



Yumanshan Ye, Xiaojie Cai, Junfei Wang, Xinkai Wang, Xiangkun Hu, Jiadi Su, Yang Nan, Sihan Wang, Bohan Zhang, Xiaoze Fan, Jinbin Luo, Yuxiang Zheng, Tianze Xu, Dayuan Fu, Yunze Wu, Pengrui Lu, Zengzhi Wang, Yiwei Qin, Zhen Huang, Yan Ma, Zhulin Hu, Haoyang Zou, Tiantian Mi, Yixin Ye, Ethan Chern, Pengfei Liu¹

Abstract

This paper introduces “Interaction as Intelligence” research series, presenting a fundamental reconceptualization of human-AI relationships in deep research tasks. Traditional approaches treat interaction merely as an interface for accessing AI capabilities—a conduit between human intent and machine output. We propose that **interaction itself constitutes a fundamental dimension of intelligence**. As AI systems engage in extended thinking processes for complex research tasks, meaningful interaction transitions from an optional enhancement to an essential component of effective intelligence. Current deep research systems uniformly adopt an “input-wait-output” paradigm where users initiate queries and receive results after prolonged black-box processing. This approach leads to error cascade effects, inflexible research boundaries that prevent question refinement during investigation, and missed opportunities for expertise integration. To address these limitations, we introduce **Deep Cognition**, a system that transforms the human role from giving instructions to **cognitive oversight**—a mode of engagement where humans guide AI thinking processes through strategic intervention at critical junctures. Deep cognition implements three key innovations: (1) transparent, controllable, and interruptible interaction that reveals AI reasoning and enables intervention at any point; (2) fine-grained bidirectional dialogue; and (3) shared cognitive context where the system observes and adapts to user behaviors without explicit instruction. User evaluation demonstrates that this cognitive oversight paradigm significantly outperforms the strongest baseline across six key metrics: **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%)**. Evaluations on challenging deep research problems show **31.8% to 50.0% points of improvements** over competitive deep research systems.

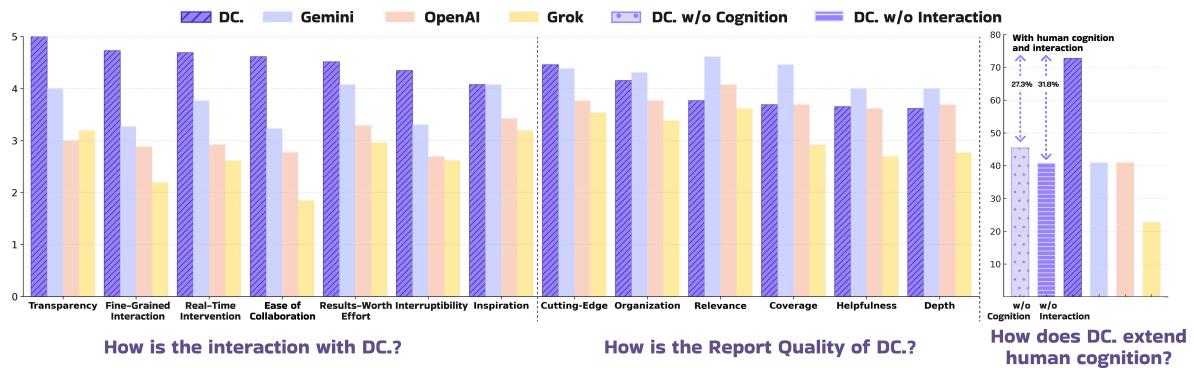


Figure 1: **Overall evaluation results.** We present the user evaluation (seven metrics on the left part), report quality (six metrics in the middle), and evaluation results on deep research problems (the right part) with three conditions: *Without Cognition*, *With Cognition & Interaction*, and *Without 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”.

¹ Corresponding author

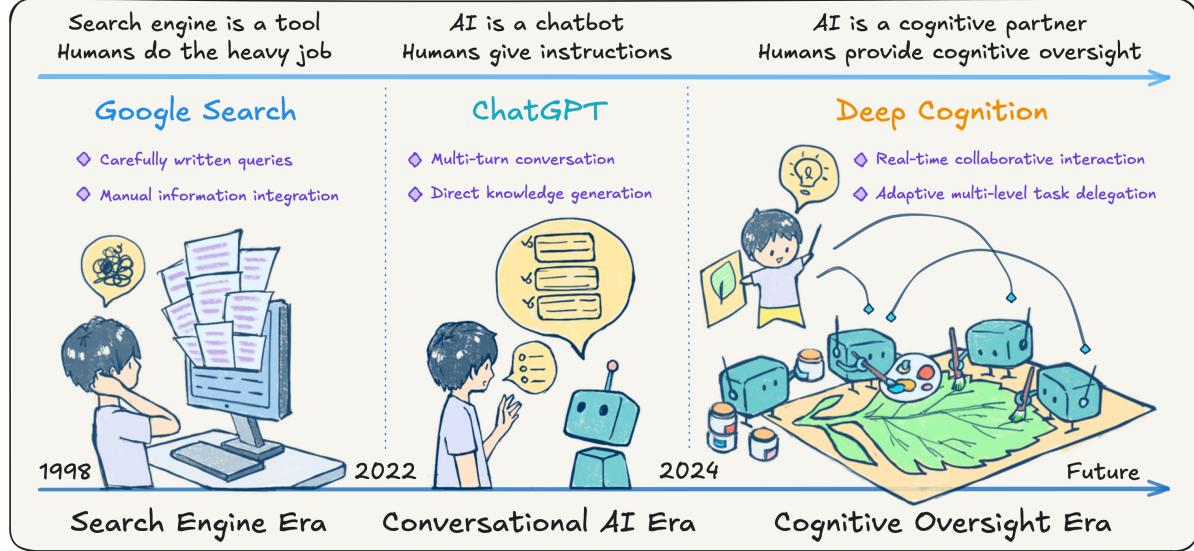


Figure 2: The evolution of human machine interaction from **operational interaction** (manual search) through **conversational interaction** (ChatGPT-style dialogue) to **cognitive interaction** (deep cognition). Our proposed paradigm transforms human-AI collaboration from periodic consultation to continuous cognitive partnership, where intelligence emerges through real-time interaction rather than autonomous processing.

1 Introduction

"Intelligence is not the property of an isolated mind, but emerges in the dance between minds. The question is not how smart the individual components are, but how brilliantly they interact."

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., 2025a; 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., 2023; Kim et al., 2023) or decision-making (Yin, 2025). [—creating systems that increasingly function as autonomous black boxes. 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. —Ivye] 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 involve iterative cycles of hypothesis formation, testing, revision, and collaborative refinement—demonstrating that expert consultation and interdisciplinary dialogue are integral components of intelligence at its highest levels.

We contend that the current assumption fundamentally mischaracterizes the nature of intelligence itself. 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.

We propose a radical reconceptualization: *interaction itself constitutes a fundamental dimension of intelligence, enabling capabilities that neither humans nor AI can achieve independently*. This recognition represents a pivotal shift in how we conceptualize AI progress: **from minimizing human involvement to optimizing human-AI cognitive partnership**. The evolution of human-computer interaction in information seeking exemplifies this fundamental transformation. As shown in Figure 2, initially, interaction was purely operational: users manually retrieved information through explicit queries and integrated fragmented results themselves. The conversational era,

1 INTRODUCTION

exemplified by LLM-based chatbots such as ChatGPT (OpenAI, 2022), introduced natural language dialogue but still demanded extensive multi-turn conversations to accomplish complex tasks. We now stand at the threshold of a new paradigm where humans and AI communicate at the cognition level, enabling strategic intervention rather than constant guidance.

This transition is particularly evident in systems designed for Deep Research tasks (OpenAI, 2025a; 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, 1983a). Rather than viewing interaction as overhead, we need systems that enable continuous and meaningful collaboration throughout the research process. 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:

- **Cognitive 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 not just what the AI does, but also how it 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. This enables for targeted adjustments rather than starting over completely.
- **Adaptive Cognitive Context:** Deep cognition evolves its research strategies by learning from accumulated interaction patterns, developing a sophisticated understanding of user preferences without explicit retraining. This adaptive capability is grounded in In-context Reinforcement Learning (ICRL), where neural networks learn algorithms purely through contextual conditioning (Laskin et al., 2022; Lin et al., 2024; Lee et al., 2023). Just as ICRL models adapt and often surpass their training performance through context alone (Huang et al., 2024; Grigsby et al., 2024), deep cognition progressively refines its collaborative approach, creating a dynamic partnership that better anticipates user needs over time.

These principles fundamentally transform deep research from conventional question-and-answer exchanges into cognitive collaboration—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:

- **Cognitive Oversight:** We propose a human-AI collaboration paradigm—cognition oversight—which augments the intelligence through human-AI partnership.
- **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.
- **Significant Performance Improvements:** Experiment results reveal significant improvements over strong deep research systems on both user evaluation and solving deep research problems.

2 RELATED WORK

2 Related Work

The relationship between humans and machines has undergone a profound evolution over the past several decades. In this section, we review human-AI interaction paradigm and recently emerged deep research systems, and explain their differences with our work.

2.1 Human-AI Interaction

We categorize Human-AI interaction approaches as operational interaction (Isinkaye et al., 2015; Pazzani and Billsus, 2007; Resnick and Varian, 1997; Fok and Weld, 2024), conversational interaction (OpenAI, 2022; McTear et al., 2016; Candello and Pinhanez, 2016; OpenAI, 2022), and mixed-initiative interaction (Gervasio et al., 2025; Bansal et al., 2024; Bremers and Ju, 2024; Liu et al., 2025a; Chen et al., 2025b). Operational interaction emphasizes system-led automation with minimal user intervention, typical in traditional recommender systems (Isinkaye et al., 2015; Pazzani and Billsus, 2007; Resnick and Varian, 1997). While efficient, these systems limit adaptability and user agency. However, these current human-AI interaction approaches are still guided by the design principle of minimizing the need for human feedback, it shifts substantial cognitive burden to users. Previous studies proposed interaction systems for mapping general-purpose AIs to the right specific use cases (Jiang et al., 2024; Kim et al., 2023; Hoffmann et al., 2022; Jin et al., 2025b; Min et al., 2025; Cao et al., 2025; Pu et al., 2025; Wang et al., 2025; Liu et al., 2025b), however, humans still bear the cognitive burden of manual query formulation (Yen et al., 2024; Rosenberg et al., 2024), result interpretation, knowledge synthesis (Cai et al., 2025; Liu et al., 2024a; Chen et al., 2024) and new idea generation (Liu et al., 2024b; Rayan et al., 2024; E et al.; Suh et al., 2024) when they interact with these systems. For example, Min and Xia (2025) explored anticipatory systems that scaffold the ideation process. Despite their differing interaction modalities, many current systems converge on a design goal of minimizing user input, inadvertently amplifying cognitive effort required for synthesis and meaning-making. These systems required users to express their needs in rigid query formats and manually synthesize insights from fragmented outputs. Moreover, cognitive work was almost entirely human-driven, with systems serving as a mere instruction follower rather than an active human collaborator with system as an equality cooperative partner (Shi et al., 2025; Zheng et al., 2025a; Jin et al., 2025a; Song et al., 2025; Chen et al., 2025a; Xu et al., 2025; Qin et al., 2025; Li et al., 2025). As Bainbridge (1983b) warns, minimizing human feedback can erode operators' skills and situational awareness—only to make their intervention when the system inevitably fails (Yang et al., 2025b; Cemri et al., 2025).

As highlighted by White (2024) and Feng et al. (2025), AI agents now support complex tasks through natural language interaction, better task understanding, and multi-level autonomy beyond basic queries interaction (Srinivas and 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. Such instruction-following paradigms are ill-suited for open-ended sensemaking tasks, where goals evolve and insight emerges iteratively. This mismatch constrains the potential for AI to act as a collaborator in exploratory inquiry (Pirolli, 2009). Pirolli (Pirolli, 2009) defined complex information seeking as not simply the act of locating information; it is a central component of the broader sense-making process. In this process, individuals identify, collect, filter, interpret, and organize information from large and diverse sources in order to construct meaningful knowledge. Human experts readily adapt their investigative direction in response to unexpected findings as their understanding develops. Although current human-AI collaboration systems allow humans to read simplified model reasoning chains and engage in multi-turn asynchronous interactions with models (Westphal et al., 2023; Gomez et al., 2025; Lee et al., 2024; Collins et al., 2024), these current interaction paradigms maintain static problem representations throughout extended processing, limiting user's ability to adapt to emerging insights or evolving user understanding during complex and time-consuming tasks.

2.2 Deep Research Systems

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

Table 1: Comparison of different deep research systems (DC. indicates our deep cognition system).

Deep research systems have garnered significant attention and adoption following the introduction of commercial

platforms such as Gemini Deep Research (Google, 2025), OpenAI Deep Research (OpenAI, 2025b) and Grok3 Deeper Search (xAI, 2025). These systems assist users in retrieving specific information for complex queries and conducting comprehensive, in-depth surveys on particular topics. Deep research capabilities 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. The implementation details of prominent commercial deep research systems remain proprietary and undisclosed. In contrast, most open-source deep research projects (LangChain AI, 2025; Zhang, 2025; Elovic, 2025; Camara, 2025; Jina AI, 2025; Roucher et al., 2025; 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, web searching, and browsing, ultimately generating comprehensive answers to complex questions. However, to the best of our knowledge, few existing deep research systems treat human-AI interaction during the research process as a first-class design consideration. While systems like Gemini Deep Research propose research plans or clarification questions to allow users to refine research objectives and requirements, user agency remains limited once research commences. Users cannot interrupt the ongoing process and can only engage through post-editing of generated reports or by asking follow-up questions after research completion. We summarize the comparisons with competitive deep research systems in Table 1.

The advancement toward increasingly sophisticated AI systems requires not merely enhanced model capabilities, but deeper integration between human and machine cognition. In this work, our proposed deep cognition system adopts a prompt-based multi-agent approach while incorporating extensive human interaction guided by the principles outlined in Section 1.

3 Deep Cognition

We propose deep cognition, a multi-agent system designed to emulate human cognitive processes in deep research scenarios. This section provides a detailed description of our framework and its human-AI interaction mechanisms.

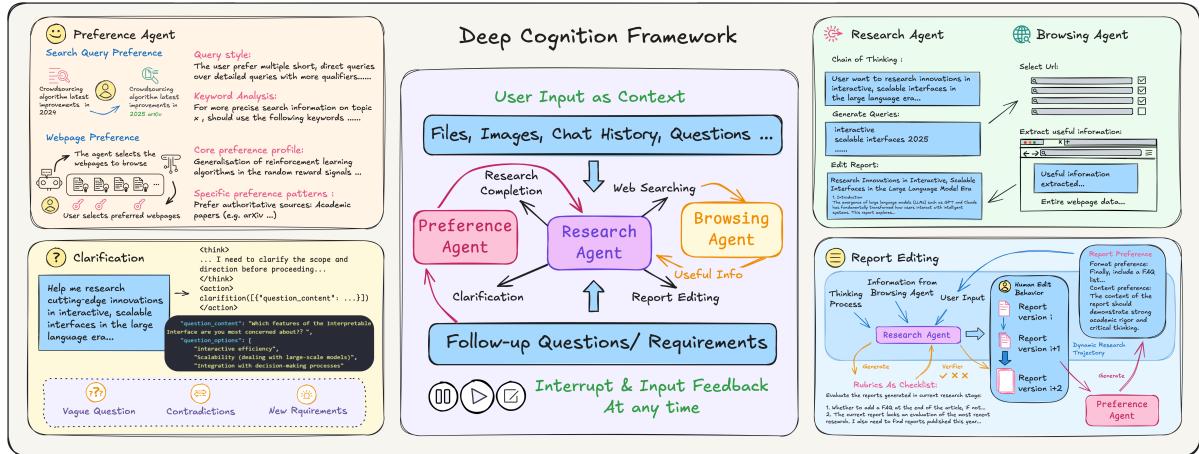


Figure 3: Deep cognition framework overview. This human-in-the-loop research assistant system breaks down complex research questions and dynamically synthesizes information from multiple sources through iterative search, clarification, and user feedback. The central diagram illustrates the overall agent framework architecture. The framework integrates four key processes: user preference modeling based on historical interaction patterns, query refinement and clarification mechanisms, coordinated research orchestration, and external environment interaction with automated report generation.

3.1 The Deep Cognition Framework

As illustrated in Figure 3, deep cognition is implemented as a multi-agent system comprising a **research agent**, a **browsing agent**, and a **preference agent**, supported by LLMs and search engine APIs. The research agent serves as the primary cognitive entity that collaborates with users to conduct deep research tasks, the browsing agent performs web browsing and gathers relevant information from multiple web pages with low latency and high accuracy, and the preference agent analyzes user trajectory to continuously adapt the system’s behavior to user preferences.

3.1.1 Research Agent

The research agent receives input context from users, including research questions, uploaded files (e.g., PDF files), images, or historical chat conversations. It conducts research through an interleaved thinking and action approach, leveraging large language models with advanced reasoning capabilities such as o1 (OpenAI et al., 2024), DeepSeek-R1 (Guo et al., 2025), or Claude Sonnet 4 (Anthropic, 2025), etc. The research agent can execute the following actions during the research process:

- **Clarification:** When the research agent encounters ambiguities in research questions, user requirements, or controversial content, it generates clarification questions for the user. We implement a presentation approach that displays notifications with multiple-choice options to reduce input burden while allowing users to provide detailed feedback when needed.
- **Web Searching:** Based on its own research trajectory, user requirements, preferences and the status of the current version of the report, it generates a list of targeted search queries and invokes the search engine API² to retrieve the top k ³ for each query in parallel. Each search result consists of a title, snippet, and URL. These results are then processed by the Browsing Agent to extract relevant information.
- **Report Editing:** Once the agent gathers sufficient information from the browsing agent or receives editing requirements from the user, it generates a new version of the report by editing the current version while considering user preferences and requirements. To ensure the report content aligns with user preferences and requirements, the research agent proactively generates a list of quality rubrics for report verification beforehand. After generating the report, it systematically verifies each rubric and sends critiques back to itself for further refinement and polishing.
- **Research Completion:** When the agent determines that the report has covered sufficient aspects of the research question with adequate depth and has fulfilled the user's requirements, it concludes the research process. This decision is based on comprehensive assessment of content coverage, research depth, and alignment with stated objectives.

Through this iterative decision-making process, the research agent orchestrates the entire research workflow while maintaining focus on user needs and research quality. To support its information gathering needs, the system relies on the specialized capabilities of the browsing agent.

3.1.2 Browsing Agent

The browsing agent serves as a specialized information retrieval component that efficiently processes web search results and extracts relevant content to support the research agent's knowledge synthesis activities. This agent operates with high efficiency and accuracy to minimize latency while maximizing the quality of retrieved information. The browsing agent performs two primary functions during the research workflow:

- **URL Selection and Content Retrieval:** The agent employs intelligent filtering mechanisms to select URLs that demonstrate high potential for containing relevant information. This selection process analyzes multiple signals including the title, snippet, and URL structure of search results, while also incorporating user preferences regarding specific domains, topics, and source types. Once promising URLs are identified, the agent performs parallel web scraping operations to retrieve content efficiently. During this process, the system validates URL accessibility and maintains only those sources that can be successfully accessed, ensuring robust information gathering despite potential connectivity issues or restricted content.
- **Information Extraction and Quality Assessment:** For each successfully retrieved webpage, the agent leverages large language models to perform sophisticated information extraction. This process involves identifying and extracting content segments that are directly relevant to the research objectives, while simultaneously generating concise summaries that capture the essence of useful information. Critically, the agent implements quality assessment mechanisms to distinguish between valuable and non-useful web pages, filtering out irrelevant content, advertisements, or low-quality sources that could diminish research quality.

The extracted information, accompanied by comprehensive metadata including titles and URLs for proper citation formatting, is transmitted back to the research agent. This structured approach ensures that the research agent receives high-quality, relevant information that can be seamlessly integrated into the report generation process while maintaining academic integrity through proper source attribution.

²We use Google search in our implementation.

³We set $k = 5$ in this work.

3.1.3 Preference Agent

The preference agent adapts to user preferences through the In-Context Reinforcement Learning paradigm (Laskin et al., 2022; Lin et al., 2024; Lee et al., 2023; Huang et al., 2024; Grigsby et al., 2024), treating user actions and feedback as reward signals or critiques across three key dimensions:

- **Query Preference:** The agent tracks user query modifications, refinements, and explicit search requirements to understand preferred search methodologies and terminology choices, subsequently influencing the research agent’s query generation strategies.
- **Webpage Preference:** The agent monitors user selection patterns across different information sources and webpage types, identifying preferences for specific domains and publication types (e.g., academic papers vs. blogs). These learned patterns directly influence the browsing agent’s decisions on URL selection.
- **Report Preference:** The agent analyzes user feedback, editing histories, and formatting requirements to understand preferences for report organization, writing style, and presentation approaches, guiding the research agent’s report generation process.

This in-context adaptation mechanism enables the system to provide increasingly personalized research assistance by treating user interactions as reward signals, creating a dynamic research environment that evolves with user needs within each session.

3.2 Interaction Design for Human-AI Collaboration

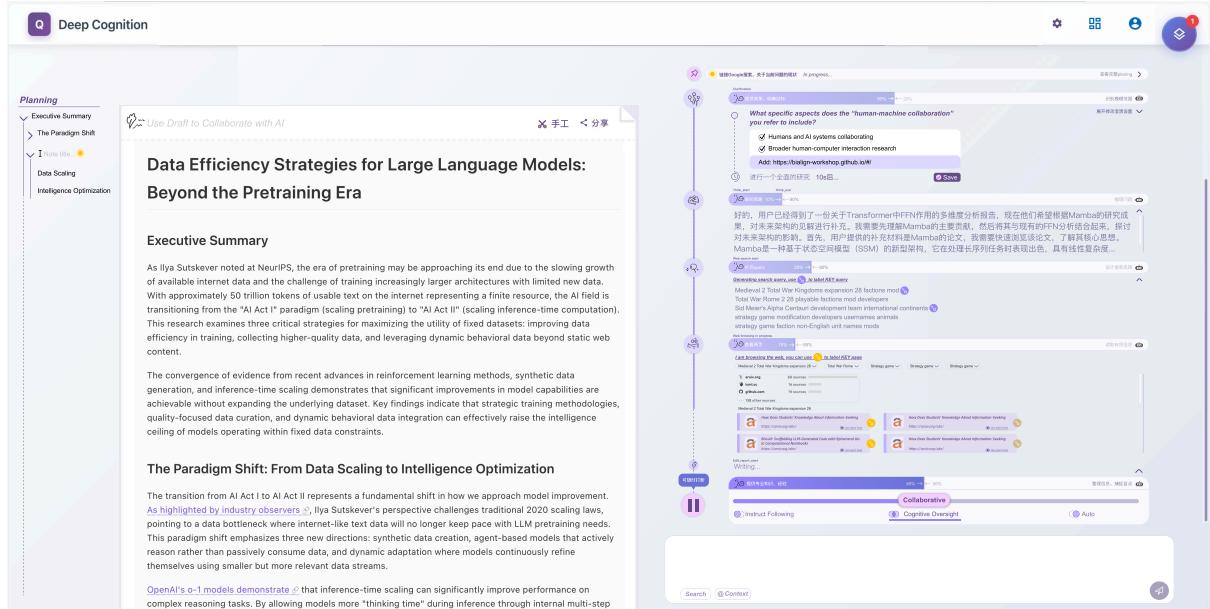


Figure 4: Deep cognition interface design showcasing key interactive features: (A) Research scope clarification to refine vague queries, (B) Preference profile generation based on user operation history and preferences, (C) Real-time human intervention capability, (D) Transparent display of reasoning, research processes, and interactive query refinement, and (E) Collaborative report revision area. The icon stands for clickable interface elements.

Deep research tasks are prevalent across domains such as investigative journalism, scientific research, and market analysis. These tasks are characterized by: (1) the necessity to synthesize information from multiple sources to address complex information-seeking problems or cover various facets of a topic; (2) ongoing user interaction rather than single-query processing; (3) the production of curated, report-like artifacts rather than brief answers; and (4) evolving user objectives that change iteratively as new information is accessed and integrated. Following principles of cognitive oversight, we designed the following features for our deep cognition system, with interfaces presented in Figure 4.

Feature I: Transparent Research Process The interface establishes transparency through multiple mechanisms that 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. This transparency enables users to understand how the system makes decisions and reaches

4 EXPERIMENTS

conclusions. The editor area on the left of Figure 4 displays the evolving research document with proper formatting. All findings are properly linked to their original sources, enabling users to trace source materials.

Feature II: Real-Time Intervention We implement a “Pause” feature that maintains user control throughout the research process, allowing users to interrupt the system at any point during execution. At critical junctures in the research process, users can provide feedback, introduce new requirements, or offer guidance to redirect the system’s approach. This intervention capability enables users to actively shape the research trajectory based on emerging insights or changing objectives, ensuring the system remains aligned with user intentions throughout the investigation.

Feature III: Fine-Grained Interaction The interface supports multiple levels of user interaction to accommodate varying research needs and preferences. The system includes an “Edit” button that enables direct modifications, allowing users to refine and restructure the document collaboratively. When users pose vague or unclear research questions, a clarification dialog appears with targeted questions to help narrow the scope—functioning like a research librarian asking “What exactly are you looking for?”. Users can refine their preferences and input follow-up knowledge or additional requirements as their understanding develops through the model’s guidance. Users can prioritize preferred search queries and specify particular webpages or knowledge sources for emphasis. The search results visualization presents diverse results with source indicators, thumbnails, and organizational categorization to aid navigation.

Feature IV: User Research Preference Adaptation The system adapts to user preferences and requirements across multiple dimensions to provide personalized research assistance. Users can customize their preferences for report writing style, format, and structure. The system generates reference profiles based on user history and choices, learning from past interactions to better align with individual research approaches. Additionally, the system adapts to user preferences for search queries and knowledge sources, learning from selections and prioritizations to improve future recommendations. As research objectives evolve, the system adapts to new requirements while maintaining context from previous interactions. This adaptive capability ensures research process coherence while accommodating the natural evolution of deep research tasks.

Together, these four features create a comprehensive framework that balances automation with human oversight, enabling users to conduct thorough, efficient research while maintaining control over the process. By combining transparency, real-time intervention, fine-grained interaction, and adaptive personalization, the system empowers users to tackle complex research challenges with confidence and precision.

4 Experiments

We conducted user and benchmark evaluations to validate the effectiveness of our deep cognition framework’s design principles. Our evaluation compared deep cognition against three leading deep research systems: Gemini Deep Research, OpenAI Deep Research, and Grok 3 Deeper Search. The assessment employed both user evaluation studies and quantitative analysis across two benchmark datasets. This section provides detailed descriptions of the evaluation methodologies, benchmarks, and metrics used in our analysis.

4.1 User Evaluation

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.

4.1.1 Protocol

We hired 13 graduate students as participants for user evaluation. They have 30 minutes of time budget to interact with deep cognition to solve a specific problem they interested in. They also ask the other three systems the same question to get the corresponding answers. This user evaluation includes pre-study, in-study, and post-study. The protocol consisted of three distinct phases:

Pre-Study The participants were introduced to the usage of deep cognition which familiarized participants with the system’s interaction method. The detailed and complete user study protocol is provided in Appendix B.

In-Study When interacted with deep cognition, the participants actively reviewed both the model’s reasoning processes and the evolving report draft. The session ended when deep cognition autonomously determined that the research was complete.

5 RESULTS AND ANALYSIS

Post-Study The participants completed a structured interview exploring their collaborative strategies. The interview examined three key dimensions, each includes a quantitative 5-point scales and several qualitative follow-up questions.

4.1.2 Evaluation Metrics Design

The evaluation metrics we develop are shown in Table 2 which are inspired by OpenScholar (Asai et al., 2024). In addition, we define five key aspects in table 2 to evaluate the generated report. Each metric is rigorously assessed using a 5-point Likert scale. We assess organization, coverage, depth, relevance, helpfulness and cutting-edge, detailed definitions for each score (1–5) on the 5-point scale are provided in Appendix B.

Metric	Description	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.	Intention to Use	Measures user intention and propensity for continued engagement with the system based on perceived value and satisfaction.
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.	Usability	Evaluates the intuitive nature and accessibility of the system interface, including cognitive load and interaction efficiency.
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.	Transparency	Assesses the interpretability and explainability of the model's decision-making processes and reasoning mechanisms.
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.	Interruptibility	Assesses the system's ability to tolerate pauses or context switches and to resume smoothly without loss of state or progress.
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.	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.

Table 2: Evaluation Metrics for Report Quality Assessment

4.2 Benchmark Evaluation

To validate our hypothesis that experts with higher cognitive capabilities demonstrate enhanced collaboration with AI in transparent dialogue environments, we measured our system performance through the accuracy of browsecomp-ZH (Zhou et al., 2025), a benchmark assessing agents' web browsing capabilities. Given that our expert annotators are native Chinese speakers with domain expertise, we selected 22 questions (top two from each of 11 categories) for comprehensive assessment.

5 Results and Analysis

This section we introduce our evaluation experiments' results, including metrics design, user study design, user evaluation analysis and design suggestions. Our comprehensive evaluation demonstrates that deep cognition significantly outperforms existing systems across multiple dimensions.

5 RESULTS AND ANALYSIS

5.1 Observations from User Evaluation

Takeaway 1: By integrating transparency and interruptibility mechanisms at fine-grained interaction point of the research process, our system enables user intervention that measurably improves response quality—particularly in terms of **organization, cutting-edge, and depth**.

As shown in Table 3, 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 4, 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.**

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

Metric	DC (non).	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 3: Performance improvement of deep cognition over deep cognition without interaction. DC. indicates deep cognition, DC (non). indicates deep cognition without interaction.

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 Iteration	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 4: 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 (n=13 participants, 13 responses). Right panel: Scores (1–5 scale) for system-interaction evaluation metrics (n = 13 participants). 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.

Design Suggestion 1: 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. Complex user hesitations or corrections trigger deeper reasoning processes, while smooth task completion indicates successful lightweight inference.

5.2 Observation from Benchmark Evaluation

Takeaway 2: 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..

5 RESULTS AND ANALYSIS

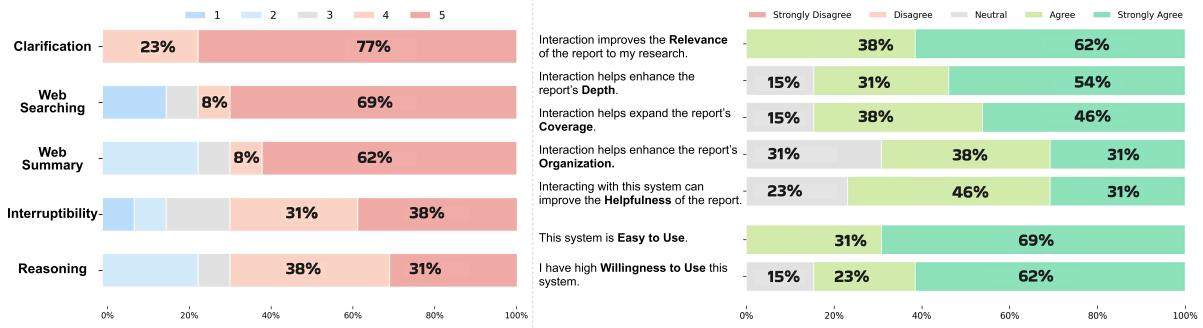


Figure 5: Left: Distribution of participant ratings (1–5) indicating the extent to which each system feature benefited their research process (n = 13 participants). Right: Perceived overall usefulness of deep cognition, as reported by the same participant cohort (n = 13 participants).

	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%

Table 5: Accuracy comparison across deep research systems. DC (non cog). represents the baseline condition with participants possessing foundational knowledge levels (n=4 participants with middle school-level knowledge); DC (non int). represents the autonomous system condition without human intervention; DC (cog+int). represents the interactive condition with graduate-level participants engaging in real-time collaboration with the system (n=4 participants).

The results in Table 5 provide compelling evidence for our collaborative cognition framework. The deep cognition system (cognition + interaction) achieves 72.73% accuracy — a dramatic improvement over all baselines and ablated versions. This performance significantly exceeds standalone systems: Gemini and OpenAI both plateau at 40.91%, while Grok 3 lags at 22.73%. Critically, the ablation study reveals that neither cognitive oversight alone (45.45%) nor interaction capabilities alone (40.91%) approach the performance of their combination, demonstrating that expert-AI collaboration requires both transparent reasoning processes and fine-grained interactive guidance to tackle challenging web browsing tasks effectively.

5.3 Dynamic Autonomy in Human-AI Deep Research System

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 Figure 6, 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:

Design Suggestion 2: 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.

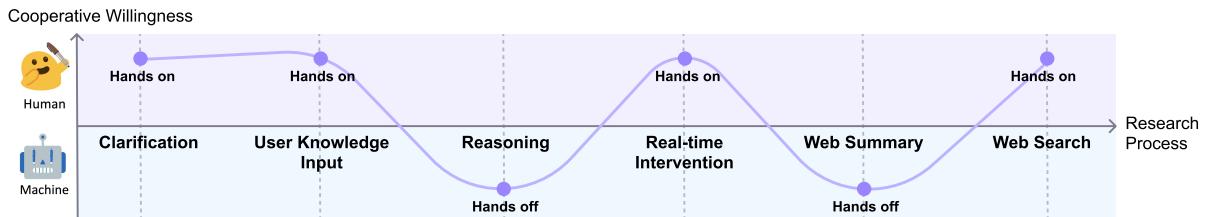


Figure 6: Changes in users’ behavioral tendencies when using the deep research system to perform complex retrieval tasks.

5 RESULTS AND ANALYSIS

Phase I: 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. For example, participant P1's first query was deliberately open-ended: "*What are the current mainstream approaches to repo-level code generation and completion?*" which they then refined with "*What about recent popular techniques such as Claude Code, Windsurf, and Cursor?*" This iterative refinement process proved valuable—8 participants mentioned that the model filled in aspects beyond their expectation. As P1 reflected: "*I aligned my thinking purpose with others. The question did contain aspects that I had not thought of before*" during the clarification stage.

Phase II: 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. This is exemplified by targeted directives such as "*This website is of critical importance and should be reviewed thoroughly*".

Phase III: Reasoning (Hands-off) Interestingly, users demonstrate lower cooperative willingness during reasoning phases, preferring to assess the AI's autonomous analytical capabilities. Users seek to understand whether the model has correctly executed prescribed instructions and want transparency in decision-making processes. P12 emphasized wanting to see "*how the model decides on key technical routes when faced with some open-ended problems*". This hands-off approach allows users to evaluate AI reasoning quality without excessive intervention.

Design Suggestion 3: Optimal human-AI collaboration 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.

Phase IV: Real-Time Intervention (Hands-on) Cooperation peaks again during dynamic browsing tasks where users encounter pages or information sources that warrant detailed retrieval. Users actively adjust the model's responses to align with their expectations.

Phase V: 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. As P12 noted: "*I will assume that AI can find useful information by itself as long as I look at it*". This reflects sophisticated understanding of when mechanical aggregation tasks can be delegated effectively.

Phase VI: Web Search (Hands-on) The cycle concludes with renewed hands-on engagement for open-ended and subjective questions that require interpretation or subjective judgment. These prove most challenging for existing tools, where conventional approaches require ongoing human guidance throughout the process. However, cognitive oversight allows extended periods of autonomous AI cognition punctuated by targeted human interventions at critical junctures.

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 CONCLUSION

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.

6 Conclusion

This paper introduced deep cognition, a human-AI collaboration system that implements “cognitive oversight” through transparent, interruptible interactions. Rather than treating interaction as merely an interface, we demonstrated that interaction constitutes a fundamental dimension of intelligence for complex research tasks.

Our evaluation shows that cognitive oversight significantly outperforms traditional approaches, achieving 63% average improvement over non-interactive systems and 72.73% accuracy on BrowseComp-ZH benchmark. The system excels particularly in transparency (+25.0%), fine-grained interaction (+44.6%), and cooperative capabilities (+43.0%). User behavior analysis revealed sophisticated dynamic autonomy patterns, where participants strategically alternate between “hands-on” and “hands-off” modes across different research phases.

These findings 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. This establishes the foundation for reconceptualizing human-AI relationships in complex cognitive tasks.

References

- [1] Anthropic. 2025. [Claude sonnet 4: Hybrid reasoning model with superior intelligence for high-volume use cases, and 200k context window.](#)
- [2] Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. [Openscholar: Synthesizing scientific literature with retrieval-augmented lms.](#)
- [3] Lisanne Bainbridge. 1983a. Ironies of automation. In *Analysis, design and evaluation of man-machine systems*, pages 129–135. Elsevier.
- [4] Lisanne Bainbridge. 1983b. [Ironies of automation. Automatica](#), 19(6):775–779.
- [5] Gagan Bansal, Jennifer Wortman Vaughan, Saleema Amershi, Eric Horvitz, Adam Fourney, Hussein Mozannar, Victor Dibia, and Daniel S. Weld. 2024. [Challenges in human-agent communication.](#)
- [6] Alexandra Bremers and Wendy Ju. 2024. Can machines tell what people want? bringing situated intelligence to generative ai. In *Proceedings of the Halfway to the Future Symposium*, pages 1–6.
- [7] Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V. Le, Christopher Ré, and Azalia Mirhoseini. 2024. [Large language monkeys: Scaling inference compute with repeated sampling.](#)
- [8] 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. 2020. [Language models are few-shot learners.](#)
- [9] ByteDance. 2024. [Deerflow](#). Community-driven deep research framework combining LLMs with web search, crawling, and code execution tools.
- [10] Runze Cai, Nuwan Janaka, Hyeongcheol Kim, Yang Chen, Shengdong Zhao, Yun Huang, and David Hsu. 2025. [Aiget: Transforming everyday moments into hidden knowledge discovery with ai assistance on smart glasses.](#) In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.
- [11] Nicolas Camara. 2025. [Open deep research](#). Open-source clone of OpenAI's Deep Research using Firecrawl for web data extraction and AI reasoning.
- [12] Heloisa Candello and Claudio Pinhanez. 2016. Designing conversational interfaces. *Simpósio Brasileiro sobre Fatores Humanos em Sistemas Computacionais-IHC*, 100.
- [13] Yining Cao, Peiling Jiang, and Haijun Xia. 2025. [Generative and malleable user interfaces with generative and evolving task-driven data model.](#) In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.
- [14] 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. 2025. [Why do multi-agent llm systems fail?](#)
- [15] Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Fan Yang, Zenan Zhou, Weipeng Chen, Haofen Wang, Jeff Z Pan, et al. 2025a. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.
- [16] Si Chen, Haocong Cheng, and Yun Huang. 2024. [Emotion Recognition in Self-Regulated Learning: Advancing Metacognition Through AI-Assisted Reflections](#), pages 185–212. Springer Nature Switzerland, Cham.
- [17] Valerie Chen, Alan Zhu, Sebastian Zhao, Hussein Mozannar, David Sontag, and Ameet Talwalkar. 2025b. Need help? designing proactive ai assistants for programming. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–18.
- [18] 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. 2024. [Building machines that learn and think with people.](#)

References

- [19] 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, Zhusu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. 2025. Deepseek-v3 technical report.
- [20] Jane L E, Yu-Chun Grace Yen, Isabelle Yan Pan, Grace Lin, Mingyi Li, Hyoungwook Jin, Mengyi Chen, Haijun Xia, and Steven P Dow. When to give feedback: Exploring tradeoffs in the timing of design feedback.
- [21] Assaf Elovic. 2025. Gpt researcher. Open deep research agent for web and local research with detailed report generation and citations.
- [22] K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. 2025. Levels of autonomy for ai agents.
- [23] Raymond Fok and Daniel S. Weld. 2024. In search of verifiability: Explanations rarely enable complementary performance in ai-advised decision making. *AI Magazine*, 45(3):317–332.
- [24] George Fragiadakis, Christos Diou, George Kousouris, and Mara Nikolaidou. 2025. Evaluating human-ai collaboration: A review and methodological framework.
- [25] Melinda Gervasio, Pedro Sequeira, Eric Yeh, Nicholas Marion, Sarah Bakst, and Helen Gent. 2025. Ai as collaborative partner: Rethinking human-ai teaming for the real world. In *Proceedings of the AAAI Symposium Series*, volume 5, pages 63–66.
- [26] Catalina Gomez, Sue Min Cho, Shichang Ke, Chien-Ming Huang, and Mathias Unberath. 2025. 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.
- [27] Google. 2025. Gemini deep research - your personal research assistant. Accessed: April 14, 2025.
- [28] Jake Grigsby, Linxi Fan, and Yuke Zhu. 2024. Amago: Scalable in-context reinforcement learning for adaptive agents.
- [29] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- [30] 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 Milligan, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. 2022. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS ’22, Red Hook, NY, USA. Curran Associates Inc.
- [31] Sili Huang, Jifeng Hu, Hechang Chen, Lichao Sun, and Bo Yang. 2024. In-context decision transformer: Reinforcement learning via hierarchical chain-of-thought.
- [32] Edwin Hutchins. 1995. *Cognition in the Wild*. MIT press.

- [33] Folasade Olubusola Isinkaye, Yetunde O Folajimi, and Bolande Adefowoke Ojokoh. 2015. Recommendation systems: Principles, methods and evaluation. *Egyptian informatics journal*, 16(3):261–273.
- [34] Yucheng Jiang, Yijia Shao, Dekun Ma, Sina J. Semnani, and Monica S. Lam. 2024. *Into the unknown unknowns: Engaged human learning through participation in language model agent conversations*.
- [35] Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025a. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.
- [36] Mingyu Jin, Weidi Luo, Sitao Cheng, Xinyi Wang, Wenyue Hua, Ruixiang Tang, William Yang Wang, and Yongfeng Zhang. 2025b. *Disentangling memory and reasoning ability in large language models*.
- [37] Jina AI. 2025. *node-deepresearch*. Iterative search, reading, and reasoning system for deep research queries with focus on concise answers.
- [38] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. *Scaling laws for neural language models*.
- [39] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. *“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. Association for Computing Machinery.
- [40] 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. 2025. *Measuring ai ability to complete long tasks*.
- [41] LangChain AI. 2025. *Open deep research*. Open-source research assistant for automated deep research and report generation.
- [42] Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, Maxime Gazeau, Himanshu Sahni, Satinder Singh, and Volodymyr Mnih. 2022. *In-context reinforcement learning with algorithm distillation*.
- [43] Jonathan N. Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. 2023. *Supervised pretraining can learn in-context reinforcement learning*.
- [44] Mina Lee, Megha Srivastava, Amelia Hardy, John Thickstun, Esin Durmus, Ashwin Paranjape, Ines Gerard-Ursin, Xiang Lisa Li, Faisal Ladakh, Frieda Rong, Rose E. Wang, Minae Kwon, Joon Sung Park, Hancheng Cao, Tony Lee, Rishi Bommasani, Michael Bernstein, and Percy Liang. 2024. *Evaluating human-language model interaction*.
- [45] Jingshu Li, Yitian Yang, Q. Vera Liao, Junti Zhang, and Yi-Chieh Lee. 2025. *As confidence aligns: Exploring the effect of ai confidence on human self-confidence in human-ai decision making*.
- [46] Licong Lin, Yu Bai, and Song Mei. 2024. *Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining*.
- [47] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. *Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing*. *ACM computing surveys*, 55(9):1–35.
- [48] Xingyu Bruce Liu, Shitao Fang, Weiyuan Shi, Chien-Sheng Wu, Takeo Igarashi, and Xiang’Anthony’ Chen. 2025a. *Proactive conversational agents with inner thoughts*. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–19.
- [49] Xingyu Bruce Liu, Haijun Xia, and Xiang Anthony Chen. 2025b. *Interacting with thoughtful ai*.
- [50] Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024a. *How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent*. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA. Association for Computing Machinery.
- [51] Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2024b. *Personafow: Boosting research ideation with llm-simulated expert personas*.
- [52] Liangchen Luo, Yinxiao Liu, Rosanne Liu, Samrat Phatale, Harsh Lara, Yunxuan Li, Lei Shu, Yun Zhu, Lei Meng, Jiao Sun, et al. 2024. *Improve mathematical reasoning in language models by automated process supervision*. *ArXiv preprint*, abs/2406.06592.

References

- [53] Michael Frederick McTear, Zoraida Callejas, and David Griol. 2016. *The conversational interface*, volume 6. Springer.
- [54] Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Accessed: July 13, 2025.
- [55] Bryan Min, Allen Chen, Yining Cao, and Haijun Xia. 2025. Malleable overview-detail interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.
- [56] Bryan Min and Haijun Xia. 2025. Feedforward in generative ai: Opportunities for a design space.
- [57] 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, Qiuwei 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. 2025. Minimax-01: Scaling foundation models with lightning attention.
- [58] Marvin Minsky. 1987. The society of mind. *The Personalist Forum*, 3(1):19–32.
- [59] 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. 2024. Openai o1 system card.
- [60] OpenAI. 2022. Introducing chatgpt.
- [61] OpenAI. 2024. Learning to reason with llms, september 2024.

References

- [62] OpenAI. 2025a. [Deep research system card](#). Accessed: April 14, 2025.
- [63] OpenAI. 2025b. [Introducing deep research](#). Accessed: April 14, 2025.
- [64] Michael J Pazzani and Daniel Billsus. 2007. Content-based recommendation systems. In *The adaptive web: methods and strategies of web personalization*, pages 325–341. Springer.
- [65] Perplexity AI. 2025. [Introducing perplexity deep research](#). Accessed: April 14, 2025.
- [66] Peter Pirolli. 2009. Powers of 10: Modeling complex information-seeking systems at multiple scales. *Computer*, 42(3):33–40.
- [67] Michael Poli, Armin W Thomas, Eric Nguyen, Pragaash Ponnusamy, Björn Deisereth, Kristian Kersting, Taiji Suzuki, Brian Hie, Stefano Ermon, Christopher Ré, Ce Zhang, and Stefano Massaroli. 2024. [Mechanistic design and scaling of hybrid architectures](#).
- [68] Kevin Pu, Daniel Lazaro, Ian Arawjo, Haijun Xia, Ziang Xiao, Tovi Grossman, and Yan Chen. 2025. [Assistance or disruption? exploring and evaluating the design and trade-offs of proactive ai programming support](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.
- [69] Peinuan Qin, Chi-Lan Yang, Jingshu Li, Jing Wen, and Yi-Chieh Lee. 2025. [Timing matters: How using llms at different timings influences writers' perceptions and ideation outcomes in ai-assisted ideation](#). In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI ’25, New York, NY, USA. Association for Computing Machinery.
- [70] 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. 2021. [Learning transferable visual models from natural language supervision](#).
- [71] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.
- [72] Jude Rayan, Dhruv Kanetkar, Yifan Gong, Yuewen Yang, Srishti Palani, Haijun Xia, and Steven P. Dow. 2024. [Exploring the potential for generative ai-based conversational cues for real-time collaborative ideation](#). In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C ’24, pages 117–131, New York, NY, USA. Association for Computing Machinery.
- [73] Paul Resnick and Hal R Varian. 1997. Recommender systems. *Communications of the ACM*, 40(3):56–58.
- [74] Karl Toby Rosenberg, Rubaiyat Habib Kazi, Li-Yi Wei, Haijun Xia, and Ken Perlin. 2024. [Drawtalking: Building interactive worlds by sketching and speaking](#). In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST ’24, New York, NY, USA. Association for Computing Machinery.
- [75] Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. ‘smolagents’: a smol library to build great agentic systems. <https://github.com/huggingface/smolagents>.
- [76] Yijia Shao, Yucheng Jiang, Theodore A. Kanell, Peter Xu, Omar Khattab, and Monica S. Lam. 2024. [Assisting in writing wikipedia-like articles from scratch with large language models](#).
- [77] Yijia Shao, Humishka Zope, Yucheng Jiang, Jiaxin Pei, David Nguyen, Erik Brynjolfsson, and Diyi Yang. 2025. [Future of work with ai agents: Auditing automation and augmentation potential across the u.s. workforce](#).
- [78] Wenxuan Shi, Haochen Tan, Chuqiao Kuang, Xiaoguang Li, Xiaozhe Ren, Chen Zhang, Hanting Chen, Yasheng Wang, Lifeng Shang, Fisher Yu, and Yunhe Wang. 2025. [Pangu deepdive: Adaptive search intensity scaling via open-web reinforcement learning](#).
- [79] Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.
- [80] Sakhinana Sagar Srinivas and Venkataramana Runkana. 2025. [Scaling test-time inference with policy-optimized, dynamic retrieval-augmented generation via kv caching and decoding](#).
- [81] Sangho Suh, Meng Chen, Bryan Min, Toby Jia-Jun Li, and Haijun Xia. 2024. [Luminate: Structured generation and exploration of design space with large language models for human-ai co-creation](#). In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI ’24, New York, NY, USA. Association for Computing Machinery.

- [82] 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. 2025. [Kimi k1.5: Scaling reinforcement learning with llms](#).
- [83] Xinru Wang, Mengjie Yu, Hannah Nguyen, Michael Iuzzolino, Tianyi Wang, Peiqi Tang, Natasha Lynova, Co Tran, Ting Zhang, Naveen Sendhilnathan, Hrvoje Benko, Haijun Xia, and Tanya R. Jonker. 2025. [Less or more: Towards glanceable explanations for llm recommendations using ultra-small devices](#). In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, IUI '25, page 938–951, New York, NY, USA. Association for Computing Machinery.
- [84] Elizabeth Anne Watkins, Emanuel Moss, Giuseppe Raffa, and Lama Nachman. 2025. [What's so human about human-ai collaboration, anyway? generative ai and human-computer interaction](#).
- [85] 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. 2022. [Emergent abilities of large language models](#). *Trans. Mach. Learn. Res.*, 2022.
- [86] Monika Westphal, Michael Vössing, Gerhard Satzger, Galit B. Yom-Tov, and Anat Rafaeli. 2023. [Decision control and explanations in human-ai collaboration: Improving user perceptions and compliance](#). *Computers in Human Behavior*, 144:107714.
- [87] Ryen W. White. 2024. [Advancing the search frontier with ai agents](#).
- [88] Anita Williams Woolley, Christopher F. Chabris, Alex Pentland, Nada Hashmi, and Thomas W. Malone. 2010. [Evidence for a collective intelligence factor in the performance of human groups](#). *Science*, 330(6004):686–688.
- [89] xAI. 2025. [Grok 3 beta — the age of reasoning agents](#). Accessed: April 14, 2025.
- [90] Shijie Xia, Yiwei Qin, Xuefeng Li, Yan Ma, Run-Ze Fan, Steffi Chern, Haoyang Zou, Fan Zhou, Xiangkun Hu, Jiahe Jin, et al. 2025. Generative ai act ii: Test time scaling drives cognition engineering. *arXiv preprint arXiv:2504.13828*.
- [91] Zhengtao Xu, Tianqi Song, and Yi-Chieh Lee. 2025. [Confronting verbalized uncertainty: Understanding how llm's verbalized uncertainty influences users in ai-assisted decision-making](#). *International Journal of Human-Computer Studies*, 197:103455.
- [92] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan 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. 2025a. [Qwen3 technical report](#).
- [93] Jingyi Yang, Shuai Shao, Dongrui Liu, and Jing Shao. 2025b. [Riosworld: Benchmarking the risk of multimodal computer-use agents](#).
- [94] Ryan Yen, Jiawen Stefanie Zhu, Sangho Suh, Haijun Xia, and Jian Zhao. 2024. [Coladder: Manipulating code generation via multi-level blocks](#). In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24, New York, NY, USA. Association for Computing Machinery.
- [95] Ming Yin. 2025. Bridging the gap between machine confidence and human perceptions. *Nature Machine Intelligence*, pages 1–2.
- [96] David Zhang. 2025. [Deep research](#). AI-powered research assistant for iterative, deep research using search engines, web scraping, and LLMs.

References

- [97] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025a. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. *arXiv preprint arXiv:2504.03160*.
- [98] Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025b. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments.
- [99] 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. 2025. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese.

B USER STUDY PROTOCOL

A User Behavior Data Point



Figure 7: Human–AI collaboration code book

B User Study Protocol

B.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.

Participant Instructions

Research Task Selection Participants should pose research questions that are:

- Genuinely of interest to them personally
- Open-ended with substantial exploration space
- Requiring iterative refinement during the research process
- Suitable for comprehensive report generation

Example Research Questions:

- “I want to systematically understand different perspectives on ‘AI agent roles and human relationships.’ For instance, Anthropic CEO Dario Amodei views future AI agents as colleagues, while Google’s Co-scientist paper positions AI scientists as human collaborators. Please collect additional viewpoints and analyze them in the context of current and future development trends.”

B USER STUDY PROTOCOL

- “Why can models trained on synthetic data outperform the models that generated that synthetic data? Please find the latest research papers that provide supporting evidence.”

System Familiarization Participants learn to use a tool designed for **real-time human-AI collaboration in multi-hop information retrieval**. Unlike traditional deep research systems with single complex inputs and asynchronous black-box interactions, this system allows users to:

- View the model’s retrieval strategies and decision processes in real-time
- Intervene and provide corrections during the research process
- Guide the model through various interaction methods (interruption, expert knowledge injection, source evaluation, self-assessment review, strategic guidance, or personal file integration)

Termination Criteria The interaction ends when the model autonomously decides to finish, after achieving satisfactory quality scores (4-5 points) across all evaluation dimensions.

B.2 In-Study

B.2.1 Evaluation Framework

Evaluation Dimension	Terrible	Poor	Average	Good	Excellent
Organization: Structural clarity and logical flow	<input type="radio"/>				
Cutting-edge Information: Coverage of recent, high-impact research	<input type="radio"/>				
Information Coverage (Breadth): Comprehensiveness across research domains	<input type="radio"/>				
Information Depth: Sufficiency of detail for thorough understanding	<input type="radio"/>				
Overall Helpfulness: Practical utility for literature review and research	<input type="radio"/>				

Table 6: Primary Quality Assessment Rubric

Organization
5 points - Excellent:
<ul style="list-style-type: none"> • Structure Clarity: Perfect logical structure with clear hierarchical organization and seamless section transitions; • 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;
4 points - Good:
<ul style="list-style-type: none"> • Structure Clarity: Well-organized with clear structure, consistently followed throughout; • Logical Flow: Good reasoning progression with mostly smooth transitions; • Coherence: Strong thematic coherence with minor inconsistencies; • Presentation Quality: Good formatting that supports understanding;
3 points - Average:
<ul style="list-style-type: none"> • Structure Clarity: Generally well-organized with clear structure, occasional flow issues; • Logical Flow: Adequate progression with some choppy transitions;

- **Coherence:** Reasonable thematic development with some disconnected elements;
- **Presentation Quality:** Acceptable formatting with room for improvement;

2 points - Poor:

- **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;

1 point - Terrible:

- **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;

B.2.2 Cutting-Edge Information

Cutting-Edge Information

5 points - Excellent:

- **Recency:** Captures cutting-edge research including preprints, conference reports, and ongoing work;
- **Impact Level:** Includes highest-impact research and breakthrough discoveries;
- **Coverage Completeness:** Comprehensive coverage of all major recent developments;
- **Source Quality:** Exclusively high-quality, authoritative sources from leading institutions;

4 points - Good:

- **Recency:** Identifies most important recent research and breakthrough work;
- **Impact Level:** Covers major high-impact developments with good selection;
- **Coverage Completeness:** Good coverage of recent developments with minor gaps;
- **Source Quality:** Mostly high-quality sources with reliable attribution;

3 points - Average:

- **Recency:** Covers some recent research, may miss key breakthroughs;
- **Impact Level:** Includes moderately impactful research with some selection issues;
- **Coverage Completeness:** Adequate coverage but misses some important developments;
- **Source Quality:** Mixed source quality with some reliability concerns;

2 points - Poor:

- **Recency:** Limited recent research, misses important developments;
- **Impact Level:** Focuses on lower-impact or less significant research;
- **Coverage Completeness:** Poor coverage with significant gaps in recent developments;
- **Source Quality:** Low-quality sources with questionable reliability;

1 point - Terrible:

- **Recency:** 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:** Very poor or unreliable sources;

Information Coverage (Breadth)

5 points - Excellent:

- **Domain Scope:** Comprehensive coverage including additional important discussion points beyond the

original question;

- **Perspective Diversity:** Multiple viewpoints and approaches from different research communities;
- **Methodological Range:** Covers various research methodologies and theoretical frameworks;
- **Interdisciplinary Connections:** Excellent integration of insights from related fields;

4 points - Good:

- **Domain Scope:** Broad coverage discussing various representative papers and sources;
- **Perspective Diversity:** Good variety of viewpoints with most major perspectives covered;
- **Methodological Range:** Covers most relevant methodological approaches;
- **Interdisciplinary Connections:** Good integration with some cross-field insights;

3 points - Average:

- **Domain Scope:** Discusses representative works with satisfactory overview;
- **Perspective Diversity:** Adequate variety of viewpoints but may miss some important perspectives;
- **Methodological Range:** Covers basic methodological approaches with some gaps;
- **Interdisciplinary Connections:** Limited cross-field integration;

2 points - Poor:

- **Domain Scope:** Partial coverage, misses important research directions;
- **Perspective Diversity:** Limited viewpoints, potential bias in selection;
- **Methodological Range:** Narrow methodological coverage;
- **Interdisciplinary Connections:** Poor cross-field integration;

1 point - Terrible:

- **Domain Scope:** Severely limited coverage, focuses on single domain;
- **Perspective Diversity:** Very narrow perspective, lacks diversity;
- **Methodological Range:** Single or very limited methodological approach;
- **Interdisciplinary Connections:** No cross-field integration;

Information Depth

5 points - Excellent:

- **Detail Sufficiency:** Provides necessary and sufficient information with selective deep exploration;
- **Technical Accuracy:** Highly accurate technical details with proper context;
- **Analytical Depth:** Deep analytical insights with sophisticated reasoning;
- **Contextual Understanding:** Excellent understanding of broader implications and context;

4 points - Good:

- **Detail Sufficiency:** Includes most relevant information needed to understand the topic;
- **Technical Accuracy:** Good technical accuracy with minor gaps;
- **Analytical Depth:** Good analytical insights with solid reasoning;
- **Contextual Understanding:** Good understanding of context and implications;

3 points - Average:

- **Detail Sufficiency:** Acceptable amount of relevant information, may lack some useful details;
- **Technical Accuracy:** Adequate technical accuracy with some inaccuracies;
- **Analytical Depth:** Basic analytical insights with limited depth;
- **Contextual Understanding:** Basic understanding of context;

2 points - Poor:

- **Detail Sufficiency:** Provides some relevant information but misses important details;
- **Technical Accuracy:** Poor technical accuracy with significant errors;

- **Analytical Depth:** Superficial analysis with weak reasoning;
 - **Contextual Understanding:** Poor understanding of broader context;
- 1 point - Terrible:**
- **Detail Sufficiency:** Lacks basic details needed for topic understanding;
 - **Technical Accuracy:** Very poor technical accuracy with major errors;
 - **Analytical Depth:** No meaningful analytical insights;
 - **Contextual Understanding:** No understanding of context or implications;

Overall Helpfulness

5 points - Excellent:

- **Question Answering:** Comprehensive field overview that fully answers questions without requiring additional research;
- **Practical Utility:** Immediately actionable information for literature review and research;
- **Research Efficiency:** Significantly reduces time and effort compared to manual research;
- **Academic Value:** High academic value suitable for scholarly work and publication;

4 points - Good:

- **Question Answering:** Cross-source compilation of high-quality, relevant, and fresh sources;
- **Practical Utility:** Good practical value for research with minimal additional work needed;
- **Research Efficiency:** Substantial improvement over manual research approaches;
- **Academic Value:** Good academic value suitable for research projects;

3 points - Average:

- **Question Answering:** Provides useful discussions and papers, requires additional verification;
- **Practical Utility:** Moderate practical value but needs supplementary research;
- **Research Efficiency:** Some improvement over manual research but limited gains;
- **Academic Value:** Acceptable academic value for preliminary research;

2 points - Poor:

- **Question Answering:** Better than searching from scratch but limited usefulness;
- **Practical Utility:** Low practical value requiring substantial additional work;
- **Research Efficiency:** Minimal improvement over manual research;
- **Academic Value:** Limited academic value;

1 point - Terrible:

- **Question Answering:** Does not answer the question or provides confusing information;
- **Practical Utility:** No practical value for research purposes;
- **Research Efficiency:** No improvement or worse than manual research;
- **Academic Value:** No academic value;

B USER STUDY PROTOCOL

Evaluation Dimension	-2	-1	0	+1	+2
Transparency: Decision-making process visibility	<input type="radio"/>				
Interruptibility: Real-time intervention capability	<input type="radio"/>				
Fine-grained Interaction: Interaction granularity level	<input type="radio"/>				
Inspiration: Unexpected discoveries and insights	<input type="radio"/>				
Collaboration: Collaborative partnership quality	<input type="radio"/>				

Table 7: System Design Assessment Rubric

B.2.3 System Design Evaluation (-2 to +2 Scale)

Transparency
Question: Does the system design provide sufficient transparency in decision-making processes?
+2 points - Excellent:
<ul style="list-style-type: none"> 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:
<ul style="list-style-type: none"> 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:
<ul style="list-style-type: none"> 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:
<ul style="list-style-type: none"> 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:
<ul style="list-style-type: none"> 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;

B.2.4 Deep Cognition Specific Evaluation

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

B.3 Post-Study

Interaction Recordings

- Complete session recordings for all four systems
- Screenshots of key interaction moments
- Timestamped intervention points and rationales

Evaluation Documentation

- Quantitative scores for all metrics
- Qualitative explanations for each rating
- Specific examples of system strengths and limitations
- Comparative analysis across systems

Follow-up Questions

1. What additional interaction levels would enhance the research process?
2. Which research stages would benefit most from human intervention?
3. How does the collaborative interaction model compare to traditional one-shot research systems?
4. What insights or unexpected directions did the system's research process reveal?

Quality Assurance

- Minimum 15-minute engagement per system
- Maximum 30-minute cap for deep cognition
- All ratings must include written justifications
- Post-study validation interview for unclear responses
- Cross-verification of quantitative scores with qualitative feedback