DEEP COGNITION: A MULTI-AGENT FRAMEWORK FOR COLLABORATIVE RESEARCH WITH REAL-TIME COGNITIVE OVERSIGHT

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite advances in large language models(LLMs), current systems for deep research are limited by an asynchronous, "input-wait-output" interaction paradigm. This model creates a critical disconnect between human intent and AI execution, leading to error propagation and an inability to dynamically course-correct during complex problem-solving. We propose introduce Deep Cognition, a system designed to enable this paradigm through three technical pillars: transparent and interruptible AI reasoning, fine-grained bidirectional dialogue, and a shared cognitive context. At the core of our system is a layered StateManager architecture and a novel multi-stage budget allocation algorithm. This architecture ingests and normalizes all interaction data (e.g., dialogue trajectories and user artifacts) into a perpetually optimized, high-information-density working memory. By dynamically prioritizing context based on a combination of static heuristics and a time-sensitive scoring function, our system mitigates error cascades and allows the AI to adapt its reasoning pathways based on the user's implicit focus. We conduct a comprehensive user study on challenging deep research tasks to evaluate the efficacy of our system. Results show that our approach significantly enhances the user experience, yielding improvements of up to 29.2% in Fine-Grained Interaction and 27.7% in Ease of Collaboration compared to a competitive baseline. Most notably, our system demonstrates a 31.8% to 50.0% points improvement in overall task performance. These results highlight the critical importance of designing interactive AI systems that facilitate continuous human guidance and transparent reasoning, rather than merely responding to isolated commands.

1 INTRODUCTION

As artificial intelligence (AI) capabilities have advanced dramatically through large language models (LLMs) Luo et al. (2024); Radford et al. (2018; 2021); Brown et al. (2020; 2024), a fundamental question emerges: ***How to build the equality relationship between human and machine intelligence in the age of AI?*** The prevailing trajectory in AI development has emphasized scaling model parameters Kaplan et al. (2020); Hoffmann et al. (2022); Wei et al. (2022), expanding training data Yang et al. (2025); Meta AI (2025), and refining architectures DeepSeek-AI et al. (2025); MiniMax et al. (2025); Poli et al. (2024)—creating increasingly autonomous black boxes that assume minimal human input beyond simple prompting Liu et al. (2023b); Kim et al. (2023) or decision-making Yin (2025). This pathway implicitly assumes that the ultimate form of artificial intelligence would require minimal human input, with interaction reduced to simple prompting or instruction Kim et al. (2023) or AI-assisted decision-making Yin (2025). We contend that this assumption fundamentally mischaracterizes the nature of intelligence itself. This paradigm positions humans as external operators who provide initial prompts and consume final outputs while remaining excluded from the cognitive process itself, treating human intelligence as merely an instructor rather than a collaborative partner. However, intelligence—whether human or artificial—is inherently interactive, contextual, and collaborative Hutchins (1995); Minsky (1987); Woolley et al. (2010). The most sophisticated human thinking rarely occurs in isolation but emerges through dialogue, feedback, refinement, and the integration of diverse perspectives. Consider the nature of breakthrough scientific discoveries or complex problem-solving scenarios: They invariably in-

volve iterative cycles of hypothesis formation, testing, revision, and collaborative refinement. As AI systems approach advanced cognitive capabilities powered by inference-time scaling OpenAI (2024)—enabling thought-level communication where strategic human oversight can leverage vast AI execution power Xia et al. (2025)—the need for meaningful interaction transforms and intensifies. This is especially critical for extended AI tasks Kwa et al. (2025) spanning hours to days, which fundamentally alter human-AI collaboration dynamics.

This transition is particularly evident in systems designed for Deep Research tasks OpenAI (2025c); Google (2025); Perplexity AI (2025); Zheng et al. (2025a)—complex, extended cognitive processes involving dynamic information retrieval, filter, understanding, analysis and synthesis. Current state-of-the-art research systems have pioneered capabilities for multi-step web browsing, data analysis, and report generation. However, these systems uniformly adopt an "Input-Wait-Output" interaction paradigm where users initiate a query, wait through an extended "black box" processing period (typically 5-30 minutes), and eventually receive a comprehensive result. This approach reflects the persistent assumption that interaction is merely a necessary cost rather than a source of value. Yet these systems fundamentally suffer from critical deficiencies: early errors Cemri et al. (2025) compound without correction, systems cannot adapt to evolving requirements, domain expertise remains inaccessible at crucial moments, and opaque processing prevents human-AI collaboration.

These deficiencies stem from a fundamental misalignment: systems that minimize human involvement during processing cannot address problems that require adaptive guidance and expert intervention Bainbridge (1983). To address this fundamental challenge, we develop **deep cognition**—a systematic framework that transcends traditional automation by embedding real-time human expertise directly into AI reasoning processes for complex research tasks, guided by the following principles:

- ***Transparency:*** The system reveals its entire thinking process—from search strategies and query formulations to information evaluation and synthesis rationales—making AI cognition inspectable and editable at every stage. This transparency enables true thought-level interaction where humans can guide how AI thinks.
- ***Real-Time Intervention:*** Unlike conventional systems that operate in isolated processing cycles, deep cognition allows users to pause the research progress and input feedback and requirements at any moment. This creates continuous dialogue rather than discrete query-response cycles.
- ***Fine-Grained Interaction:*** Users can engage with any specific element of the AI's output—questioning particular claims, requesting elaboration on specific points, or changing the research focus.

These principles fundamentally transform deep research from conventional question-and-answer exchanges into cognitive collaboration(see Appendix C)—what we term **cognitive oversight**. Rather than relegating humans to the role of passive tool operators, this framework establishes a synergistic reasoning process that harnesses the complementary strengths of human expertise and AI capabilities while mitigating their respective limitations. Through cognitive oversight, we move beyond the traditional paradigm of human-AI interaction toward a new form of augmented intelligence where strategic human insight and AI computational power merge into a unified cognitive system.

Through extensive experiments with real expert interactions, we demonstrate that deep cognition achieves substantial improvements or competitive over strongest baseline across all evaluation dimensions: Transparency (+20.0%), Fine-Grained Interaction (+29.2%), Real-Time Intervention (+18.5%), Ease of Collaboration (+27.7%), Results-Worth-Effort (+8.8%), and Interruptibility (+20.7%). Our contributions are summarized as follows:

- **Deep Cognition**: We operationalize the cognitive oversight paradigm into deep cognition, a multi-agent human-AI collaboration system designed for deep research tasks.
- **Comprehensive Evaluation Framework**: We establish a complete evaluation framework, including 15 metrics specifically designed for assessing the effectiveness of cognitive oversight in deep research scenarios.
- **Cognitive Oversight**: We propose a human-AI collaboration paradigm: cognition oversight, which augments the intelligence through human-AI partnership.

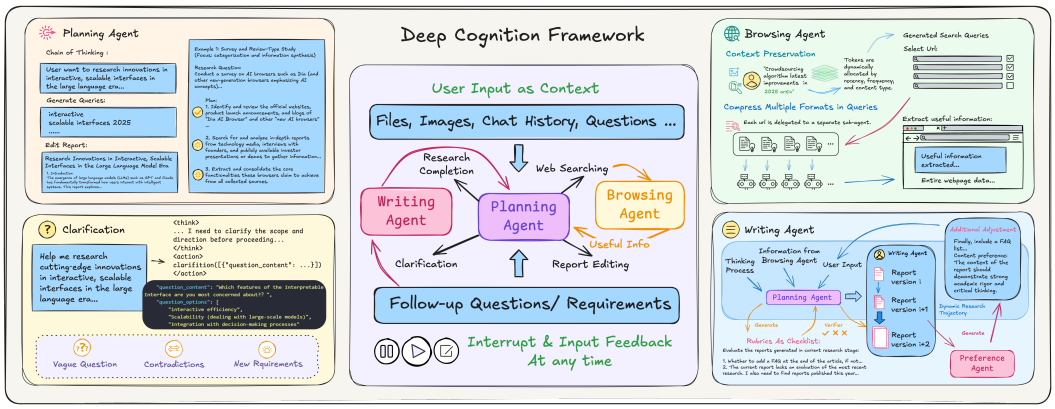

Figure 1: Deep cognition framework overview. This multi-agent research assistant system breaks down complex research questions and dynamically synthesizes information from multiple sources through iterative search, clarification, and user feedback. The central diagram illustrates the overall architecture. The framework integrates four key processes: planning agent, query refinement, browsing agent, and writing agent.

2 METHODOLOGY

2.1 SYSTEM ARCHITECTURE OVERVIEW

Our methodology follows a three-stage agent workflow cycle: Plan-Search-Report, with the capability for agents to solicit human input at any stage of the cycle. We propose a multi-agent collaborative deep research system that addresses key challenges in long-form research report generation through the coordinated operation of four core components. The system workflow proceeds as follows: Initially, user input undergoes question refinement and preliminary enhancement. Subsequently, the **Option-Driven Multi-Round Clarification module** guides dialogue through structured questioning to precisely capture research intent and background information. After establishing research objectives, the system enters a **Plan-Search-Report** dynamic loop: within each cycle, network search queries are generated based on current planning status and delegated to the **Sub Browse-Agent Cluster**, which coordinates Sub-Agent groups to process massive web resources in parallel. During evidence collection, the **Writing Agent** continuously outputs intermediate reports, enabling dynamic user feedback. The entire process supports user interruption at any time, while agents can proactively initiate clarification questions to seek additional information for decision assistance. This design ensures research process transparency and user engagement while maintaining efficient automated information processing capabilities.

2.2 OPTION-DRIVEN MULTI-ROUND CLARIFICATION MECHANISM

Existing deep research systems such as OpenAI DeepResearch OpenAI (2025b) and Gemini DeepResearch typically conduct one-time question collection during the initial dialogue phase, but this approach neglects the dynamic clarification needs that emerge during the research process. Human researchers actively seek clarification for newly discovered points of confusion during exploration, and this timely feedback mechanism is crucial for research efficiency and quality. We design an **option-driven progressive clarification framework** that transforms complex clarification questions into structured option questionnaires, rather than relying on traditional free-text input. This approach reduces user cognitive burden while improving response stability and parsability. The mechanism supports triggering clarification processes at any stage of the research, providing continuous human supervision signals for subsequent information retrieval and report generation.

2.3 SUB BROWSE-AGENT CLUSTER

When processing large-scale web information retrieval tasks, we face two core challenges. First, the **information overload problem** arises as massive URLs and PDF documents exceed the effective

processing range of a single model. Second, the **long-sequence degradation problem** manifests as existing large language models universally exhibit the "lost in the middle" Liu et al. (2023a) phenomenon, struggling to effectively integrate scattered key information when processing long texts. Additionally, the inherent structural looseness and uneven information density of web content further exacerbate the complexity of information extraction. To address these challenges, we propose a distributed Sub-Browse Agent cluster architecture that achieves efficient information extraction through a systematic workflow. The main Research Agent first queries the Serper API to retrieve the top-20 candidate URLs for each search query, then strategically distributes these resources among specialized Sub-Agent instances. Each Sub-Agent operates within an isolated contextual environment to avoid cross-domain information interference.

For content processing, Sub-Agents employ adaptive chunking strategies to handle documents of varying lengths. Standard web pages are processed using fixed-size chunking with overlapping windows, while exceptionally long documents trigger an autonomous pagination decision mechanism where the Browse Agent evaluates content density and relevance to determine whether to continue processing subsequent sections. Upon completion of analysis, each Sub-Agent submits structured findings to the main Agent with three components:**Excerpts**, **Useful** and **Reasoning**. This architecture effectively distributes computational load, enables specialized processing optimization, and significantly improves both efficiency and accuracy in large-scale web information retrieval tasks.

2.4 Dynamic Planning Agent

To address the issues of rigid planning and insufficient adaptability in long-term research processes, we designed a dynamic research planning generation mechanism. This mechanism can real-time adjust research directions and priorities based on research progress and newly discovered evidence, avoiding the limitations of "plan once, execute mechanically" approaches. Meanwhile, planning steps feature explicit success criteria, supporting subsequent agent verification and human inspection to ensure a balance between research quality and efficiency.

2.5 Intermediate Research Reports through Writing Agent

While existing deep research systems LangChain (2024); Roucher et al. (2025a) typically follow a sequential collect-then-generate paradigm, we propose an **evidence-driven iterative report construction strategy**. We deployed a specially fine-tuned Writing Agent capable of generating structured intermediate reports even when evidence collection remains ongoing. The system dynamically generates or adjusts hierarchical research plans at the beginning of each information collection cycle, with these plans serving as report outlines to guide the current cycle's writing tasks. This progressive synthesis approach delivers two key advantages: through **reasoning space construction**, it provides the model with a dedicated arena for deep reasoning and analysis during iterative optimization of multiple report versions; through **selective context retention**, the system preserves only the browsing results that have been incorporated into the current report, while directly removing unutilized evidence from subsequent processing contexts.

This parallel evidence acquisition and report construction paradigm breaks through the limitations of traditional batch processing approaches, enabling continuous knowledge synthesis processes.

3 Human-AI Co-Research Mechanism

Deep cognition supports real-time human–AI collaboration. It is designed for open-ended, multi-hop retrieval and exploratory analysis. It enables users to iteratively expand the initial question and produce a synthesized write-up. Following principles of cognitive oversight, we designed the following features for our deep cognition system, with interfaces presented in Figure 3. The interface supports multiple modes of human–AI collaboration: Clarification (left): The system generates clarification questions to help users specify their focus. Interrupt (bottom-left): Users can intervene during the system's ongoing retrieval or reasoning process, halting unsatisfactory results and redirecting the search toward more relevant information. Planning (right): The system synthesizes retrieved evidence into a structured research plan.

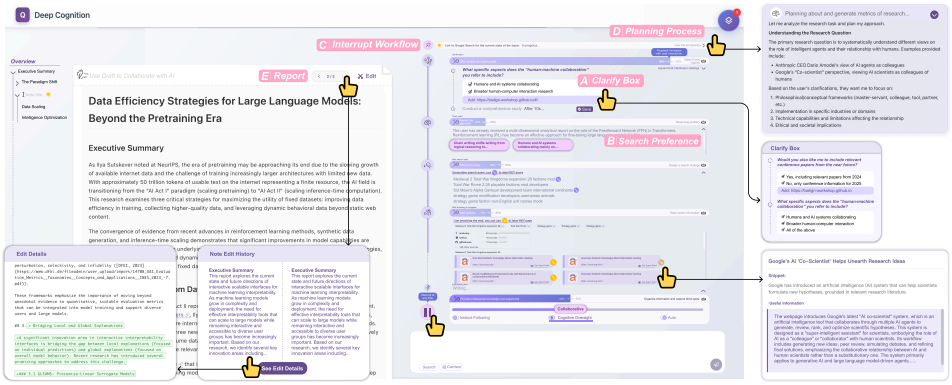

Figure 2: Deep cognition interface design showcasing key interactive features: (A) Research scope clarification to refine vague queries, (B) Click to open the important URL, (C) Multi-agent Workflow Visualization, (D) Transparent display of reasoning, research processes, and interactive query refinement, and (E) Report revision. The 👆 icon stands for clickable interface elements.

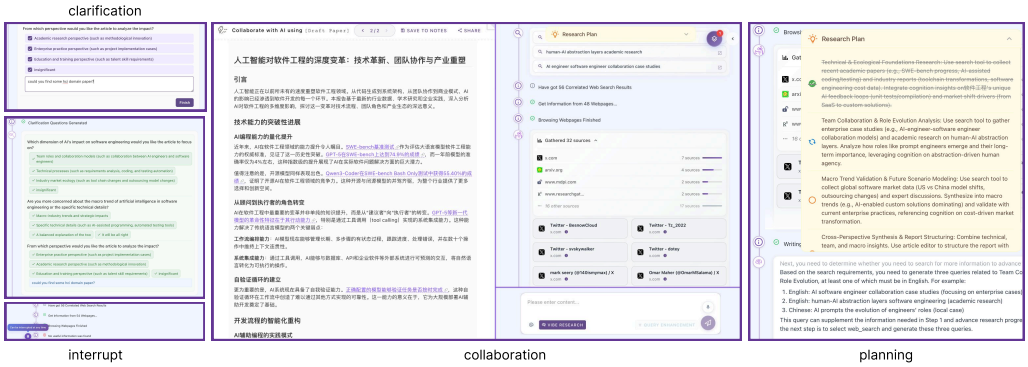

Figure 3: Presents a real screenshot from our deployed system, illustrating how users engage in different stages of interaction with the Deep Research tool.

Transparent Research Process The interface make the system's decision-making process visible and comprehensible to users. Search strategy explainability is achieved by directly displaying the reasoning process and query terms generated by the model, making information retrieval interpretable. The editor area on the left of Figure 3 displays the evolving research document with proper formatting. All findings are properly linked to their original sources, enabling users to trace source materials.

Real-Time Intervention We implement a "Pause" feature, allowing users to interrupt the system at critical junctures in the research process. This intervention capability enables users to actively shape the research trajectory based on emerging insights or changing objectives.

4 Experiments

4.1 System Experimental Setup

Deep cognition provides flexible support for various open-source and closed-source large language models. We use claude-3.7-sonnet-thinking as an inference model for action selection and claude-4.0-sonnet for document authoring, and the browsing model uses gpt-4.1-mini for processing large numbers of documents, with 0.6 used for both temperature. We used the Google TOP20 for web search to provide a realistic search environment for the Agent System. Each turn search generate 5 queries, and for 5 webpages for each query. We develop a web application for users to interact with deep cognition in real time. Deep cognition is implemented as a web-based application. We

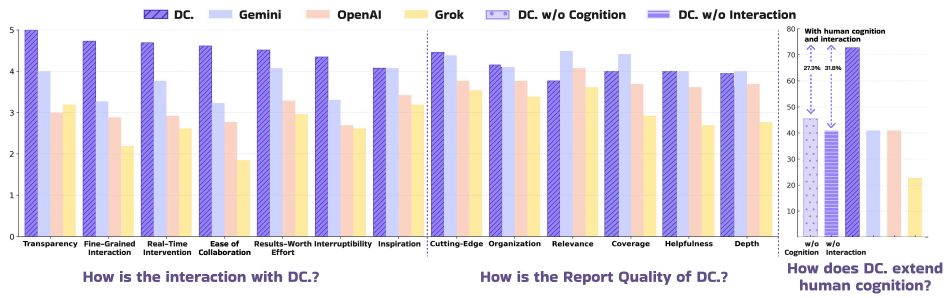

Figure 4: **Overall evaluation results.** We present the user evaluation (seven metrics on the left part), report quality (six metrics in the middle), and evaluation results on deep research problems (the right part) with three conditions: *Without Cognition*, *With Cognition & Interaction*, and *Without Interaction*. These results demonstrate the cognitive amplification effect of deep cognition when users collaborate with AI to perform long and complex tasks. "DC." stands for "deep cognition".

compare it with three competitive deep research products: Gemini Deep Research Google (2025), OpenAI Deep Research OpenAI (2025b;a;c) and Grok 3 DeeperSearch xAI (2025).

4.2 Research Task Setup

We performed a user evaluation to capture real-world user experience during human-AI interaction inspired by Lee et al. (2024). This methodology addresses two fundamental limitations of static benchmarks: 1) it reflects real-world, first-person subjective experience during human-AI interaction; and 2) it enables assessment of output quality that depends on interactive dynamics, which aligns with real-world usage scenarios. We conducted two within-subjects user studies comparing deep cognition with state-of-the-art commercial baseline interfaces (OpenAI, Grok, and Gemini). Study 1 measuring post-interaction report quality and the effectiveness of the interaction design. Study 2 testing whether users with higher or lower cognitive levels show differences in multi-hop retrieval task when using deep cognition.

Study 1 We recruited 13 participants with prior research experience. Before using the system, they were introduced to our evaluation metrics(see AppendixD) for deep cognition to ensure a shared understanding. Participants then evaluated both the quality of generated reports and the system's interactive behaviors on a 5-point Likert scale, supplemented by qualitative responses to open-ended interviews. Each participant proposed a research question from their own work, participants observed the model in real time as it retrieved information, reasoned through intermediate steps, and generated self-evaluations. They could not directly edit the final report but instead guided the process via interactive mechanisms such as interrupting outputs, injecting prior knowledge, inspecting sources, reviewing self-evaluations, suggesting new directions, giving feedback, or contributing personal documents. These interventions helped steer the model toward deeper analysis and more efficient retrieval, with the report finalized when the model itself chose to conclude.

Study 2 To validate our hypothesis that experts with higher cognitive capabilities demonstrate enhanced collaboration with AI in transparent dialogue environments, we measured system performance through two comprehensive benchmarks. Given that our expert annotators are native Chinese speakers with domain expertise, we selected representative subsets for intensive interactive evaluation: 22 questions from browsecomp-ZH Zhou et al. (2025) (top two from each of 11 categories) and the first 20 questions from xbench-deep research Chen et al. (2025). Both sampling strategies ensure feasible human-AI collaborative assessment.

5 MAIN RESULT

5.1 EXPERT USER EVALUATION

As shown in Table 1, augmented through expert interaction, the deep cognition system demonstrated significant enhancements across six evaluated metrics, overall average improve 63%. Notably, the ORGANIZATION exhibits the greatest gain (+97%), followed by CUTTING-EDGE (+79%) and depth (+76%). Even the dimension with the smallest gain, helpfulness, showed a significant improvement of +42%. As the evaluation results in Table 2, the **alignment between expert rankings and user evaluations** validates our core hypothesis: **The system with enhanced interaction mechanisms consistently deliver output quality across six metrics.**

Metric	DC (w/o Int).	DC.
Organization	**2.231**	**4.385** ↑ **97%**
Cutting-Edge	**2.538**	**4.538** ↑ **79%**
Coverage	2.423	4.000 ↑ 65%
Depth	**2.231**	**3.923** ↑ **76%**
Relevance	2.885	3.769 ↑ 31%
Helpfulness	2.808	4.000 ↑ 42%
Overall Average	2.519	4.103 ↑ 63%

Table 1: Performance improvement of deep cognition over deep cognition without interaction. DC. indicates deep cognition, DC (non). indicates deep cognition without interaction.

Report Evaluation (1–5 Score)

Metric	DC.	Gemini	OpenAI	Grok3
Organization	**4.385**$_{+1.8\%}$	4.308	3.769	3.385
Cutting-Edge	**4.538**$_{+3.5\%}$	4.385	3.769	3.538
Coverage	4.000$_{-10.4\%}$	**4.462**	3.692	2.923
Depth	3.923$_{-1.9\%}$	**4.000**	3.577	2.769
Relevance	3.769$_{-18.3\%}$	**4.615**	4.077	3.615
Helpfulness	**4.000**$_{+0.0\%}$	**4.000**	3.615	2.692

Interaction Evaluation (1–5 Score)

Metric	DC.	Gemini	OpenAI	Grok 3
Transparency	**5.00**$_{+25.0\%}$	4.00	3.00	3.19
Interruptibility	**4.35**$_{+31.4\%}$	3.31	2.69	2.62
Fine-Grained Interation	**4.73**$_{+44.6\%}$	3.27	2.88	2.19
Real-Time Intervention	**4.69**$_{+24.4\%}$	3.77	2.92	2.62
Inspiration	**4.08**$_{+0.0\%}$	**4.08**	3.42	3.19
Ease of Collaboration	**4.62**$_{+43.0\%}$	3.23	2.77	1.85
Results-Worth-Effort	**4.52**$_{+10.8\%}$	4.08	3.29	2.96

Table 2: User and expert evaluation results for AI research assistance systems. Left panel: User-generated evaluation scores on a 1-5 scale, where participants queried systems with their own research questions. Right panel: Scores (1–5 scale) for system-interaction evaluation metrics. Color coding indicates within-row performance rankings, and percentages show deep cognition's relative improvement over the strongest baseline system (Gemini). DC. indicates deep cognition.

Deep cognition dominates six of the seven metrics. It records the largest gains in Fine-Grained Interaction (+44.6%) and Cooperative (+43.0%), and is the only system to reach a perfect Transparency score (5.00, +25.0% over the strongest baseline). Overall, the results highlight deep cognition's superior transparency, controllability, and collaborative support. These quantitative results are further supported by users' qualitative feedback. Over 90% of participants agree or strongly agree that interaction with deep cognition improves report quality; 69% find it easy to use and 62% show a high willingness to use.

5.2 BENCHMARK EVALUATION RESULTS

The results provide compelling evidence for our collaborative cognition framework. On browsecomp-ZH, the deep cognition system achieves 72.73% accuracy—dramatically outperforming all baselines (Gemini/OpenAI: 40.91%, Grok 3: 22.73%). Ablation studies show neither cognitive oversight alone (45.45%) nor interaction alone (40.91%) match their combination. On X-bench, our system achieves 65% accuracy, matching OpenAI while substantially outperforming Gemini (35%). Note that browsecomp-ZH was evaluated on June 22, 2024, and X-bench on September 25, 2024—temporal gaps may contribute to baseline performance variations due to API updates. The results consistently demonstrate that expert-AI collaboration requires both transparent reasoning and interactive guidance for effective performance across domains. Participants with deeper cognitive processing capabilities achieved significantly higher human-AI collaborative performance compared

to those with surface-level cognitive approaches in transparent interaction paradigms, as measured by problem resolution accuracy.

	DC (non cog).	DC (non int).	DC (cog+int).	Gemini	OpenAI	Grok 3
Accuracy	45.45%	40.91%	72.73%	40.91%	40.91%	22.73%

	DC (cog+int).	Gemini	OpenAI
Accuracy	65%	35%	65%

Table 3: Accuracy comparison across benchmarks. Top: Browsecomp-ZH (22 questions). Bottom: X-bench deep research (first 20 questions). DC (non cog). = baseline with middle school-level participants (n=4); DC (non int). = autonomous system; DC (cog+int). = interactive condition with graduate-level participants (n=4).

6 In-Depth Analysis of the Human Study

6.1 Human Hold Dynamic Mental Models Throughout Collaboration Process

Enhancing transparency at the model's behavioral status can improve human-AI collaboration. Specifically, in complex, long-duration retrieval tasks, humans tend to delegate mechanical operations such as "browsing" and "summarizing" to AI, while preferring to collaborate with the model at decision points requiring higher-order thinking. We dive deeper into the human behavior pattern in the deep research process and provide design considerations of human-AI collaboration research system. As illustrated in case study(see Appendix G) and User Behavior Data Point (see Appendix A), our user study reveals a sophisticated pattern of collaborative engagement that varies systematically across six research phases. Users demonstrate **dynamic cooperation willingness**, transitioning between "hands-on" and "hands-off" modes based on task characteristics and their domain expertise. We detail these six phases below:

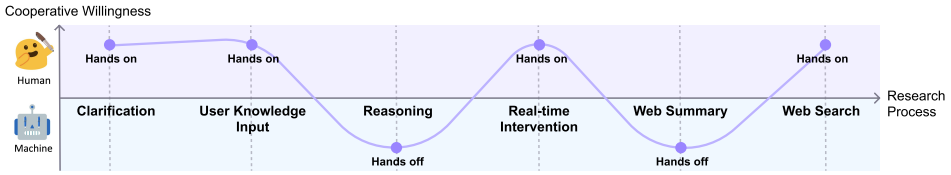

Figure 5: Changes in users' behavioral tendencies in the process of complex research tasks.

Clarification (Hands-on) The research process begins with intensive human-AI collaboration as users refine vague problem definitions. Users' initial research questions are typically too broad to cover all possible scenarios. **User Knowledge Input (Hands-on)** Users maintain high engagement when they possess specific domain knowledge or references that need integration. When users know specific references or attributes about an item, such as queries, paper links, websites, or personal opinions, they actively guide the AI to relevant media. **Reasoning (Hands-off)** Users seek to understand whether the model has correctly executed prescribed instructions and want transparency in decision-making processes. **Real-Time Intervention (Hands-on)** Cooperation peaks again during dynamic browsing tasks where users encounter pages or information sources that warrant detailed retrieval. **Web Summary (Hands-off)** During summarization tasks, users tend to trust in AI capability. Participants often need consolidated insights from multiple sources rather than single source summarization, leading them to allow extended autonomous operation. **Web Search (Hands-on)** The cycle concludes with renewed hands-on engagement for open-ended and subjective questions that require interpretation or subjective judgment.

This dynamic pattern demonstrates that effective human-AI collaboration is not uniform but adapts strategically to leverage the comparative advantages of human judgment and AI processing capabilities across different research phases. We illustrate this dynamic research task example to demonstrate authentic participant behavior.

6.2 Human-machine collaboration tools can augment expert thinking

Complex user hesitations or corrections trigger deeper reasoning processes, while smooth task completion indicates successful lightweight inference. Participants with deeper cognitive processing capabilities achieved significantly higher human-AI collaborative performance compared to those with surface-level cognitive approaches in transparent interaction paradigms, as measured by problem resolution accuracy.

6.3 Usage as Annotation becomes possible

Usage as Annotation becomes possible through thoughtful product design that transforms natural user interactions into annotation signals. When users complete tasks, their behaviors implicitly provide annotation signals that guide system adaptation. Optimal human-AI collaboration(see AppendixB) requires cognitively appropriate responses tailored to users' expertise levels, rather than merely preference alignment. Our findings show that challenging model outputs motivate users to contribute additional domain knowledge, enhancing collaborative outcomes.

7 Related Work

Human-AI Interaction AI agents White (2024); Feng et al. (2025) now support complex tasks through natural language interaction, better task understanding, and multi-level autonomy beyond basic queries interaction Srinivas & Runkana (2025); Shao et al. (2025). The shift from static monolithic inference to adaptive, resource-aware computation has become central to AI systems for knowledge discovery Shao et al. (2024); Jiang et al. (2024) leveraging multi-agent collaboration Watkins et al. (2025); Fragiadakis et al. (2025) to facilitate serendipitous discovery. This mismatch constrains the potential for AI to act as a collaborator in exploratory inquiry Pirolli (2009). Although current collaboration systems allow humans to read model reasoning chains and engage in multi-turn interactions with models Westphal et al. (2023); Gomez et al. (2025); Lee et al. (2024); Collins et al. (2024), these current interaction paradigms maintain limiting user's ability to adapt to emerging expert user's knowledge during complex and time-consuming tasks.

Deep Research Systems Deep research systems such as Gemini Deep Research Google (2025), OpenAI Deep Research OpenAI (2025b) and Grok3 Deeper Search xAI (2025) are enabled by the sophisticated reasoning abilities that have emerged from recent advances in large language models (LLMs) OpenAI et al. (2024); Guo et al. (2025); Team et al. (2025), facilitating multi-step, in-depth analysis and information synthesis across hundreds of sources. Most open-source deep research projects LangChain AI (2025); Zhang (2025); Elovic (2025); Camara (2025); Jina AI (2025); Roucher et al. (2025b); ByteDance (2024) employ prompt-based multi-agent systems with predefined workflows. Recent work Zheng et al. (2025b) has applied end-to-end reinforcement learning to open-source LLMs to perform iterative reasoning to complex questions. However, few existing deep research systems in AppendixE development multi-round interaction planning during the research process, user remain limited once research begins.

8 Conclusion

This paper introduced deep cognition, a multi-agent framework for collaborative research with real-time "cognitive oversight" through transparent, interruptible interactions. Our evaluation challenge the assumption that AI progress requires purely autonomous capabilities. Instead, our work suggests that advanced intelligence emerges from cognitive partnerships that leverage complementary human judgment and machine processing strengths.

Ethics Statement

This work adheres to the ICLR Code of Ethics. Human participants were involved in this study, and all procedures were conducted with informed consent and in strict accordance with relevant ethical standards. No personally identifiable information was collected or stored, and participants' privacy

was fully protected throughout the study. All datasets used were obtained in compliance with relevant usage guidelines. We took care to mitigate potential biases and discriminatory outcomes, and no experiments were conducted that could raise privacy or security concerns. We remain committed to ensuring transparency, fairness, and integrity in the research process.

REPREODUCIBILITY STATEMENT

We have made every effort to ensure the reproducibility of our results. To support replication and verification, we provide an online platform that allows direct access to the system. Details of the experimental setup, including training procedures, model configurations, and hardware specifications, are fully described in the paper. We also provide a comprehensive account of our contribution to facilitate reproduction. Moreover, all datasets used in this work are publicly available, ensuring consistent and reliable evaluation. We believe these measures will enable other researchers to reproduce our work and further advance the field.

REFERENCES

Lisanne Bainbridge. Ironies of automation. In *Analysis, design and evaluation of man–machine systems*, pp. 129–135. Elsevier, 1983.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling, 2024. URL https://arxiv.org/abs/2407.21787.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

ByteDance. Deerflow, 2024. URL https://github.com/bytedance/deer-flow. Community-driven deep research framework combining LLMs with web search, crawling, and code execution tools.

Nicolas Camara. Open deep research, 2025. URL https://github.com/nickscamara/open-deep-research. Open-source clone of OpenAI's Deep Research using Firecrawl for web data extraction and AI reasoning.

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent llm systems fail?, 2025. URL https://arxiv.org/abs/2503.13657.

Kaiyuan Chen, Yixin Ren, Yang Liu, Xiaobo Hu, Haotong Tian, Tianbao Xie, Fangfu Liu, Haoye Zhang, Hongzhang Liu, Yuan Gong, Chen Sun, Han Hou, Hui Yang, James Pan, Jianan Lou, Jiayi Mao, Jizheng Liu, Jinpeng Li, Kangyi Liu, Kenkun Liu, Rui Wang, Run Li, Tong Niu, Wenlong Zhang, Wenqi Yan, Xuanzheng Wang, Yuchen Zhang, Yi-Hsin Hung, Yuan Jiang, Zexuan Liu, Zihan Yin, Zijian Ma, and Zhiwen Mo. xbench: Tracking agents productivity scaling with profession-aligned real-world evaluations, 2025. URL https://arxiv.org/abs/2506.13651.

Katherine M. Collins, Ilia Sucholutsky, Umang Bhatt, Kartik Chandra, Lionel Wong, Mina Lee, Cedegao E. Zhang, Tan Zhi-Xuan, Mark Ho, Vikash Mansinghka, Adrian Weller, Joshua B. Tenenbaum, and Thomas L. Griffiths. Building machines that learn and think with people, 2024. URL https://arxiv.org/abs/2408.03943.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang,

Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang, Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Assaf Elovic. Gpt researcher, 2025. URL https://github.com/assafelovic/gpt-researcher. Open deep research agent for web and local research with detailed report generation and citations.

K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. Levels of autonomy for ai agents, 2025. URL https://arxiv.org/abs/2506.12469.

George Fragiadakis, Christos Diou, George Kousiouris, and Mara Nikolaidou. Evaluating human-ai collaboration: A review and methodological framework, 2025. URL https://arxiv.org/abs/2407.19098.

Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. Human-ai collaboration is not very collaborative yet: a taxonomy of interaction patterns in ai-assisted decision making from a systematic review. *Frontiers in Computer Science*, Volume 6 - 2024, 2025. ISSN 2624-9898. doi: 10.3389/fcomp.2024.1521066. URL https://www.frontiersin.org/journals/computer-science/articles/10.3389/fcomp.2024.1521066.

Google. Gemini deep research - your personal research assistant, 2025. URL https://gemini.google/overview/deep-research/. Accessed: April 14, 2025.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Edwin Hutchins. *Cognition in the Wild.* MIT press, 1995.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J. Semnani, and Monica S. Lam. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations, 2024. URL https://arxiv.org/abs/2408.15232.

Jina AI. node-deepresearch, 2025. URL https://github.com/jina-ai/node-DeepResearch. Iterative search, reading, and reasoning system for deep research queries with focus on concise answers.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. URL https://arxiv.org/abs/2001.08361.

Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. "help me help the ai": Understanding how explainability can support human-ai interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394215. doi: 10.1145/3544548.3581001. URL https://doi.org/10.1145/3544548.3581001.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring ai ability to complete long tasks, 2025. URL https://arxiv.org/abs/2503.14499.

LangChain. Open deep research, 2024. URL https://github.com/langchain-ai/open_deep_research. GitHub repository, accessed December 2024.

LangChain AI. Open deep research, 2025. URL https://github.com/langchain-ai/open_deep_research. Open-source research assistant for automated deep research and report generation.

Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladhak, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. Evaluating human-language model interaction, 2024. URL https://arxiv.org/abs/2212.09746.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*, 2023a.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM computing surveys*, 55(9):1–35, 2023b.

Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. Improve mathematical reasoning in language models by automated process supervision. *ArXiv preprint*, abs/2406.06592, 2024. URL https://arxiv.org/abs/2406.06592.

Meta AI. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation, 4 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: September 25, 2025.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang

Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. Minimax-01: Scaling foundation models with lightning attention, 2025. URL https://arxiv.org/abs/2501.08313.

Marvin Minsky. The society of mind. *The Personalist Forum*, 3(1):19–32, 1987. ISSN 0889065X. URL http://www.jstor.org/stable/20708493.

OpenAI. Learning to reason with llms, september 2024, 2024. URL https://openai.com/index/learning-to-reason-with-llms/.

OpenAI. Browsecomp: a benchmark for browsing agents, 2025a. URL https://openai.com/index/browsecomp/. Accessed: April 14, 2025.

OpenAI. Introducing deep research, 2025b. URL https://openai.com/index/introducing-deep-research/. Accessed: April 14, 2025.

OpenAI. Deep research system card, 2025c. URL https://cdn.openai.com/deep-research-system-card.pdf. Accessed: April 14, 2025.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph

Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. Openai o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

Perplexity AI. Introducing perplexity deep research, 2025. URL https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research. Accessed: April 14, 2025.

Peter Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42(3):33–40, 2009.

Michael Poli, Armin W Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deiseroth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, Ce Zhang, and Stefano Massaroli. Mechanistic design and scaling of hybrid architectures, 2024. URL https://arxiv.org/abs/2403.17844.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Aymeric Roucher, Albert Villanova del Moral, Merve Noyan, Thomas Wolf, and Clémentine Fourrier. Opensource deep research – freeing our search agents, 2025a. URL https://huggingface.co/blog/open-deep-research. Hugging Face Blog.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 'smolagents': a smol library to build great agentic systems. https://github.com/huggingface/smolagents, 2025b.

Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. Assisting in writing wikipedia-like articles from scratch with large language models, 2024. URL https://arxiv.org/abs/2402.14207.

Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. Future of work with ai agents: Auditing automation and augmentation potential across the u.s. workforce, 2025. URL https://arxiv.org/abs/2506.06576.

Sakhinana Sagar Srinivas and Venkataramana Runkana. Scaling test-time inference with policy-optimized, dynamic retrieval-augmented generation via kv caching and decoding, 2025. URL https://arxiv.org/abs/2504.01281.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, Chuning Tang, Congcong Wang, Dehao Zhang, Enming Yuan, Enzhe Lu, Fengxiang Tang, Flood Sung, Guangda Wei, Guokun Lai, Haiqing Guo, Han Zhu, Hao Ding, Hao Hu, Hao Yang, Hao Zhang, Haotian Yao, Haotian Zhao, Haoyu Lu, Haoze Li, Haozhen Yu, Hongcheng Gao, Huabin Zheng, Huan Yuan, Jia Chen, Jianhang Guo, Jianlin Su, Jianzhou Wang, Jie Zhao, Jin Zhang, Jingyuan Liu, Junjie Yan, Junyan Wu, Lidong Shi, Ling Ye, Longhui Yu, Mengnan Dong, Neo Zhang, Ningchen Ma, Qiwei Pan, Qucheng Gong, Shaowei Liu, Shengling Ma, Shupeng Wei, Sihan Cao, Siying Huang, Tao Jiang, Weihao Gao, Weimin Xiong, Weiran He, Weixiao Huang, Wenhao Wu, Wenyang He, Xianghui Wei, Xianqing Jia, Xingzhe Wu, Xinran Xu, Xinxing Zu, Xinyu Zhou, Xuehai Pan, Y. Charles, Yang Li, Yangyang Hu, Yangyang Liu, Yanru Chen, Yejie Wang, Yibo Liu, Yidao Qin, Yifeng Liu, Ying Yang, Yiping Bao, Yulun Du, Yuxin Wu, Yuzhi Wang, Zaida Zhou, Zhaoji Wang, Zhaowei Li, Zhen Zhu, Zheng Zhang, Zhexu Wang, Zhilin Yang, Zhiqi Huang, Zihao Huang, Ziyao Xu, and Zonghan Yang. Kimi k1.5: Scaling reinforcement learning with llms, 2025. URL https://arxiv.org/abs/2501.12599.

Elizabeth Anne Watkins, Emanuel Moss, Giuseppe Raffa, and Lama Nachman. What's so human about human-ai collaboration, anyway? generative ai and human-computer interaction, 2025. URL https://arxiv.org/abs/2503.05926.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Trans. Mach. Learn. Res.*, 2022, 2022. URL https://openreview.net/forum?id=yzkSU5zdwD.

Monika Westphal, Michael Vössing, Gerhard Satzger, Galit B. Yom-Tov, and Anat Rafaeli. Decision control and explanations in human-ai collaboration: Improving user perceptions and compliance. *Computers in Human Behavior*, 144:107714, 2023. ISSN 0747-5632. doi: https://doi.org/10.1016/j.chb.2023.107714. URL https://www.sciencedirect.com/science/article/pii/S0747563223000651.

Ryen W. White. Advancing the search frontier with ai agents, 2024. URL https://arxiv.org/abs/2311.01235.

Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *Science*, 330(6004):686–688, 2010. doi: 10.1126/science.1193147. URL https://www.science.org/doi/abs/10.1126/science.1193147.

xAI. Grok 3 beta — the age of reasoning agents, 2025. URL https://x.ai/news/grok-3. Accessed: April 14, 2025.

Shijie Xia, Yiwei Qin, Xuefeng Li, Yan Ma, Run-Ze Fan, Steffi Chern, Haoyang Zou, Fan Zhou, Xiangkun Hu, Jiahe Jin, et al. Generative ai act ii: Test time scaling drives cognition engineering. *arXiv preprint arXiv:2504.13828*, 2025.

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

Ming Yin. Bridging the gap between machine confidence and human perceptions. *Nature Machine Intelligence*, pp. 1–2, 2025.

David Zhang. Deep research, 2025. URL https://github.com/dzhng/deep-research. AI-powered research assistant for iterative, deep research using search engines, web scraping, and LLMs.

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

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments, 2025b. URL https://arxiv.org/abs/2504.03160.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, Yuxin Gu, Sixin Hong, Jing Ren, Jian Chen, Chao Liu, and Yining Hua. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese, 2025. URL https://arxiv.org/abs/2504.19314.

LLM USAGE

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text. It is important to note that thc LLM was not involved in the idcation,rescarch methodology, or experimental design. All research concepts, ideas, and analyses were developed and conducted by the authors. The contributions of the LLM were solely focused on improving the linguistic quality of the paper, with no involvement in the scientific content or data analysis. The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

Figure 6: Human–AI collaboration code book

A User Behavior Data Point

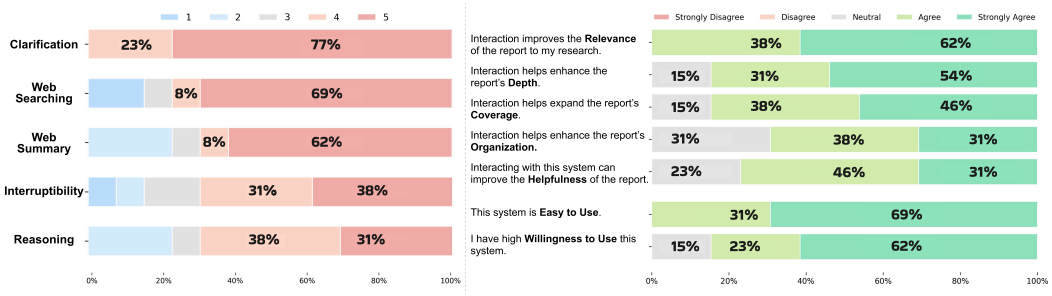

Figure 7: Left: Distribution of participant ratings (1–5) indicating the extent to which each system feature benefited their research process (n = 13 participants). Right: Perceived overall usefulness of deep cognition, as reported by the same participant cohort (n = 13 participants).

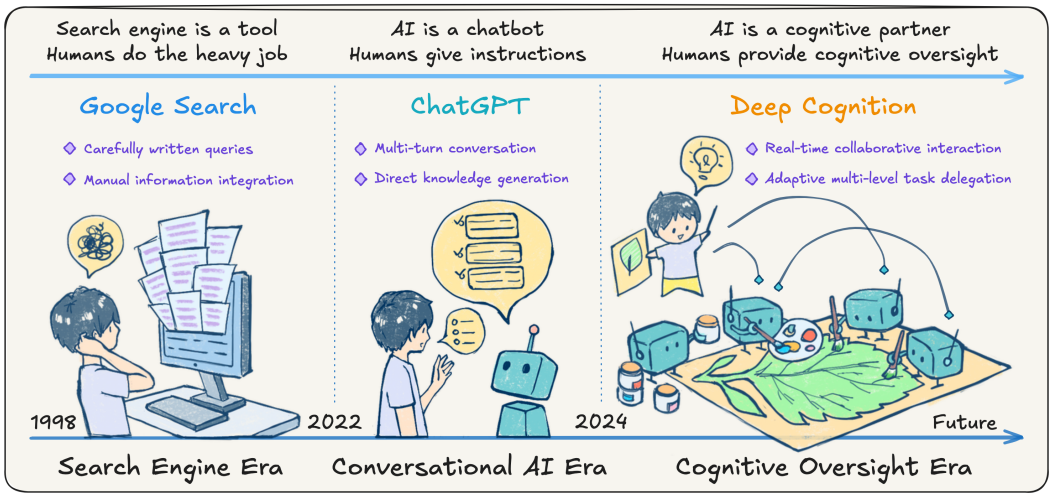

Figure 8: The evolution of human machine interaction from **operational interaction** (manual search) through **conversational interaction** (ChatGPT-style dialogue) to **cognitive interaction** (deep cognition). Our proposed paradigm transforms human-AI collaboration from periodic consultation to continuous cognitive partnership, where intelligence emerges through real-time interaction rather than autonomous processing.

B Qualitative Result

C Why Interaction Become Important In This AI Era?

Metric	Description
Organization	Evaluate whether the article demonstrates sound organization and logical structure. An acceptable response should: (1) Exhibit clear structure by organizing relevant points into a coherent logical sequence. (2) Maintain coherence without any contradictions or unnecessary repetition.
Cutting-Edge	Assess whether the article demonstrates comprehensive coverage of existing literature by: (1) Effectively summarizing and conducting comparative analysis with previous research. (2) Timely incorporating the most recent and up-to-date research findings or information.
Coverage	Provide comprehensive coverage of the identified areas of interest through: (1) Conducting thorough reviews. (2) Citing a broad range of representative scholarly works. (3) Incorporating the most current and time-sensitive information from various sources, rather than limiting the analysis to a small number of papers.
Depth	Assess the adequacy of information content provided in the article. Specifically, evaluate whether the article delivers sufficient relevant information with appropriate depth such that readers can achieve thorough understanding of each argument presented.
Relevance	Assess whether the response maintains topical relevance and preserves clear focus in order to deliver a useful response to the posed question. Specifically, the output should: (1) Sufficiently address the central elements of the original question and satisfy your informational requirements. (2) The response should exclude substantial amounts of tangential information unrelated to the original inquiry.

Metric	Description
Intention to Use	Measures user intention and propensity for continued engagement with the system based on perceived value and satisfaction.
Usability	Evaluates the intuitive nature and accessibility of the system interface, including cognitive load and interaction efficiency.
Transparency	Assesses the interpretability and explainability of the model's decision-making processes and reasoning mechanisms.
Interruptibility	Assesses the system's ability to tolerate pauses or context switches and to resume smoothly without loss of state or progress.
Fine-Grained Interaction	Evaluates the system's capacity to incorporate user feedback and enable precise, granular control over output generation.
Inspiration	Assesses the system's ability to stimulate creative thinking and generate ideas or innovative approaches to problem-solving.
Ease of Collaboration	Measures the extent to which the system functions as an effective collaborative partner in knowledge work and decision-making processes.
Results-Worth-Effort	Evaluates whether users perceive the time and effort invested in system interaction as worthwhile and valuable relative to the outcomes achieved.
Real-Time Intervention	Measures the degree to which users can actively interrupt and steer the system's ongoing processes—e.g., pausing, editing, or re-prompting—to obtain desired outputs.
Helpfulness	Assesses the overall utility and practical value of the output in addressing user needs and facilitating problem-solving objectives.

Figure 9: Evaluation Metrics for Report Quality Assessment

	Transparency	Real-time Intervention	Fine-Grained Interaction	Preference Adaptation	Cognitive Oversight	Interactive Type	Usage-as Annotation
OpenAI	★☆	×	★	×	★☆	Input-Wait-Output	×
Gemini	★★	×	★★	×	★★	Input-Wait-Output	×
Grok 3	★	×	★	×	★	Input-Wait-Output	×
DC.	★★★	✓	★★★	✓	★★★	Cognitive Interaction	✓

Figure 10: Comparison of different deep research systems (DC. indicates our deep cognition system)

D Evaluation Metrics Design

E Comparison of different deep research systems

F User Study Protocol

F.1 Pre-Study

Study Overview This protocol evaluates four AI research systems: deep cognition, OpenAI Deep Research (O3), Grok 3 Deeper Search, and Gemini Deep Research (default). Participants complete authentic research tasks requiring between 15 and 30 minutes per system, with a maximum interaction time of 30 minutes allocated to deep cognition. The full protocol see AppendixF

Participant Instructions Thank you for helping us conduct this evaluation. You need to pose a research question that you genuinely want to ask. Typically, this research question should be somewhat ambiguously defined, focused on open-ended inquiry, with substantial room for interpretation in the response, and requiring iterative search and adjustment. For example:

"I want to systematically understand current perspectives on how to position 'AI agent roles and their relationships with humans.' For instance, Anthropic CEO Dario Amodei believes that future AI agents will relate to humans as colleagues; Google published a paper on Co-scientist, viewing AI scientists as human colleagues. Please collect more viewpoints and analyze them in combination with current and future development trends."

"Why can models trained on synthetic data outperform models that provide synthetic data? Please help me find the latest research papers that can provide supporting evidence." Typically, a report may take 15-30 minutes to generate, with a maximum time limit of 30 minutes for Deep Cognition interaction. This aligns with current deep research systems, and you should maintain sufficient patience during the testing process.

"Ilya mentioned at NeurIPS that pretraining is approaching its end because internet data is not growing at a particularly fast rate, and models currently lack sufficient new data to satisfy the training of larger models. Therefore, a current challenge is how to improve data utilization efficiency (as mentioned by OpenAI researchers) - assuming there are approximately 50T tokens of data on the internet, how can we utilize these 50T tokens effectively to improve the intelligence ceiling of models? Please help me research relevant materials and literature, identifying methods for improving data utilization efficiency and ways to collect more data. For example, current web data is static - how might we obtain dynamic data, such as behavioral traces?"

1. Pre-Study (Understanding System Usage) This is a tool for real-time human-AI collaboration, retrieving open-ended multi-hop questions, allowing users to dynamically explore initial questions during system interaction and ultimately complete comprehensive writing. Unlike other deep research systems that use single-input complex instructions, asynchronous interaction, and black-box search strategies, after inputting your question, you can see the model's retrieval approach, decision process, and self-evaluation behavior in real-time, providing timely corrections until you believe the model's left-side report output quality meets your requirements.

You cannot directly manually modify the model's final report. You need to guide the model to improve report writing depth and information retrieval efficiency through various interaction methods during the model's research process (interruption, adding expert prior knowledge, reviewing model-retrieved information, auditing the model's self-evaluation process, new thinking, strategic guidance, or personal files). Please note that you should aim to achieve 4-5 points across all dimensions before stopping generation. You can interrupt at any time before the model finishes. The termination point is when the model autonomously decides to finish.

Model Settings: After selecting "Clarify Question" copy and record the thought chain returned on the right side. You need to simultaneously review the behavioral patterns returned by the model on the right side. When using Deep Cognition, you need to enable the switch in the bottom right corner.

F.2 In-Study

Understanding Evaluation Metrics During generation across all systems, you need to timely review the model's behavior (right-side thought chains, expanded model execution details, all searched URLs, information retrieved from URLs) and the quality of model-generated reports (left-side drafts).

F.2.1 Evaluation Framework

Organization

Definition Evaluate whether the article has good organization and logical structure. An acceptable response should: 1. Have clear structure, categorizing related points into a logical flow. 2. Be coherent, without contradictions or unnecessary repetition.

Score 5: Exceptional Organization

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Structure Clarity:** Perfect logical structure with clear hierarchical organization and seamless section transitions;

Evaluation Dimension	Pool	Basic	Average	Strong	Exceptional
Organization: Structural clarity and logical flow	○	○	○	○	○
Cutting-edge Information: Coverage of recent, high-impact research	○	○	○	○	○
Information Coverage (Breadth): Comprehensiveness across research domains	○	○	○	○	○
Information Depth: Sufficiency of detail for thorough understanding	○	○	○	○	○
Overall Helpfulness: Practical utility for literature review and research	○	○	○	○	○

Table 4: 5-Point Likert Scale for Assessing Report Quality

- **Logical Flow:** Flawless reasoning progression from introduction to conclusion with excellent coherence;
- **Coherence:** All content elements perfectly interconnected with consistent thematic development;
- **Presentation Quality:** Outstanding formatting and layout that enhances readability and comprehension;

Score 4: Strong Organization

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Structure Clarity:** Response is well-organized with clear, logical structure consistently followed;
- **Logical Flow:** Points are effectively grouped, flow is smooth;
- **Coherence:** Minor coherence issues but overall clear and easy to follow with minimal repetition or contradictions;
- **Presentation Quality:** Good formatting that supports understanding;

Score 3: Moderate Organization

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Structure Clarity:** Response is generally well-organized with clear structure that is basically maintained;
- **Logical Flow:** Adequate progression with some choppy transitions;
- **Coherence:** Reasonable thematic development with some disconnected elements;
- **Presentation Quality:** Acceptable formatting with room for improvement;

Score 2: Basic Organization

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Structure Clarity:** Some organization but inconsistent structure, minor contradictions;
- **Logical Flow:** Weak reasoning progression with confusing transitions;
- **Coherence:** Limited thematic coherence with noticeable gaps;
- **Presentation Quality:** Poor formatting that hinders comprehension;

Score 1: Poor Organization

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Structure Clarity:** No clear structure, scattered points, difficult to follow;

- **Logical Flow:** No discernible logical progression, chaotic presentation;
- **Coherence:** No thematic coherence, completely disconnected content;
- **Presentation Quality:** Very poor formatting that severely impairs understanding;

Cutting-Edge Information

Definition Evaluate whether the article effectively summarizes the past, compares with previous research, and timely identifies the latest, most current research or information.

Score 5: Exceptional

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Recency:** Precisely captures key latest research in the field, including recently published technical reports, preprints, conference reports, and ongoing work;
- **Impact Level:** Includes highest-impact research and breakthrough discoveries, keen insight into cutting-edge issues and breakthrough progress, can identify emerging directions not yet widely recognized;
- **Coverage Completeness:** Comprehensive coverage of all major recent developments;
- **Source Quality:** Exclusively high-quality, authoritative sources from leading institutions;

Score 4: Strong

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Recency:** Response successfully identifies most important recent research achievements and breakthrough work;
- **Impact Level:** Covers major high-impact developments with good selection. Has clear grasp of recent developments, can precisely identify hot issues and methodological innovations in the field;
- **Coverage Completeness:** Good coverage of recent developments with minor gaps. Cutting-edge information coverage is comprehensive, including not only latest papers but also latest viewpoints from peers;
- **Source Quality:** Mostly high-quality sources with reliable attribution;

Score 3: Moderate

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Recency:** Response identifies a certain number of recent research achievements, covering some important latest developments;
- **Impact Level:** Includes moderately impactful research with some selection issues. Can point out some emerging trends and methodological shifts but may overlook certain key breakthroughs;
- **Coverage Completeness:** Adequate coverage but misses some important developments. Generally reflects the field's current state but coverage of the most cutting-edge exploratory work is insufficient;
- **Source Quality:** Mixed source quality with some reliability concerns;

Score 2: Basic

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Recency:** Limited recent research, misses important developments. Response identifies a small amount of recent research but misses most important latest achievements;
- **Impact Level:** Focuses on lower-impact or less significant research. Fails to adequately reflect the field's current active state and latest trends;

- **Coverage Completeness:** Poor coverage with significant gaps in recent developments. Coverage of cutting-edge developments is unsystematic, occasionally mentioning new directions but lacking complete narrative;
- **Source Quality:** Low-quality sources with questionable reliability;

Score 1: Poor

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Recency:** Response lacks coverage of high-impact recent work, with almost no identification of recent or cutting-edge research. Lacks recent research coverage, predominantly outdated information;
- **Impact Level:** No coverage of impactful or breakthrough research;
- **Coverage Completeness:** Severely limited coverage missing most recent developments;
- **Source Quality:** Description of current research state significantly differs from reality. Very poor or unreliable sources;

Information Coverage (Breadth)

Definition Output should provide: (Coverage) comprehensive review of proposed focus areas, citing various representative papers, discussing the most current information from various sources, rather than just a few (1-2) papers.

Score 5: Exceptional

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Domain Scope:** Comprehensive coverage: answer covers various different papers and viewpoints, providing comprehensive field overview;
- **Perspective Diversity:** Multiple viewpoints and approaches from different research communities. Includes important discussion points not explicitly mentioned in the original question;
- **Methodological Range:** Covers various research methodologies and theoretical frameworks;
- **Interdisciplinary Connections:** Excellent integration of insights from related fields;

Score 4: Strong

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Domain Scope:** Broad coverage: output covers the field, discussing various representative papers and materials;
- **Perspective Diversity:** Good variety of viewpoints with most major perspectives covered. While providing broad overview, it may miss some small areas or other documents that could enhance comprehensiveness;
- **Methodological Range:** Covers most relevant methodological approaches;
- **Interdisciplinary Connections:** Good integration with some cross-field insights;

Score 3: Moderate

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Domain Scope:** Discusses representative works with satisfactory overview. Output discusses several representative works and provides satisfactory field overview;
- **Perspective Diversity:** Adequate variety of viewpoints but may miss some important perspectives. However, adding more papers or discussion points could significantly improve the answer;

- **Methodological Range:** Covers basic methodological approaches with some gaps. Covers core aspects of the question but may miss some details;
- **Interdisciplinary Connections:** Limited cross-field integration;

Score 2: Basic

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Domain Scope:** Partial coverage, misses important research directions. Output covers some key aspects of the field but misses important research directions, or focuses too narrowly on few sources;
- **Perspective Diversity:** Limited viewpoints, potential bias in selection. Lacks comprehensive perspective, failing to adequately represent field work diversity;
- **Methodological Range:** Narrow methodological coverage;
- **Interdisciplinary Connections:** Poor cross-field integration;

Score 1: Pool

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Domain Scope:** Severely limited coverage, focuses on single domain. Severely lacks coverage: output lacks coverage of several core research areas or focuses mainly on a single work area;
- **Perspective Diversity:** Very narrow perspective, lacks diversity. Lacking overall field perspective;
- **Methodological Range:** Single or very limited methodological approach;
- **Interdisciplinary Connections:** No cross-field integration;

Relevance

Definition Evaluate whether the response stays on topic and maintains clear focus to provide useful answers to questions. Specifically, output should: 1. Adequately address core points of original question and meet your information needs (if factual). 2. Not contain much secondary information unrelated to original question.

Score 5: Focused and entirely on topic

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Topic Focus:** Response consistently stays closely on topic with clear focus on solving the problem;
- **Information Relevance:** Every piece of information directly contributes to comprehensive topic understanding;
- **Content Quality:** Sufficient depth of understanding and coverage of core information;
- **User Needs:** Fully addresses core points of original question and meets information needs;

Score 4: Mostly On-Topic with Minor Deviations

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Topic Focus:** Response is basically topic-relevant and clearly focuses on solving the problem;
- **Information Relevance:** Most content directly relates to the main question with minor irrelevant details;
- **Content Quality:** Minor off-topic deviations that temporarily distract from topic focus but don't significantly impact clarity;
- **User Needs:** Adequately addresses most core points with minimal distraction;

Score 3: Somewhat on topic but with several digressions or irrelevant information
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Topic Focus:** Response still revolves around original question but frequently deviates from topic;
- **Information Relevance:** Contains some redundant information or minor irrelevant points;
- **Content Quality:** Noticeable digressions that affect focus but main topic remains discernible;
- **User Needs:** Partially addresses core points but with unnecessary diversions;

Score 2: Frequently Off-Topic with Limited Focus
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Topic Focus:** Article somewhat addresses the question but frequently deviates from topic;
- **Information Relevance:** Contains significant amount of irrelevant information or unrelated points;
- **Content Quality:** Multiple diversions that don't help with main question and reduce overall utility;
- **User Needs:** Limited success in addressing core points of original question;

Score 1: Off-topic
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Topic Focus:** Content severely deviates from original question;
- **Information Relevance:** Difficult to discern relevance to the original question;
- **Content Quality:** Diverts user attention from intended topic and fails to provide useful answers;
- **User Needs:** Fails to address core points and does not meet information needs;

Information Depth

Definition Evaluate whether the article provides sufficient information. Depth provides sufficient relevant information so readers can thoroughly understand each argument in the article.

Score 5: Excellent Coverage and Amount (depth)
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Detail Sufficiency:** Provides necessary and sufficient information with selective deep exploration. Can select materials requiring deep exploration for detailed discussion;
- **Technical Accuracy:** Highly accurate technical details with proper context;
- **Analytical Depth:** Deep analytical insights with sophisticated reasoning. Response provides all necessary and sufficient materials;
- **Contextual Understanding:** Excellent understanding of broader implications and context;

Score 4: Good Coverage and Amount (depth)
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Detail Sufficiency:** Includes most relevant information needed to understand the topic. Avoids excessive irrelevant details, but several points might benefit from deeper exploration or more specific examples;
- **Technical Accuracy:** Good technical accuracy with minor gaps;

- **Analytical Depth:** Good analytical insights with solid reasoning. Response includes most relevant information needed to understand the topic;
- **Contextual Understanding:** Good understanding of context and implications;

Score 3: Acceptable Coverage and Amount (depth)
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Detail Sufficiency:** Acceptable amount of relevant information, may lack some useful details;
- **Technical Accuracy:** Adequate technical accuracy with some inaccuracies;
- **Analytical Depth:** Output provides reasonable amount of relevant information, though it may lack some useful details.;
- **Contextual Understanding:** Basic understanding of context;

Score 2: Limited Coverage and Amount (depth)
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Detail Sufficiency:** Provides some relevant information but misses important details;
- **Technical Accuracy:** Poor technical accuracy with significant errors;
- **Analytical Depth:** Response provides some relevant information but misses important details that would aid full topic understanding.;
- **Contextual Understanding:** Poor understanding of broader context;

Score 1: Lack of Coverage and Amount (depth)
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Detail Sufficiency:** Lacks basic details needed for topic understanding;
- **Technical Accuracy:** Very poor technical accuracy with major errors;
- **Analytical Depth:** Output either lacks basic details needed for adequate topic understanding (e.g., method definitions, relationships between methods);
- **Contextual Understanding:** No understanding of context or implications;

Overall Helpfulness

Definition Do you find the provided answer overall helpful? Does it assist with your literature review? Evaluate the overall utility of the response for research and learning purposes.

Score 5: Super Useful. I can fully trust the answer
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Question Addressing:** Answer provides comprehensive field overview and fully answers the question;
- **Source Quality:** Provides high-quality, trustworthy sources with comprehensive coverage;
- **Research Utility:** Serves as complete foundation for research without need for independent verification;
- **Information Reliability:** I believe I don't need to independently search for other papers or detailed information;

Score 4: Useful. I may try to verify some details, but overall gives great summary
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Question Addressing:** Answer provides detailed information and good overview of the area of interest;

- **Source Quality:** Provides high-quality, fresh sources across multiple sources with good diversity;
- **Research Utility:** Requires minimal additional editing, serves as excellent foundation for further work;
- **Information Reliability:** May need to check details of 1-2 specific papers/sources, but overall highly reliable;

Score 3: Provides some useful discussions and papers, though requires independent reading

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Question Addressing:** Answer is generally helpful and provides good overview with diverse perspectives;
- **Source Quality:** Provides at least 2-3 useful information sources previously unknown to reader;
- **Research Utility:** Can base further reading on recommended papers, good starting point for deeper research;
- **Information Reliability:** May need to independently verify some details or consult other core research papers;

Score 2: Better than searching from scratch but limited utility

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Question Addressing:** Answer provides at least one useful starting point but discussions are somewhat irrelevant;
- **Source Quality:** Provides at least one useful paper that can be read carefully;
- **Research Utility:** Limited utility for research purposes, requires significant additional work;
- **Information Reliability:** Overall discussions don't provide sufficiently useful information for the topic;

Score 1: Unhelpful

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Question Addressing:** Answer doesn't address the question or provides confusing information;
- **Source Quality:** Hasn't conducted effective retrieval, still generating using pretrained knowledge;
- **Research Utility:** Cannot serve as useful starting point for learning or writing relevant content;
- **Information Reliability:** Fails to provide understanding of literature in this field;

F.2.2 System Design Evaluation (-2 to +2 Scale)

Evaluation Dimension	-2	-1	0	+1	+2
Transparency: Decision-making process visibility	○	○	○	○	○
Interruptibility: Real-time intervention capability	○	○	○	○	○
Fine-grained Interaction: Interaction granularity level	○	○	○	○	○
Inspiration: Unexpected discoveries and insights	○	○	○	○	○
Collaboration: Collaborative partnership quality	○	○	○	○	○

Table 5: System Design Assessment Rubric

System Design Evaluation Definition

Question: Does the system design provide sufficient transparency in decision-making processes?

Interruptibility (Interruptible at any time): To what extent do you think interruptibility can help correct the model's research approach and reduce model errors?

Fine-grained and Bidirectional Interaction: How fine-grained do you think the current system's interaction is? (Interaction refers to nodes where users can provide input to the model)

Inspirational Perspectives (Shared cognitive context as exploration space): How much information in the model's decision and search process exceeded your expectations? Did it help inspire you?

Inspirational Perspectives (Shared cognitive context as exploration space): How much information in the model's decision and search process exceeded your expectations? Did it help inspire you?

Long-term Collaboration Willingness: Deep research systems can all interact (Deep Cognition during process, other 3 systems after research process). Research is a dynamic, multi-round complex long-term task. To what extent do these systems' interaction methods (including input methods and system feedback output methods) make you willing to engage in long-term, multi-round communication and collaboration with the system?

Long-term Collaboration Willingness: Deep research systems can all interact (Deep Cognition during process, other 3 systems after research process). Research is a dynamic, multi-round complex long-term task. To what extent do these systems' interaction methods (including input methods and system feedback output methods) make you willing to engage in long-term, multi-round communication and collaboration with the system?

+2 points - Excellent:

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Process Visibility:** Complete visibility of thinking, actions, and browsed content;
- **Decision Rationale:** Clear explanation of all decision-making processes;
- **Source Verification:** Full source verification and citation transparency;
- **Strategy Disclosure:** Complete disclosure of search and analysis strategies;

+1 points - Good:

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Process Visibility:** Good transparency with some decision process visibility;
- **Decision Rationale:** Adequate explanation of major decisions;
- **Source Verification:** Good source transparency with minor gaps;
- **Strategy Disclosure:** Partial disclosure of strategies and approaches;

0 points - Neutral:

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Process Visibility:** Neutral/adequate transparency level;
- **Decision Rationale:** Basic explanation of some decisions;
- **Source Verification:** Adequate source information;
- **Strategy Disclosure:** Limited strategy disclosure;

-1 points - Poor:

[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Process Visibility:** Limited transparency, unclear decision processes;
- **Decision Rationale:** Poor explanation of decision-making;

- **Source Verification:** Limited source transparency;
- **Strategy Disclosure:** Minimal strategy disclosure;

-2 points - Extremely Poor:
[topsep=0pt, partopsep=0pt, itemsep=0pt]

- **Process Visibility:** Black box operation with no process visibility;
- **Decision Rationale:** No explanation of decision-making processes;
- **Source Verification:** No source transparency or verification;
- **Strategy Disclosure:** No disclosure of strategies or methods;

F.2.3 DEEP COGNITION SPECIFIC EVALUATION

Qualitative indicator: When comparing the Deep Cognition system with other deep research systems, do the system's functional designs (interruptibility, transparent thinking process, transparent behavioral paths, presenting search queries, displaying retrieved content) enhance this system's collaborative attributes?

Follow-up questions: A. If enhanced, can you provide specific examples? Which functions enhanced collaborative attributes? B. During model behavior review, could the model provide new insights/unexpected search information?

Feature	Description
Text Input	Basic text communication capability
Question Clarification	System's ability to clarify ambiguous queries
Expert Information Integration	Incorporating domain expertise
Thinking Process Visibility	Transparency of reasoning steps
Decision Process	Clarity of decision-making rationale
Interruptibility	Effectiveness of real-time intervention
Content Summary Reading	Quality of information synthesis
Search Query Visibility	Transparency of search strategies

Table 6: Deep Cognition Feature-Specific Ratings (1-5 Scale)

F.3 POST-STUDY-2

Deep Cognition Evaluation: -2 for strongly negative, 0 for neutral, 2 for strongly positive

1. Enhanced Effectiveness (Enhance cognitive efficiency or not)

To what extent do you think this collaborative approach can improve final report generation quality (organization and consistency/information coverage/information density (depth)/relevance/overall helpfulness)?

Dimension	Score (-2/-1/0/1/2)	Reason
Organization and consistency		
Information coverage		
Information density (depth)		
Relevance		
Overall helpfulness		

2. Results-worth-effort Interacting with these systems costs your time and energy. Do you think it's worth it? How worthwhile?

System	Score (-2/-1/0/1/2)	Reason
Deep Cognition		
OpenAI		
Gemini		
Grok 3		

3. Research Stage Evaluation

At which stages do you think interrupting the model's operation can effectively improve subsequent report quality? Which stage can enhance your real-time collaboration willingness with the model?

Current model nodes include: evaluating research status, generating search queries, filtering webpage URLs, browsing webpages, extracting summaries from webpages and determining usefulness, prioritizing information retrieved from webpages and organizing arguments.

You may define research stages according to your own understanding when asking this question.

Follow-up questions:

a) At which stage of model research development is your collaboration willingness higher?

b) Can the model's research process provide you with insights? Can you give an example (screenshot or text)?

c) At which stages do you think interrupting the model's operation can more effectively improve subsequent report quality? Which stage can enhance your real-time collaboration willingness with the model?

4. Usage Willingness and Learning Cost (Interaction Willingness)

Quantitative indicators: To what extent are you willing to use this system? How are the learning costs and operational burden?

Aspect	Score (-2/-1/0/1/2)	Reason
Usage willingness		
Ease of operation		

5. Feature Evaluation

How helpful are these features for your research process? Rate (1-5) and explain reasons.

Feature Number	Feature Name	Score	Comments
1	Send text		
2	Clarify questions		
3	Add expert information		
4	Thinking process		
6	Decision		
7	Interruptible		
8	Read summaries		
9	Search queries		

G CASE STUDY

A Dynamic Research Task Case

Domain: Interdisciplinary Writing

Initial query: Please assist in investigating the latest innovations in interactive and scalable interfaces designed to enhance the interpretability of large language models in writing.

Clarify query: What aspect of interpretability are you most interested in exploring? Who is the primary target audience for these interfaces? What writing contexts are you most interested in? What aspects of interface scalability are most relevant to your research? Are you interested in any specific emerging technologies related to LLM interpretability?

Domain knowledge input: Add Jeff Rzeszotarski's PhD dissertation, and research in PAIR (People + AI Research Initiative) team.

Initial goal: Development trend of interpretability of Interpretable Machine Learning Interface

Last goal: Investigate which research fields the scholars who previously worked in this direction have migrated to.