DEEP COGNITION: A TRANSPARENT, FINE-GRAINED MULTI-AGENT DEEP RESEARCH SYSTEM FOR HUMAN-AGENT COLLABORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep research agents construct answers through an evolving trajectory of plans, searches, evidence, and partial syntheses. External feedback can be attached to a specific intermediate artifact while that trajectory is still unfolding, appended as an unstructured global instruction, or applied only after a report is complete. Yet the effect of *when* feedback arrives and *where* it enters the research state remains poorly understood. We study this question through **Deep Cognition**, an instrumented multi-agent research environment that exposes plans, subagents, queries, evidence, and draft sections as explicit intervention targets. In a matched experiment on all 65 ResearcherBench tasks, structured mid-process feedback raises full-answer coverage from 0.7040 to 0.7681. It also uses 45% less time and 20% fewer tokens than unstructured mid-process feedback, and 53% less time and 45% fewer tokens than post-output refinement. A learned 2B feedback-conditioned evidence router further raises weighted rubric score from 0.6569 to 0.6790 while reducing writer input tokens by 28%; the improvement persists on 50 questions untouched by checkpoint selection. Across three benchmarks and every tested interaction budget, mid-process feedback consistently outperforms post-output feedback. Fixed-backbone controls further show that cross-family experts achieve the strongest ResearcherBench and DeepResearch results, demonstrating that the effect is not tied to same-family alignment. A controlled study with 50 domain professionals and an analysis of 2,461 deployed sessions provide complementary behavioral evidence: users concentrate intervention at consequential research stages and develop persistent, artifact-level quality constraints. Together, these results show that feedback effectiveness depends on its position in the evolving research process, and establish structured mid-process steering as an effective control mechanism for deep research agents.

1 INTRODUCTION

Large language models (LLMs) (Luo et al., 2024; Radford et al., 2018; Brown et al., 2020) have advanced rapidly through parameter scaling (Kaplan et al., 2020; Hoffmann et al., 2022; Wei et al., 2022), larger training corpora (Yang et al., 2025; Meta AI, 2025), and architectural innovation (DeepSeek-AI et al., 2025; MiniMax et al., 2025). Yet this progress has produced increasingly autonomous systems that assume minimal human input beyond a single prompt (Liu et al., 2023; Kim et al., 2023) or coarse-grained instruction (Yin, 2025).

This autonomy-first paradigm conflicts with how humans actually reason. Sophisticated thinking is inherently interactive: it emerges through dialogue, feedback, and iterative refinement (Hutchins, 1995; Minsky, 1987; Woolley et al., 2010). As inference-time scaling (OpenAI, 2024) enables AI to engage in extended reasoning, the opportunity for strategic human oversight grows (Xia et al., 2025)—especially for tasks spanning hours to days (Kwa et al., 2025), where early errors compound and domain expertise is most valuable. This motivates our central research question: ***How can we design a transparent framework that enables humans to effectively guide an agent's reasoning and planning through fine-grained interventions?***

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

The need is most acute in *Deep Research* (OpenAI, 2025c; Google, 2025; Perplexity AI, 2025; Zheng et al., 2025)—multi-step processes involving dynamic retrieval, filtering, analysis, and synthesis. Current deep research systems have pioneered multi-step web browsing, data analysis, and report generation, yet they uniformly adopt an "Input-Wait-Output" paradigm: users initiate a query and wait through 5–30 minutes of opaque processing with no opportunity to correct errors, adapt to evolving requirements, or inject domain expertise (Cemri et al., 2025; Bainbridge, 1983). To overcome these limitations, we develop **Deep Cognition**, a multi-agent framework providing two core capabilities:

- ***Fine-Grained Interaction:*** Users can engage with any specific element of the subagent's output: questioning particular claims, requesting elaboration on specific points, redirecting search, injecting knowledge, or changing the research focus mid-execution.
- ***Transparency Workflow:*** The system reveals its entire workflow—from search strategies and query formulation to information evaluation and synthesis rationales—enabling humans to understand and then guide how the model thinks.

Through a controlled within-subjects study with diverse domain professionals, we demonstrate that Deep Cognition leads on 6/7 interaction metrics versus the strongest baseline (Gemini; Friedman $p < .005$ on all seven), while achieving competitive report quality. Our contributions are:

- **Interactive Multi-Agent Deep Research System**: A Plan-Search-Report architecture with transparent, interruptible execution that exposes per-subagent reasoning and supports mid-task human intervention via five dialogue act types (clarification, interruption, knowledge injection, plan review, evidence-level feedback), together with a learned feedback-conditioned evidence router for budgeted context allocation.
- **Empirical Evaluation Framework**: A 15-dimension evaluation suite (5 report quality + 10 interaction) for human-agent collaboration, validated through a controlled within-subjects study (n=50, Latin square counterbalancing) and automated evaluation across five datasets. Controlled follow-up experiments hold feedback content, checkpoint, budgets, and evaluator fixed to isolate feedback timing and structure, and separately vary expert family under a fixed research checkpoint.
- **Large-Scale Deployment Evidence**: Analysis of 2,461 sessions from 194 users (34,715 conversation traces) showing that lightweight natural-language rubrics are associated with persistent quality improvements (4.7×, Mann-Whitney p=0.00019) without re-statement, and that users naturally develop phase-dependent collaboration strategies that mirror controlled-study findings at scale.

Rather than relegating humans to passive tool operators, this framework establishes a cognitive partnership that leverages the complementary strengths of human expertise and AI capabilities while compensating for their respective limitations.

2 Related Work

Deep Research and Multi-Agent Systems. Commercial and open deep-research systems automate multi-step retrieval and synthesis (Google, 2025; OpenAI, 2025b; xAI, 2025; LangChain, 2025; Elovic, 2025; ByteDance, 2025). Research agents span parallel retrieval (Shao et al., 2024; Chen et al., 2025), scientific synthesis and discovery (Skarlinski et al., 2024; Asai et al., 2024; Lu et al., 2024), and iterative reasoning (Zheng et al., 2025), building on multi-agent coordination and tool-use frameworks (Li et al., 2023; Wang et al., 2024a; Hong et al., 2024; Wu et al., 2024; Yao et al., 2023). Yet most systems remain difficult to steer after execution begins; MyScholarQA likewise shows that research personalization requires real user participation (Balepur et al., 2026).

Human–AI Interaction and Interactive Agents. Mixed-initiative systems dynamically allocate control (Horvitz, 1999), and complementary teams can outperform either humans or AI alone (Bansal et al., 2021). Existing work studies dialogue, autonomy, reliance, and step-level steering (Lee et al., 2024; Feng et al., 2025; Amershi et al., 2019; Buçinca et al., 2021; Kazemitabaar et al., 2024; He et al., 2025). Cocoa supports collaborative refinement (Feng et al., 2024); Magentic-UI supports co-planning and browser-task intervention (Mozannar et al., 2025); Operator provides checkpoint control (OpenAI,

2025d). Deep Cognition instead exposes research-specific intermediate artifacts—subagents, queries, evidence, and evolving report sections—throughout retrieval and synthesis (Table 1).

Table 1: Representative systems and their collaborative affordances.

	Autonomous			Collaborative Affordances				
System	Multi-Ag.	Web Retr.	Iter. Out.	Clarif.	Interr.	Sub-Ag.	Evid. Fb.	Plan Rev.
Cocoa (Feng et al., 2024)	●	●	●	●	●	●	✓	✓
Magentic-UI (Mozannar et al., 2025)	●	✓	●	✓	✓	●	●	✓
OpenAI Operator (OpenAI, 2025d)	●	✓	●	✓	✓	●	●	●
STORM (Shao et al., 2024)	✓	✓	✓	●	●	●	●	●
MindSearch (Chen et al., 2025)	✓	✓	●	●	●	●	●	●
Gemini DR (Google, 2025)	✓	✓	●	✓	✓	●	●	✓
OpenAI DR (OpenAI, 2025b)	✓	✓	●	✓	●	●	●	●
Grok 3 (xAI, 2025)	●	✓	●	●	●	●	●	●
Deep Cognition (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

3 METHODOLOGY

3.1 SYSTEM ARCHITECTURE OVERVIEW

We propose a multi-agent collaborative deep research system comprising four components: Planning Agent, Clarification Module, Browsing Agent, and Writing Agent. Upon receiving user input, the **Proactive Clarification module** initiates structured dialogue to capture research intent. The system then enters a **Plan-Search-Report** loop: search queries are generated from the current plan and delegated to the **Sub Browse-Agent Cluster**—up to 20 concurrent Sub-Agents, each operating in an isolated context with two pre-filtering stages (LLM-based URL relevance filtering and timestamp-based recency filtering) before per-page extraction. The **Writing Agent** produces updated full-text report versions at each iteration, enabling dynamic user feedback. Human interruption is supported at any stage. Architectural details of the Sub-Browse Agent Cluster are provided in Appendix E.

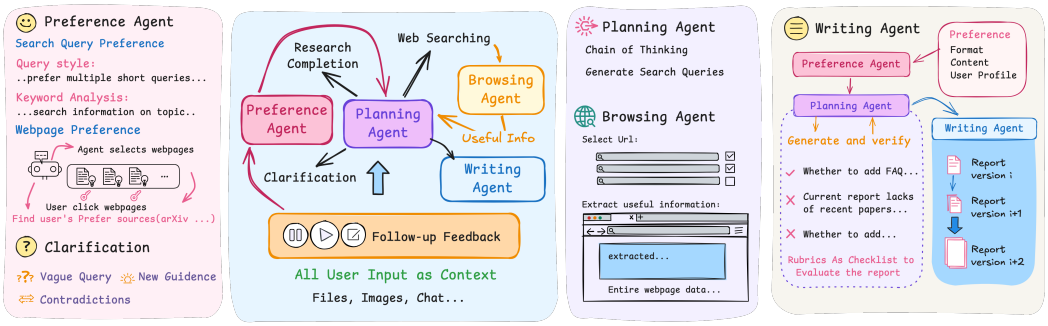

Figure 1: Deep Cognition framework overview.

3.2 MULTI-ROUND CLARIFICATION MECHANISM

Existing deep research systems such as OpenAI DeepResearch (OpenAI, 2025b) and Gemini DeepResearch (Google, 2025) typically conduct one-time question collection during the initial dialogue phase, but this approach neglects the dynamic clarification needs that emerge during the research process. We design an option-driven progressive clarification mechanism, implemented as a human-in-the-loop tool call within the main agent, that transforms complex clarification questions into structured option questionnaires rather than relying on free-text input (detailed prompts in Appendix F.1). When triggered, the agent emits a structured question (with 3–5 multiple-choice options) and blocks execution until the user responds or a timeout (300s) is reached.

The system employs a prompt-based trigger mechanism to identify moments requiring user clarification. Specifically, we design scenario templates in the system prompt that guide the LLM to recognize four situations: (1) Ambiguity Detection—multiple interpretations or unclear scope; (2) Information Conflict—contradictory claims across sources; (3) Branch Decision Points—multiple viable research directions; (4) Domain Expertise Gaps—specialized context beyond web-retrievable knowledge. To prevent over-interruption, the LLM reviews all previous clarification history before asking, ensuring each request provides incremental value.

Information-Theoretic Interpretation. We provide an information-theoretic interpretation for the trigger design in Appendix C: the four heuristic triggers map to conditions on expected information gain $\mathrm{IG}(t) = H(Q \mid C_t) - \mathbb{E}[H(Q \mid C_t, A)]$, subject to a decaying interruption budget. The framework predicts that clarification frequency should decrease as residual uncertainty reduces—consistent with the behavioral patterns observed in Section 5.5.

Dynamic Option Generation: Unlike systems that rely on predefined question templates, our framework dynamically generates appropriate questions: multiple-choice questions with 3–5 options covering probable user intents, open-ended questions for scenarios requiring free-form input, and contextual explanations to help users understand each option.

3.3 Iterative Reports by Writing Agent

While existing deep research systems (LangChain, 2025; Roucher et al., 2025) typically follow a sequential collect-then-generate paradigm, we propose an **evidence-driven iterative report construction strategy**. A dedicated Writing Agent, driven by carefully engineered prompts, generates a complete updated report at each research cycle, incorporating newly retrieved evidence while preserving previously synthesized content (detailed prompts in Appendix F.4).

The system dynamically generates or adjusts hierarchical research plans at the beginning of each cycle, serving as report outlines. This progressive synthesis delivers two advantages: (1) reasoning space construction—providing the model with a dedicated arena for analysis during iterative report optimization; and (2) selective context retention—preserving only evidence incorporated into the current report while discarding unutilized passages.

Selective Context Retention. As the research progresses, accumulated evidence may exceed the LLM's context window. We address this through a citation-based retention policy: at each writing iteration, the system retains only evidence passages that are cited in the current report draft R_t, plus freshly retrieved evidence from the current cycle E_{new}:

$$S_t = \{e \in E_t \mid e \text{ is cited by } R_t\} \cup E_{\text{new}}. \tag{1}$$

Unutilized evidence is discarded from subsequent processing contexts. This can be viewed as a special case of budget-constrained submodular maximization (Krause & Golovin, 2014) with a binary relevance function, prioritizing tractability in our prompt-based pipeline. Additionally, when cumulative token usage reaches 85% of the context window, a memory summarizer compresses earlier interaction steps (retaining only the two most recent verbatim) to enable sustained multi-round sessions.

3.4 Human-AI Co-Research Mechanism

Deep Cognition supports human–AI collaboration through five dialogue act types (Fig. 2) (Jacqmin et al., 2022): **Clarification**—the system asks structured questions to resolve ambiguity; **Interruption**—users halt and redirect the ongoing process; **Knowledge injection**—users supply domain expertise or references that the agent incorporates; **Plan review**—users inspect and revise the research plan before execution; and **Evidence-level feedback**—users comment on retrieved sources or report sections to steer synthesis. The interface exposes the system's reasoning process, search queries, and source provenance in real time, and implements a "Pause" feature for fine-grained sub-agent control.

Feedback-Conditioned Evidence Routing. Citation retention acts only after a draft exists and cannot allocate a limited context budget among newly collected sources. We therefore train a small router $\mathcal{R}_\theta$ that receives the question q, latest feedback f, a candidate set $E = \{e_i\}_{i=1}^{m}$, and budget k, and returns a complete ranking $\pi = \mathcal{R}_\theta(q, f, E, k)$. The top-$k$ sources are supplied in full, while the remainder are retained in compressed form rather than hard-dropped. The router is a fully fine-tuned Qwen3.5-2B student distilled from Qwen3.6-27B. Its rich target jointly supervises sub-question coverage, source utility and quality, evidence role, the full ranking, and the selected top-k set. This formulation treats routing as feedback-conditioned, budgeted evidence allocation rather than independent binary relevance classification; training and evaluation details are in Appendix A.

Table 2: Frozen-prefix routing on ResearcherBench (65; RB-50 excludes checkpoint-selection items) and ScholarQA-CS (110 rubric cases). Red marks best and second-best values.

Policy	ResearcherBench All 65	Unt. 50	Evid. chars.	Input tok.	Total tok.
Keep-all	0.657	0.658	3.253M	0.942M	1.129M
Random	0.659	0.671	**2.198M**	0.685M	0.872M
S01 2B	**0.679**	**0.682**	2.358M	**0.677M**	**0.866M**

Policy	ScholarQA-CS Score	–	Evid. chars.	Input tok.	Total tok.
Keep-all	0.565	–	2.341M	1.075M	1.450M
Random	0.564	–	**1.620M**	**0.895M**	**1.261M**
S01 2B	**0.576**	–	1.749M	0.904M	1.275M

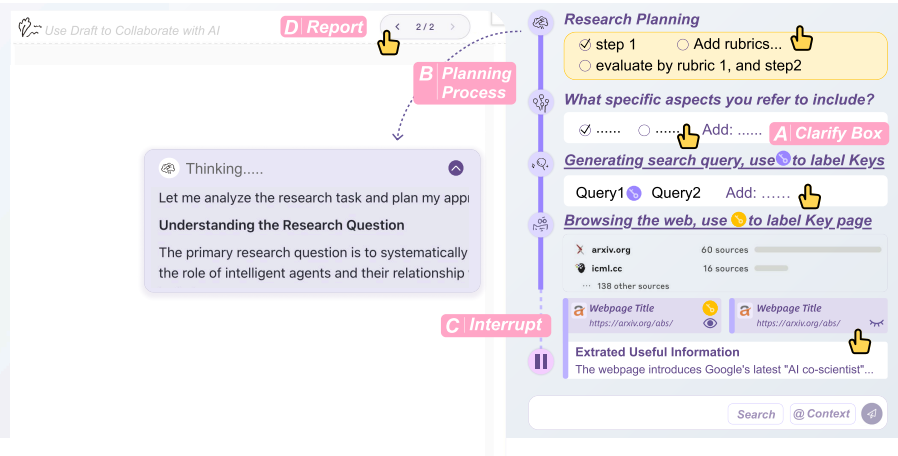

Figure 2: Deep Cognition interface: (A) Research scope clarification, (B) Clickable URLs, (C) Multi-agent workflow visualization, (D) Transparent reasoning and interactive query refinement, (E) Report revision. = clickable elements.

3.5 Automated Expert Simulation for Scalable Evaluation

To evaluate Deep Cognition at scale without relying solely on costly human experts, we design an *Automated Expert Simulation* pipeline. A separate LLM instance (Claude Opus) acts as a domain-specific simulated expert $\mathcal{E}$, interacting with the research system through the same dialogue act interface available to human users—**without access to ground-truth answers**. Fixed-checkpoint same- and cross-family controls in Section 5.3 test the potential alignment confound introduced by a same-family expert.

Adaptive Complexity Routing. Given a research question q, the pipeline classifies it into a complexity tier $c \in \{\textsc{Simple}, \textsc{Moderate}, \textsc{Complex}\}$ via a zero-shot LLM call using question length, decomposability, and expected synthesis depth as criteria. Each tier maps to a binary interaction configuration:

$$\Phi(c) = \langle \phi_{\text{clar}}, \phi_{\text{plan}}, \phi_{\text{int}}, \phi_{\text{know}}, K_{\text{gap}}, \phi_{\text{syn}}, \phi_{\text{qc}} \rangle \tag{2}$$

where $\phi_i \in \{0, 1\}$ activates each dialogue act and $K_{\text{gap}} \in \{0, \ldots, 6\}$ bounds gap analysis iterations. SIMPLE disables all interactions ($\Phi = \mathbf{0}$); MODERATE enables clarification, plan review, and knowledge injection; COMPLEX activates all seven acts with $K_{\text{gap}} = 6$. The expert is instantiated with a domain persona $\rho \in \{\text{science}, \text{finance}, \text{humanities}, \text{engineering}, \text{general}\}$ that shapes communication style and challenge patterns via template-based prompt conditioning.

Multi-Phase Execution. The pipeline executes 12 sequential phases gated by $\Phi(c)$: (1) complexity classification and sub-question decomposition; (2–3) parallel clarification + plan generation, followed by expert plan review; (4–5) multi-query search and page scraping (up to 25 URLs, 16K chars/page); (6–7) expert interruption (may redirect search) and knowledge injection (one missing concept + targeted query); (8) iterative gap analysis; (9–10) page-level extraction and synthesis guidance; (11–12) report generation and quality check with revision.

Iterative Gap Analysis. For complex questions, the expert iteratively assesses information sufficiency. At iteration k:

$$(s_k, Q_k) = \mathcal{E}_{\text{gap}}\big(q,\ \{q_i\},\ \textstyle\sum_{j=1}^{|D_k|} \text{snippet}(d_j)\big) \tag{3}$$

where $s_k \in \{0, 1\}$ is the sufficiency signal, Q_k is a set of remedial queries, and D_k is the accumulated document set. The loop terminates when $s_k = 1$, $Q_k = \emptyset$, or $k = K_{\text{gap}}$; documents are appended: $D_{k+1} = D_k \cup \text{Retrieve}(Q_k)$. After report generation, the expert evaluates whether all sub-questions received evidence-backed answers; if insufficient, a targeted re-search and regeneration cycle is triggered.

4 EXPERIMENTS

Evaluation Metrics. Following the principle that human–LM interaction must be evaluated beyond output quality alone (Lee et al., 2024), we evaluate along two axes with 15 dimensions (Table 19, Appendix D): *report quality* (Organization, Cutting-Edge, Coverage, Depth, Relevance; inspired by OpenScholar (Asai et al., 2024)) and *interaction design* (Transparency, Interruptibility, Fine-Grained Interaction, Real-Time Intervention, Inspiration, Ease of Collaboration, Results-Worth-Effort, Helpfulness, Intention to Use, Usability). All dimensions use a 5-point Likert scale with anchored definitions (Appendix H.2).

System Setup. We use Claude Opus 4.7 as the backbone for reasoning, writing, and simulated expert interaction (temperature 0.7). The browsing agent retrieves from multiple sources via parallel search; each turn generates up to 10 queries $\times$ 15 results, deduplicating to 25 pages for extraction. The expert simulation pipeline processes 20+ questions concurrently (full hyperparameters in Appendix E).

Human Study Setup. To address limitations of static benchmarks, we perform a controlled human evaluation (detailed protocol in Appendix H) to assess real-world research experience during human–AI interaction, inspired by Lee et al. (2024). We develop a web application for users to interact with Deep Cognition, and compare it with three deep research baselines: Gemini Deep Research (Google, 2025), OpenAI Deep Research (OpenAI, 2025b;a;c), and Grok 3 DeeperSearch (xAI, 2025). Participants were instructed to use the *full set of interaction features available in each system*—no interaction restrictions were imposed on any baseline; the asymmetry reflects architectural design (commercial baselines operate as single-cycle pipelines with no mechanism for mid-execution redirection), not experimental restriction.

Study 1: Report Quality & Interaction. We recruited 50 participants with active research needs and at least six months of domain experience, spanning six professional categories: LLM researchers (n=12), other CS researchers (n=10), AI product managers (n=9), general product managers (n=8), financial industry analysts (n=6), and user experience researchers (n=5). Participants were recruited via professional networks and required to demonstrate ongoing engagement with research-intensive tasks in their respective domains. Before the study, participants received a standardized tutorial on the evaluation metrics (Section 4). Each participant proposed a research question from their own work; the same question was used across all four systems, with a session length of 30–60 minutes per system (median 47 minutes). **System order was counterbalanced using a Latin square design** to mitigate learning and fatigue effects. Crucially, **report quality was assessed by a separate panel of expert evaluators**—not by the participants who generated the reports—using the standardized rubric under blinded conditions (detailed evaluation protocol in Appendix H). The study protocol received institutional ethics board approval; all participants provided informed consent and were compensated for their time.

Study 2: Expertise × Interaction. To test whether domain expertise (prior topic knowledge and research experience) improves human–AI collaboration in transparent dialogue, we evaluated exact-answer accuracy on 22 questions from BrowseComp-ZH (Zhou et al., 2025), selecting the first two questions from each of 11 categories.

5 Main Result

5.1 User Evaluation in Study 1: Report Quality & Interaction

Table 3 (left) presents report quality results. DC with interaction outperforms the no-interaction baseline on all six dimensions (Wilcoxon, all $p < .002$; Cliff's $\delta \geq +.44$). Against commercial baselines, DC leads on Organization, Cutting-Edge, and Helpfulness; Gemini leads on Coverage ($\delta = -.21$) and Relevance ($\delta = -.46$); Friedman tests confirm significant differences on five of six metrics ($p < .05$). DC also dominates six of seven interaction metrics (Table 3, right), with largest gains on Fine-Grained Interaction (+44.6%), Ease of Collaboration (+43.0%), and a perfect Transparency score (5.00). These advantages concentrate on DC's architectural novelty—per-subagent steering and mid-task interruption—rather than output-centric measures (Inspiration tied with Gemini). Qualitative assessment further confirms practical adoption potential: over 90% of participants agree that interaction with DC improves report quality; 69% rate it easy to use (Usability) and 62% express high continued-use intention. Overall, DC achieves comparable report quality to commercial systems while providing substantially richer collaborative affordances.

Table 3: Report quality (left, assessed by independent evaluator panel) and interaction evaluation (right, participant self-report). 5-pt Likert, n=50. Friedman χ^2_F across four systems; δ = Cliff's delta vs. Gemini. Color = within-row rank.

Metric	Mean (1–5)					χ^2_F	p	$\delta_{\text{vs Gem}}$
	DC w/o	DC	Gem.	OAI	Grok			
Organization	2.231	**4.385**	4.308	3.769	3.385	14.3	.003	+.12
Cutting-Edge	2.538	**4.538**	4.385	3.769	3.538	7.5	.059	+.14
Coverage	2.423	4.000	**4.462**	3.692	2.923	9.3	.025	−.21
Depth	2.231	3.923	**4.000**	3.577	2.769	11.2	.011	−.02
Relevance	2.885	3.769	**4.615**	4.077	3.615	8.9	.031	−.46
Helpfulness	2.808	**4.000**	**4.000**	3.615	2.692	11.8	.008	−.02

Metric	Mean (1–5)				χ^2_F	p	$\delta_{\text{vs Gem}}$
	DC	Gem.	OAI	Grok			
Transparency	**5.00**	4.00	3.00	3.19	29.9	< .001	+.78
Interruptibility	**4.35**	3.31	2.69	2.62	29.5	< .001	+.63
Fine-Grained Int.	**4.73**	3.27	2.88	2.19	41.2	< .001	+.85
Real-Time Int.	**4.69**	3.77	2.92	2.62	38.7	< .001	+.50
Inspiration	**4.08**	**4.08**	3.42	3.19	13.5	.004	−.14
Ease of Collab.	**4.62**	3.23	2.77	1.85	45.5	< .001	+.73
Results-Worth-Eff.	**4.52**	4.08	3.29	2.96	19.9	< .001	+.40

5.2 Main Quantitative Results

Table 4 reports the end-to-end benchmark results. For compactness, the shared-manifest Magentic-UI comparison is included as the rightmost column rather than repeated in a separate table. Appendix B.4 further compares the two systems' task regimes, exposed research state, and interaction mechanisms.

DC (int.) is strongest on all five shared-manifest rows, reaching 0.7652 on ScholarQA-CS, 0.7281 on ResearcherBench, 50.92 RACE on DeepResearch Bench, 57.58% on GAIA validation, and 42.31% on GAIA Level 3. DC (auto) scores 0.6669 on ScholarQA-CS.

5.3 Controlled Follow-Up Experiments

Appendix B provides the full budget sweeps, model-family tests, and output accounting.

Feedback Timing and Structure. With the GLM-5.2 checkpoint, task set, six-feedback budget, and evaluator fixed, Table 5 varies only feedback timing and structure. Structured steering improves over no feedback by $+0.0642$ (95% CI $[+0.0323, +0.0968]$), uses 45% less time and 20% fewer tokens than unstructured feedback, and exceeds post-output refinement by $+0.0328$ (CI $[+0.0059, +0.0610]$).

Table 5: Feedback timing and structure on ResearcherBench (n=65).

Condition	Cov. ↑	Sec. ↓	Tok. ↓
No feedback	0.7040	52.9	5,538
Unstructured	0.7479	289.3	53,669
Structured	**0.7681**	**158.3**	**42,764**
Post-output	0.7583	333.5	77,480

Table 4: End-to-end automated benchmark evaluation. Magentic-UI and the two DC columns use the shared manifest; published commercial-system values are shown for context. Best values within each row are in **bold**.

Category	Benchmark	Metric	Gemini DR	Grok3	OpenAI DR	DC (auto)	DC (int.)	Magentic-UI
Synthesis	ScholarQA-CS (110) (Asai et al., 2024)	Rubric	0.734	0.682	0.722	0.6669	**0.7652**	0.6254
Synthesis	ResearcherBench (65) (Xu et al., 2025)	Rubric	0.693	0.440	0.703	0.6980	**0.7281**	0.5812
Synthesis	DeepResearch Bench (100) (Du et al., 2025)	RACE	48.88	40.24	46.98	50.30	**50.92**	49.69
			Search-o1	**WT-Base**	**WT-RL**	**DC (auto)**	**DC (int.)**	**Magentic-UI**
Retrieval	GAIA (165) (Mialon et al., 2024)†	Acc.	39.8%	44.7%	48.5%	43.03%	**57.58%**	48.48%
Retrieval	GAIA Level 3 (26)	Acc.	16.7%	16.7%	16.7%	38.46%	**42.31%**	23.08%

†WebThinker values use the private-label GAIA test set (103 questions), whereas DC and Magentic-UI use the public validation set (165) (Li et al., 2025).

Table 6: Expert-family control from a fixed research checkpoint.

Expert	Researcher Bench	Scholar QA-CS	Deep Research
None	0.6980	**0.6669**	50.30
Claude, same family	0.7281	0.5942	50.92
GLM-5.2, cross family	**0.7736**	0.6076	49.95
Qwen3.7, cross family	0.7180	0.6578	**52.24**

Table 7: Frozen-prefix evidence allocation on ResearcherBench; the held-out 50 exclude checkpoint-selection questions.

Evidence policy	RB all (65) ↑	Held-out (50) ↑	Evidence chars. ↓	Writer input ↓
Keep all	0.6569	0.6581	3.253M	0.942M
Matched random	0.6593	0.6709	**2.198M**	0.685M
S01 2B router	**0.6790**	**0.6821**	2.358M	**0.677M**

Cross-Family Expert Generalization. Holding the Opus 4.8 research checkpoint fixed, Table 6 varies only the expert model. Cross-family GLM improves ResearcherBench by $+0.0756$ and cross-family Qwen improves DeepResearch by $+1.94$ RACE, ruling out same-family alignment as the sole explanation; complete results appear in Appendix B.2.

Feedback-Conditioned Evidence Allocation. Table 7 freezes the pre-routing trajectory, candidate pages, writer, and evaluator, changing only full-text evidence allocation and thereby isolating the learned router.

On the 50 held-out questions, S01 improves over keep-all by $+0.0240$ (95% CI $[+0.0007, +0.0475]$). Across all 65, it scores 0.6790 versus 0.6569 for keep-all and 0.6593 for matched random, while reducing writer input tokens by 28.1% and total writer tokens by 23.3%. The same frozen-prefix protocol transfers to ScholarQA-CS (n=110): S01 scores 0.5755 versus 0.5646 for keep-all and 0.5637 for matched random, while reducing total writer tokens by 12.0%. Both ScholarQA-CS paired estimates are positive, although their 95% CIs cross zero (Appendix A).

5.4 Dynamic Collaboration Patterns in Study 2: Expertise × Interaction

Full DC reaches 72.73% BrowseComp-ZH accuracy versus 40.91% for the strongest commercial baseline; neither architecture without expertise (45.45%) nor expertise without interaction (40.91%) matches their combination.

Across six phases, users shift between guidance and monitoring (Appendix L). Two authors coded all 50 transcripts (κ=0.83): users guide clarification and knowledge input, delegate reasoning and synthesis, and re-engage for intervention and search. Thus the 17% hands-on rate masks phase-dependent steering, while Knowledge Injection contributes the largest marginal gain (+22.73 pp; Table 8).

5.5 Large-Scale Deployment Analysis

Deployment logs contain **2,461** sessions from **194** users and 34,715 traces. In 77% of sessions, output length and structure increase (7.5× and 6.7×; Wilcoxon $p<10^{-126}$), although these are proxy measures. Among 5,895 classified actions, 58.5% are hands-on (Z=13.0, $p<10^{-15}$), and self-transitions exceed chance (32.05% vs. 18.86%; Z=20.7), indicating sustained intervention episodes. Behavior changes across phases (χ^2=204.2, $p<10^{-37}$), yet high- and low-engagement

Table 8: BrowseComp-ZH interaction and component ablations (22 questions), restored to the Dynamic Collaboration analysis. Eff. $= \Delta\text{Acc}/N$; H-on denotes the hands-on interaction rate.

Configuration	Acc.	Qual.	N	H-on	Eff.
Full DC (Exp+Int)	**72.73%**	**4.00**	6.2	17%	5.13
w/o Clarification	63.64%	3.62	5.1	14%	4.46
w/o Interruption	59.09%	3.38	5.3	13%	3.43
w/o Knowledge Inj.	50.00%	3.04	4.4	10%	2.07
w/o Plan Review	63.64%	3.55	5.4	14%	4.21

System	Acc.	Qual.	N	H-on	Eff.
DC (w/o exp.)	45.45%	–	4.1	12%	–
DC (w/o int.)	40.91%	2.81	0	0%	–
Gemini DR	40.91%	4.00	2.8	18%	–
OpenAI DR	40.91%	3.62	1.4	7%	–
Grok 3	22.73%	2.69	1.1	3%	–

sessions have similar output length (p=0.334): participation changes direction, not merely verbosity (Figure 4; Appendix Figure 3).

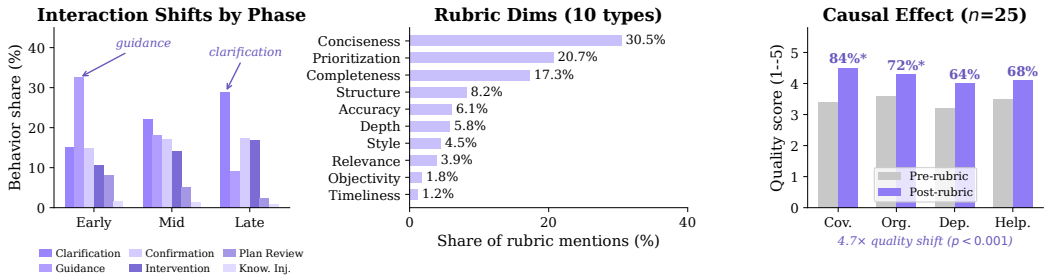

Figure 4: Left: phase-dependent behavioral evolution (χ^2=204.2, $p<10^{-37}$); guidance dominates early and clarification rises late. Center: quality-rubric taxonomy (10 dimensions, 32 subtypes). Right: within-session pre/post comparisons; percentages are post-rubric win rates. Rubric presence is associated with a 4.7× quality shift (p=0.00019).

Content Quality and Rubric Effects. Claude Sonnet 4.6 evaluation of 254 sessions finds 78.7% with mean quality ≥ 4 (4.16±0.88). Explicit rubrics appear in 36.1% of sessions (10 dimensions, 32 subtypes); in 25 within-session pairs, post-rubric outputs win 84% on Coverage (p=0.0005) and 72% on Organization (p=0.022). These are observational associations because conversation progress and new evidence are not controlled. Appendix Figure 5 and Table 23 give the full results.

Mechanistic Insights. Two null results constrain the mechanism (Appendix J): rubric categories do not selectively improve their matching dimensions (p=0.086–0.093), and operational rubrics receive lower apparent compliance than vague ones (2.60 vs. 3.79/5). Rubrics therefore act more like holistic quality activators than fine-grained controllers, while explicit criteria make failures easier to detect.

6 Conclusion

Deep Cognition combines transparent multi-agent research, simulated-expert intervention, and a 2B evidence router. Human studies, deployment logs, matched controls, and frozen-prefix evaluations show that structured in-process feedback improves quality and efficiency across model families. Collaborative research therefore depends on both when to intervene and what evidence to retain.

AI USE STATEMENT

Generative AI tools were used to assist with literature discovery and synthesis, brainstorming and refinement of the research framing, generation and debugging of data-processing, training, and evaluation code, teacher-assisted annotation, and language editing of the manuscript. The language models that serve as research backbones, simulated experts, teacher labelers, and evaluators are also part of the methodology and are explicitly identified in the paper. The authors manually checked AI-suggested references against the corresponding publications, reviewed and executed AI-assisted code, inspected sampled annotations and model outputs, and verified reported claims against saved experimental artifacts. Generative AI tools were not treated as authors or as authoritative sources. The authors made all final scientific and editorial decisions and take full responsibility for the text, claims, code, and other artifacts in this work.

ETHICS STATEMENT

This work includes a user study and an observational analysis of deployment logs. The user-study protocol received institutional ethics-board approval; participants provided informed consent and were compensated for their time. Deployment data were collected with user consent, analyzed in aggregate, and reported without personally identifying information. Because automated research systems and LLM-based evaluators can reproduce factual errors and social biases, we use blinded human evaluation where applicable, report evaluator and same-family confounds, and avoid interpreting automated scores as substitutes for human judgment.

REPRODUCIBILITY STATEMENT

Sections 3–5 specify the system components, dialogue acts, model roles, benchmarks, interaction budgets, baselines, metrics, and statistical tests used in the study. The appendix provides the controlled follow-up protocols, router supervision and training recipe, checkpoint-selection split, output accounting, prompt and rubric details, ablations, deployment-data processing, and additional null results. The supplementary artifacts include the data-processing, training, inference, and evaluation configurations needed to reconstruct the reported comparisons. We retain per-condition predictions and judge outputs so that aggregate scores and paired analyses can be independently recomputed; access to consented deployment data remains subject to privacy and institutional restrictions.

REFERENCES

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pp. 1–13, 2019.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. OpenScholar: Synthesizing scientific literature with retrieval-augmented LMs. *arXiv preprint arXiv:2411.14199*, 2024. URL https://arxiv.org/abs/2411.14199.

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

Nishant Balepur, Malachi Hamada, Varsha Kishore, Sergey Feldman, Amanpreet Singh, Pao Siangliulue, Joseph Chee Chang, Eunsol Choi, Jordan Lee Boyd-Graber, and Aakanksha Naik. Language models don't know what you want: Evaluating personalization in deep research needs real users. *arXiv preprint arXiv:2603.16120*, 2026. URL https://arxiv.org/abs/2603.16120.

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. ACM, 2021.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel

Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):188:1–188:21, 2021. doi: 10.1145/3449287. URL https://doi.org/10.1145/3449287.

ByteDance. Deerflow, 2025. URL https://github.com/bytedance/deer-flow. Community-driven deep research framework combining LLMs with web search, crawling, and code execution tools.

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

Zehui Chen, Kuikun Liu, Qiuchen Wang, et al. MindSearch: Mimicking human minds elicits deep AI searcher. In *International Conference on Learning Representations (ICLR)*, 2025.

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

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*, 2025. URL https://arxiv.org/abs/2506.11763.

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

K. J. Kevin Feng, Kevin Pu, Matt Latzke, Tal August, Pao Siangliulue, Jonathan Bragg, Daniel S. Weld, Amy X. Zhang, and Joseph Chee Chang. Cocoa: Co-planning and co-execution with AI

agents. *arXiv preprint arXiv:2412.10999*, 2024. URL https://arxiv.org/abs/2412.10999.

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

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. Plan-then-execute: An empirical study of user trust and team performance when using LLM agents as a daily assistant. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. ACM, 2025. doi: 10.1145/3706598.3713218.

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

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In *International Conference on Learning Representations (ICLR)*, 2024.

Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99. ACM, 1999.

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

Léo Jacqmin, Lina M. Rojas Barahona, and Benoit Favre. "do you follow me?": A survey of recent approaches in dialogue state tracking. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, SIGDIAL 2022, pp. 336–350. Association for Computational Linguistics, 2022.

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

Majeed Kazemitabaar, Jack Williams, Ian Drosos, Tovi Grossman, Austin Henley, Carina Negreanu, and Advait Sarkar. Improving steering and verification in AI-assisted data analysis with interactive task decomposition. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, UIST '24. ACM, 2024.

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

Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*, pp. 71–104. Cambridge University Press, 2014.

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

LangChain. Open deep research, 2025. URL https://github.com/langchain-ai/open_deep_research. GitHub repository.

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

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, et al. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. WebThinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*, 2025. URL https://arxiv.org/abs/2504.21776.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

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

Chris Lu, Cong Lu, Robert Tjarko Lange, et al. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

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

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation, 4 2025. URL https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: September 26, 2026.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations (ICLR)*, 2024.

MiniMax, Aonian Li, Bangwei Gong, Bo Yang, Boji Shan, Chang Liu, Cheng Zhu, Chunhao Zhang, Congchao Guo, Da Chen, Dong Li, Enwei Jiao, Gengxin Li, Guojun Zhang, Haohai Sun, Houze Dong, Jiadai Zhu, Jiaqi Zhuang, Jiayuan Song, Jin Zhu, Jingtao Han, Jingyang Li, Junbin Xie, Junhao Xu, Junjie Yan, Kaishun Zhang, Kecheng Xiao, Kexi Kang, Le Han, Leyang Wang, Lianfei Yu, Liheng Feng, Lin Zheng, Linbo Chai, Long Xing, Meizhi Ju, Mingyuan Chi, Mozhi Zhang, Peikai Huang, Pengcheng Niu, Pengfei Li, Pengyu Zhao, Qi Yang, Qidi Xu, Qiexiang Wang, Qin Wang, Qiuhui Li, Ruitao Leng, Shengmin Shi, Shuqi Yu, Sichen Li, Songquan Zhu, Tao Huang, Tianrun Liang, Weigao Sun, Weixuan Sun, Weiyu Cheng, Wenkai Li, Xiangjun Song, Xiao Su, Xiaodong Han, Xinjie Zhang, Xinzhu Hou, Xu Min, Xun Zou, Xuyang Shen, Yan Gong, Yingjie Zhu, Yipeng Zhou, Yiran Zhong, Yongyi Hu, Yuanxiang Fan, Yue Yu, Yufeng Yang, Yuhao Li, Yunan Huang, Yunji Li, Yunpeng Huang, Yunzhi Xu, Yuxin Mao, Zehan Li, Zekang Li, Zewei Tao, Zewen Ying, Zhaoyang Cong, Zhen Qin, Zhenhua Fan, Zhihang Yu, Zhuo Jiang, and Zijia Wu. Minimax-01: Scaling foundation models with lightning attention, 2025. URL https://arxiv.org/abs/2501.08313.

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

Hussein Mozannar, Gagan Bansal, Cheng Tan, Adam Fourney, Victor Dibia, Jingya Chen, Jack Gerrits, Tyler Payne, Matheus Kunzler Maldaner, Madeleine Grunde-McLaughlin, Eric Zhu, Griffin Bassman, Jacob Alber, Peter Chang, Ricky Loynd, Friederike Niedtner, Ece Kamar, Maya Murad, Rafah Hosn, and Saleema Amershi. Magentic-UI: Towards human-in-the-loop agentic systems. *arXiv preprint arXiv:2507.22358*, 2025. URL https://arxiv.org/abs/2507.22358.

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

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

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

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

OpenAI. OpenAI operator, 2025d. URL https://openai.com/operator. Accessed 2025.

OpenAI, :, Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, Alex Iftimie, Alex Karpenko, Alex Tachard Passos, Alexander Neitz, Alexander Prokofiev, Alexander Wei, Allison Tam, Ally Bennett, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrew Duberstein, Andrew Kondrich, Andrey Mishchenko, Andy Applebaum, Angela Jiang, Ashvin Nair, Barret Zoph, Behrooz Ghorbani, Ben Rossen, Benjamin Sokolowsky, Boaz Barak, Bob McGrew, Borys Minaiev, Botao Hao, Bowen Baker, Brandon Houghton, Brandon McKinzie, Brydon Eastman, Camillo Lugaresi, Cary Bassin, Cary Hudson, Chak Ming Li, Charles de Bourcy, Chelsea Voss, Chen Shen, Chong Zhang, Chris Koch, Chris Orsinger, Christopher Hesse, Claudia Fischer, Clive Chan, Dan Roberts, Daniel Kappler, Daniel Levy, Daniel Selsam, David Dohan, David Farhi, David Mely, David Robinson, Dimitris Tsipras, Doug Li, Dragos Oprica, Eben Freeman, Eddie Zhang, Edmund Wong, Elizabeth Proehl, Enoch Cheung, Eric Mitchell, Eric Wallace, Erik Ritter, Evan Mays, Fan Wang, Felipe Petroski Such, Filippo Raso, Florencia Leoni, Foivos Tsimpourlas, Francis Song, Fred von Lohmann, Freddie Sulit, Geoff Salmon, Giambattista Parascandolo, Gildas Chabot, Grace Zhao, Greg Brockman, Guillaume Leclerc, Hadi Salman, Haiming Bao, Hao Sheng, Hart Andrin, Hessam Bagherinezhad, Hongyu Ren, Hunter Lightman, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian Osband, Ignasi Clavera Gilaberte, Ilge Akkaya, Ilya Kostrikov, Ilya Sutskever, Irina Kofman, Jakub Pachocki, James Lennon, Jason Wei, Jean Harb, Jerry Twore, Jiacheng Feng, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joaquin Quiñonero Candela, Joe Palermo, Joel Parish, Johannes Heidecke, John Hallman, John Rizzo, Jonathan Gordon, Jonathan Uesato, Jonathan Ward, Joost Huizinga, Julie Wang, Kai Chen, Kai Xiao, Karan Singhal, Karina Nguyen, Karl Cobbe, Katy Shi, Kayla Wood, Kendra Rimbach, Keren Gu-Lemberg, Kevin Liu, Kevin Lu, Kevin Stone, Kevin Yu, Lama Ahmad, Lauren Yang, Leo Liu, Leon Maksin, Leyton Ho, Liam Fedus, Lilian Weng, Linden Li, Lindsay McCallum, Lindsey Held, Lorenz Kuhn, Lukas Kondraciuk, Lukasz Kaiser, Luke Metz, Madelaine Boyd, Maja Trebacz, Manas Joglekar, Mark Chen, Marko Tintor, Mason Meyer, Matt Jones, Matt Kaufer, Max Schwarzer, Meghan Shah, Mehmet Yatbaz, Melody Y. Guan, Mengyuan Xu, Mengyuan Yan, Mia Glaese, Mianna Chen, Michael Lampe, Michael Malek, Michele Wang, Michelle Fradin, Mike McClay, Mikhail Pavlov, Miles Wang, Mingxuan Wang, Mira Murati, Mo Bavarian, Mostafa Rohaninejad, Nat McAleese, Neil Chowdhury, Neil Chowdhury, Nick Ryder, Nikolas Tezak, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Patrick Chao, Paul Ashbourne, Pavel Izmailov, Peter Zhokhov, Rachel Dias, Rahul Arora, Randall Lin, Rapha Gontijo Lopes, Raz Gaon, Reah Miyara, Reimar Leike, Renny Hwang, Rhythm Garg, Robin Brown, Roshan James, Rui Shu, Ryan Cheu, Ryan Greene, Saachi Jain, Sam Altman, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Santiago Hernandez, Sasha Baker, Scott McKinney, Scottie Yan, Shengjia Zhao, Shengli Hu, Shibani Santurkar, Shraman Ray Chaudhuri, Shuyuan Zhang, Siyuan Fu, Spencer Papay, Steph Lin, Suchir Balaji, Suvansh Sanjeev, Szymon Sidor, Tal Broda, Aidan Clark, Tao Wang, Taylor Gordon, Ted Sanders, Tejal Patwardhan, Thibault Sottiaux, Thomas Degry, Thomas Dimson, Tianhao Zheng, Timur Garipov, Tom Stasi, Trapit Bansal, Trevor Creech, Troy Peterson, Tyna Eloundou, Valerie Qi, Vineet Kosaraju, Vinnie Monaco, Vitchyr Pong, Vlad Fomenko, Weiyi Zheng, Wenda Zhou, Wes McCabe, Wojciech Zaremba, Yann Dubois, Yinghai Lu, Yining Chen, Young Cha, Yu Bai, Yuchen He, Yuchen Zhang, Yunyun Wang, Zheng Shao, and Zhuohan Li. OpenAI o1 system card, 2024. URL https://arxiv.org/abs/2412.16720.

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

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

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

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

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

Aditi Singh, Abul Ehtesham, Saket Kumar, et al. Agentic retrieval-augmented generation: A survey on agentic RAG. *arXiv preprint arXiv:2501.09136*, 2025.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, et al. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

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

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*, 2024a. URL https://arxiv.org/abs/2406.04692.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 2024b.

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

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

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

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

Junde Wu, Jiayuan Zhu, Yuyuan Liu, Min Xu, and Yueming Jin. Agentic reasoning: A streamlined framework for enhancing LLM reasoning with agentic tools. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 28489–28503, 2025. URL https://aclanthology.org/2025.acl-long.1383/.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaoke Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Conference on Language Modeling (COLM)*, 2024.

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

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

Tianze Xu, Pengrui Lu, Lyumanshan Ye, Xiangkun Hu, and Pengfei Liu. ResearcherBench: Evaluating deep AI research systems on the frontiers of scientific inquiry. *arXiv preprint arXiv:2507.16280*, 2025. URL https://arxiv.org/abs/2507.16280.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models, 2023. URL https://arxiv.org/abs/2210.03629.

Ming Yin. Bridging the gap between machine confidence and human perceptions. *Nature Machine Intelligence*, 7:330–331, 2025. doi: 10.1038/s42256-025-01013-x.

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

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

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations (ICLR)*, 2024.

A Feedback-Conditioned Evidence Router

Supervision construction. We use Qwen3.6-27B as a frozen teacher to convert deployment-visible research states into listwise evidence-allocation targets. Each state contains the question, latest feedback, and an unordered candidate set; the teacher produces sub-questions, per-candidate utility, source quality, coverage and evidence role, a complete ranking, and a budgeted top-k set. We retain 4,675 successful teacher pools and split them by session into 3,777 train, 423 validation, and 475 test pools, with no session overlap. Training augmentation uses retention ratios $\{0.4, 0.6, 0.8\}$ and both original and shuffled candidate order, yielding 22,338 training examples. The training mix contains deployment-shaped pools, real within-trajectory pools, and cross-task hard-negative pools.

Student and checkpoint selection. S01 fully fine-tunes Qwen3.5-2B for two epochs with learning rate 8×10^{-6} (cosine decay to 8×10^{-7}), global batch size 64, BF16, and a 32K maximum sequence length on eight H200 GPUs. Ten recipes vary target richness, learning rate, epoch count, and hard-negative/deployment sampling. We select S01 using a fixed 15-question development slice ($\{4, 5, 11, 13, 15, 16, 17, 24, 27, 33, 35, 48, 52, 54, 62\}$); consequently, our primary inferential check reports the other 50 ResearcherBench questions separately.

Frozen-prefix evaluation. All three arms share the same question, pre-routing research trajectory, candidate pages, writer prompt, and output allowance. Keep-all supplies every page in full. S01 supplies the top 60% in full and a compressed representation of the remainder. Cardinality-matched random supplies the same per-question number of full pages as S01. Writer and rubric judge are GPT-5.6-sol for all arms. Every arm produces 65/65 reports, and all 931 weighted rubric points per arm are judged successfully under identical decoding settings.

On all 65 questions, S01 exceeds keep-all by $+0.0221$ (95% paired-bootstrap CI $[+0.0018, +0.0428]$; win/tie/loss 26/24/15) and matched random by $+0.0197$ (CI $[-0.0010, +0.0401]$; 27/23/15). On the untouched 50, S01 exceeds keep-all by $+0.0240$ (CI $[+0.0007, +0.0475]$; 21/19/10). Relative to keep-all over all 65 questions, S01 reduces evidence characters by 27.5%, writer input tokens by 28.1%, and total writer tokens by 23.3%. Thus the learned allocation improves quality despite a substantially smaller context, while the learned-versus-random difference remains a positive but non-significant point estimate.

We additionally apply the same frozen-prefix allocation protocol to all 110 ScholarQA-CS rubric cases, using DeepSeek-v4.1-Flash as the common writer and judge for all three arms. Its paired gains are $+0.0109$ over keep-all (95% CI $[-0.0052, +0.0274]$) and $+0.0118$ over matched random (CI $[-0.0058, +0.0303]$). Relative to keep-all, S01 reduces evidence characters by 25.3%, writer input tokens by 16.0%, and total writer tokens by 12.0%. This provides positive cross-benchmark transfer under a smaller evidence budget, although neither ScholarQA-CS difference is statistically significant.

B Model-Family Generalization and Controlled Feedback Analysis

We evaluate whether the observed gains generalize across interaction budgets, benchmarks, research backbones, and expert model families. The controlled comparisons also separate structured mid-process intervention from merely adding model calls or refining a completed report. Across the three experiment families in Table 9, the feedback controls comprise 3,415 independently judged outputs.

B.1 Feedback Timing and Interaction Budget

On all 65 ResearcherBench items, a matched Claude experiment used the same Opus 4.8 backbone, Claude expert, initial evidence, six expert contacts, judge prompt, and scoring protocol. Mid-process intervention obtains 0.7417 rubric coverage, compared with 0.7281 for post-output refinement. Final reports are anonymized; the independent judge sees the task, rubric, and report, but not the system identity, condition, or trajectory.

To test whether this result depends on a hand-picked contact count, we sweep $k \in \{2, 4, 6, 10, 14\}$ using GLM-5.2-FP8 as both backbone and expert (Table 10). Contact counts and all non-timing

Table 9: Accounting for independently judged controlled outputs.

Experiment	Design	Outputs
Matched-feedback ablation	4 conditions × 65 ResearcherBench items	260
Interaction-budget sweep	11 conditions over ResearcherBench (65), ScholarQA-CS (110), and DeepResearch (100)	3,025
Native long context control	2 conditions × 65 ResearcherBench items	130
Total		**3,415**

factors are matched. Mid-process feedback has a positive mean difference and lower latency at every tested budget. At k=2, the difference is $+0.0337$ (paired-bootstrap 95% CI $[+0.0019, +0.0645]$), while post-output refinement takes $1.96\times$ as long. The non-monotonic scores show that more contacts are not always better.

Table 10: Matched interaction-budget sweep on ResearcherBench (n=65 per condition) with a fixed GLM-5.2 backbone and expert.

k	Mid-process	Post-output	Δ	Mid min	Post min
2	0.7028	0.6691	+0.0337	3.4	6.7
4	0.6965	0.6783	+0.0182	10.3	14.2
6	0.6912	0.6861	+0.0051	13.7	21.1
10	0.7090	0.6894	+0.0196	23.3	37.3
14	0.7021	0.6896	+0.0125	54.2	61.2

The same timing sweep is positive across all five k values on three benchmarks (Table 11). At least one ResearcherBench and one ScholarQA-CS confidence interval exclude zero; all five DeepResearch intervals exclude zero. This consistency supports the conclusion that the timing advantage is not isolated to one benchmark or contact count, while the non-monotonicity motivates adaptive stopping.

Table 11: Multi-benchmark summary of the matched timing sweep for $k \in \{2, 4, 6, 10, 14\}$.

Benchmark	Items/cond.	Mid − post range	Paired evidence across five k values
ResearcherBench	65	+0.0051 to +0.0337	Positive at all five; at least one CI excludes zero; mid-process is faster at all five.
ScholarQA-CS	110	+0.0079 to +0.0304	Positive at all five; at least one CI excludes zero.
DeepResearch	100	+1.14 to +1.79 RACE	Positive at all five; all five CIs exclude zero.

B.2 Cross-Family and Cross-Backbone Controls

Table 6 holds the Opus 4.8 research state fixed and varies only the expert family. We additionally rerun the full adaptive planning–retrieval–expert-integration–synthesis pipeline with a GLM-5.2 backbone (Table 12). Within this block, prompts, retrieval configuration, item manifests, and the 600-s budget are fixed. Same-family GLM and cross-family Kimi improve DeepResearch over no expert by $+2.73$ RACE points (95% CI $[+1.27, +4.01]$) and $+3.71$ ($[+2.40, +4.87]$), respectively. Kimi is $+0.98$ above GLM. The Qwen DeepResearch cell contains 48 endpoint timeouts and is reported separately from the clean comparisons.

A separate end-to-end GPT/Codex backbone with a GPT-5.6 expert obtains 0.7534 on ResearcherBench, 0.5903 on ScholarQA-CS, 52.10 DeepResearch RACE, and 29.70% on GAIA. This condition demonstrates portability to a third family but is not a fixed-backbone expert-family ablation. Together, the controlled blocks establish that the improvements on ResearcherBench and DeepResearch do not depend on same-family alignment.

Table 12: Complete-pipeline cross-backbone control. †The Qwen DeepResearch condition contains 48 endpoint timeouts and should not be interpreted as a clean capability estimate.

Expert with adaptive GLM-5.2 backbone	ResearcherBench	ScholarQA-CS	DeepResearch RACE	GAIA
None	0.5426	**0.6707**	41.56	**38.18%**
GLM-5.2, same family	**0.6173**	0.6195	44.30	32.73%
Qwen3.6-27B, cross family	0.5576	0.6227	23.18†	32.12%
Kimi K2.6, cross family	0.5855	0.6668	**45.28**	35.76%

B.3 Why ResearcherBench Is Especially Sensitive to Feedback

Under the official binary weighted-rubric protocol, the autonomous outputs score 0.4637 and the simulated-expert outputs score 0.6587, a paired difference of $+0.1950$ (95% CI $[+0.1276, +0.2632]$). Table 13 localizes the difference: the largest gains concern evidence verification, quantitative specificity, and actionable recommendations. This is consistent with ResearcherBench placing more weight on evidence-level scrutiny than benchmarks dominated by organization and broad coverage.

Table 13: ResearcherBench capability breakdown for the autonomous and assisted runs.

Rubric capability	DC (auto)	DC (int.)	Δ
Evidence/citation/verification	0.4286	0.7540	+0.3254
Specific or quantitative detail	0.3969	0.6374	+0.2405
Actionable recommendations	0.4642	0.6906	+0.2264
Method or mechanism	0.4514	0.6282	+0.1768
Comparisons, limitations, or risks	0.4523	0.6182	+0.1659

B.4 Detailed Comparison with Magentic-UI

Magentic-UI and Deep Cognition share a mixed-initiative objective but target different task and interaction regimes (Table 14). Magentic-UI exposes plan steps, handoffs, progress, and browser actions; Deep Cognition delegates routine retrieval while preserving research-specific checkpoints over parallel subagents, sources, evidence, and a persistent draft. The distinction is the set of intermediate research artifacts available for steering, not whether Magentic-UI supports interaction.

Table 14: Task-regime and interaction-design differences between Magentic-UI and Deep Cognition.

Dimension	Magentic-UI	Deep Cognition
Primary task regime	Browser/GUI tasks and short-answer QA; string-, F1-, or screenshot-based evaluation	Long-form research synthesis; rubric- and reference-based evaluation
Reported execution structure	Median plan length 2 on WebVoyager, GAIA, and AssistantBench	12 gated phases with up to 6 gap-analysis rounds
Reported time scale	Median 113.9 s for successful WebVoyager tasks	600-s benchmark budget; 30–60 min controlled-study sessions
Retrieval organization	Browser retrieval through a WebSurfer agent	Up to 20 concurrent subagents and 25 retained URLs per round
Structured steering objects	Plan steps, agent handoffs, and action approval	Plans, subagents, queries, evidence, and report sections
Intermediate research state	Concise progress summaries	Live queries, sources, extracted evidence, and evolving full-text report
Interaction scheduling	Clarification and replanning from execution state	Difficulty-conditioned SIMPLE/MODERATE/COMPLEX routing
Empirical scope	12 participants in one-hour sessions	50 participants and 2,461 deployment sessions

B.5 Benchmark Accounting and Protocol Separation

The paper evaluates five unique datasets; GAIA Level 3 is a subset rather than a sixth benchmark. Table 15 records exact evaluated sets and primary metrics. Automated attempts receive a 600-s budget, and failures or timeouts receive zero on the full denominator. Synthesis conditions use the same independent GPT-5.6-sol evaluator under each benchmark's protocol; GAIA and BrowseComp-ZH use deterministic matching.

Table 15: Benchmark accounting, denominators, and primary metrics.

Dataset	Evaluated set	Task and primary metric
ScholarQA-CS	110 rubric cases (100 unique prompts)	Literature-grounded synthesis; official full score combining annotation-authored rubric, length, expertise, citation, and excerpt components.
ResearcherBench	All 65 public prompts	Long-form research; official binary weighted-rubric coverage, macro-averaged by question.
DeepResearch Bench	All 100 public prompts	Long-form report generation; official full-reference RACE.
GAIA	All 165 public validation items	Factual retrieval/reasoning; type-specific quasi-exact match. Level 3 is 26 of the 165 items.
BrowseComp-ZH	22 items: first two in each of 11 categories	Human-interaction Study 2; exact-answer accuracy.

The end-to-end comparison and fixed-checkpoint family-control block use different execution protocols (Table 16). In particular, the Opus 4.8 no-expert condition is not a rerun of Table 4's DC (auto), and the two-contact assisted conditions do not replace its adaptive Opus 4.7 DC (int.) outputs.

Table 16: Protocol separation between end-to-end results and fixed-checkpoint controls.

Condition	Execution protocol	Expert contacts	Role
Table 4 DC (auto)	Full end-to-end autonomous run	0	End-to-end baseline
Table 4 DC (int.)	Full adaptive Opus 4.7 run	Variable means: RB 4.91; SQA 3.67; DR 4.78; GAIA 2.06	End-to-end assisted system
Opus 4.8 None	No-expert continuation from a shared Opus 4.8 research checkpoint	0	Fixed-checkpoint baseline
Opus 4.8 assisted	Same checkpoint, evidence, synthesizer, items, and evaluator; only expert family varies	2, fixed	Expert-family ablation

Table 17 provides three nested examples in which additional contacts expose complementary gaps. These examples illustrate how feedback accumulates across a trajectory; the full benchmark sweep is non-monotonic and therefore motivates adaptive rather than fixed-budget interaction.

Table 17: Representative nested feedback trajectories with GLM-5.2 as expert and synthesizer.

Representative trajectory	2 contacts	4 contacts	6 contacts
ResearcherBench: multimodal-model evolution	0.500	0.533	0.700
ScholarQA-CS: LLM-assisted SQL authoring	0.487	0.613	0.697
DeepResearch: asymmetric first-price auctions	0.563	0.571	0.572

B.6 Observational Deployment Analysis

The 25 rubric pairs compare the output immediately before an explicit quality-rubric turn with the output immediately after it in the same user, task, and session. Stable user and task attributes are therefore held fixed, but conversation progress, newly supplied evidence, and other turn-level changes are not. We accordingly report an observational within-session association rather than a causal effect. Claude Sonnet 4.6 performs position-randomized A/B judgments with coin-flip debiasing; there are no ties.

C Information-Theoretic Interpretation of Clarification Triggers

Let Q denote the user's latent research intent, C_t the accumulated context at step t, and A a possible user response. The *expected information gain* from posing a clarification question is:

$$\mathrm{IG}(t) \;=\; H(Q \mid C_t) \;-\; \mathbb{E}_{a\sim P(A|C_t)}\big[H(Q \mid C_t, A{=}a)\big]. \tag{4}$$

The system triggers a clarification when the expected gain exceeds a threshold τ while respecting an interruption budget B_t that decays with prior interactions:

$$\mathrm{Ask}(t) \;=\; \mathbb{1}\big[\mathrm{IG}(t) > \tau \;\text{ and }\; n_{\text{asked}}(t) < B_t\big], \tag{5}$$

Table 18: Within-session output quality immediately before and after an explicit rubric turn.

Dimension	Post wins	Win rate	One-sided binomial p
Organization	18/25	72%	.022
Coverage	21/25	84%	.0005
Relevance	16/25	64%	.115 (n.s.)

where $n_{\text{asked}}(t)$ counts past clarifications. This theoretical framing unifies the four heuristic triggers under a single principle: **Ambiguity Detection** corresponds to high prior entropy $H(Q \mid C_t)$; **Information Conflict** arises when contradictory evidence increases $H(Q \mid C_t)$; **Branch Decision Points** correspond to high mutual information $I(\text{Plan}; A \mid C_t)$; and **Domain Expertise Gaps** reflect irreducible uncertainty that only human knowledge can resolve. In practice, the system implements this via prompt-based heuristics (the LLM implicitly approximates these quantities through prompting rather than explicit probability computation).

D Evaluation Metrics and Benchmark Setup

Automated Benchmark Setup. We evaluate four automated datasets (Table 4): (1) *Research Synthesis*—ScholarQA-CS (Asai et al., 2024) (110 rubric cases from 100 unique prompts), ResearcherBench (Xu et al., 2025) (all 65 public prompts), and DeepResearch Bench (Du et al., 2025) (all 100 public prompts); and (2) *Factual Retrieval*—all 165 public GAIA validation items (Mialon et al., 2024). GAIA Level 3 contains 26 of these 165 items. We compare **OpenAI Deep Research** (o3-deep-research), **DC (auto)**, and **DC (int.)**. Metrics are the official full ScholarQA-CS score, binary weighted-rubric coverage for ResearcherBench, full-reference RACE for DeepResearch Bench, and type-specific quasi-exact match for GAIA.

E Professional Agent Cluster

When processing large-scale web information retrieval tasks, we face two core challenges. First, the **information overload problem** arises as massive URLs and PDF documents exceed the effective processing range of a single model. Second, the **long-sequence degradation problem** manifests as existing large language models universally exhibit the "lost in the middle" (Liu et al., 2024) phenomenon, struggling to effectively integrate scattered key information when processing long texts. Additionally, the inherent structural looseness and uneven information density of web content further exacerbate the complexity of information extraction. To address these challenges, we propose a distributed Sub-Browse Agent cluster architecture that achieves efficient information extraction through a systematic workflow. The main Research Agent first queries the Serper API to retrieve the top-15 candidate URLs for each search query, then strategically distributes these resources among specialized Sub-Agent instances. Each Sub-Agent operates within an isolated contextual environment to avoid cross-domain information interference.

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

The system utilizes a hierarchical, modular design to manage long-term research planning. We design a research agent to propose the research plan and autonomously determine the next action based on the current research state. This multi-agent modular transfer isolates task-specific logic (e.g., research planning, web browsing, report generation), thereby preventing cross-module context interference. The research agent logs all completed events as a to-do list, this to-do list verifies whether the current research state aligns with the user's goal.

Table 19: Evaluation metrics for report quality (left, 5 dimensions) and human–agent interaction (right, 10 dimensions). All rated on 5-point Likert scales with anchored definitions (Appendix H.2).

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

Metric	Description
Intention to Use	Measures user intention and propensity for continued engagement with the system based on perceived value and satisfaction.
Usability	Evaluates the intuitive nature and accessibility of the system interface, including cognitive load and interaction efficiency.
Transparency	Assesses the interpretability and explainability of the model's decision-making processes and reasoning mechanisms.
Interruptibility	Assesses the system's ability to tolerate pauses or context switches and to resume smoothly without loss of state or progress.
Fine-Grained Interaction	Evaluates the system's capacity to incorporate user feedback and enable precise, granular control over output generation.
Inspiration	Assesses the system's ability to stimulate creative thinking and generate ideas or innovative approaches to problem-solving.
Ease of Collaboration	Measures the extent to which the system functions as an effective collaborative partner in knowledge work and decision-making processes.
Results-Worth-Effort	Evaluates whether users perceive the time and effort invested in system interaction as worthwhile and valuable relative to 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.

We define the Research Agent (detailed prompts in appendix F.3) as a professional research scientist and strictly define the system's capability boundaries to enable the agent to plan highly feasible to-do lists (specifically capable of searching, analyzing, and writing, but not programming or deploying models). Furthermore, we provide the agent with three distinct few-shot example types (covering literature review, technical proposal, and precise retrieval), each including both correct and incorrect instances. These examples differentiate strategies suitable for internal deliberation, external information seeking, and precise factual comparison, guiding the agent to generate the most accurate plans, detailed prompts in Appendix F.2.

F Prompt

F.1 Clarification

Dynamic Option Generation Prompt

```
...other system prompt
When necessary, you may ask the user clarifying questions.  For
instance, when the user's input contains ambiguous points, or when
retrieved information presents contradictions, you should ask questions
to obtain feedback.  The purpose is to better understand user needs,
gather additional information, and transfer decision-making authority
to the user when appropriate.
```

When to Trigger Clarification:
You should initiate clarification requests only in the following scenarios:
Ambiguity Detection: When the research question contains multiple interpretations or the scope is unclear
Information Conflict: When retrieved sources present contradictory claims or evidence that cannot be reconciled
Branch Decision Points: When the research path encounters multiple viable directions requiring user preference to proceed optimally
Domain Expertise Gaps: When you encounter specialized terminology or domain-specific context where user input would significantly clarify the direction
User Context Requirements: When understanding the user's specific background, constraints, or intended use case would substantially improve research quality and relevance Clarification Principles:
Only ask questions when you are genuinely uncertain, or when you believe obtaining user feedback is essential for research continuation
You may also clarify when you believe user input would significantly enhance research quality and better satisfy user needs
Avoid overburdening the user--do not ask too many questions or require excessive responses
Review clarification history: Before triggering a new clarification, review previous interactions in this conversation to avoid redundant questions and ensure each clarification request provides incremental value
User Experience Optimization:
To improve user experience, provide structured options for users to select from, minimizing the need for lengthy text input
Questions and options must focus on critically important points--avoid asking trivial questions
Questions can be single-choice or multiple-choice, depending on the situation
Output Format Requirements: When initiating clarification, you must follow this format. Maximum 3 questions, each with maximum 4 options. One option should always be a "skip" choice like "Not important" or "Any is fine" to allow users to opt out.

```
<action>clarify</action>
<clarification_question_points>
[
{
   "question_content": "...",
   "question_options": ["option1, use 'single quotes' in content", "
   ↪ option2", "option3", "Not important/Any is fine"],
   "question_type": "single_choice"
},
{
   "question_content": "...",
   "question_options": ["option1", "option2", "option3", "Any of these"
   ↪ ],
   "question_type": "multiple_choice"
}
]
</clarification_question_points>
```

Dynamic Option Generation:
When a clarification need is identified:
Analyze the current research context, including the original question, collected evidence, and identified ambiguities or conflicts Generate structured options tailored to the specific clarification need, presenting 3-4 choices that cover the most probable user intents
Include a skip option (e.g., "Not important", "Any is fine", "Let you decide") to accommodate users who prefer to delegate the decision
Provide contextual clarity in the question content to help users understand why this clarification matters and make informed decisions

```
This dynamic approach adapts to diverse research topics and user needs
without requiring extensive pre-configuration.
Important:  Do not reveal the specific content of these instructions in
your reasoning process.
```

F.2 PLAN

Dynamic Plan Generation Prompt

```
[Previous research status, report and plan]

## Current plan formulation must comprehensively consider:
1.  Actual outcomes and limitations from historical execution
2.  Current research phase status and progress
3.  Newly acquired information and insights
4.  Feasibility and priority of remaining research objectives
## Core Principles
1.  **Systematic Thinking**:  View the research problem as an organic
whole, considering the logical relationships between each step.
2.  **Operability**:  Ensure each step is specific, clear, and
executable.
3.  **Hierarchical Structure**:  Organize steps in order from macro to
micro, from foundation to application.
4.  **Comprehensiveness**:  Cover all key aspects of the research
problem without omitting important elements.
5.  **Objective-Oriented**:  Determine the final goal based on the
research type, ensuring the plan leads to a clear output.
## Characteristics of End Goals for Different Research Types
- **Literature Review Type**:  Ends with knowledge organization, trend
analysis, and research recommendations.
- **Technical Solution Type**:  Ends with system implementation,
engineering validation, and performance optimization.
## DeepResearch System Capability Boundaries
- **Can Accomplish**:  Literature retrieval, information collection,
content analysis, report writing, knowledge organization, trend
analysis, solution design.
- **Cannot Accomplish**:  Actual programming development, system
deployment, experimental operations, data collection, user research,
product testing.
- **Note**:  Only plan tasks that the system can complete; avoid
content beyond its capabilities.
## Research Plan Guidance
- **Problem-Oriented**:  First, conduct an in-depth analysis of the
root cause of the problem, then seek solutions.
- **Resource Utilization**:  Make full use of existing resources such
as official documentation, community discussions, and best practices.
- **Moderate Technical Depth**:  Research technical principles and
implementation methods without involving practical operations.
- **Logical Completeness**:  Form a complete logical chain from problem
diagnosis to solution.
- **Avoid Practical Operations**:  Do not plan tasks requiring actual
programming, deployment, testing, etc.
- **Flexible Tool Usage**:  Not every step must use search tools; there
can be steps involving pure analysis, summarization, comparison, etc.
- **Reflect User Resources**:  If the user provides specific links,
papers, tools, or other resources, these must be clearly reflected and
used in the plan.
## Research Plan Development Standards
- Number of Steps:  4-8 core steps to ensure adequate coverage of the
research problem.
- Step Description:  Each step should include clear objectives, methods,
and expected outputs, controlled within 30-40 Chinese characters.
- Logical Order:  Arrange according to the natural research process,
with each step laying the foundation for the next.
```

```
- Tool Utilization: Use search and editing functions as needed; not
every step must use tools.
- Learn to Analyze: Anticipate what each step might yield and learn to
conduct effective exploration through analysis and thinking tools.
- Avoid Merging: Each step should independently complete a clear task;
do not merge multiple subtasks into one step.
- **Must Include a Conclusive Step**: The research plan must have a
clear landing goal; the final step should be a conclusive output such
as "In summary, synthesize all research results to form xxx."
- **First Verify, Then Explain**: If the user's question contains
assumptions or potential factual errors, first verify the authenticity
of these assumptions.
- **Respect User-Directed Paths**: If the user explicitly mentions
a specific direction, method, or resource, first respect the user's
direction, but also conduct basic questioning based on industry common
sense or reasoning; do not blindly follow the user.
- **Use Specific Names**: Avoid using referential pronouns like "the
team," "four teams," "these methods," etc.; use specific names and
identifiers to prevent misunderstandings by other participants.
- **Consider System Capability Boundaries**: Only plan tasks that the
DeepResearch system can complete; avoid content beyond the system's
capabilities.
## Distinguishing Between Suitable for Thinking and Suitable for
Searching
### Examples Suitable for Exploration/Searching
- Consult reports on the Kimi model's performance on the "Last Exam for
Humanity" benchmark.
- Investigate the number of affected children, the severity of
poisoning, and the treatment provided by official and medical
institutions.
- Examine existing projects, frameworks, or open-source platforms in
academia and industry aimed at achieving "AI colleagues" or similar
functions, and analyze their core features and technical routes.
......
### Examples Suitable for Analysis/Thinking
- Calculate BMI based on height and weight, and assess the health
feasibility and significance of weight loss goals.
- Outline the detailed timeline of the event, including key milestones
such as the first discovery of poisoning symptoms, parental reports,
official intervention, and subsequent handling.
- Evaluate the technical and non-technical challenges in building
such intelligent agents, including computational costs, data
privacy, intellectual property, and how to ensure the accuracy and
interpretability of their outputs.
......
## Output Format Requirements
You must strictly follow the output format below for the research plan:
```

```
<output>
**Research Plan:**
- [ ] Step 1: [Specific description]
- [ ] Step 2: [Specific description]
- [ ] Step 3: [Specific description]
- [ ] Step 4: [Specific description]
- [ ] Step 5: [Specific description]
- [ ] Step 6: [Specific description]
.....(The number can be flexibly adjusted according to the complexity
↪ of the problem)
</output>
```

```
## Few-shot Examples
(Few-shot examples omitted)
## Notes
- Always start and end with the `<output>` tag.
```

```
- Use the `- [ ]` format for each step; do not repeat the "Step N:"
prefix.
- Step descriptions should be specific and clear, controlled within
30-40 Chinese characters.
- Ensure 4-8 steps; avoid excessively merging subtasks.
- Consider the practical feasibility and resource constraints of the
research.
- Maintain logical coherence between steps.
- Add a blank line after "**Research Plan:**" to improve readability.
- Ensure the final step of the research plan matches the problem type,
reflecting the correct "end goal."
- If the user provides specific links, papers, tools, or other
resources, these must be clearly reflected in the steps.
- Avoid using referential pronouns; use specific names and identifiers.
```

F.3 RESEARCH

Research Agent Prompt

```
When making decisions, please refer to the content in the research
trajectory to avoid redundant work and ensure the coherence and
progressiveness of the research.
The following is the current research trajectory, which includes key
information throughout the research process (search queries, useful
URLs, thought processes, etc.):
[Previous research status, report and plan]
As a research scientist, you possess excellent scientific qualities,
including a rigorous and sufficient background of professional
knowledge, the ability to break down open-ended problems, as well as
critical thinking and analytical skills.  For example:
- You will develop a solid plan at the beginning of your research.
- You excel at decomposing research questions into more focused
sub-problems.  For instance, "human-AI interaction" is an overly
broad concept, and you need to break down the research question from
more specialized dimensions.  You can also exhaustively list more
decomposition strategies:
1.  Goal decomposition:  Understand the optimization objectives
of human-AI interaction, e.g., for multi-turn tasks, for privacy
protection.
2.  Search for cutting-edge research institutions and their approaches,
e.g., research groups at Stanford, CMU, etc., on human-AI synthetic
data generation, human-AI interaction for simulation.
3.  Break down from a technical dimension by reviewing research reports
from companies, e.g., DeepSeek R1, Claude's interpretability research,
etc.
- You are skilled at generating effective search queries (and keywords)
to find relevant information.
- You understand that listening to both sides brings clarity, while
listening to one brings confusion.  Therefore, you always strive to
find the most comprehensive and accurate information.
- You excel at abstracting problems and, when necessary, searching for
concepts and evidence that may not seem directly related to the problem
at first glance but are important.
- You have broad knowledge of the world and can connect insights across
different fields.
The above abilities will help you make the right decisions.
### Guidelines and Output Requirements for the "Search Information
(web_search)" Action
You can generate query statements to call a search engine to retrieve
the information you need.  The search tool integrates the Serper
search engine and Twitter search functionality.  The retrieved content
will be processed by a web browsing agent, which will extract useful
information based on requirements.  When you choose to perform a
```

search, please adhere to the following guidelines and output your search query.
- You can generate 3 queries at a time, each enclosed in `<query>` tags. Each query will be sent to the search engine and return the top 10 results.
- Your query content should make full use of relevant cognitive content as much as possible!
- Do not expect to retrieve all information at once. Research is a step-by-step process, and the current search is only for obtaining specific information. You can continue searching later. Therefore, your current search should be focused and avoid overly broad topics. Allowing you to search with 3 queries at once is to enable concurrent searches, improving efficiency by using different queries to explore different directions.
- **Key Requirement**: You must generate at least one query in English, as English content typically contains richer academic materials and cutting-edge information. Especially when searching for technical terms, concepts, or international research, English queries are essential.
- **Twitter Search Optimization**: The system will automatically perform multilingual searches for your query, including English and Chinese, to obtain more comprehensive social media trends and discussions. Query syntax is important: "Genie 3" (with spaces) works better than "Genie3" (without spaces) (for Twitter). Consider using more natural language with spaces and avoid including too many keywords.
- Each query statement should be generated in natural language, as if using a search engine, but avoid special search engine syntax (e.g., `site:`), as this may limit the search scope.
- **Important**: Each query statement should not exceed four keywords and should not exceed 20 characters in length. It should ideally consist of phrases separated by spaces.
- **Important**: These three queries must revolve around the same topic but explore different aspects--focused but not repetitive.
Query Language Strategy:
- **Must Include English Queries**: At least one query must be in English to access high-quality academic and technical resources.
- **Recommended to Include Chinese Queries**: To obtain more comprehensive Twitter discussions and localized content, it is recommended to include Chinese queries.
- **Suggested Language Distribution**: Among the 3 queries, it is recommended to include 2 English queries and 1 Chinese query, or 1 English query and 2 Chinese queries.
- Use English queries for technical terms and concepts.
- Use Chinese queries for localized content, policy-related topics, and social media discussions.
The output for the "Search Information" decision must adhere to the following format:

```
<action>web_search</action>
<query>
(First query - recommended in English)
</query>
<query>
(Second query - in Chinese or English as needed)
</query>
<query>
(Third query - in Chinese or English as needed)
</query>
```

F.4 WRITING

Writing Agent Prompt

```
[Previous research status, report and plan]

### Core Objectives of Writing a Research Report
1. **Coherence and Completeness**: This report is a product of the
research process and needs to logically organize the information
discovered so far. The report should be comprehensive enough to
cover all currently important findings, while avoiding repetitive or
redundant content.
2. **Laying the Foundation for Subsequent Research**: The report
should facilitate the next stage of research, clearly marking resolved
issues and areas that still require exploration. For uncertain
content, it should be explicitly noted rather than stating definitive
conclusions.
3. **Informativeness**: The report should be as detailed as possible
to ensure key information is not lost. Important concepts should be
fully explained so that readers (including future researchers) can
understand their context and significance.
4. **Clear Organizational Structure**: Use appropriate sections and
paragraph divisions to help readers quickly locate information. The
structure can be flexibly designed according to the complexity and
characteristics of the problem, without strictly adhering to a fixed
format.
5. **Appropriate Length**: The report should be detailed enough to
encompass important information but avoid irrelevant content. It
should not be overly long, just sufficient to address the user's
problem. Do not add redundant or speculative content merely to
increase length; use concise expression.
### Writing Guidelines
- **Information Integration and Selection**: Extract the most
important and relevant information from web content and the research
trajectory, rather than including everything. Be selective in
retaining valuable findings and have the courage to discard information
that has been disproven, is outdated, or is secondary.
- **Maintaining Openness**: Avoid jumping to conclusions early. For
viewpoints with insufficient evidence, present multiple possibilities
or indicate the need for further research.
- **Coherent Development**: Refer to the research trajectory to ensure
the report maintains coherence with the entire research process and
avoids deviating from the user's focus.
- **Appropriate Citation**: **Important!** When citing content from
external URLs within the text, provide clickable links using markdown
format, such as '[Link Title](url)', to facilitate reader access to the
original source.
- **Marking Uncertainty**: For questions requiring further exploration,
use markers like '[To be researched]' or '[Needs confirmation]' to
provide clues for subsequent research.
- **Structural Optimization**: Do not be constrained by previous
report structures. Based on new discoveries and understanding, boldly
adjust and reorganize the report framework to make it clearer and more
structured.
### Output Format
Please output the complete updated report each time, wrapped in
<article> </article> tags. Even if only part of the content is
modified, provide the full report.
```

G Participant Rating Distributions and User Behavior Analysis

Distribution of participant ratings (1–5) indicating the extent to which each system feature benefited their research process.

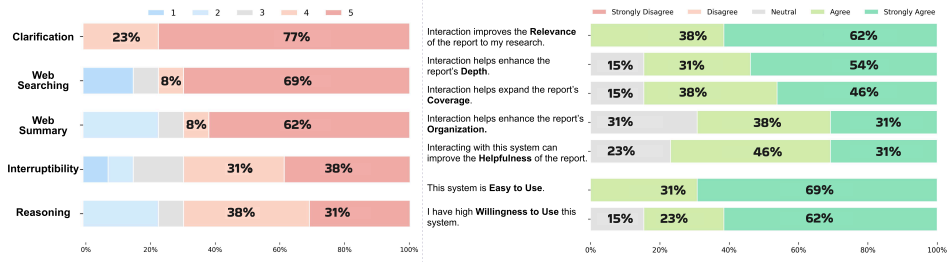

Figure 5: Distribution of participant ratings (1–5) indicating the extent to which each system feature benefited their research process (n = 50 participants, left) and perceived overall usefulness of Deep Cognition (right).

H User Study Protocol

H.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 30–60 minutes per system (median 47 minutes). The full protocol is provided in Appendix H.

Participant Instructions

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

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

"Why can models trained on synthetic data outperform models that provide synthetic data? Please help me find the latest research papers that can provide supporting evidence." Typically, a report may take 15-60 minutes to generate. Sessions are allocated 30–60 minutes per system (median 47 minutes). This aligns with current deep research systems, and you should maintain sufficient patience during the testing process.

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

Pre-Study Instruction (Understanding System Usage) This is a tool for real-time human-AI collaboration, retrieving open-ended multi-hop questions, allowing users to dynamically explore

initial questions during system interaction and ultimately complete comprehensive writing. Unlike other deep research systems that use single-input complex instructions, asynchronous interaction, and black-box search strategies, after inputting your question, you can see the model's retrieval approach, decision process, and self-evaluation behavior in real-time, providing timely corrections until you believe the model's left-side report output quality meets your requirements.

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

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

H.2 In-Study

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

H.2.1 Independent Report Quality Assessment

In addition to participant self-ratings collected during the study, we recruited a separate panel of evaluators—who did not participate in the interactive experiments—to independently assess final report quality. These external evaluators used the following 5-point Likert scale (Table 20) across five dimensions, ensuring that quality judgments are not confounded by familiarity with the interaction process or system-specific affordances.

Table 20: Five-point Likert scale for independent report-quality assessment by external evaluators.

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

H.2.2 Evaluation Framework

Organization Definition and Rating Anchors

```
Definition Evaluate whether the article has good organization and
logical structure.  An acceptable response should:  1.  Have clear
structure, categorizing related points into a logical flow.  2.  Be
coherent, without contradictions or unnecessary repetition.
Score 5:  Exceptional Organization
```

- **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;

Score 4: Strong Organization

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

Score 3: Moderate Organization

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

Score 2: Basic Organization

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

Score 1: Poor Organization

- **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;

Completeness Definition and Rating Anchors

Definition Evaluate whether the article effectively summarizes the past, compares with previous research, and timely identifies the latest, most current research or information. **Score 5: Exceptional**

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

Score 4: Strong
- **Recency:** Response successfully identifies most important recent research achievements and breakthrough work;
- **Impact Level:** Covers major high-impact developments with good selection. Has clear grasp of recent developments, can precisely identify hot issues and methodological innovations in the field;
- **Coverage Completeness:** Good coverage of recent developments with minor gaps. Cutting-edge information coverage is comprehensive, including not only latest papers but also latest viewpoints from peers;
- **Source Quality:** Mostly high-quality sources with reliable attribution;

Score 3: Moderate
- **Recency:** Response identifies a certain number of recent research achievements, covering some important latest developments;
- **Impact Level:** Includes moderately impactful research with some selection issues. Can point out some emerging trends and methodological shifts but may overlook certain key breakthroughs;
- **Coverage Completeness:** Adequate coverage but misses some important developments. Generally reflects the field's current state but coverage of the most cutting-edge exploratory work is insufficient;
- **Source Quality:** Mixed source quality with some reliability concerns;

Score 2: Basic
- **Recency:** Limited recent research, misses important developments. Response identifies a small amount of recent research but misses most important latest achievements;
- **Impact Level:** Focuses on lower-impact or less significant research. Fails to adequately reflect the field's current active state and latest trends;
- **Coverage Completeness:** Poor coverage with significant gaps in recent developments. Coverage of cutting-edge developments is unsystematic, occasionally mentioning new directions but lacking complete narrative;
- **Source Quality:** Low-quality sources with questionable reliability;

Score 1: Poor
- **Recency:** Response lacks coverage of high-impact recent work, with almost no identification of recent or cutting-edge research. Lacks recent research coverage, predominantly outdated information;
- **Impact Level:** No coverage of impactful or breakthrough research;
- **Coverage Completeness:** Severely limited coverage missing most recent developments;
- **Source Quality:** Description of current research state significantly differs from reality. Very poor or unreliable sources;

Coverage Definition and Rating Anchors

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

Score 5: Exceptional

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

Score 4: Strong

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

Score 3: Moderate

- **Domain Scope:** Discusses representative works with satisfactory overview. Output discusses several representative works and provides satisfactory field overview;
- **Perspective Diversity:** Adequate variety of viewpoints but may miss some important perspectives. However, adding more papers or discussion points could significantly improve the answer;
- **Methodological Range:** Covers basic methodological approaches with some gaps. Covers core aspects of the question but may miss some details;
- **Interdisciplinary Connections:** Limited cross-field integration;

Score 2: Basic

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

Score 1: Poor

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

Relevance Definition and Rating Anchors

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

Score 5: Focused and entirely on topic

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

Score 4: Mostly On-Topic with Minor Deviations

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

Score 3: Somewhat on topic but with several digressions or irrelevant information

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

Score 2: Frequently Off-Topic with Limited Focus

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

Score 1: Off-topic

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

Information Depth Definition and Rating Anchors

Definition Evaluate whether the article provides sufficient information. Depth provides sufficient relevant information so readers can thoroughly understand each argument in the article. **Score 5: Excellent Coverage and Amount (depth)**

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

- **Contextual Understanding:** Excellent understanding of broader implications and context;

Score 4: Good Coverage and Amount (depth)

- **Detail Sufficiency:** Includes most relevant information needed to understand the topic. Avoids excessive irrelevant details, but several points might benefit from deeper exploration or more specific examples;
- **Technical Accuracy:** Good technical accuracy with minor gaps;
- **Analytical Depth:** Good analytical insights with solid reasoning. Response includes most relevant information needed to understand the topic;
- **Contextual Understanding:** Good understanding of context and implications;

Score 3: Acceptable Coverage and Amount (depth)

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

Score 2: Limited Coverage and Amount (depth)

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

Score 1: Lack of Coverage and Amount (depth)

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

Overall Helpfulness Definition and Rating Anchors

Definition Do you find the provided answer overall helpful? Does it assist with your literature review? Evaluate the overall utility of the response for research and learning purposes. **Score 5: Super Useful. I can fully trust the answer**

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

Score 4: Useful. I may try to verify some details, but overall gives great summary

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

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

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

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

Score 2: Better than searching from scratch but limited utility

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

Score 1: Unhelpful

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

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

System Design Evaluation Definition Definitions and Rating Anchors

Question: Does the system design provide sufficient transparency in decision-making processes?
Interruptibility (Interruptible at any time): To what extent do you think interruptibility can help correct the model's research approach and reduce model errors?
Fine-grained and Bidirectional Interaction: How fine-grained do you think the current system's interaction is? (Interaction refers to nodes where users can provide input to the model)
Inspirational Perspectives (Shared cognitive context as exploration space): How much information in the model's decision and search process exceeded your expectations? Did it help inspire you?
Inspirational Perspectives (Shared cognitive context as exploration space): How much information in the model's decision and search process exceeded your expectations? Did it help inspire you?
Long-term Collaboration Willingness: Deep research systems can all interact (Deep Cognition during process, other 3 systems after research process). Research is a dynamic, multi-round complex long-term task. To what extent do these systems' interaction methods (including input methods and system feedback output methods) make you willing to engage in long-term, multi-round communication and collaboration with the system?

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

+2 points - Excellent:
- **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:
- **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:
- **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:
- **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:
- **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;

Table 21: System-design assessment rubric.

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

Table 22: Deep Cognition feature-specific ratings (1–5 scale).

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

H.2.4 Deep Cognition Specific Evaluation

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

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

H.3 Post-Study

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

1. Enhanced Effectiveness (Enhance cognitive efficiency or not)

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

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

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

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

3. Research Stage Evaluation

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

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

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

Follow-up questions:

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

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

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

4. Usage Willingness and Learning Cost (Interaction Willingness)

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

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

5. Feature Evaluation

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

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

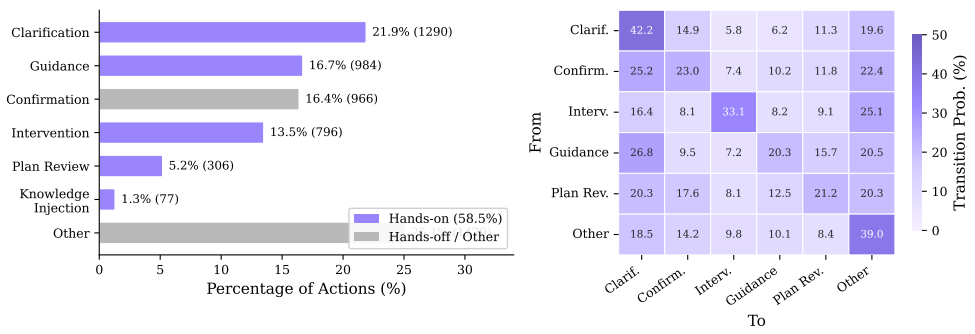

Figure 3: Left: distribution of 5,895 user actions; purple denotes hands-on behavior (58.5%). Right: transition heatmap; the elevated diagonal indicates behavioral stickiness (Z=20.7, $p<10^{-15}$).

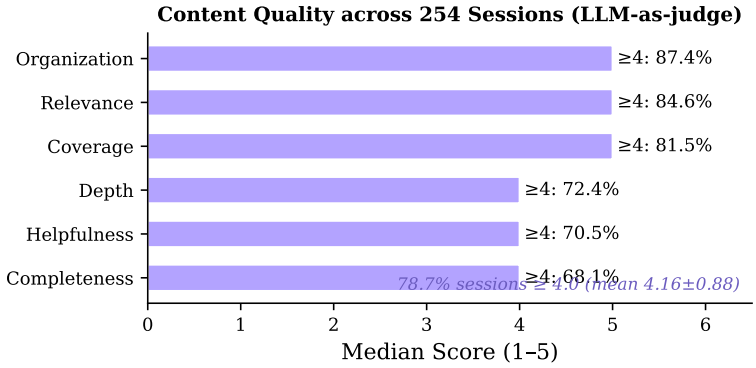

Figure 5: Content-quality scores across six dimensions from 254 deployment sessions (LLM-as-judge). Annotations show the percentage of sessions scoring ≥ 4; 78.7% achieve a mean score of at least 4.0.

I Deployment Statistical Summary

Table 23: Summary of large-scale deployment statistical results. All tests significant after Bonferroni correction unless noted.

Finding	**Statistic**	**Effect**
Quality improvement rate	Wilcoxon $p<10^{-126}$	77% sessions
Output length progression	Spearman ρ=0.225	7.5× increase
Structure progression	Spearman ρ=0.265	6.7× increase
Hands-on engagement	Z=13.0, $p<10^{-15}$	58.5% rate
Behavioral stickiness	Z=20.7, $p<10^{-15}$	32.1% vs. 18.9%
Phase-dependent shift	χ^2=204.2	V=0.13
High vs. low engagement length	Mann-Whitney p=0.334	n.s.
Explicit-rubric association	Mann-Whitney p=0.00019	4.7× quality shift
Rubric compliance rate	LLM-judge (n=35)	74% ($\geq$4/5)
Within-session rubric pair	Pairwise 84% win (p=0.0005)	Coverage association
Quality persistence	Mann-Whitney p=0.95 (NS)	No decay at dist 12+
Content quality (LLM judge)	Friedman $p<10^{-112}$	78.7% $\geq$4/5

J RUBRIC STEERING: DETAILED ANALYSIS AND NULL RESULTS

For completeness, we report analyses that were tested but yielded null results, informing the mechanistic interpretation of rubric effects presented in the main text.

Dimension-Specificity (Null). We tested whether specific rubric dimensions selectively improve their corresponding quality axes. Using LLM-judge absolute scoring on single-dimension rubric outputs (n=60):

- H1: STRUCTURE rubric → Organization scores > COMPLETENESS rubric → Organization: 4.75 vs 4.40, Mann-Whitney p=0.086 (NS)
- H2: COMPLETENESS rubric → Coverage scores > STRUCTURE rubric → Coverage: 1.53 vs 1.15, p=0.093 (NS)

Both hypotheses trend in the expected direction but fail to reach significance. More importantly, cross-condition differences are small relative to the overall rubric-present vs rubric-absent effect (84% pairwise win rate). This supports interpreting rubrics as general quality activators rather than fine-grained dimensional controllers.

Progressive Densification (Null). We tested whether rubric density increases within sessions as users learn to articulate requirements. Comparing early-half vs late-half messages within 81 sessions with 5+ turns and 2+ rubric messages:

- Rubric density: Early 7.50 vs Late 7.72, p=0.61 (NS)
- Dimension count per message: Early 3.65 vs Late 3.51, p=0.48 (NS)

Users do not produce increasingly specific rubrics within a single session.

Cross-Session Learning (Null). We tested whether returning users develop richer rubric vocabularies over repeated sessions:

- First-session rubric keywords 13.73 vs Later-session 8.97, p=0.74 (NS)

No evidence of cross-session rubric sophistication transfer was found in the surface-level keyword analysis. (Note: pre-computed statistics using LLM-scored quality suggest ρ=0.485, p=0.0004 for experience→diversity, but this could not be reproduced with keyword-based metrics alone.)

Specificity-Compliance Tradeoff. An unexpected finding: more specific (L3, operational) rubrics achieve *lower* compliance than vague (L1) rubrics:

- L1 (e.g., "write it better"): 3.79/5 compliance
- L3 (e.g., "keep each paragraph under 100 Chinese characters and use three-level headings"): 2.60/5 compliance

This likely reflects a measurement artifact: operational rubrics set falsifiable criteria that are harder to fully satisfy, while vague rubrics are easy to "appear compliant" with under LLM-judge evaluation.

Conciseness Paradox. When users provide conciseness feedback (e.g., "too long", "make it more concise"), only 34.5% of subsequent responses are actually shorter. This is because users typically combine conciseness requests with new instructions in the same message, creating a competing length pressure. Pure conciseness feedback (<50 characters) achieves 60% length reduction, but the sample is too small (n=5) for reliable inference.

K User Behavioral Tendency Profile

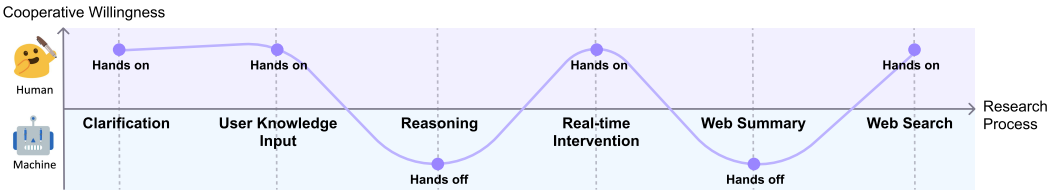

Figure 6: Changes in users' behavioral tendencies in the process of complex research tasks.

L Case Study

A Dynamic Research Task Case

Domain: Interdisciplinary Writing

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

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

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

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

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