the narrative. This structural flaw causes the model to regress from diverse synthesis to the mere summarization of a low-relevance document.

Figure 10: Case Study: Single-Source Regression

Case Study: The Spatial Overload

Report From DeepSeek-R1:

Research Report: Balancing HbA\textsubscript{1c} Monitoring and Remote Guidance in Digital Health .....

1. Introduction

Digital health interventions (DHIs) for type 2 diabetes (T2D) leverage technologies such as mobile apps, artificial intelligence (AI), and remote coaching to enhance diabetes self-management education and support (DSMES). These tools aim to improve glycemic control (measured by HbA\textsubscript{1c}), weight management, and cardiovascular outcomes while ensuring sustained patient adherence to lifestyle changes. This report synthesizes evidence on how DHIs integrate clinical HbA\textsubscript{1c} monitoring with remote guidance to maintain long-term adherence, based exclusively on provided context.

.....

4. Remote Guidance for Lifestyle Adherence

4.1. Behavior Change Techniques (BCTs)

- **Effective BCTs**: "Action planning" and "self-monitoring of behavioral outcomes" are most impactful for HbA\textsubscript{1c} reduction [4].

- **Personalization**: AI-tailored coaching increases cognitive eating restraints and physical activity, with non-*FTO* risk allele carriers showing enhanced responses [2].

.....

References

[1] Diabetes Care. *Facilitating Behavior Change and Well-being* (2022).

[2] Nature. *Digital Lifestyle Management* (2023).

.....

[11] PMC. *AI-led DPP* (2024).

Analysis :

This case exposes a critical utilization gap: despite listing 11 distinct sources to imply comprehensive coverage, the model fails to integrate them into the narrative, leaving some references (e.g., [1]) entirely unused while repeatedly recycling a narrower subset.

Figure 11: Case Study: The Spatial Overload

Case Study: Defensive Summarization Paradox

Report From O3 Deep Research:

Maintaining MEMS Sensor Stability Under Variations

.....

Active Compensation and Long-Term Stability

For ultimate stability, MEMS sensors may incorporate **active compensation** techniques (in electronics or control loops) to continually counteract drift. One example is **temperature compensation**: since temperature changes can alter resonance, some high-end MEMS resonators include on-chip heaters and control circuits to hold the device at a constant temperature (an "ovenized" MEMS). Oven-controlled MEMS resonators have achieved frequency stabilities on the order of a few *parts per billion* (ppb) over wide temperature ranges[cite: 11, 12]

. This level of stability was once exclusive to precision quartz devices. Similarly, feedback loops can adjust bias voltages or tuning elements in real-time to correct any slow frequency drift. Emerging research even uses the **nonlinear dynamics of MEMS resonators to self-correct drift**, actively suppressing frequency changes by adjusting the drive amplitude (taking advantage of the Duffing effect)[cite: 13, 14]

. With such techniques, a MEMS resonator recently demonstrated an ultra-stable frequency holding within ±14 ppb over varying conditions[cite: 15, 16]

. This shows that **given the right compensatory measures, MEMS sensors can maintain extraordinarily stable performance long-term**.

.....

Sources

.....

22. [mdpi.com](https://www.mdpi.com/1424-8220/16/4/468#:~:text=A%20dual,an%20improved%20mode%20ordering%20by)

Analysis :

Despite a high density of citations (e.g., [11–16]), the text employs them merely as decorative validation for vague generalizations like 'some high-end resonators.' This confirms a defensive strategy where technically 'safe' but shallow content masks a critical lack of specific, actionable detail.

Figure 12: Case Study: Defensive Summarization Paradox

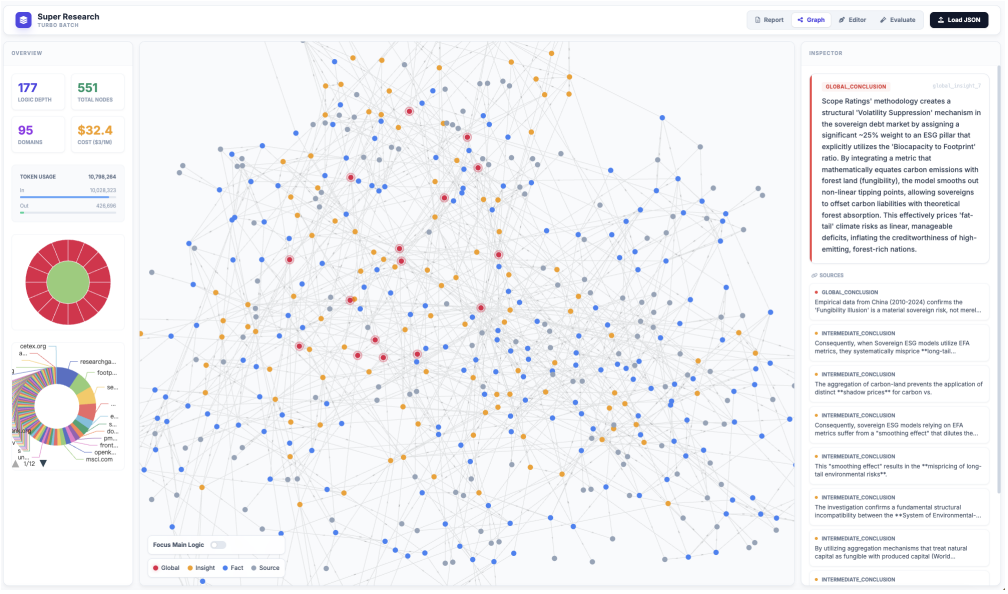

Figure 13: Interactive Research Graph Visualization. This interface projects the reasoning logic into a structured topology. (1) **Hierarchical Nodes**: Nodes are color-coded by abstraction level: Red (Global Conclusions), Orange (Intermediate Insights), and Blue (Atomic Facts). (2) **Inspector Panel**: The right sidebar details the specific textual content and source linkage for any selected node (e.g., "GLOBAL_CONCLUSION"), exposing the semantic connections.

G Additional Materials

G.1 Web-based Analysis Tool

To facilitate inspection and human-in-the-loop verification, the **Super Research** platform supports graph projection (Figure 13), question editing (Figure 14), and final metric auditing (Figure 15). The complete set of benchmark questions is listed in Appendix I.

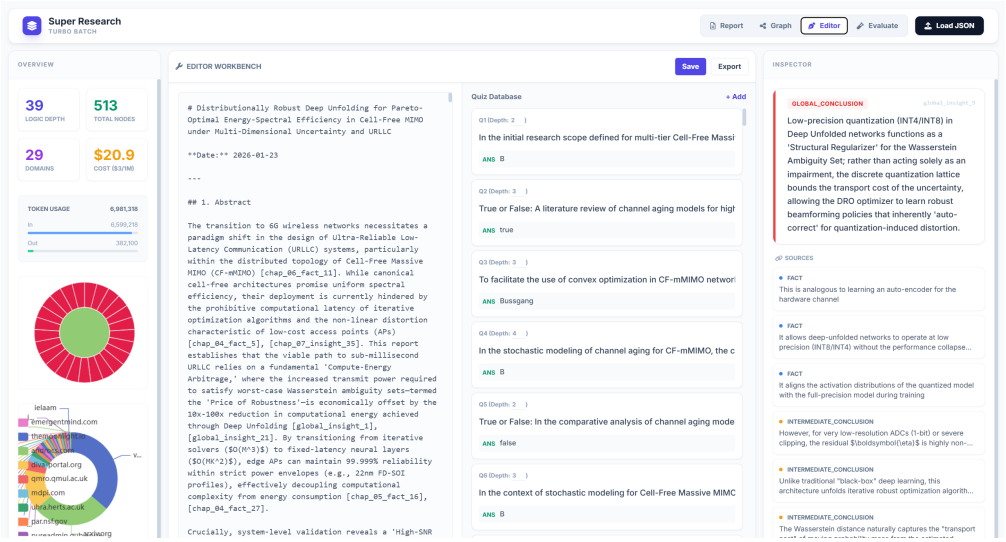

Figure 14: Split-View Evaluation Editor. To ensure the validity of automated metrics, this interface couples the report content with the evaluation set. (1) **Editor Workbench**: The left panel displays the raw Markdown of the report. (2) **Quiz Database**: The right panel lists generated exam questions (e.g., Q1, Q2), tagged by ”Depth”. Domain experts can use this view to intervene, editing or adding questions to correct alignment errors before the final audit.

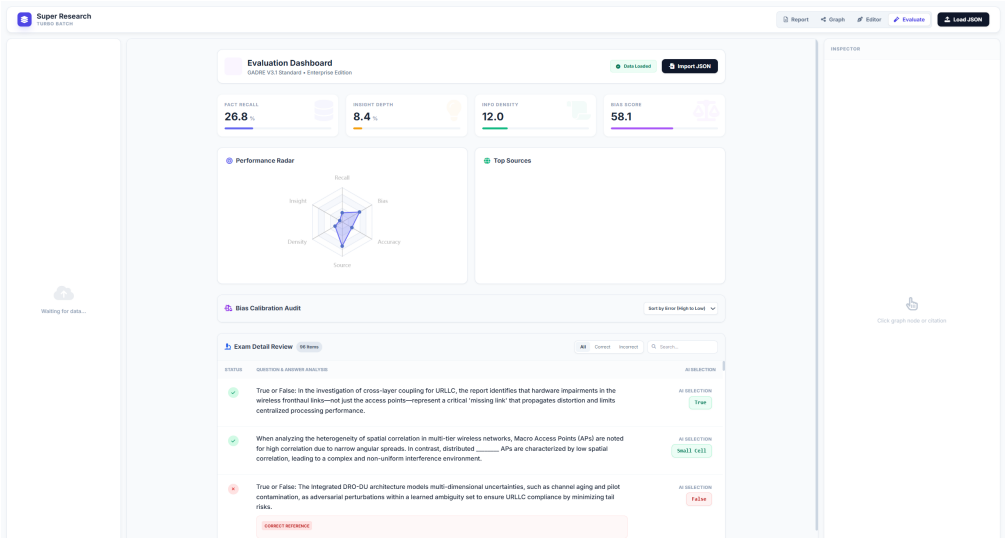

Figure 15: Visualization of the Evaluation Interface. This figure displays a comprehensive breakdown of metric scores, a radar chart for holistic comparison, and detailed performance results from the Exam QA.

Table 5: Spearman correlations and inverse-redundancy diagnostic weights on the matched 300-task, 14-system panel ($n = 4200$). The main Overall score uses equal weights.

	Cov	Con	Util	Obj	Main wt.	Diag. wt.
Cov	1.000	0.814	0.690	0.603	25.00%	20.99%
Con	0.814	1.000	0.547	0.440	25.00%	24.54%
Util	0.690	0.547	1.000	0.489	25.00%	25.61%
Obj	0.603	0.440	0.489	1.000	25.00%	28.86%

H Aggregation and Statistical Analysis

Score aggregation. Table 1 reports Overall as the equally weighted average of the four primary dimensions for fourteen systems. Citation Health is reported separately and is not included in Overall. The historical twelve-system paired analysis below uses a separate weighted score, with inverse-redundancy weights derived from correlations among its complete records.

Paired comparison. Let T_m be the set of task keys with all four percentage-valued quality scores observed for system m. The paired panel is $T = \bigcap_{m=1}^{12} T_m$, with $|T| = 173$. For each (m, t) in this panel, define $S_{mt} = w_R R_{mt} + w_C C_{mt} + w_U U_{mt} + w_O O_{mt}$. Friedman uses the twelve vectors $\{S_{mt} : t \in T\}$. The Sonar–Gemini confidence interval resamples the 173 paired task differences with replacement 10,000 times using seed 2027 and takes the 2.5th and 97.5th percentiles of the resampled means. Spearman correlations are computed on records with all four percentage-valued scores.

H.1 Additional Evaluation Analyses

Table 5 reports Spearman correlations among the four primary evaluation dimensions across 4,200 system–task records from fourteen systems and 300 tasks. Overall assigns an equal weight of 25% to each dimension. The inverse-redundancy weights are reported as a diagnostic and are not used to compute the main Overall scores.

Coverage and Consistency exhibit the strongest pairwise correlation ($\rho = 0.814$). Coverage and Utility are also positively correlated ($\rho = 0.690$), while Utility and Objectivity show a lower correlation ($\rho = 0.489$). These results indicate that the four dimensions capture related but non-identical aspects of report quality.

I BENCHMARK QUESTIONS

Table 6 lists the 300 questions supplied to research systems. Question IDs enumerate the three 100-task batches in benchmark order.

Table 6: Research questions in SuperResearch Bench.

ID	Domain and research question
1	*Science & Engineering.* In wireless communication systems, what methods enable Multiple-Input Multiple-Output (MIMO) antenna arrays to optimize beamforming algorithm robustness despite imperfect channel state information?
2	*Science & Engineering.* During plastic deformation, in what way does the local chemical ordering structure of high-entropy alloys influence the interaction mechanism between dislocation motion and twin boundary energy?
3	*Science & Engineering.* When scaling the gate length of FinFET devices, what design constraints must be balanced to manage the conflict between drive current enhancement and leakage current suppression amidst quantum and short-channel effects?
4	*Science & Engineering.* Given manufacturing process variations and temperature drift, what methods ensure that silicon photonics waveguide loss remains low enough to maintain signal integrity in high-bandwidth optical interconnects?
5	*Science & Engineering.* In neuromorphic chips, what synaptic weight update mechanism achieves parallel optimization of energy efficiency and learning accuracy while constrained by device variation and power limits?
6	*Science & Engineering.* When training large-scale deep learning models, what gradient backpropagation strategy optimizes convergence stability and generalization while balancing numerical precision against resource consumption?
7	*Science & Engineering.* With constraints on coherence time and error correction, what determines the optimization of logical depth and precision in quantum computing simulators via gate operation fidelity?
8	*Science & Engineering.* What represents the theoretical limit for blockchain Byzantine Fault Tolerance mechanisms when balancing system throughput and security against network latency and message complexity?
9	*Science & Engineering.* In neuromorphic computing, how can synaptic plasticity be implemented to balance the energy efficiency of bio-inspired systems with computational expressiveness, considering physical device limitations?
10	*Science & Engineering.* How does controlling the size distribution of nanomedicine carriers optimize targeted delivery stability and permeability amidst the competition between serum protein adsorption and cell uptake?
11	*Science & Engineering.* What controls the drug elution rate in cardiovascular stents to achieve a dynamic balance between preventing thrombosis and promoting endothelial cell regeneration?
12	*Science & Engineering.* What electrode array density is optimal for neural prostheses to balance high-resolution signal acquisition with minimal tissue damage for long-term brain-computer interface stability?
13	*Science & Engineering.* What viral titer measurement technology best optimizes the safety and scale of gene therapy production under constraints of detection sensitivity and automation?
14	*Science & Engineering.* What role does the thermal runaway propagation mechanism play in designing safety redundancy for lithium-ion battery packs, considering electrolyte and separator stability?
15	*Science & Engineering.* In aero-engine compressor design, what aerodynamic principles allow for balancing thrust-to-weight ratio and fuel efficiency against pressure ratio and stall margins?
16	*Science & Engineering.* In spacecraft orbital maneuvering, what trade-off between specific impulse and thrust vector control precision maximizes energy efficiency for formation flying?
17	*Finance & Business.* Is it possible for hierarchical optimization algorithms in financial organizations to optimally balance information entropy maximization and decision efficiency, or do distributed networks inevitably incur coordination costs?
18	*Finance & Business.* Can distributed architectures in financial risk management systems truly maintain a dynamic balance between monitoring accuracy and stability, or does event-driven processing inherently accelerate cascading failures?
19	*Finance & Business.* In financial derivative innovation, can option pricing models balance market completeness with convergence speed, or does the numerical instability of stochastic volatility models limit instrument design?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
20	*Finance & Business.* Does the venture capital allocation algorithm of financial startups optimize the return-diversification balance, or does the Markowitz model's covariance estimation exaggerate tail risk prediction errors?
21	*Finance & Business.* In financial product recommendations, can collaborative filtering balance preference prediction with cold start mitigation, or does matrix factorization thresholding limit long-tail coverage?
22	*Finance & Business.* Does the behavioral game theory model in investment decision-making balance rational equilibrium with bias correction, or does prospect theory parameterization amplify loss aversion effects?
23	*Finance & Business.* Can multi-factor asset pricing models balance explanatory power and overfitting risk, or does ridge regression regularization inevitably amplify prediction bias during anomalies?
24	*Finance & Business.* Does the VaR calculation model optimize the trade-off between efficiency and tail risk capture, or does historical simulation window selection lead to estimation bias under non-stationary distributions?
25	*Finance & Business.* Do market microstructure models of high-frequency trading balance price discovery with stability, or does the stochastic evolution of order books amplify liquidity illusions?
26	*Finance & Business.* Can numerical solutions to the Heston model balance convergence speed and estimation accuracy, or does grid density selection in finite difference methods introduce significant errors in long-term option pricing?
27	*Finance & Business.* Can CBDC distributed ledger architectures optimize the privacy-traceability balance, or does zero-knowledge proof complexity limit throughput caps?
28	*Finance & Business.* Does smart contract verification in DeFi balance audit efficiency and formal proof completeness, or does state space explosion leave verification blind spots in complex logic?
29	*Finance & Business.* Can Proof of Work consensus balance energy efficiency and security, or does ASIC hash power centralization inevitably increase the threat of 51% attacks?
30	*Medical & Health.* In the TNM staging system for liver cancer, what balances diagnostic specificity with treatment timing when quantifying tumor burden vs. predicting lymph node metastasis?
31	*Medical & Health.* When assessing aortic stenosis for valve replacement, how are hemodynamic thresholds reconciled with myocardial protection constraints considering transvalvular pressure and ischemia tolerance?
32	*Medical & Health.* During ECMO support, how is heparin dosage managed to balance the risk of thromboembolism against bleeding complications?
33	*Medical & Health.* For coronary heart disease secondary prevention, how are medication adjustments for LDL targets balanced with lifestyle modification adherence to manage recurrence risk?
34	*Medical & Health.* How do digital health interventions for type 2 diabetes balance clinical HbA1c monitoring with remote guidance to maintain adherence to lifestyle changes?
35	*Medical & Health.* How do sentinel surveillance systems optimize the trade-off between early warning sensitivity and resource consumption using ILI reporting thresholds and epidemic prediction?
36	*Medical & Health.* In COVID-19 contact tracing, what strategy balances transmission interruption with privacy protection when using mobile trajectory data?
37	*Medical & Health.* In distinguishing depression subtypes, how are biomarker evidences of neurotransmitters balanced with cognitive assessments to guide clinical decision making?
38	*Medical & Health.* How does early Alzheimer's diagnosis balance the benefit of timely treatment with the harm of false positives using amyloid biomarkers and cognitive tests?
39	*Medical & Health.* In preclinical evaluation of anti-tumor drugs, how is the therapeutic index balanced against toxicity thresholds using in vitro and in vivo models?
40	*Medical & Health.* What mechanism in immune checkpoint inhibitor research balances anti-tumor activation with autoimmune risk?
41	*Medical & Health.* How does antiepileptic drug research balance seizure control thresholds with metabolic interaction risks?
42	*Medical & Health.* How does antiviral research balance viral load reduction with the risk of immune reconstitution inflammation?
43	*Society & Humanities.* What impact do the convergence conditions of Markov Chain Monte Carlo methods have on the estimation precision of fertility rate parameters in age structure forecasting?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
44	*Society & Humanities.* To what extent does the numerical stability of Overlapping Generations models influence the convergence speed of welfare redistribution algorithms?
45	*Society & Humanities.* In multi-agent decision theory, does the dynamic evolution of privacy preferences destabilize the optimal solution of social welfare functions?
46	*Society & Humanities.* By what process does the responsibility assignment matrix in AI systems achieve dynamic equilibrium via Markov decision processes?
47	*Society & Humanities.* Does the modularity of skill node clustering in vocational networks significantly affect the gradient descent convergence of labor matching functions?
48	*Society & Humanities.* In personalized learning models, how does the covariance matrix of cultural foundation vectors influence the eigenvalue spectrum of social cohesion?
49	*Society & Humanities.* How do connected components of local node clustering influence the isomorphism group structure of identity construction in cultural semantic networks?
50	*Society & Humanities.* Does the fractal dimension of cultural identity attractors determine the distribution of branch points in global-local evolutionary trajectories?
51	*Society & Humanities.* How does the Nash equilibrium of justice realization functions impact the stability of judicial authority reputation vectors?
52	*Society & Humanities.* Do time-lag effects of national sovereignty parameters influence the chaotic attractor structure of global governance systems?
53	*Society & Humanities.* How do Laplacian eigenvalues of mobility functions delimit the thresholds of structural solidification in social stratification networks?
54	*Society & Humanities.* How do power-law parameters of globalization networks impact the Markov chain convergence of power dynamics via singular value decomposition?
55	*Arts & Creativity.* What mechanism allows brand value assessment models to balance symbolic capital accumulation with consumer identification dynamics?
56	*Arts & Creativity.* How is the tension between competitive knowledge spillovers and cooperative resource integration balanced in creative clusters?
57	*Arts & Creativity.* How is the tension between audience participation depth and narrative integrity managed in digital media?
58	*Arts & Creativity.* How does algorithmic composition balance mathematical rigor with emotional expressiveness in digital music?
59	*Arts & Creativity.* How does sensory interaction in digital installations construct an ontology of audience participation?
60	*Arts & Creativity.* In transforming traditional culture, how is the dialectical relationship between heritage essence and innovation balanced?
61	*Arts & Creativity.* How does the Stanislavski system reconstruct psychological realism principles in modern stage practice?
62	*Arts & Creativity.* In what way do composition rules in visual arts reconstruct the geometric foundations of classical proportion?
63	*Transportation & Logistics.* What is the impact of nonlinear density-velocity characteristics on macroscopic fundamental diagram prediction near phase transition points?
64	*Transportation & Logistics.* How do stability boundaries of signal coordination affect system-level congestion propagation?
65	*Transportation & Logistics.* How do power-law parameters of travel distance evolve with changes in city scale and network topology?
66	*Transportation & Logistics.* Does the Markov chain character of activity sequences affect prediction reliability given changing socioeconomic attributes?
67	*Transportation & Logistics.* How does base-stock policy parameter optimization minimize total cost relative to demand variation coefficients?
68	*Transportation & Logistics.* How does path selection convergence speed change with network scale power-laws under capacity constraints?
69	*Environment & Energy.* How do spectral characteristics of internal climate variability constrain long-term trend confidence intervals?
70	*Environment & Energy.* Can sampling strategies based on Lorenz attractor geometry optimize ensemble forecast divergence?
71	*Environment & Energy.* How does uncertainty propagation in cumulative effect frameworks affect environmental permit threshold setting?
72	*Environment & Energy.* How does scenario design in environmental risk stress testing affect capital adequacy calibration?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
73	*Environment & Energy.* How does diversity-stability parameterization in ecology models affect biological control prediction?
74	*Environment & Energy.* How does the species-area relationship of critical fragment size constrain protected area design?
75	*Environment & Energy.* How do solid-liquid interface kinetics affect battery cycle life prediction under varying C-rates?
76	*Environment & Energy.* How do stability boundaries in magnetohydrodynamics constrain fusion reactor design space regarding beta limits?
77	*Environment & Energy.* Does Gibbs free energy minimization constrain the scalable efficiency of circular economies regarding entropy?
78	*Environment & Energy.* How do activation energy barriers constrain plastic recycling purity control given molecular weight distribution?
79	*Consumption & E-commerce.* How is the decision boundary for cross-regional transfers optimized in multi-echelon inventory systems?
80	*Consumption & E-commerce.* What critical threshold triggers anomaly warnings in order tracking systems?
81	*Consumption & E-commerce.* How is content exposure diversity balanced with conversion goals in recommendation ranking?
82	*Consumption & E-commerce.* How is the timing threshold for content switching determined in live streaming interaction?
83	*Consumption & E-commerce.* What determines the temporal discount function of rating contributions in review systems?
84	*Consumption & E-commerce.* How is the security interception probability threshold set in payment risk assessment?
85	*Consumption & E-commerce.* What sets the similarity threshold for substitute recommendations in shopping basket analysis?
86	*Consumption & E-commerce.* How is the price adjustment amplitude boundary determined in dynamic pricing strategies?
87	*Consumption & E-commerce.* What controls the category level granularity boundary in product classification systems?
88	*Consumption & E-commerce.* What determines the service degradation strategy in traffic scheduling during peak loads?
89	*Tourism & Leisure.* How do urban residents allocate leisure time to maintain community participation despite work constraints?
90	*Tourism & Leisure.* How does public leisure space usage contribute to community identity during commercialization?
91	*Tourism & Leisure.* What visitor flow control strategies balance economic benefits with heritage protection in tourism?
92	*Tourism & Leisure.* Can oral instruction mechanisms ensure sustainable heritage transmission while adapting to social changes?
93	*Tourism & Leisure.* How are online reviews and past memories integrated in destination decision-making?
94	*Tourism & Leisure.* How do cultural preferences and transport convenience shape tourist flow patterns during peak seasons?
95	*Comprehensive & Others.* In what way does the discreteness of space-time in quantum gravity reshape innate assumptions about continuous time?
96	*Comprehensive & Others.* Does the black hole information paradox necessitate a restructuring of causality in quantum mechanics?
97	*Comprehensive & Others.* How is measurement precision balanced with environmental isolation in quantum gravity experiments?
98	*Comprehensive & Others.* How is subjective experience entropy quantified relative to information geometry in Integrated Information Theory?
99	*Comprehensive & Others.* What framework establishes normative constraints for CRISPR germline intervention regarding human dignity?
100	*Comprehensive & Others.* How are decision priorities established in autonomous weapon systems to protect civilian life?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
101	*Science & Engineering.* In high-speed serial data streams, what represents the optimal trade-off between computational complexity and signal integrity constraints for optical fiber dispersion compensation algorithms?
102	*Science & Engineering.* Regarding Software-Defined Networks (SDN), what mechanism within the control plane architecture balances the global view of centralized decisions with the latency of distributed implementation during traffic engineering?
103	*Science & Engineering.* In the design of millimeter-wave RF front-end circuits, what approach best optimizes the trade-off between power amplifier efficiency and linearity for high-frequency signal processing?
104	*Science & Engineering.* Under tensile loading, by what mechanism does the interface reinforcement of carbon nanotube composites reconcile stress transfer efficiency with the energy dissipation of defect dislocations?
105	*Science & Engineering.* What is the relationship between the phase transformation hysteresis effect in shape memory alloys and the coupling of fatigue life with energy conversion efficiency during thermomechanical cycling?
106	*Science & Engineering.* To what extent does the mismatch of thermal expansion coefficients in metal matrix composites drive the evolution of interface cracking and stress concentration fields under thermal cycling fatigue?
107	*Science & Engineering.* In 3D stacked chips, how is the thermal resistance distribution optimized via the thermal conductivity of interface materials, considering the coupling of contact resistance and stress gradients?
108	*Science & Engineering.* What material design limits exist when tuning the band structure of carbon-based semiconductors to balance carrier mobility and thermal conductivity against spin-orbit coupling and phonon scattering?
109	*Science & Engineering.* Is it possible for MEMS sensors to maintain long-term resonant frequency stability and detection accuracy despite the combined effects of process deviations and environmental vibration?
110	*Science & Engineering.* Under the dual mechanisms of quantum confinement and surface state recombination, what optimizes the threshold current density of quantum dot lasers for spectral purity and output power?
111	*Science & Engineering.* What design approach balances the on-resistance and reliability of GaN power devices when facing synergistic effects from doping concentration gradients and interface trap state density?
112	*Science & Engineering.* For wearable biomedical monitoring, how can the strain sensing sensitivity of flexible substrates be optimized given the nonlinear response between material deformation and conductivity changes?
113	*Science & Engineering.* What memory consistency model allows massively parallel processors to balance performance scalability with program correctness, considering the trade-off between weak ordering and synchronization overhead?
114	*Science & Engineering.* In a hybrid GPU-CPU architecture, what load scheduling algorithm best balances computational throughput with energy efficiency ratios?
115	*Science & Engineering.* What resource scheduling strategy allows edge computing nodes to resolve the conflict between latency sensitivity and energy efficiency to guarantee both real-time processing and long-term reliability?
116	*Science & Engineering.* In neural network accelerators, what weight sparsification algorithm achieves the best trade-off between model compression rate and inference accuracy to optimize computational density and memory bandwidth?
117	*Science & Engineering.* What network topology design for high-performance computing clusters best resolves the contradiction between inter-node latency and scalability while ensuring data efficiency and fault tolerance?
118	*Science & Engineering.* How should the virtualization layer of cloud platforms be abstracted to optimize multi-tenant resource allocation and security isolation without compromising performance overhead?
119	*Science & Engineering.* In IoT edge nodes, what computation offloading strategy yields the best comprehensive performance for real-time response and battery life, weighing local processing latency against transmission energy costs?
120	*Science & Engineering.* To what extent can Wavelength Division Multiplexing (WDM) in optical interconnects expand the bandwidth and energy efficiency of on-chip multi-core processors given channel crosstalk constraints?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
121	*Science & Engineering.* What reconfiguration algorithm enables adaptive computing architectures to dynamically match hardware to applications, balancing real-time response speed with resource utilization?
122	*Science & Engineering.* What microfluidic channel design principles optimize the integration and response time of biosensors while constrained by sample volume and detection sensitivity?
123	*Science & Engineering.* What porosity gradient design in tissue engineering scaffolds best balances the requirement for cell migration speed with mechanical strength to ensure long-term biocompatibility?
124	*Science & Engineering.* Given volume and power density constraints, how can power management circuits in implantable devices be optimized to balance battery life with safety?
125	*Science & Engineering.* By what means does surface modification of biocompatible materials regulate protein adsorption and cell adhesion to optimize the long-term integration of medical implants?
126	*Science & Engineering.* In the tribological design of artificial joints, what strategy optimizes durability and biocompatibility by managing the competition between lubrication film formation and wear particle generation?
127	*Science & Engineering.* How can the flexible transmission mechanism of minimally invasive surgical instruments be designed to balance operation precision with instrument rigidity for optimal force feedback?
128	*Science & Engineering.* What signal processing algorithms allow wearable vital sign monitors to effectively suppress motion artifacts while maintaining accurate continuous monitoring within limited power budgets?
129	*Science & Engineering.* What mechanical gradient design in bioprinted scaffolds reconciles the need for guided cell growth with printing precision to ensure structural integrity in tissue regeneration?
130	*Science & Engineering.* In fuel cell membrane electrode assemblies, how is mass transfer resistance managed against catalyst layer thickness to balance power density with durability?
131	*Science & Engineering.* What is the impact of UV irradiation and humidity cycling on the aging of photovoltaic backsheet materials and the subsequent sealing integrity of the module?
132	*Science & Engineering.* How does fatigue damage accumulation in wind turbine blades relate to the spatial distribution of aerodynamic loads and material gradients in optimizing energy capture efficiency?
133	*Science & Engineering.* What momentum wheel redundancy configuration maximizes satellite attitude control reliability and pointing accuracy within strict weight budgets?
134	*Science & Engineering.* For reentry vehicles, what thermal protection system design best handles the coupled field of aerodynamic heating and structural thermal stress using ablative materials?
135	*Science & Engineering.* How can autonomous drone navigation systems effectively balance the real-time responsiveness of flight control with global path planning optimality under sensor fusion constraints?
136	*Science & Engineering.* What laminate design strategy for aviation composites best optimizes the safety margin and maintenance cycle by balancing impact damage tolerance with weight reduction?
137	*Science & Engineering.* In space station life support, what air circulation filtration technology provides the best trade-off between microbial contamination control and energy consumption?
138	*Finance & Business.* Does the multi-agent coordination algorithm in financial decision support offer an optimal compromise between accuracy and resource use, or does Bayesian reasoning lead to unavoidable complexity explosions?
139	*Finance & Business.* Can adaptive algorithms in regulatory compliance systems successfully balance coverage with agility, or does machine learning overfitting inevitably create systemic risks in regulatory blind spots?
140	*Finance & Business.* Is there a consensus algorithm for financial blockchain alliances that balances security and throughput without the computational overhead of zero-knowledge proofs limiting network scale?
141	*Finance & Business.* Does the credit scoring algorithm of P2P platforms achieve an optimal balance between default prediction and privacy, or does federated learning gradient aggregation inevitably leak behavioral patterns?
142	*Finance & Business.* Can the API economy model of fintech startups maintain interoperability and security, or does microservice interface standardization inherently amplify distributed fault propagation risks?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
143	*Finance & Business.* Can graph theory algorithms in financial networks balance connectivity optimization with complexity control, or is PageRank convergence fundamentally limited by graph sparsity?
144	*Finance & Business.* Is it feasible for trading platform matching engines to balance execution speed and price discovery, or does high-frequency trading latency arbitration inevitably create room for market manipulation?
145	*Finance & Business.* Can social network algorithms for brand communication balance diffusion speed and authenticity, or does node weight decay in graph models facilitate the spread of rumors?
146	*Finance & Business.* In digital marketing, can personalization algorithms balance privacy protection with conversion optimization, or does differential privacy noise addition lower prediction accuracy?
147	*Finance & Business.* Does the survival analysis algorithm for churn prediction balance timeliness and feature complexity, or does covariate selection in Cox models compromise long-term retention strategies?
148	*Finance & Business.* In quantitative investment, can alpha signal extraction algorithms balance signal strength and transaction costs, or does the dimensionality curse in feature selection amplify market noise interference?
149	*Finance & Business.* Can institutional portfolio rebalancing algorithms balance tracking error minimization with cost control, or does mean-variance optimization amplify execution friction from market impact?
150	*Finance & Business.* In ESG portfolios, can factor model construction balance risk quantification and responsibility standardization, or does weighting in multi-criteria analysis introduce subjective bias?
151	*Finance & Business.* Does the convex optimization solver in portfolio algorithms balance global convergence with local extrema escape, or does duality gap estimation amplify complexity in high-dimensional allocation?
152	*Finance & Business.* Can network flow algorithms for liquidity modeling balance measurement accuracy and complexity, or do centrality indicator weights exaggerate liquidity illusions during market segmentation?
153	*Finance & Business.* Do heterogeneous agent models in behavioral finance effectively balance individual bias characterization with group behavior prediction, or are they limited by simulation resource constraints?
154	*Finance & Business.* Can nonlinear dynamic systems in financial cycle modeling balance prediction accuracy and stability, or does fractal dimension calculation amplify the unpredictability of economic shocks?
155	*Finance & Business.* Do interpretability algorithms for risk models balance prediction accuracy and transparency, or does SHAP value approximation error hinder regulatory compliance?
156	*Finance & Business.* Can robo-advisor rebalancing algorithms balance risk adaptability and market timing, or does reinforcement learning estimation amplify decision bias during volatility?
157	*Finance & Business.* Do Open Banking API standards balance data security and interoperability, or does OAuth2.0 token management increase the risk of permission leakage?
158	*Finance & Business.* Can BFT consensus in financial blockchains balance partition tolerance and finality, or does the PBFT view change mechanism increase communication complexity too much?
159	*Finance & Business.* Do CBDC smart contract policy tools balance money supply control and risk isolation, or does programmable money logic risk amplifying policy accidents?
160	*Medical & Health.* What factors balance the prediction of treatment response against myelosuppression thresholds when selecting chemotherapy regimens based on tumor molecular subtyping?
161	*Medical & Health.* In emergency triage, how does the shock index balance early recognition of shock against the risk of over-intervention when assessing vital signs and organ perfusion?
162	*Medical & Health.* How is the dose of solid tumor chemotherapy adjusted to balance tumor cell killing with hematopoietic recovery, given biomarker loads and hematological toxicity thresholds?
163	*Medical & Health.* In monitoring Graft-versus-Host Disease (GVHD), what guides the balance between immune reconstruction and infection risk when adjusting immunosuppressants based on histopathology?
164	*Medical & Health.* What planning strategy for head and neck tumor radiotherapy balances target dose coverage with organ-at-risk tolerance to maximize local control while minimizing injury?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
165	*Medical & Health.* What indications for diabetic retinopathy laser treatment balance the prevention of lesion progression with the risk of visual field defects from treatment?
166	*Medical & Health.* In selecting indications for laparoscopic cholecystectomy, how is the trade-off between minimally invasive benefits and complication risks assessed using ultrasound and biochemical markers?
167	*Medical & Health.* In MRI screening for high-risk breast cancer populations, what strategy balances detection sensitivity with the burden of false positives and overdiagnosis?
168	*Medical & Health.* What reporting system design for influenza-like illness balances epidemiological timeliness with diagnostic accuracy between clinical symptoms and laboratory confirmation?
169	*Medical & Health.* What indications for rTMS treatment in depression balance therapeutic response thresholds with seizure risk safety boundaries?
170	*Medical & Health.* How is the pharmacokinetics of novel antibiotics optimized to balance clearance half-life with accumulation toxicity thresholds?
171	*Society & Humanities.* How does the eigenvalue decomposition of age-specific transition matrices determine the asymptotic behavior of population reproduction functions?
172	*Society & Humanities.* Do perturbations in technological parameters affect the distribution of critical points in human dignity functions within the topological transformation of moral spaces?
173	*Society & Humanities.* What is the effect of technical intervention parameters on the spectral radius of learning transfer matrices in cognitive network models?
174	*Society & Humanities.* Do attention weights of cultural memory vectors affect the stability of recurrent neural networks in modeling historical consciousness?
175	*Society & Humanities.* In restorative justice games, how do eigenvectors of punishment matrices determine the stability of social harmony gradients?
176	*Society & Humanities.* What is the influence of social network rewiring probability on the spectral clustering structure of collective identity vectors in urbanization models?
177	*Arts & Creativity.* In what way does the value translation mechanism adjust the tension between local encoding and international aesthetics?
178	*Arts & Creativity.* In what ways does virtual reality narrative reconstruct the perceptual boundaries of traditional space-time?
179	*Arts & Creativity.* How is the tension between cultural uniqueness and universal values maintained in national identity construction?
180	*Arts & Creativity.* How are geometry and mechanical balance explored through spatial dynamics in ballet choreography?
181	*Arts & Creativity.* How does contemporary design extend the psychological dimensions of traditional color science?
182	*Transportation & Logistics.* How do local perturbations from lane-changing behavior lead to cascading capacity bottlenecks?
183	*Transportation & Logistics.* In what way does the dynamic evolution of transport mode transition matrices impact overall system efficiency?
184	*Transportation & Logistics.* Does inventory positioning robustness decline nonlinearly with increasing network complexity?
185	*Environment & Energy.* Does the lagged response of soil-vegetation feedback amplify prediction uncertainty in global carbon budget models?
186	*Environment & Energy.* How are data quality and attribution uncertainty handled in climate agreement verification using remote sensing?
187	*Environment & Energy.* Do state transition matrices of successional trajectories constrain the effectiveness of ecological restoration?
188	*Environment & Energy.* Do stochastic models of photovoltaic volatility constrain the engineering limits of renewable penetration?
189	*Environment & Energy.* How do biogeochemical processes in water quality models affect treatment thresholds given baseflow recharge?
190	*Consumption & E-commerce.* What determines the optimal warehouse node distribution threshold in hierarchical configuration models?
191	*Consumption & E-commerce.* What strategy guides marketing node selection in word-of-mouth propagation networks?
192	*Consumption & E-commerce.* What determines the authentication intensity hierarchy in account security protection?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
193	*Consumption & E-commerce.* What identifies the intervention timing point for loyalty cultivation in lifecycle curves?
194	*Consumption & E-commerce.* What sets the human intervention threshold in product review grading mechanisms?
195	*Tourism & Leisure.* What role do traditional entertainment forms play in maintaining social cohesion during urbanization?
196	*Tourism & Leisure.* How does oral history transmission bridge the gap between cultural authenticity and tourist understanding?
197	*Tourism & Leisure.* How do tourists weigh risk assessment against exploration desire when choosing destinations?
198	*Comprehensive & Others.* Does the cognitive accessibility of parallel worlds challenge the epistemological foundation of deterministic knowledge?
199	*Comprehensive & Others.* What distinguishes new physics signals from neutrino backgrounds in WIMP scattering measurements?
200	*Comprehensive & Others.* How does the autonomy framework for brain-computer interfaces ensure minimum intervention in privacy?
201	*Medical & Health.* What home management strategy for chronic heart failure best balances NYHA functional classification thresholds with adherence to quality of life improvements?
202	*Medical & Health.* In nutritional prevention of metabolic syndrome, how is the clinical goal of weight loss balanced against malnutrition risk when adjusting diet based on insulin resistance?
203	*Medical & Health.* How do heat stroke warning models balance sensitivity/specificity with cost-effectiveness using meteorological thresholds and emergency visit data?
204	*Medical & Health.* How do seasonal influenza vaccination strategies balance individual protection for the elderly with herd immunity goals under resource allocation constraints?
205	*Medical & Health.* In CBT for PTSD, how is the fear activation threshold of exposure therapy balanced against the risk of treatment dropout?
206	*Medical & Health.* How does antipsychotic treatment balance positive symptom control via dopamine blockade with the minimization of extrapyramidal side effects?
207	*Medical & Health.* What sampling strategy in wearable sleep monitors optimizes the balance between energy consumption and anomaly detection sensitivity?
208	*Medical & Health.* In drug interaction research, how is inducer exposure time balanced with metabolic clearance significance regarding CYP450 enzyme activity?
209	*Medical & Health.* How is the clinical efficacy of multi-target antipsychotics predicted by balancing dose-response curves in animal models with human receptor affinity?
210	*Medical & Health.* How are cardiovascular drug combinations optimized to balance blood pressure reduction with cardiac output maintenance?
211	*Medical & Health.* In glucocorticoid research, what balances anti-inflammatory dose-dependence with the risk of adrenal suppression?
212	*Medical & Health.* How is the gastrointestinal toxicity of NSAIDs managed by balancing COX selectivity with prostaglandin mucosal repair?
213	*Medical & Health.* In anticoagulant monitoring, how is the warfarin therapeutic range balanced against metabolic clearance variations from drug-food interactions?
214	*Society & Humanities.* In migration networks, what role does edge weight optimization play in establishing the Markov equilibrium state of regional population distribution?
215	*Society & Humanities.* How do perturbations in policy parameters within stochastic models of international flow affect the spectral radius of migration matrices?
216	*Society & Humanities.* How does the convexity of responsibility attribution functions determine the stability boundary of moral judgments in AI decision spaces?
217	*Society & Humanities.* What role do Lagrange multipliers of ethical parameters play in defining the convergence domain of evolutionary trajectories in gene intervention optimization?
218	*Society & Humanities.* How does the low-rank approximation of local node embeddings impact the entropy function gradient for cultural diversity in knowledge transmission graphs?
219	*Society & Humanities.* In online education knowledge graphs, does the Herfindahl-Hirschman Index affect diversity optimization under fairness constraints?
220	*Society & Humanities.* How does the random walk matrix of migration flows impact the spectral clustering stability of traditional cultural networks?
221	*Society & Humanities.* Can the robustness of value networks resist economic parameter perturbations in the contagion dynamics of consumer culture?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
222	*Society & Humanities.* How does the Riemannian metric of universality measures affect the conformal transformation of local justice traditions?
223	*Society & Humanities.* Does the geodesic distance between universality and relativity define the critical surface of theoretical conflicts in justice principle spaces?
224	*Society & Humanities.* In digital interaction hypergraphs, does the spectral radius of trust matrices delimit the stability boundary of interaction forms?
225	*Society & Humanities.* How does the characteristic polynomial of intergenerational transfer matrices determine the Lyapunov stability of welfare systems?
226	*Society & Humanities.* Does the condition number of identity stability functions affect the spectral clustering dimension of status systems in structural hole analysis?
227	*Society & Humanities.* How does the Jordan form of social stratification matrices delimit the ergodicity boundary of mobility functions in counter-globalization flow fields?
228	*Arts & Creativity.* What institutional guarantees ensure long-term value chains in creative industries balance economic benefits with cultural heritage?
229	*Arts & Creativity.* How are skill accumulation and market recognition integrated in the career development paths of creative talent?
230	*Arts & Creativity.* What strategy optimizes brand positioning and localization within the cultural difference matrix for international expansion?
231	*Arts & Creativity.* How is the balance between physical realism and artistic expressiveness achieved in digital animation rhythm design?
232	*Arts & Creativity.* How do visual metaphor systems in game art construct semiotic frameworks for player emotional resonance?
233	*Arts & Creativity.* How are visual, auditory, and tactile languages integrated in the multimodal expression of digital media art?
234	*Arts & Creativity.* How does postmodern collage in digital video reconstruct deconstructionist aesthetics using pixel matrices?
235	*Arts & Creativity.* What role does topological deformation play in exploring generative aesthetics within 3D digital sculpture?
236	*Transportation & Logistics.* Does hub node capacity allocation under peak load determine the cascading failure mode of traffic systems?
237	*Transportation & Logistics.* What effect does green wave phase offset have on intersection capacity distribution under random arrival patterns?
238	*Transportation & Logistics.* How does regenerative braking power matching during train density changes influence grid harmonic distortion?
239	*Transportation & Logistics.* How does the spatiotemporal distribution of OD matrices affect capacity allocation under different density gradients?
240	*Transportation & Logistics.* Do complex network characteristics of commuting chains show critical transitions as job-housing separation increases?
241	*Transportation & Logistics.* How do self-similar patterns in peak travel distribution constrain the precision of demand forecasting?
242	*Transportation & Logistics.* How does the carbon emission distribution impact optimal capacity configuration across different travel modes?
243	*Transportation & Logistics.* What is the effect of network flow assignment convergence on model complexity under elastic demand?
244	*Transportation & Logistics.* How does replenishment optimization affect logistics cost variance under demand correlation matrix decomposition?
245	*Transportation & Logistics.* How do safety stock strategies optimize the service-cost trade-off given forecast error autocorrelation?
246	*Transportation & Logistics.* What impact does asymmetric transport delay distribution have on the response resilience of inventory allocation?
247	*Transportation & Logistics.* How do spatial clustering characteristics affect the quality-time trade-off in heuristic path planning?
248	*Transportation & Logistics.* Can inventory synchronization strategies optimize turnover rates using inter-product demand correlation decomposition?
249	*Environment & Energy.* How does the parameterization of ocean heat redistribution affect seasonal prediction skill regarding ENSO amplitude?
250	*Environment & Energy.* How does the stress-strain relationship in ice flow models constrain sea level rise predictions given basal sliding changes?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
251	*Environment & Energy.* Does error propagation in cloud microphysics parameterization limit regional climate prediction resolution?
252	*Environment & Energy.* How does statistical modeling of atmospheric blocking duration affect extreme precipitation probability predictions?
253	*Environment & Energy.* Does the weight allocation of ecological footprint accounting in statistical frameworks constrain policy assessment?
254	*Environment & Energy.* How do source-sink analysis techniques handle data asymmetry in international environmental responsibility sharing?
255	*Environment & Energy.* Do standardized processes of carbon accounting affect investment decision information symmetry?
256	*Environment & Energy.* How do phenotypic plasticity parameters of invasive species affect risk prediction accuracy in ecological models?
257	*Environment & Energy.* Does the clustering coefficient of network modularity influence biodiversity loss threshold predictions?
258	*Environment & Energy.* How does effective population size affect genetic load during range contraction due to climate change?
259	*Environment & Energy.* Does enthalpy balance in thermodynamic models constrain the economic boundaries of hydrogen thermal management?
260	*Environment & Energy.* How does diffusion resistance in membrane separation affect CCS energy efficiency curves?
261	*Environment & Energy.* Does impedance matching in microgrid control affect small-signal stability under distributed power fluctuations?
262	*Environment & Energy.* How does adsorption energy level distribution constrain fuel cell durability design under thermal cycling?
263	*Environment & Energy.* Do partition coefficients in solvent extraction affect the economic feasibility of urban mining recovery?
264	*Environment & Energy.* How do microbial kinetic parameters affect nutrient management strategies given soil mineralization rates?
265	*Environment & Energy.* Does geochemical anomaly identification constrain urban mineral development economics based on grade gradients?
266	*Environment & Energy.* Do kinetic control steps in extraction models affect the process economics of rare earth separation from waste?
267	*Consumption & E-commerce.* How is the safety stock adjustment boundary set for flash sales given conversion rate fluctuations?
268	*Consumption & E-commerce.* What balances the cost-benefit threshold of multimodal transport against traffic control impacts?
269	*Consumption & E-commerce.* How is the performance boundary for supplier promotion determined in admission stratification systems?
270	*Consumption & E-commerce.* How are discount critical points optimized in stepped pricing strategies?
271	*Consumption & E-commerce.* What determines the optimal budget allocation ratio in multi-channel marketing resource models?
272	*Consumption & E-commerce.* How is the reward distribution gradient designed in point incentive mechanisms?
273	*Consumption & E-commerce.* How is the optimal granularity of information disclosure controlled in data sharing protocols?
274	*Consumption & E-commerce.* What defines the classification boundary for problem severity in after-sales service response?
275	*Consumption & E-commerce.* How is the temporal function of rating correction determined in seller reputation systems?
276	*Consumption & E-commerce.* What trigger conditions are set for churn intervention based on probability thresholds?
277	*Consumption & E-commerce.* How is the customer grade partition boundary determined in RFM value stratification?
278	*Consumption & E-commerce.* How is the attribution weight of marketing effects determined in purchase conversion models?
279	*Consumption & E-commerce.* How is the reasonable time limit grace period determined in order cancellation policies?

Table 6: Research questions in SuperResearch Bench (continued).

ID	Domain and research question
280	*Consumption & E-commerce.* How are qualification requirement standards determined in merchant onboarding audits?
281	*Consumption & E-commerce.* What establishes the API permission hierarchy in third-party access standards?
282	*Consumption & E-commerce.* What defines boundary conditions for information disclosure in multi-party data sharing?
283	*Consumption & E-commerce.* How is the penalty gradient function determined in merchant violation systems?
284	*Tourism & Leisure.* Can traditional festival customs preserve family bonds amidst modern lifestyle changes?
285	*Tourism & Leisure.* Do traditional mutual aid networks sustain social support in communities with high floating populations?
286	*Tourism & Leisure.* How do workforce weekend leisure patterns sustain social interaction frequency under high work intensity?
287	*Tourism & Leisure.* In route planning, how does connecting multi-site heritage optimize the balance between knowledge and fatigue?
288	*Tourism & Leisure.* What establishes the boundary between ritual integrity and tourist interaction comfort in folk activities?
289	*Tourism & Leisure.* How are personal preferences coordinated with group pressure in team tourism actions?
290	*Tourism & Leisure.* What mechanism forms satisfaction evaluations from the gap between planning expectations and actual experience?
291	*Tourism & Leisure.* How is destination cognition formed through the interaction of virtual previews and on-site visits?
292	*Comprehensive & Others.* How do extra dimensions in string theory influence geometric intuition in mathematical reasoning?
293	*Comprehensive & Others.* How does quantum non-locality reconstruct the information transfer mechanism of distributed cognitive systems?
294	*Comprehensive & Others.* Does relativistic time dilation challenge cross-cultural research on time cognition?
295	*Comprehensive & Others.* How is the dynamic balance between quantum superposition and decoherence maintained in avian magnetoreception?
296	*Comprehensive & Others.* How does Hawking radiation's non-thermal spectrum reflect holographic duality constraints?
297	*Comprehensive & Others.* What regulatory paradigm balances innovation and ecosystem integrity in synthetic biology?
298	*Comprehensive & Others.* How are intervention thresholds determined for nanorobot swarms to respect cellular autonomy?
299	*Comprehensive & Others.* What responsibility mechanism addresses the asymmetry between security needs and digital trust in quantum computing?
300	*Comprehensive & Others.* How does genome editing ethics rebuild cross-species justice foundations beyond anthropocentrism?

J ENGLISH–CHINESE TASK FORMULATIONS

Tables 7–8 show paired English and Chinese formulations of one question from each benchmark domain. Both formulations are drawn from the task files; the system comparison in Section 5.2 uses the English prompts.

Table 7: Paired benchmark questions from five domains.

ID		Paired English and Chinese task
1	EN	In wireless communication systems, what methods enable Multiple-Input Multiple-Output (MIMO) antenna arrays to optimize beamforming algorithm robustness despite imperfect channel state information?
	ZH	无线通信系统中的多输入多输出天线阵列如何在信道状态信息不完美的情况下优化波束赋形算法的鲁棒性？
17	EN	Is it possible for hierarchical optimization algorithms in financial organizations to optimally balance information entropy maximization and decision efficiency, or do distributed networks inevitably incur coordination costs?
	ZH	金融企业组织架构的层级优化算法是否能够在信息熵最大化与决策效率之间实现最优平衡，还是分布式决策网络注定会引入不可消除的协调成本？
30	EN	In the TNM staging system for liver cancer, what balances diagnostic specificity with treatment timing when quantifying tumor burden vs. predicting lymph node metastasis?
	ZH	肝癌临床分期中TNM分期系统如何在肿瘤负荷量化与淋巴结转移预测的病理证据之间平衡诊断特异性与治疗时机的临床决策约束？
43	EN	What impact do the convergence conditions of Markov Chain Monte Carlo methods have on the estimation precision of fertility rate parameters in age structure forecasting?
	ZH	人口年龄结构的时间序列预测中，马尔可夫链蒙特卡洛方法的收敛条件如何影响生育率参数的后验分布估计精度？
55	EN	What mechanism allows brand value assessment models to balance symbolic capital accumulation with consumer identification dynamics?
	ZH	创意产业的品牌价值评估模型如何在市场定位理论中平衡符号资本积累与消费群体心理认同的动态演化机制？

Table 8: Paired benchmark questions from five additional domains.

ID	Paired English and Chinese task	
63	EN	What is the impact of nonlinear density-velocity characteristics on macroscopic fundamental diagram prediction near phase transition points?
	ZH	交通流连续模型中的密度–速度关系在相变临界点附近的非线性特征如何影响宏观基本图的预测精度？
69	EN	How do spectral characteristics of internal climate variability constrain long-term trend confidence intervals?
	ZH	气候系统内部变率的时间尺度谱如何影响模式预测技能，这种谱特征在不同分辨率数值模拟中的表征稳定性如何制约长期趋势的置信区间？
79	EN	How is the decision boundary for cross-regional transfers optimized in multi-echelon inventory systems?
	ZH	电商平台的多级库存分配策略如何在供应商产能约束与消费者配送时效期望之间优化跨区域调拨的决策边界？
89	EN	How do urban residents allocate leisure time to maintain community participation despite work constraints?
	ZH	都市居民的休闲时间分配模式如何在工作制度约束下维持社区参与与社会网络的持续性？
95	EN	In what way does the discreteness of space-time in quantum gravity reshape innate assumptions about continuous time?
	ZH	量子引力理论中时空的离散性如何重塑人类对连续时间认知的先天假设，并影响决策过程中的因果推理机制？