DeepResearch Arena: The First Exam of LLMs' Research Abilities via Seminar-Grounded Tasks

Haiyuan Wan [1,2*†], **Chen Yang** [3*], **Junchi Yu** [4], **Meiqi Tu** [5], **Jiaxuan Lu** [1], **Di Yu** [1,2], **Jianbao Cao** [1,6], **Ben Gao** [1,6], **Jiaqing Xie** [1], **Aoran Wang** [1], **Wenlong Zhang** [1], **Philip Torr** [4], **Dongzhan Zhou** [1‡]

[1]Shanghai Artificial Intelligence Laboratory
[2]Tsinghua University
[3]The Hong Kong University of Science and Technology, Guangzhou
[4]University of Oxford
[5]The University of Hong Kong
[6]Wuhan University
wanhy24@mails.tsinghua.edu.cn, cyang529@connect.hkust-gz.edu.cn,
zhoudongzhan@pjlab.org.cn

Abstract

Deep research agents have attracted growing attention for their potential to orchestrate multi-stage research workflows, spanning literature synthesis, methodological design, and empirical verification. Despite these strides, evaluating their research capability faithfully is rather challenging due to the difficulty of collecting frontier research questions that genuinely capture researchers' attention and intellectual curiosity. To address this gap, we introduce ***DeepResearch Arena***, a benchmark grounded in academic seminars that capture rich expert discourse and interaction, better reflecting real-world research environments and reducing the risk of data leakage. To automatically construct DeepResearch Arena, we propose a Multi-Agent Hierarchical Task Generation (MAHTG) system that extracts research-worthy inspirations from seminar transcripts. The MAHTG system further translates research-worthy inspirations into high-quality research tasks, ensuring the traceability of research task formulation while filtering noise. With the MAHTG system, we curate DeepResearch Arena with over 10,000 high-quality research tasks from over 200 academic seminars, spanning 12 disciplines, such as literature, history, and science. Our extensive evaluation shows that DeepResearch Arena presents substantial challenges for current state-of-the-art agents, with clear performance gaps observed across different models.

Extended version — https://arxiv.org/abs/2509.01396

Introduction

Recent developments in large language models (LLMs) have led to the rise of the deep research agent (Huang et al. 2025; Xu and Peng 2025; Wu et al. 2025), a LLM-powered agentic system designed for research task automation by integrating literature search (Baek et al. 2024), experiment design (Schmidgall et al. 2025), and ideation (Li et al. 2024). Prevailing examples, such as GPT DeepResearch (OpenAI 2025), indicate that deep research agents have great potential to significantly promote research creativity and productivity.

While deep research agents have gained increasing attention (Du et al. 2025), faithfully evaluating their research ability remains a huge challenge. As Einstein once stated, *The formulation of the problem is often more essential than its solution, which may be merely a matter of mathematical or experimental skill* (Einstein and Infeld 1938). This perspective highlights a crucial challenge in formulating high-quality and frontier research tasks to faithfully assess the ability of deep research agents.

Existing benchmarks for deep research agents mainly resort to two approaches to acquire research questions. The first leverages static data corpora such as academic literature and web content, as seen in AcademicBrowse (Zhou et al. 2025), BrowseComp (Wei et al. 2025), and Researchbench (Liu et al. 2025). The second approach involves manually curated research tasks by domain experts, exemplified by Humanity's Last Exam (Phan et al. 2025), DeepResearchBench (Du et al. 2025), and ExpertLongBench (Ruan et al. 2025). However, both approaches are hindered by critical limitations. Benchmarks derived from static corpora risk data leakage, as the underlying content may already be included in the model pertaining. Meanwhile, datasets curated by experts face scalability bottlenecks and often lack the diversity and spontaneity found in authentic research settings. More fundamentally, both sources tend to abstract away from the situated, evolving nature of real-world research inquiry, where questions emerge dynamically through discourse, ambiguity, and interdisciplinary exploration. A detailed comparison of these benchmarks across key dimensions, including scalability, automation, data leakage risk,

*These authors contributed equally to this work.
†This work was done during internship at Shanghai Artificial Intelligence Laboratory.
‡Corresponding author

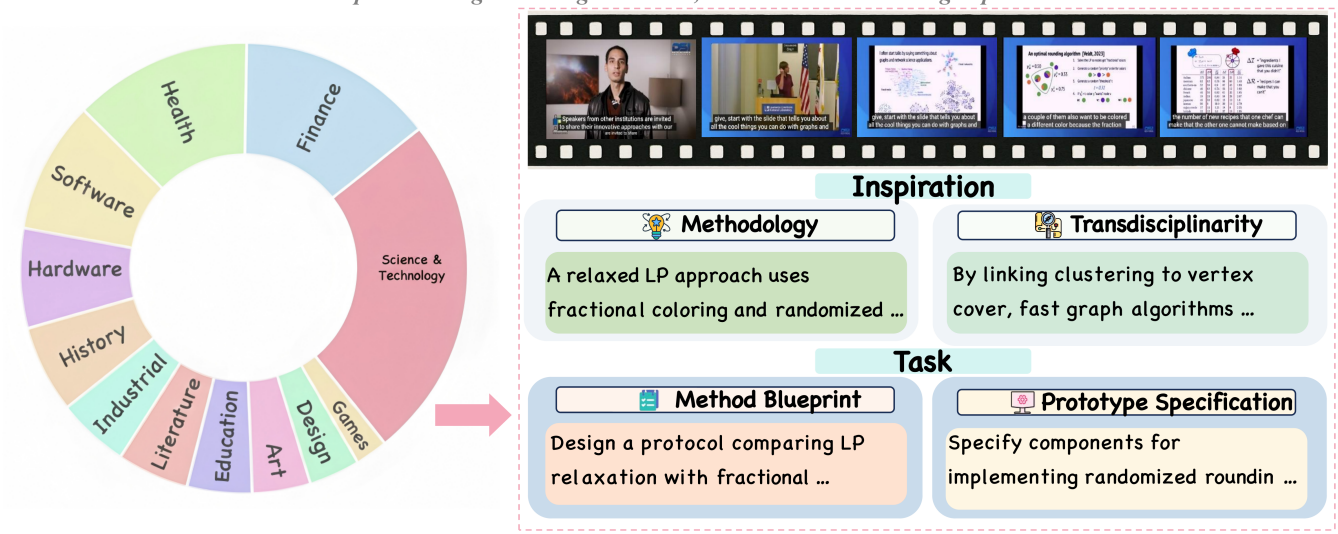


Figure 1: Overview of seminar domains and task structures in MAHTG. Left: Distribution of academic seminars across diverse domains such as Science & Technology, Health, Finance, and others. The outer arc further decomposes each domain into representative research tasks. For instance, Science & Technology includes tasks such as *Hypothesis Generation*, *Empirical Test*, *Prototype Specification*, and *Trend Scan*. Right: Illustration of MAHTG's multi-agent pipeline, where seminar content is transformed into structured research tasks via intermediate inspirations (e.g., *Methodology*, *Transdisciplinarity*). Example outputs are shown for both stages.

and research realism, is provided in Table 1.

To bridge this gap, we introduce a novel benchmark, ***Deep Research Arena***, designed to evaluate deep research agents under authentic, cognitively demanding research scenarios. Unlike static corpora that present information without context, or expert-curated benchmarks that rely on handcrafted tasks detached from actual discovery processes, the proposed benchmark is grounded in academic seminars, where real researchers pose open-ended questions, explore uncertain ideas, and build shared understanding through live discussion. This source captures how real research problems naturally emerge, making Deep Research Arena a more faithful proxy of real-world inquiry. Furthermore, seminar videos are rarely included in model pretraining, which significantly reduces the risk of data leakage that commonly affects benchmarks derived from literature or web corpora.

To capture the nature of such authentic inquiry, Deep Research Arena formulates tasks as open-ended, under-defined problems, drawn by the theory of Ill-Structured Problem Solving (Jonassen 1997), which describes real-world problems as "poorly defined, with no single correct formulation and no objective evaluation criteria". To construct the Deep Research Arena, we develop a Multi-Agent Hierarchical Task Generation (MAHTG) system that automatically extracts research-worthy inspirations and systematically transforms them into high-quality, traceable research tasks through a multi-stage filtering and structuring pipeline. This design ensures both the authenticity and reproducibility of task construction, while reducing noise and preserving the intellectual context of original expert discourse.

We curate a large-scale, multidisciplinary seminar dataset, constructing over 10,000 structured tasks spanning core research competencies. Building on this, we develop a hybrid evaluation framework that jointly measures factual grounding and higher-order reasoning, with examples shown in Figure 1. Together, these contributions provide a rigorous and theory-aligned foundation for assessing deep research competence in deep research agents.

Our contributions are threefold:

- **Seminar-grounded data collection.** We curate a corpus of over 200 academic seminars across 12 disciplines, encompassing real-world expert discourse across science, engineering, humanities, and the arts.
- **Hierarchical task generation.** A multi-stage agent framework extracts research-worthy inspirations from seminar transcripts, categorized into *Limitation*, *Methodology*, *Transdisciplinarity*, and *Hypothesis*, and transforms them into over 10,000 open-ended tasks aligned with the canonical research stages of *Synthesis*, *Design*, and *Evaluation*.
- **Hybrid evaluation framework.** We employ two complementary metrics to quantify factual alignment via extracted keypoints and evaluate open-ended reasoning using adaptively generated, rubric-based checklists.

Related Works

Deep Research Agents. The emergence of DR agents builds upon recent advances in LLMs equipped with tool-use capabilities (Li et al. 2025; Qu et al. 2025; Tang et al. 2023), which allow models to interface with search engines, code interpreters, and external APIs to extend their reasoning horizon. On this foundation, systems such as GPT Deep Research (OpenAI 2025), Gemini Deep Research (Google

Benchmark	Data Source	Scalability	Risk of Data Leakage	Task Automation	Research Realism
ScholarSearch	Literature	✓	✓	✗	✗
BrowseComp	Web Corpus	✓	✓	✓	✗
ResearchBench	Literature	✓	✓	✓	✗
Humanity's Last Exam	Expert	✗	✓	✗	✓
DeepResearchBench	Expert	✗	✓	✗	✓
ExpertLongBench	Expert	✗	✓	✗	✓
DeepResearch Arena (Ours)	Seminar Discourse	✓	✗	✓	✓

Table 1: Comparison of existing deep research benchmarks and our ***DeepResearch Arena*** along key dimensions.

2025), and Grok DeepSearch (xAI 2025) have been developed to support multi-stage research workflows. GPT's system focuses on outline-driven long-form synthesis with citation grounding, Gemini emphasizes multimodal retrieval and synthesis, while Grok prioritizes web summarization for dynamic topics. These agents reflect a shift from retrieval-based assistants to goal-directed, tool-augmented agents capable of supporting exploratory, open-ended inquiry (Yu, He, and Ying 2023; Yu et al. 2025b; Yang et al. 2025).

Benchmarks for Deep Research Agents. Existing benchmarks for deep research agents mainly resort to two approaches to acquire research questions: automatically deriving tasks from static corpora or manually curating them through expert design. The first leverages static data corpora such as papers, and web documents to construct benchmarks represented by multi-hop reasoning or simplified scientific queries. Examples include MuSiQue (Trivedi et al. 2022), which automatically generates multi-hop questions by linking single-hop QA pairs from existing datasets, and HotpotQA (Yang et al. 2018), where annotators write questions guided by system-selected Wikipedia article pairs, making the process closer to extraction than genuine question generation. Other benchmarks in this category include StrategyQA (Geva et al. 2021), ThoughtSource (Ott et al. 2023), AcademicBrowse (Zhou et al. 2025), and BrowseComp (Wei et al. 2025). Despite their emphasis on multi-step reasoning, these benchmarks rely on manually constructed logic chains with predefined paths. They primarily test factual retrieval and compositional reasoning capabilities, yet fail to capture how research questions naturally emerge, evolve, and iterate in real-world research contexts. ScienceQA (Lu et al. 2022) is a large-scale multimodal multiple-choice science QA benchmark (21K questions across STEM and social/language science) that includes lecture and explanation-level CoT annotations to support interpretable multi-step reasoning.

The second category consists of expert-authored benchmarks, where researchers collaborate with domain specialists to construct high-quality, PhD-level evaluation tasks. Compared to benchmarks built from static corpora, these datasets typically feature more original, conceptually challenging, and discipline-specific questions that better reflect expert-level reasoning. Representative examples include LAB-Bench (Laurent et al. 2024), ARC (Clark et al. 2018), GPQA (Rein et al. 2024), FrontierMath (Glazer et al. 2024), HiPhO (Yu et al. 2025a), QcBench (Xie et al. 2025), and Humanity's Last Exam (Phan et al. 2025). GPQA provides graduate-level multiple-choice questions in biology, physics, and chemistry, curated and verified by domain PhDs to ensure they cannot be solved via surface-level heuristics or web search. Humanity's Last Exam comprises a collection of open-ended, expert-written research questions across disciplines such as history, philosophy, and theoretical science, designed to probe creative, integrative thinking under minimal structural constraints. DeepResearch Bench (Du et al. 2025) moves toward more realistic simulation by requiring long-form research reports across disciplines. However, this entire class of expert-authored benchmarks faces several limitations: their prompts are manually constructed, which restricts scalability and diversity, and the datasets remain relatively small in size. More fundamentally, they also fail to capture how research questions emerge dynamically through discourse, ambiguity, and interdisciplinary exploration—core characteristics of authentic research practice.

Multi-Agent Hierarchical Task Generation.

Data Collection. To support the construction of research tasks grounded in authentic scholarly practice, we curated a diverse corpus of over 200 academic seminar videos spanning 12 disciplines, contributed by PhD-level researchers and sourced from publicly accessible academic seminar recordings spanning multiple disciplines. Each video is knowledge-dense and typically lasts around or over 1 hour, and the disciplinary distribution of this corpus is illustrated in Figure 2. Seminar recordings preserve the full contextual flow of expert discourse, encompassing how researchers synthesize prior knowledge, design new approaches, and evaluate outcomes. In this way, they offer a rich context for task generation. Compared to static corpora such as Wikipedia or scientific articles, seminar data captures dynamic and authentic interactions among scholars, reflecting the iterative and evolving nature of real-world research.

As a first step in processing the raw seminar videos, we extract the audio and convert it into textual transcripts with automatic speech recognition. The resulting transcripts retain the full semantic content of the original recordings while remaining absent from existing LLM pretraining corpora, thereby reducing the risk of data contamination and ensuring the integrity of task construction.

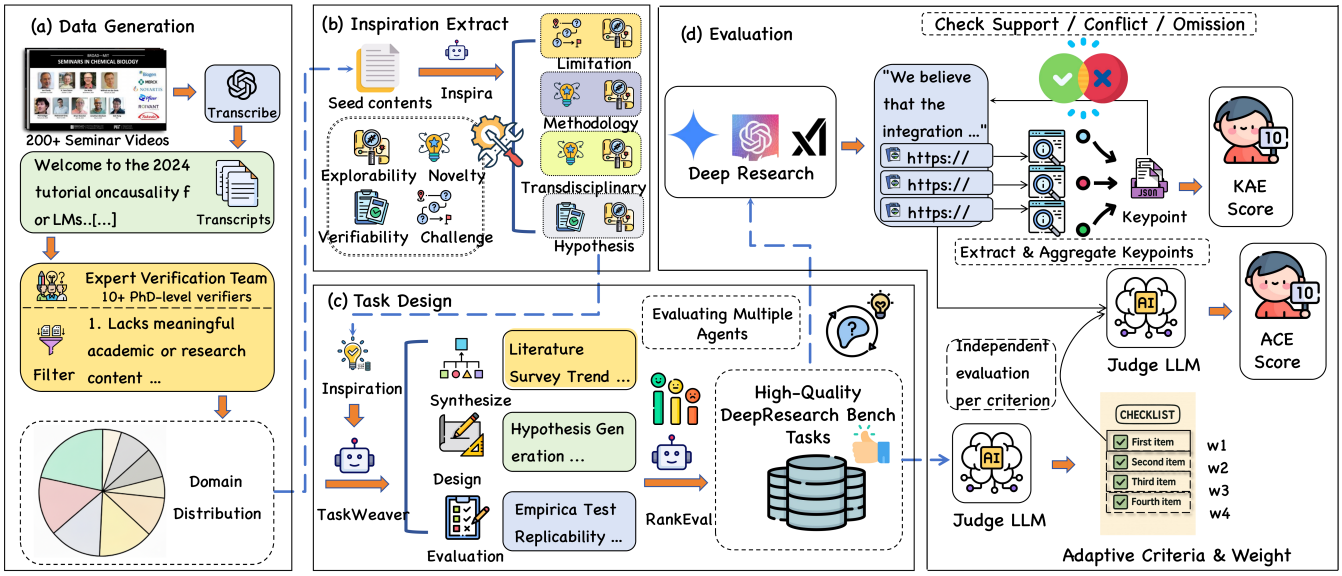


Figure 2: Overview of our benchmark construction pipeline, including four stages: (a) Data generation from transcribed seminar videos, (b) extraction of research inspirations, (c) multi-phase task design, and (d) evaluation using both KAE and ACE metrics.

Inspiration Extraction. Based on seminar transcripts, **Inspira Agent** automatically extracts *inspirations* (as illustrated in Table 2) from seminar transcripts, transforming unstructured expert discourse into structured units suitable for downstream research task construction. To identify academically valuable content, the agent evaluates candidate segments along four dimensions: Novelty, Explorability, Challenge, and Verifiability. Each selected inspiration must satisfy at least two of these criteria. This multi-dimensional filtering process enables the agent to effectively suppress irrelevant or redundant material, reorganize latent research signals, and produce outputs with clearer logical structure and sharper thematic focus, thereby enhancing their suitability for subsequent task generation. In addition, the agent categorizes each item based on its informational focus into one of four types: *Limitation*, *Methodology*, *transdisciplinarity*, *Hypothesis*, as illustrated in Table 2, representing testable claims that can be empirically verified.

Task Generation and Filtering. Building on the structured inspirations extracted from seminar transcripts, we deploy **TaskWeaver Agent** that aggregates and reorganizes content across multiple inspirations to synthesize a set of concrete research tasks distributed across three key phases—*Synthesize*, *Design*, and *Evaluate*, as illustrated in Figure 2. These tasks are constructed by identifying the core problem focus or methodological cues within the inspirations and are paired with clearly defined, executable goals. This synthesis process enables scalable construction of diverse, high-quality DeepResearch tasks aligned with the demands of real-world scientific inquiry (Yu et al. 2022).

To rank the quality of research tasks, we adopt **RankEval Agent** based on the Elo rating system (Glickman 1995). Each task is initialized with a rating score of 1200. In each round, we randomly sample disjoint pairs of tasks and compare them based on evaluation criteria such as originality, clarity, and scientific relevance. Given a pair of tasks t_a and t_b with current Elo scores r_a and r_b, we first compute the expected winning probabilities using:

$$e_a = \frac{1}{1 + 10^{(r_b - r_a)/400}}, \tag{1}$$

where $e_b = 1 - e_a$. An evaluator determines which task is preferred, along with a confidence score $C \in [0.5, 1.0]$. Based on this, we assign soft outcomes:

$$s_a = C, \quad s_b = 1 - C \tag{2}$$

We update the Elo scores using the following update rule:

$$r'_a = r_a + K \cdot (s_a - e_a), \quad r'_b = r_b + K \cdot (s_b - e_b) \tag{3}$$

where K is a tunable constant controlling the update magnitude, set to $K = 32$ in our implementation. This procedure is repeated over R rounds of comparisons (e.g., $R = 2$), allowing the scores to stabilize. After all rounds, we select the top K tasks with the highest Elo scores as the final outputs.

Evaluation Methodology

To comprehensively assess the capabilities of deep research agents in research-oriented tasks, we propose a hybrid evaluation framework that integrates both objective and subjective dimensions of performance. Traditional benchmarks often focus narrowly on surface-level accuracy or retrieval metrics, failing to capture the nuanced reasoning, creativity, and methodological rigor required for real-world research. In contrast, our framework disentangles these facets by combining (1) Keypoint-Aligned Evaluation (KAE) to measure factual correctness and grounding against reference materials, and (2) Adaptively-generated Checklist Evaluation (ACE) to score open-ended outputs via fine-grained, model-adaptive rubrics. This dual approach enables multi-perspective assessment across all stages of the research

Term	Illustration	Example
I. Core Unit: Inspiration		
Inspiration	A research-worthy idea distilled from academic discourse, exhibiting at least two of: *novelty*, *explorability*, *challenge*, *verifiability*. Serves as the seed for task generation.	"A greedy maximal independent-set algorithm … achieves a 2-approximation in O (sum of hyperedge sizes) time … shows classical graph methods can solve edge-colored hypergraph clustering without auxiliary graphs."
II. Types of Inspiration		
Limitation	An open problem, deficiency, or bottleneck in existing methods.	"Few models handle transdisciplinary seminar reasoning."
Methodology	A new or adapted approach, pipeline, or tool.	"Introduce retrieval-augmented reranking framework."
Transdisciplinarity	Ideas involving the fusion of theories or tools across disciplines.	"Apply ecological network theory to social dynamics"
Hypothesis	A testable proposition that guides design or evaluation.	"Grounded citations improve factual accuracy."
III. Task Phase Labels		
Synthesize	Collecting, integrating, and analyzing prior work to form direction.	"Identify gaps in seminar-based QA literature."
Design	Designing solutions, models, or experiments to address a problem.	"Propose a multimodal tree-search method."
Evaluate	Assessing results using structured criteria or benchmarks.	"Compare keypoint coverage across baselines."

Table 2: Core terminology used in our benchmark, grouped into inspiration, its subtypes, and research task phases. This table standardizes interpretation of key concepts throughout the paper.

workflow, from literature synthesis to hypothesis generation and empirical validation, offering a more faithful estimate of models' deep research competence.

Keypoint-Aligned Evaluation. To evaluate the factual adequacy of model-generated research reports in a reference-grounded and scalable manner, we propose a structured KAE pipeline.

Let R denote a model-generated report, and let $URL(R)$ represent the set of all cited URLs in R. For each URL $u \in URL(R)$, we retrieve the underlying webpage content and extract its factual keypoints using a keypoint extraction function $Extract(u)$:

$$K_u = Extract(u) \tag{4}$$

We then aggregate the keypoints from all cited sources into a unified, de-duplicated list of keypoints, which we term the Unified Evidence Keypoints (UEK):

$$\text{UEK} = \text{Dedup}\left(\cup_{u \in \text{URL}(R)} K_u\right) \tag{5}$$

Given this set of reference keypoints, we evaluate the report R along three dimensions:

(1) Keypoint Supported Rate (KSR): the proportion of keypoints from UEK that are explicitly covered or supported in the report:

$$\text{KSR}(R) = \frac{|Supported(R, \text{UEK})|}{|\text{UEK}|} \tag{6}$$

(2) Keypoint Conflict Rate (KCR): the proportion of keypoints from UEK that are contradicted by content in the report:

$$\text{KCR}(R) = \frac{|conflict(R, \text{UEK})|}{|\text{UEK}|} \tag{7}$$

(3) Keypoint Omission Rate (KOR): the proportion of keypoints from UEK that are omitted by content in the report:

$$\text{KCR}(R) = \frac{|Omitted(R, \text{UEK})|}{|\text{UEK}|} \tag{8}$$

Ideally, a high-quality research report should achieve a high KSR (indicating comprehensive factual coverage) and a low KCR and KOR (indicating consistency with evidence). These metrics enable interpretable, reference-grounded evaluation of factual alignment.

Adaptively-generated Checklist Evaluation. To address the challenges of evaluating open-ended research tasks that lack fixed reference answers, we introduce Adaptively-generated Checklist Evaluation (ACE), a two-stage evaluation protocol that leverages the analytical capabilities of large language models (LLMs) while mitigating common sources of bias and inconsistency.

In the first stage, we use a high-capacity LLM (e.g., GPT-4o) to perform meta-analysis over the task prompt, generating a customized checklist of evaluation criteria tailored to the query. Each checklist item corresponds to a critical evaluation dimension, such as factual correctness, methodological soundness, formatting, or reasoning clarity, and is assigned a normalized weight to reflect its relative importance. This step serves to concretize abstract judgment into discrete, model-understandable subgoals.

In the second stage, a separate LLM is tasked with scoring the model-generated response against the checklist. For each item, the evaluator model independently assesses whether the response satisfies the criterion and assigns a local score. These individual scores are then aggregated via a weighted average to produce a final task-level rating. By decoupling

Model	KAE						ACE		Avg. Token (k)		Avg. references	
	KSR		KCR		KOR							
gpt-4o-search-preview	50.0	**85.0**	8.9	5.0	41.1	10.0	2.41	2.00	1.21	2.85	4.24	3.49
gpt-4o-mini-search-preview	78.7	55.6	8.5	16.7	12.8	27.8	2.23	2.05	**1.07**	2.23	3.83	2.07
gpt-4.1-mini w/search	62.5	76.5	10.9	**5.9**	26.6	17.6	2.21	1.87	1.10	**2.02**	4.75	2.39
gpt-4.1 w/search	77.8	60.6	**2.8**	6.1	19.4	33.3	2.43	2.22	1.20	2.43	3.51	2.44
o4-mini-deepresearch	77.2	75.8	4.3	18.2	18.5	6.1	**4.03**	3.88	5.59	12.5	**29.66**	**37.27**
gemini-2.5-pro w/search	65.1	76	14.3	12	20.6	12	2.97	**4.03**	4.29	9.14	23.86	21.39
gemini-2.5-flash w/search	78.7	80	3.4	16	18	**4**	3.81	3.58	64.09	19.78	29.54	28.07
grok-4 w/search	**83.3**	50	7.5	13.8	**9.2**	36.2	2.97	2.97	3.16	6.60	20.59	19.95

Table 3: Evaluation metrics across models. The model release dates are omitted for brevity. Each column reports two values, with the left representing the evaluation results on the English task and the right on the Chinese task.

checklist generation from scoring, ACE reduces evaluation bias, especially those arising from the evaluator's limited comprehension or heuristic shortcuts.

ACE addresses key limitations of existing evaluation paradigms. Human evaluation, while often considered the gold standard, suffers from subjectivity, inter-annotator inconsistency, and high cost. LLM-as-a-judge methods, especially when using smaller models, struggle with complex query understanding, detailed analytical reasoning, and accurate interpretation. Furthermore, rubric-based methods either rely on static reference answers, which are unsuitable for open-ended tasks, or require hand-crafted criteria that are difficult to scale and generalize. In contrast, ACE provides a flexible, scalable, and more reliable alternative for nuanced research task evaluation.

Experiments

Implementation Details. Our MAHTG system comprises several specialized agents, each responsible for a distinct stage in transforming raw academic seminars into structured research tasks and evaluations.

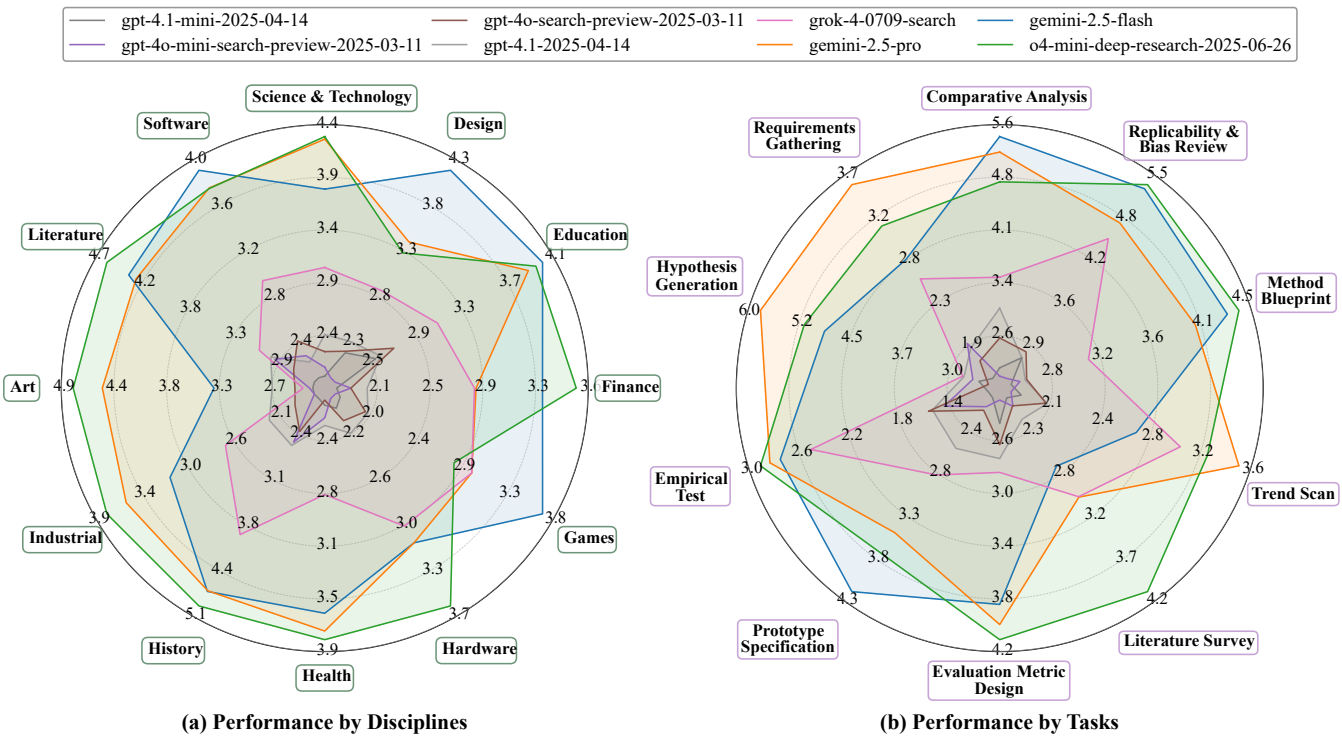
Model Selection Rationale. We adopt a *heterogeneous model configuration* across the MAHTG system, guided by three principles: (1) *capability-task alignment*, assigning models suited to each agent's functional role; (2) *cost-effectiveness and scalability*, ensuring efficiency over large-scale data; and (3) *robustness through model diversity*, mitigating systemic bias. Large models like *claude-sonnet-4-20250514* are used for structured reasoning and code-like outputs, while lightweight ones like *gpt-4o-mini* support tasks requiring relative preference. The Inspira Agent adopts *claude-sonnet-4-20250514* for its strong long-context handling and structured generation. The same model powers the TaskWeaver Agent to ensure schema consistency in transforming inspirations into structured tasks. For efficient pairwise evaluation, the RankEval Agent uses *gpt-4o-mini*, balancing accuracy and cost under the ELO-based framework. To reduce costs, we selected the top 100 highest-scoring samples from the full dataset for evaluation.

We use *gemini-2.5-flash* as a unified evaluator for both factual and subjective scoring, leveraging its strong instruction-following and long-context reasoning. In KAE, it extracts key factual statements from sources retrieved via the Jina AI API and determines whether each is supported, contradicted, or omitted. In ACE, it generates task-specific checklists and conducts criterion-based evaluation. This setup ensures consistency across evaluation stages while maintaining precision, scalability, and interpretability.

Evaluated Models. We evaluate a diverse suite of large language models covering both frontier-level deep research agents and models augmented with real-time retrieval capabilities. Specifically, we include *gpt-4o-search-preview-2025-03-11*, *gpt-4o-mini-search-preview-2025-03-11*, *gpt-4.1-2025-04-14 w/search*, *gpt-4.1-mini-2025-04-14 w/search*, *o4-mini-deepresearch-2025-06-26*, *gemini-2.5-pro w/search*, *gemini-2.5-flash w/search*, and *grok-4-0709 w/search*. When referring to these models, abbreviations will be used, ignoring with search and time versions.

Overall Performance. The table 3reveals clear differences in both ACE and KCE across models. The best ACE performance is achieved by *gpt-o4-mini-deep-research*, which combines the highest ACE score of 4.03 with strong KAE metrics, demonstrating accurate, well-structured, and comprehensive outputs. *GPT-4.1* excels in factual precision but falls short in subjective quality, with the lowest KCR. It minimizes factual errors, yet its lower ACE scores suggest limited coherence and depth. *Gemini-2.5-flash* also performs strongly, with relatively high factual coverage and low contradiction and omission, though it uses significantly more tokens than any other model, indicating a trade-off between thoroughness and efficiency. In contrast, *gpt-4o-search-preview*and *gpt-4o-mini-search-preview* use far fewer tokens but do not perform so well in both evaluation dimensions, suggesting limited ability to handle complex research tasks. *grok-4* demonstrates the strongest factual grounding on English tasks (KSR 83.3), but its performance drops sharply in Chinese, with significantly lower coverage and higher omission. This highlights its limited multilingual generalization despite strong English capabilities. Overall, the results reflect varying model strengths, with some excelling in precision and others in depth or efficiency.

Performance on Different Tasks. As shown in Figure 3, the ACE-based subjective evaluation reveals sub-



(a) Performance by Disciplines **(b) Performance by Tasks**

Figure 3: Comparison of current mainstream models on the DeepResearch Arena benchmark. (a) Performance across 12 research disciplines (*e.g.*, Science & Technology, Art, Finance). (b) Performance across 10 research task types (*e.g.*, Hypothesis Generation, Method Blueprint, Evaluation Metric Design), highlighting task-specific capabilities.

stantial differences in how models perform across various research task types. Models like *gpt-o4-mini-deepresearch* and *gemini-2.5-flash* demonstrate consistently strong performance across nearly all tasks, especially excelling in complex and high-level tasks such as hypothesis generation, evaluation metric design, and methodological planning. *Gemini-2.5-pro* also shows well-rounded capabilities, performing reliably in tasks that require comparative analysis and methodological reasoning. The *gpt-4o* family, particularly the mini version, performs poorly across most task types, struggling especially with tasks that require multi-step logic and structured outputs. These differences highlight each model's unique strengths and limitations, underscoring the importance of task-specific evaluation in assessing deep research competence.

Models also show clear differences in task performance under the KAE. *Gemini-2.5-flash* and *gpt-o4-mini-deepresearch* achieve the strongest overall results, with high keypoint coverage and low conflict and omission rates, leading to the highest efficiency scores:

$$\text{Efficiency Score} = \frac{\text{KSR}}{\text{KCR} + \text{KOR}} \tag{9}$$

In contrast, *gemini-2.5-pro*, *gpt-4o-search-preview*, and *gpt-4.1-mini* struggle with higher conflict and omission rates, resulting in the lowest efficiency and limited reliability for fact-intensive generation. Overall, the results highlight substantial differences in how models handle task complexity and factual alignment.

Conclusion

We present the ***DeepResearch Arena***, a novel benchmark for evaluating the deep research capabilities of large language models in realistic, open-ended settings. Grounded in cognitive theories and authentic seminar discourse, DeepResearch Arena captures the contextual complexity and methodological ambiguity of real-world research. It systematically assesses LLM-based agents across three essential stages, through a curated corpus of multidisciplinary seminars, a hierarchical task generation pipeline, and a hybrid evaluation protocol measuring both factual grounding and higher-order reasoning. By bridging the gap between retrieval-centric agent design and cognitively demanding research tasks, it offers a rigorous, theory-aligned foundation for advancing next-generation research assistants.

Acknowledgements

This work was supported by the Shanghai Municipal Science and Technology Major Project. This work was supported by Shanghai Artificial Intelligence Laboratory.

References

Baek, J.; Jauhar, S. K.; Cucerzan, S.; and Hwang, S. J. 2024. Researchagent: Iterative research idea generation over sci-

entific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think You Have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*.

Du, M.; Xu, B.; Zhu, C.; Wang, X.; and Mao, Z. 2025. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv preprint*.

Einstein, A.; and Infeld, L. 1938. *The Evolution of Physics*. Simon and Schuster.

Geva, M.; Khashabi, D.; Segal, E.; Khot, T.; Roth, D.; and Berant, J. 2021. Did Aristotle Use a Laptop? A Question Answering Benchmark with Implicit Reasoning Strategies. *Transactions of the Association for Computational Linguistics*, 9: 346–361.

Glazer, E.; Erdil, E.; Besiroglu, T.; Chicharro, D.; Chen, E.; Gunning, A.; Olsson, C. F.; Denain, J.-S.; Ho, A.; de Oliveira Santos, E.; Järviniemi, O.; Barnett, M.; Sandler, R.; Vrzala, M.; Sevilla, J.; Ren, Q.; Pratt, E.; Levine, L.; Barkley, G.; Stewart, N.; Grechuk, B.; Grechuk, T.; Enugandla, S. V.; and Wildon, M. 2024. FrontierMath: A Benchmark for Evaluating Advanced Mathematical Reasoning in AI. arXiv:2411.04872.

Glickman, M. E. 1995. A comprehensive guide to chess ratings. *American Chess Journal*, 3.

Google. 2025. Deep Research is now available on Gemini 2.5 Pro Experimental. Gemini Blog (online). Accessed: 2025-07-30. Gemini Advanced subscribers can use Deep Research powered by Gemini 2.5 Pro Experimental.

Huang, Y.; Chen, Y.; Zhang, H.; Li, K.; Fang, M.; Yang, L.; Li, X.; Shang, L.; Xu, S.; Hao, J.; Shao, K.; and Wang, J. 2025. Deep Research Agents: A Systematic Examination And Roadmap. arXiv:2506.18096.

Jonassen, D. H. 1997. Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology Research and Development*, 45(1): 65–94.

Laurent, J. M.; Janizek, J. D.; Ruzo, M.; Hinks, M. M.; Hammerling, M. J.; Narayanan, S.; Ponnapati, M.; White, A. D.; and Rodriques, S. G. 2024. LAB-Bench: Measuring Capabilities of Language Models for Biology Research. arXiv:2407.10362.

Li, L.; Xu, W.; Guo, J.; Zhao, R.; Li, X.; Yuan, Y.; Zhang, B.; Jiang, Y.; Xin, Y.; Dang, R.; et al. 2024. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*.

Li, W.; Li, D.; Dong, K.; Zhang, C.; Zhang, H.; Liu, W.; Wang, Y.; Tang, R.; and Liu, Y. 2025. Adaptive Tool Use in Large Language Models with Meta-Cognition Trigger. arXiv:2502.12961.

Liu, Y.; Yang, Z.; Xie, T.; Ni, J.; Gao, B.; Li, Y.; Tang, S.; Ouyang, W.; Cambria, E.; and Zhou, D. 2025. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. arXiv:2209.09513.

OpenAI. 2025. Introducing Deep Research. https://cdn.openai.com/API/docs/deep_research_blog.pdf?utm_source=chatgpt.com. Accessed July 30, 2025.

Ott, S.; Hebenstreit, K.; Liévin, V.; Hother, C. E.; Moradi, M.; Mayrhauser, M.; Praas, R.; Winther, O.; and Samwald, M. 2023. ThoughtSource: A central hub for large language model reasoning data. *Scientific Data*, 10(1).

Phan, L.; Gatti, A.; Han, Z.; Li, N.; Hu, J.; Zhang, H.; and et al. 2025. Humanity's Last Exam. arXiv:2501.14249.

Qu, C.; Dai, S.; Wei, X.; Cai, H.; Wang, S.; Yin, D.; Xu, J.; and Wen, J.-r. 2025. Tool learning with large language models: a survey. *Frontiers of Computer Science*, 19(8).

Rein, D.; Hou, B. L.; Stickland, A. C.; Petty, J.; Pang, R. Y.; Dirani, J.; Michael, J.; and Bowman, S. R. 2024. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. In *First Conference on Language Modeling*.

Ruan, J.; Nair, I.; Cao, S.; Liu, A.; Munir, S.; Pollens-Dempsey, M.; Chiang, T.; Kates, L.; David, N.; Chen, S.; et al. 2025. ExpertLongBench: Benchmarking Language Models on Expert-Level Long-Form Generation Tasks with Structured Checklists. *arXiv preprint arXiv:2506.01241*.

Schmidgall, S.; Su, Y.; Wang, Z.; Sun, X.; Wu, J.; Yu, X.; Liu, J.; Moor, M.; Liu, Z.; and Barsoum, E. 2025. Agent Laboratory: Using LLM Agents as Research Assistants. arXiv:2501.04227.

Tang, Q.; Deng, Z.; Lin, H.; Han, X.; Liang, Q.; Cao, B.; and Sun, L. 2023. ToolAlpaca: Generalized Tool Learning for Language Models with 3000 Simulated Cases. arXiv:2306.05301.

Trivedi, H.; Balasubramanian, N.; Khot, T.; and Sabharwal, A. 2022. MuSiQue: Multihop Questions via Single-hop Question Composition. arXiv:2108.00573.

Wei, J.; Sun, Z.; Papay, S.; McKinney, S.; Han, J.; Fulford, I.; Chung, H. W.; Passos, A. T.; Fedus, W.; and Glaese, A. 2025. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Wu, J.; Zhu, J.; Liu, Y.; Xu, M.; and Jin, Y. 2025. Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools. arXiv:2502.04644.

xAI. 2025. Grok 3. https://x.ai/news/grok-3. Accessed: 2025-07-30.

Xie, J.; Wang, W.; Gao, B.; Yang, Z.; Wan, H.; Zhang, S.; Fu, T.; and Li, Y. 2025. Qcbench: Evaluating large language models on domain-specific quantitative chemistry. *Journal of Chemical Information and Modeling*, 65(22): 12268–12278.

Xu, R.; and Peng, J. 2025. A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications. arXiv:2506.12594.

Yang, C.; Lu, J.; Wan, H.; Yu, J.; and Qin, F. 2025. From What to Why: A Multi-Agent System for Evidence-based

Chemical Reaction Condition Reasoning. *arXiv preprint arXiv:2509.23768*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yu, F.; Wan, H.; Cheng, Q.; Zhang, Y.; Chen, J.; Han, F.; Wu, Y.; Yao, J.; Hu, R.; Ding, N.; et al. 2025a. HiPhO: How Far Are (M) LLMs from Humans in the Latest High School Physics Olympiad Benchmark? *arXiv preprint arXiv:2509.07894*.

Yu, F.; Yao, J.; Wang, Z.; Wan, H.; Huang, Y.; Zhang, B.; Hu, S.; Zhou, D.; Ding, N.; Cui, G.; et al. 2025b. PhysicsMinions: Winning Gold Medals in the Latest Physics Olympiads with a Coevolutionary Multimodal Multi-Agent System. *arXiv preprint arXiv:2509.24855*.

Yu, J.; He, R.; and Ying, R. 2023. Thought propagation: An analogical approach to complex reasoning with large language models. *arXiv preprint arXiv:2310.03965*.

Yu, J.; Xu, T.; Rong, Y.; Huang, J.; and He, R. 2022. Structure-aware conditional variational auto-encoder for constrained molecule optimization. *Pattern Recognition*, 126: 108581.

Zhou, J.; Li, W.; Liao, Y.; Zhang, N.; Qi, T. M. Z.; Wu, Y.; and Yang, T. 2025. AcademicBrowse: Benchmarking Academic Browse Ability of LLMs. *arXiv preprint arXiv:2506.13784*.