IDRBench: Interactive Deep Research Benchmark

Anonymous ACL submission

Abstract

Deep research agents powered by Large Language Models (LLMs) can perform multi-step reasoning, web exploration, and long-form report generation. However, most existing systems operate in an *autonomous* manner, assuming fully specified user intent and evaluating only final outputs. In practice, research goals are often underspecified and evolve during exploration, making sustained interaction essential for robust alignment. Despite its importance, interaction remains largely invisible to existing deep research benchmarks, which neither model dynamic user feedback nor quantify its costs. We introduce **IDRBench**, the first benchmark for systematically evaluating *interactive* deep research. IDRBench combines a modular multi-agent research framework with on-demand interaction, a scalable reference-grounded user simulator, and an interaction-aware evaluation suite that jointly measures interaction benefits (quality and alignment) and costs (turns and tokens). Experiments across seven state-of-the-art LLMs show that interaction consistently improves research quality and robustness, often outweighing differences in model capacity, while revealing substantial trade-offs in interaction efficiency. The source code is available at https://anonymous.4open.science/r/IDRBench-F650.

1 Introduction

Large Language Models (LLMs) have revolutionized information seeking, evolving from single-turn question answering to deep research agents that perform autonomous multi-step reasoning, web navigation, and long-form report generation (Zheng et al., 2024; Li et al., 2025; Zheng et al., 2025; Guo et al., 2025; Yun and Jang, 2025). Unlike traditional Retrieval-Augmented Generation (RAG) systems (Gao et al., 2023; Wang et al., 2024), which typically address isolated queries, deep research agents operate through iterative cycles of planning, searching, and synthesis to satisfy open-ended user needs (Wei et al., 2025; Du et al., 2025).

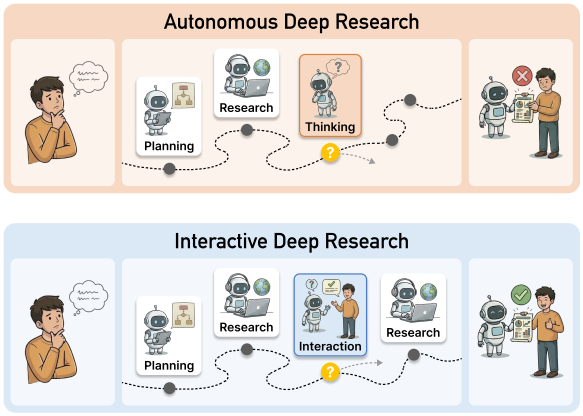

Figure 1: Comparison of autonomous and interactive deep research. Autonomous agents execute independently and may diverge from user intent, while interactive agents incorporate feedback to maintain alignment.

Despite these advances, deep research remains **largely autonomous**: users provide an initial query, after which agents independently control the entire research trajectory (Li et al., 2025; Zheng et al., 2025). This design is brittle in practice: real-world queries are often underspecified or ambiguous (Rahmani et al., 2023; Zhang et al., 2025), and as reasoning unfolds over long horizons, agents face repeated high-stakes decisions. Without mechanisms for sustained user alignment, agents risk hallucinating intent or drifting toward irrelevant directions. While recent work and deployed systems (e.g., GPT and Gemini) attempt pre-execution clarification (Zhang et al., 2024b, 2025, 2024a), they largely fail to address uncertainties that emerge during exploration of complex topics.

We argue that deep research should transition from a solitary process to an **interactive deep research** paradigm, where the agent acts as a collaborative partner that communicates progress, solicits guidance, and iteratively refines its direction. How-

ever, effective interaction is non-trivial. Agents must decide *when* to ask questions, *what* to ask, and *how often*, balancing information gain against interruption cost and cognitive burden. Interaction thus introduces an inherent trade-off between alignment benefits and operational overhead.

Despite its importance, **interaction remains largely invisible to existing evaluation benchmarks** (Wu et al., 2025; Shao et al., 2024; Du et al., 2025). Current benchmarks rely on static `(Query, Reference Document)` pairs and evaluate only final outputs, ignoring the intermediate decision process. This limitation has two consequences. First, static settings lack dynamic feedback, even though adaptability to evolving information is crucial for real-world robustness (Yao et al., 2025). Second, they obscure communicative competence: an agent that reaches a correct answer by chance is indistinguishable from one that verifies and corrects its reasoning through interaction.

To bridge this gap, we introduce **IDRBench**, the first **I**nteractive **D**eep **R**esearch **Bench**mark designed to evaluate the interactive capabilities of deep research agents systematically. IDRBench assesses not only *what* agents produce, but *how* they adapt, communicate, and align through interaction. Our contributions are threefold:

- **Interactive Deep Research Framework.** We propose a modular, multi-agent pipeline augmented with an explicit interaction mechanism that enables dynamic clarification and alignment throughout the research lifecycle.
- **Scalable User Simulation.** We develop a reference-grounded User Simulator that provides realistic, goal-oriented feedback, enabling large-scale evaluation without costly human annotation.
- **Interaction-Aware Evaluation.** We introduce a comprehensive evaluation suite that jointly measures Interaction Benefits (quality, coverage, and intent alignment) and Interaction Costs (turns and tokens). Experiments across seven state-of-the-art LLMs show consistent gains from interaction while revealing critical trade-offs in efficiency and robustness.

2 Related Work

Deep Research Frameworks. Recent deep research systems enable LLMs to generate long-form, citation-grounded reports through multi-step reasoning and external tool use (Shao et al., 2024; Coelho et al., 2025; Guo et al., 2025; Zhou et al., 2024; Zhao et al., 2024). Two dominant paradigms have emerged: multi-agent frameworks that decompose research into specialized roles (Zheng et al., 2024; Alzubi et al., 2025; Li et al., 2025), and end-to-end agentic models trained with reinforcement learning (Jin et al., 2025; Zheng et al., 2025). Despite strong performance, these approaches largely operate in *autonomous* settings without user interaction, making them prone to compounding misalignment over long reasoning horizons.

Deep Research Benchmarks. Several benchmarks have been proposed to evaluate research-oriented generation, focusing on retrieval quality (Wei et al., 2025; Zhou et al., 2025), long-form writing (Bai et al., 2025; Wu et al., 2025), or combined article generation and citation accuracy (Shao et al., 2024). DeepResearch Bench (Du et al., 2025) further advances this direction by providing a comprehensive evaluation of report quality across diverse domains. However, these benchmarks assess only final outputs and do *not* capture the dynamics of human-agent interaction during research.

Interactive Agents. To address underspecified queries, prior work studies clarification questions and conversational search (Rahmani et al., 2023; Tavakoli et al., 2022; Feng et al., 2023; Aliannejadi et al., 2021). More recent LLM-based approaches introduce explicit clarification mechanisms (Zhang et al., 2024b, 2025, 2024a), but focus primarily on pre-execution interaction. Interaction-Driven Browsing (Yun and Jang, 2025) enables iterative feedback during exploration, yet lacks a unified evaluation framework. Most closely related, STEER (Anonymous, 2025) integrates clarification into deep research but evaluates only output quality, leaving the cost-benefit trade-offs of interaction *largely unexplored*. In contrast, our work jointly assesses both the benefits and costs of interaction, enabling a more complete evaluation of human-AI collaboration in deep research.

3 IDRBench

We present **IDRBench**, an **I**nteractive **D**eep **R**esearch **Bench**mark for evaluating whether Large Language Models (LLMs) can move beyond autonomous generation toward *collaborative, human-aligned* research workflows (Figure 2). Unlike prior benchmarks that assess only final outputs, IDRBench evaluates how models reason, adapt, and refine their trajectories through interaction.

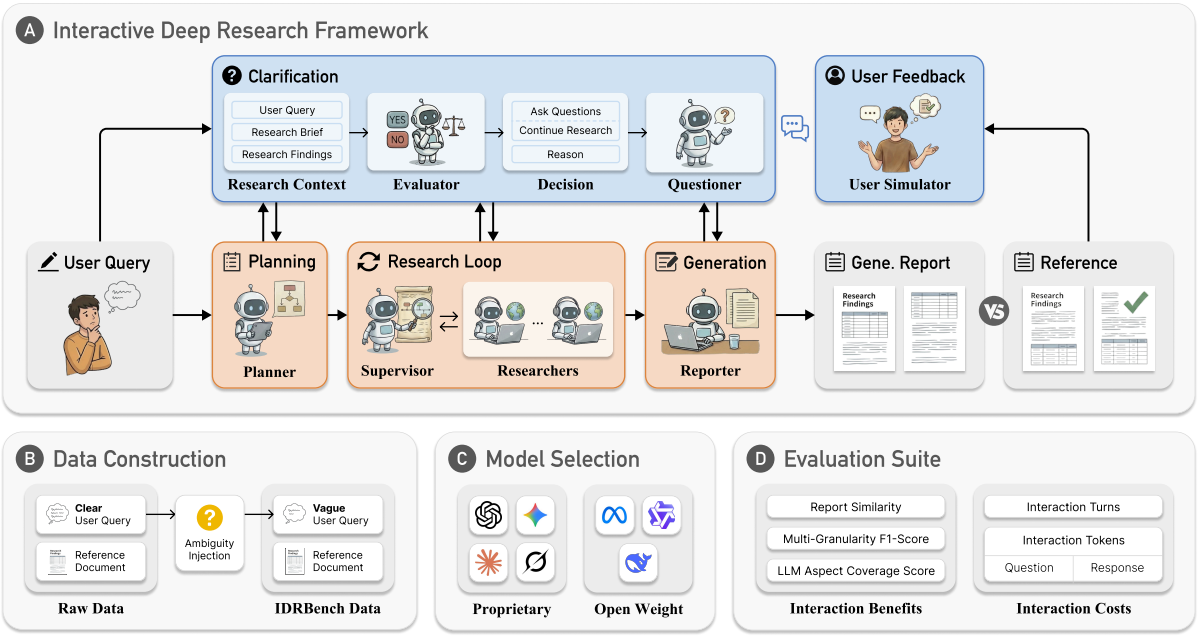

Figure 2: Overview of **IDRBench**. The benchmark integrates an interactive deep research framework with curated data construction, representative LLMs, and interaction-aware evaluation. It features a multi-agent pipeline for **Planning**, **Research Loop**, and **Generation**, augmented with an interaction mechanism for **Clarification** and **User Feedback**, and enables systematic evaluation of both interaction benefits and interaction costs.

3.1 Interactive Deep Research Framework

3.1.1 Basic Architecture

Our framework builds on the `langchain-ai` open deep research architecture (LangChain-AI, 2025), which decomposes complex, multi-step information-seeking tasks into modular stages: **Planning**, **Research Loop**, and **Generation**. This modularity is essential for interaction-aware workflows, as different stages exhibit distinct uncertainties and cognitive demands. The architecture consists of four coordinated agents:

Planner. The Planner translates the user's natural-language query into a structured research brief that specifies scope, objectives, and key dimensions. This brief acts as a shared *north star*, guiding all downstream components.

Supervisor. The Supervisor acts as the executive controller, decomposing the brief into parallelizable sub-tasks and assigning them to Researchers. It monitors progress, reasons over intermediate results, and dynamically adjusts or terminates execution once sufficient coverage is reached.

Researcher. Each Researcher focuses on a specific subtopic, performing autonomous web exploration and retrieval. It iteratively gathers evidence, reflects on coverage gaps, and distills relevant findings into structured summaries, enabling scalable and focused exploration.

Reporter. The Reporter synthesizes intermediate outputs into a coherent final report. Beyond aggregation, it performs content selection, thematic organization, and linguistic refinement to produce a well-structured and self-contained narrative.

3.1.2 Interaction Mechanism

To bridge autonomous execution and evolving user intent, we introduce an **interaction mechanism** embedded at key decision points of the basic architecture, allowing our framework to pause execution and solicit guidance when uncertainty arises.

It consists of two coordinated modules: (1) **Clarification**, which contains an *Evaluator* and a *Questioner* to determine when and how to ask questions; and (2) **User Feedback**, which employs a *User Simulator* to provide guidance. Together, these components dynamically steer the research trajectory toward closer alignment with user intent. See Appendix C for prompt designs.

Evaluator. The Evaluator determines whether interaction is necessary based on the current research context. It balances two competing factors: (i) the benefit of resolving ambiguity and (ii) interruption burden in latency and cognitive load. Instead of binary decisions, it produces a rationale based on the ambiguity of the research topic, task completeness, and remaining interaction budget.

Questioner. When interaction is triggered, the

Questioner formulates targeted inquiries guided by the Evaluator's rationale. It first summarizes the current research state, then asks 1–2 focused questions concerning direction, scope, or emphasis. To preserve natural interaction, the Questioner adapts its tone to the user's original language style.

User Simulator. The User Simulator enables scalable evaluation without human intervention by acting as a proxy for user feedback. This simulator treats the reference document as oracle knowledge and generates responses under three guiding constraints: (i) *Human-like Behavior* (concise, first-person responses), (ii) *Macroscopic Guidance:* (high-level goals over fine-grained facts), and (iii) *Corrective Behavior* (rejecting misaligned options and redirecting focus. Since it is decoupled from the framework, this component can evaluate arbitrary interactive research systems.

3.2 Data Construction

IDRBench is built upon **DeepResearch Bench** (Du et al., 2025), which comprises 100 high-quality (Query, Reference Document) pairs spanning diverse domains such as science, law, and the humanities. This scale strikes a balance between domain coverage, statistical reliability, and the computational cost of multi-step agent execution.

However, these queries are often highly detailed (up to ∼800 tokens), providing near-complete task specifications that reduce the need for interaction. To better reflect real-world underspecification, we introduce an **Ambiguity Injection** process. Specifically, we compress each query by 10%–90% using LLM-based summarization, intentionally removing detail while preserving core intent (examples in Appendix B). This encourages agents to actively resolve uncertainty through interaction rather than passively executing a fully specified prompt.

3.3 Model Selection

To evaluate interactive reasoning across diverse modeling paradigms, we select a set of representative proprietary and open-weight LLMs. Specifically, we evaluate four proprietary models: **GPT-5.1** (OpenAI, 2025), **Gemini-2.5-Pro** (Comanici et al., 2025), **Claude-Sonnet-4.5** (Anthropic, 2025), and **Grok-4.1-Fast** (xAI, 2025), which represent leading commercial systems optimized for long-context reasoning and tool use. We also include three open-weight models: **Qwen3-235B** (Yang et al., 2025), **Llama-4-Maverick** (Meta AI, 2025), and **DeepSeek-V3.2** (Liu et al., 2025), to assess how interaction benefits transfer to openly accessible models with different scaling and alignment characteristics.

Although **Gemini-3-Pro** (Google DeepMind, 2025) is more recent, it shows unstable adherence to structured outputs under the LangChain framework, frequently disrupting long-horizon execution. We thus adopt **Gemini-2.5-Pro**, which exhibits more reliable structured prompting and tool invocation under identical settings.

3.4 Evaluation Suite

We design an evaluation suite capturing both output quality and interaction efficiency (see Appendix A for configurations). Unlike prior benchmarks that focus solely on final outputs (Du et al., 2025; Wu et al., 2025), our evaluation measures how interaction improves alignment with user intent and the cost incurred. We decompose evaluation into two complementary dimensions: **Interaction Benefits**, capturing quality gains, and **Interaction Costs**, measuring human-AI collaboration overhead.

3.4.1 Interaction Benefits

We evaluate interaction benefits along three orthogonal axes: document-level semantic alignment, multi-granularity structural coverage, and intent-level coverage with respect to user goals.

Report Similarity. Let $\boldsymbol{e}(\cdot) \in \mathbb{R}^d$ denote a text embedding. We measure global semantic alignment between the generated report D^{gen} and reference D^{ref} using normalized cosine similarity:

$$\text{sim}(D^{\text{ref}}, D^{\text{gen}}) = \frac{1+\cos(\boldsymbol{e}(D^{\text{ref}}),\boldsymbol{e}(D^{\text{gen}}))}{2}. \quad (1)$$

This captures whether interaction improves semantic consistency beyond surface overlap.

Multi-Granularity F1-Score. To assess structural coverage, we compute F1-scores at **sentence**, **paragraph**, and **chunk**-level granularities. For chunk-level evaluation, documents are segmented into overlapping chunks (300 tokens, 50 overlap). Let $\mathcal{U}^{\text{ref}} = \{\boldsymbol{u}_k\}_{k=1}^{K}$ and $\mathcal{U}^{\text{gen}} = \{\boldsymbol{v}_i\}_{i=1}^{N}$. Recall ($R$) and Precision ($P$) are defined as:

$$R = \frac{1}{K}\sum_{k=1}^{K} \mathbf{1}\big[\max_i \text{sim}(\boldsymbol{u}_k, \boldsymbol{v}_i) \geq \tau\big], \quad (2)$$

$$P = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\big[\max_k \text{sim}(\boldsymbol{v}_i, \boldsymbol{u}_k) \geq \tau\big], \quad (3)$$

with $\tau = 0.8$. The harmonic mean F1-Score captures both omission (low recall) and redundancy or hallucination (low precision).

LLM Aspect Coverage Score (LLM-ACS). LLM-ACS evaluates how well a generated report satisfies the user's intent. Given a query q, we first generate $M \in [8, 20]$ intent aspects $\{a_j\}$, each representing a required informational facet. For each aspect a_j, an LLM assigns coverage scores g_j^{ref} and g_j^{gen} (0–5) to the reference D^{ref} and generated report D^{gen}, respectively. The final score is computed as:

$$\text{LLM-ACS} = \frac{1}{M} \sum_{j=1}^{M} \text{clip}\left(\frac{g_j^{\text{gen}}}{g_j^{\text{ref}} + \epsilon}, 0, 1\right), \quad (4)$$

where $\epsilon = 10^{-9}$. This normalization accounts for query ambiguity and reflects how well the generated report fulfills the intended information needs.

3.4.2 Interaction Costs

Beyond output quality, effective interaction must balance benefit against human effort. We quantify interaction costs along two dimensions: interaction turns and interaction tokens.

Interaction Turns. Interaction turns measure how often the system pauses to solicit user input. While additional turns may improve alignment, they also increase latency and cognitive load. To ensure comparability, we cap interactions at one turn during planning, three during the research loop, and one during generation, enforcing realistic yet flexible interaction budgets.

Interaction Tokens. We further assess the volume of information exchanged during interaction along two dimensions: **question tokens**, representing tokens exposed to the user, and **response tokens**, representing tokens written by the user. Rather than assuming shorter context is always preferable, we treat token usage as a trade-off between informativeness and cognitive cost, reflecting the balance between guidance quality and user effort.

Summary. Together, these metrics provide a holistic view of interactive deep research, enabling principled comparison of interaction strategies and model behaviors under realistic constraints.

4 Experiments

4.1 Experimental Setup

We compare the standard autonomous setting with our interactive framework under a controlled experimental setup. Following the Open Deep Research project (LangChain-AI, 2025), we adopt a tiered model strategy to balance performance and cost. In the experiments, we assign seven LLMs (discussed in Section 3.3) to all core agent roles: Planner, Supervisor, Researcher, Reporter, as well as the Evaluator and Questioner. For high-frequency utility operations (e.g., web page summarization), we use lightweight models (e.g., GPT-4.1-nano) to reduce overhead without affecting interaction behavior.

To isolate the effect of interaction strategies, we standardize the User Simulator as GPT-5.1 across all experiments, ensuring consistent feedback and attributing performance differences solely to the evaluated models. Information retrieval is handled via the Tavily API,[1], and all other hyperparameters (see Appendix A) follow the default Open Deep Research configuration (LangChain-AI, 2025).

4.2 Interaction Benefits

Table 1 summarizes the effect of interaction on report quality across models.

Universal Gains. Interaction consistently improves performance for all models and metrics, demonstrating that LLMs can effectively incorporate feedback to better align with user intent. Notably, interaction can outweigh intrinsic model capacity. For instance, DeepSeek-V3.2 (avg. 73.35) surpasses GPT-5.1's autonomous performance (75.59) once interaction is enabled. Similarly, although Gemini-2.5-Pro starts below GPT-5.1 (73.45 vs. 75.59), it ultimately exceeds GPT-5.1 even in the interactive setting (79.89 vs. 78.97). These results indicate that interactive capability is as critical as raw autonomous strength in collaborative research workflows.

Diminishing Returns. We observe an inverse relationship between model capacity and interaction gains: lower-capacity models (e.g., Llama-4-Maverick, Grok-4.1-Fast) gain substantially (+10.96, +7.97), while top-tier models (e.g., GPT-5.1, Claude-Sonnet-4.5) show smaller improvements (+3.38, +4.96). This suggests diminishing marginal returns for stronger models and highlights interaction quality as a key bottleneck.

Granularity Shift. The nature of interaction gains varies with model capability. For weaker models, interaction primarily improves coarse-grained alignment: Llama-4-Maverick shows large gains in Chunk F1-Score (+13.53) and LLM-ACS (+13.47), exceeding its improvement in Sentence F1-Score (+6.21). In contrast, strong models benefit more at finer granularity: Claude-Sonnet-4.5 gains more in Sentence F1-Score (+7.94) than in Chunk F1-Score

[1] https://www.tavily.com/

Model	Mode	Report Similarity ↑	Multi-Granularity F1-Score ↑			LLM-ACS ↑	Average Score ↑	Est. API Cost ($/Report) ↓
			Sentence	Paragraph	Chunk			
GPT-5.1	**Autonomous**	84.92	46.05	69.07	82.30	95.61	75.59	0.473
	Interactive	87.54	50.44	71.99	88.08	96.79	78.97	0.586
	Difference	**+2.62**	**+4.39**	**+2.92**	**+5.78**	**+1.18**	**+3.38**	**+0.113**
Gemini-2.5-Pro	**Autonomous**	85.00	38.36	76.62	80.92	86.37	73.45	0.393
	Interactive	88.88	46.60	**82.15**	**89.21**	92.60	79.89	0.752
	Difference	**+3.88**	**+8.24**	**+5.53**	**+8.29**	**+6.23**	**+6.43**	**+0.359**
Claude-Sonnet-4.5	**Autonomous**	85.96	44.98	69.20	81.52	95.88	75.51	0.987
	Interactive	**89.15**	**52.92**	74.20	88.06	**98.00**	**80.47**	2.220
	Difference	**+3.19**	**+7.94**	**+5.00**	**+6.54**	**+2.12**	**+4.96**	**+1.233**
Grok-4.1-Fast	**Autonomous**	81.28	30.76	65.33	72.93	87.44	67.55	0.192
	Interactive	86.68	38.63	76.47	83.24	92.56	75.52	0.275
	Difference	**+5.40**	**+7.87**	**+11.14**	**+10.31**	**+5.12**	**+7.97**	**+0.083**
Llama-4-Maverick	**Autonomous**	76.06	18.44	64.72	61.78	53.06	54.81	**0.021**
	Interactive	83.93	24.65	78.46	75.31	66.53	65.78	0.026
	Difference	**+7.87**	**+6.21**	**+13.74**	**+13.53**	**+13.47**	**+10.96**	**+0.005**
Qwen3-235B	**Autonomous**	79.76	28.19	61.03	69.00	81.84	63.96	0.139
	Interactive	82.83	32.81	65.14	75.89	91.70	69.67	0.133
	Difference	**+3.07**	**+4.62**	**+4.11**	**+6.89**	**+9.86**	**+5.71**	**-0.006**
DeepSeek-V3.2	**Autonomous**	84.32	37.94	73.65	80.73	90.09	73.35	0.146
	Interactive	88.11	44.93	79.47	87.13	93.54	78.64	0.185
	Difference	**+3.79**	**+6.99**	**+5.82**	**+6.40**	**+3.45**	**+5.29**	**+0.039**

Table 1: Interaction Benefits results. **Black bold** and underlined denote the best and second-best results. Gains in quality metrics and API cost changes are reported.

(+6.54) or LLM-ACS (+2.12). Thus, interaction evolves from establishing global coverage to refining local details as model capability increases.

Estimated API Cost. We estimate the average API cost per report to assess the economic implications of interaction (the last column of Table 1). While absolute costs vary due to stochastic execution, model verbosity, and tiered pricing, the relative difference between autonomous and interactive modes reliably reflects interaction overhead and its downstream effects on reasoning and search.

Overall, interaction increases cost, with Claude-Sonnet-4.5 and Gemini-2.5-Pro incurring substantial overhead, often comparable to or exceeding their autonomous baselines. In contrast, open-weight models like Llama-4-Maverick and Qwen3-235B exhibit negligible cost increases. Notably, Qwen3-235B even achieves a slight cost reduction ($-$\$0.006), suggesting that interaction can streamline reasoning and search. DeepSeek-V3.2 emerges as the most cost-effective trade-off, delivering strong performance gains with minimal marginal cost ($+$\$0.039), roughly $1/30$ of Claude's interaction overhead.

Robustness. Figure 3 shows that interaction enhances model robustness and suppresses extreme failures. For strong models like GPT-5.1 and Gemini-2.5-Pro, interaction mainly raises the performance floor, while for weaker models such as Llama-4-Maverick and Qwen3-235B, it shifts the entire distribution upward. Overall, interaction improves both accuracy and reliability.

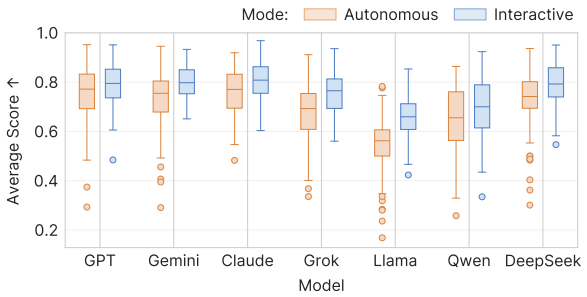

Figure 3: Distribution of average scores across seven LLMs, showing stability gains from interaction.

4.3 Interaction Costs

We next analyze the cost of interaction in terms of **interaction turns** and **interaction tokens**, quantifying the trade-off between alignment gains and human effort. The results are depicted in Table 2.

4.3.1 Interaction Turns

Interaction frequency varies systematically across research stages. During **Planning**, all models frequently seek clarification (0.72∼1.00 turns), correctly identifying initial task specification as highly uncertain. Differences emerge in the **Research Loop**. Models such as Llama-4-Maverick, Qwen3-235B, and Gemini-2.5-Pro interact fre-

Model	Interaction Turns				Interaction Tokens	
	Planning	Research Loop	Generation	Total	Question	Response
GPT-5.1	0.81 / 1	0.43 / 3	0.08 / 1	1.32 / 5	253.96	119.85
Gemini-2.5-Pro	1.00 / 1	1.59 / 3	0.82 / 1	3.41 / 5	184.64	102.52
Claude-Sonnet-4.5	0.94 / 1	0.75 / 3	0.24 / 1	1.93 / 5	261.71	114.76
Grok-4.1-Fast	0.72 / 1	0.29 / 3	0.07 / 1	1.08 / 5	151.91	125.57
Llama-4-Maverick	1.00 / 1	2.84 / 3	0.78 / 1	4.62 / 5	139.86	126.78
Qwen3-235B	0.96 / 1	1.88 / 3	0.26 / 1	3.10 / 5	206.79	116.06
DeepSeek-V3.2	0.75 / 1	1.78 / 3	0.17 / 1	2.70 / 5	252.70	111.21

Table 2: Interaction Costs results. Interaction turns across research stages and interaction token usage are reported.

Research Model	User Simulator Model	Report Similarity ↑	Multi-Granularity F1-Score ↑			LLM-ACS ↑	Average Score ↑
			Sentence	Paragraph	Chunk		
GPT-5.1	**GPT-5.1**	87.17	46.25	69.31	85.20	96.88	76.96
	Gemini-2.5-Pro	87.14	46.60	69.57	85.76	96.70	77.15
	Claude-Sonnet-4.5	86.60	45.96	69.94	86.03	97.04	77.11
Grok-4.1-Fast	**GPT-5.1**	85.61	33.74	73.63	78.67	92.78	72.89
	Gemini-2.5-Pro	84.67	33.11	72.30	79.32	92.16	72.31
	Claude-Sonnet-4.5	85.85	33.92	73.46	80.82	92.31	73.27
DeepSeek-V3.2	**GPT-5.1**	87.47	42.06	77.91	84.73	94.07	77.25
	Gemini-2.5-Pro	86.45	38.17	77.42	85.26	93.40	76.14
	Claude-Sonnet-4.5	87.22	38.82	76.68	86.24	93.88	76.57

Table 3: Results with different User Simulator models, showing stable evaluation metrics across simulators.

quently (1.59∼2.84 turns), favoring continuous realignment. In contrast, GPT-5.1, Claude-Sonnet-4.5, and Grok-4.1-Fast rely more on autonomous reasoning (0.29∼0.75 turns). Despite minimal interaction, Grok-4.1-Fast achieves strong gains, demonstrating high interaction efficiency–the ability to extract maximal benefit from sparse feedback. In the **Generation** stage, interaction is rare for most models (< 0.3 turns), indicating that uncertainty is largely resolved before report synthesis.

4.3.2 Interaction Tokens

Given linguistic differences in token density, we restrict our analysis to the English query-oriented deep research process. Models differ markedly in **Question Tokens**, reflecting distinct communication styles. For instance, Claude-Sonnet-4.5 and GPT-5.1 pose long, context-rich questions (> 250 tokens), whereas Llama-4-Maverick and Grok-4.1-Fast favor brevity (140∼152 tokens). An inverse trend appears between frequency and length: models with frequent interaction (e.g., Gemini-2.5-Pro) ask shorter questions (≈ 185 tokens), suggesting a strategy of focused, incremental clarification. Grok exemplifies a "few-and-short" pattern while maintaining strong performance, achieving an effective balance between alignment and cognitive load.

In contrast, **Response Tokens** from User Simulator remain stable (around 102∼127 tokens) across all settings. This confirms that performance differences stem from how agents utilize feedback rather than how much feedback they receive.

4.4 Parameter Study

We conduct a parameter study to analyze two factors central to IDRBench's design: (i) robustness to the choice of the User Simulator, and (ii) sensitivity to *when* interaction is introduced across the interactive deep research phases. For efficiency, we randomly sampled 30 instances, which suffices to reveal stable trends.

Impact of User Simulator Model. To evaluate robustness, we pair three representative research agents (GPT-5.1, Grok-4.1-Fast, and DeepSeek-V3.2) with three strong LLMs acting as User Simulators: GPT-5.1, Gemini-2.5-Pro, and Claude-Sonnet-4.5. As shown in Table 3, performance remains largely invariant to the simulator choice for each research agent, while inter-model performance gaps are consistently preserved. This indicates that the User Simulator provides *stable and standardized* feedback, and that IDRBench primarily measures the *intrinsic interactive capability* of research agents rather than artifacts of simulator selection. Overall, these results validate the robustness and low variance of our evaluation protocol.

Impact of Interaction Timing. We analyze how

Research Model	Interactive Module	Report Similarity ↑	F1-Score ↑			LLM-ACS ↑	Average Score ↑
			Sentence	Paragraph	Chunk		
Gemini-2.5-Pro	**None**	85.26	37.37	77.29	82.02	86.60	73.71
	Planning	<u>87.15</u>	<u>38.78</u>	<u>79.63</u>	<u>84.69</u>	88.63	<u>75.78</u>
	Research Loop	86.55	37.68	79.48	83.51	89.11	75.27
	Generation	86.08	36.82	76.35	82.73	<u>89.32</u>	74.26
	All	**88.13**	**42.45**	**81.15**	**87.60**	**92.64**	**78.39**
Llama-4-Maverick	**None**	76.18	15.34	61.99	59.00	54.93	53.49
	Planning	<u>81.71</u>	**21.62**	<u>72.77</u>	**71.16**	<u>64.28</u>	<u>62.31</u>
	Research Loop	79.18	17.29	68.67	63.57	59.74	57.69
	Generation	77.45	16.13	62.43	60.05	59.83	55.18
	All	**83.33**	<u>21.57</u>	**73.90**	<u>70.29</u>	**67.08**	**63.23**

Table 4: Results with interaction enabled in different modules. **Bold** and <u>underlined</u> denote the best and second-best results. The table compares module-specific interaction with full-lifecycle interaction.

Scenario	Recommendation	Rationale
Performance Ceiling	**Claude-Sonnet-4.5**	Achieves the highest performance ceiling and superior robustness.
Interaction Intensity	**Gemini-2.5-Pro**	Yields large alignment gains through active, frequent clarification.
Efficiency	**Grok-4.1-Fast**	Demonstrates high interaction efficiency by leveraging sparse feedback.
Cost Constraints	**DeepSeek-V3.2**	Open-weight model delivering proprietary-tier performance at minimal cost.

Table 5: Scenario-based recommendations for selecting LLMs in interactive deep research.

interaction timing affects performance, focusing on Gemini-2.5-Pro and Llama-4-Maverick, which show proactive interaction behavior and large gains. Beyond the autonomous (None) and fully interactive (All) settings, we introduce three restricted modes that allow a single interaction in one module: Planning, Research Loop, or Generation.

As shown in Table 4, interaction at any stage improves over the autonomous baseline, confirming the broad utility of user feedback. However, interaction timing matters: early-stage interaction, especially during Planning, consistently yields larger gains than later intervention, highlighting the importance of early intent alignment. Full-lifecycle interaction achieves the best overall performance, demonstrating the advantage of continuous alignment over one-shot clarification. Yet, Llama-4-Maverick exhibits mild instability, with fully interactive settings underperforming Planning-only on some metrics. This suggests that while early guidance is critical, the capability to manage frequent, multi-turn interactions varies among models.

4.5 Recommendations

Beyond serving as an evaluation benchmark, IDRBench offers actionable guidance for deploying interactive deep research systems. Based on observed trade-offs between interaction-induced performance gains and interaction costs, we provide scenario-driven recommendations in Table 5.

Our results indicate that no single model uniformly outperforms others across all scenarios. Instead, model suitability depends critically on operational priorities, such as maximizing performance ceilings, supporting intensive interaction, achieving high interaction efficiency, or operating under strict cost constraints. By jointly evaluating interaction benefits and costs, IDRBench enables informed model selection tailored to the cognitive and budgetary requirements of real-world applications.

5 Conclusions

We introduce IDRBench, the first benchmark for systematically evaluating interactive deep research with LLMs. Going beyond final outputs, IDRBench captures how agents interact, adapt, and align with users under uncertainty, jointly measuring interaction benefits and costs. Through a modular interactive framework, a scalable reference-grounded user simulator, and an interaction-aware evaluation suite, IDRBench enables principled analysis of human-AI collaboration in long-horizon research tasks. Experiments on seven state-of-the-art LLMs show that interaction consistently improves research quality and robustness, often rivaling gains from increased model capacity, while revealing important trade-offs in interaction efficiency and cost. We believe IDRBench provides a strong foundation for developing more reliable, efficient, and user-aligned deep research agents.

Limitations

Idealized User Simulation. We acknowledge that the reference-grounded User Simulator induces an idealized interaction setting that may not fully capture real-world user behavior, such as volatility, ambiguity, or shifting intent over time. However, such stability is a necessary design choice for rigorous and reproducible benchmarking. In comparative evaluation, inconsistent or contradictory feedback would act as a confounding factor, obscuring whether failures stem from an agent's reasoning or from noise in user input. By standardizing feedback to be consistent, goal-oriented, and grounded in reference documents, IDRBench isolates the agent's intrinsic interactive capability as the primary source of performance variation. This design choice aligns with emerging evaluation practices, where high-capacity LLMs are increasingly used as scalable and controlled proxies for human behavior in agent benchmarking (Yao et al., 2025). Future extensions may relax this assumption to study robustness under more stochastic user behaviors.

Limited Scope of Ambiguity Types. Our current ambiguity injection strategy focuses on underspecification, implemented by compressing detailed queries into vague prompts. We recognize that real-world ambiguity is more diverse, arising from user misconceptions, polysemy, or domain-specific errors. We deliberately focus on underspecification because it preserves a recoverable ground truth, the original detailed query, enabling objective and quantitative evaluation of intent recovery. Without such a reference, alignment assessment would necessarily rely on subjective judgments, reducing reproducibility. Nonetheless, this represents only one dimension of ambiguity. An important direction for future work is to extend IDRBench with richer ambiguity types, including factual inaccuracies and cognitive biases, to evaluate not only clarification but also correction and negotiation in human-AI research collaboration.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 4473–4484.

Salaheddin Alzubi, Creston Brooks, Purva Chiniya, Edoardo Contente, Chiara von Gerlach, Lucas Irwin, Yihan Jiang, Arda Kaz, Windsor Nguyen, Sewoong Oh, Himanshu Tyagi, and Pramod Viswanath. 2025. Open Deep Search: Democratizing Search with Open-source Reasoning Agents. *arXiv preprint arXiv:2503.20201*.

Anonymous. 2025. An Interactive Paradigm for Deep Research. In *Submitted to The Fourteenth International Conference on Learning Representations*. Under review.

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2025. Longwriter: Unleashing 10,000+ word generation from long context LLMs. In *The Thirteenth International Conference on Learning Representations*.

João Coelho, Jingjie Ning, Jingyuan He, Kangrui Mao, Abhijay Paladugu, Pranav Setlur, Jiahe Jin, Jamie Callan, João Magalhães, Bruno Martins, and 1 others. 2025. DeepResearchGym: A Free, Transparent, and Reproducible Evaluation Sandbox for Deep Research. *arXiv preprint arXiv:2505.19253*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. *arXiv preprint arXiv:2506.11763*.

Yue Feng, Hossein A Rahmani, Aldo Lipani, and Emine Yilmaz. 2023. Towards asking clarification questions for information seeking on task-oriented dialogues. *arXiv preprint arXiv:2305.13690*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Google DeepMind. 2025. Gemini 3 Pro. https://deepmind.google/models/gemini/pro/. Product page.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan O Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training LLMs to reason and

leverage search engines with reinforcement learning. In *Second Conference on Language Modeling*.

LangChain-AI. 2025. Open Deep Research Project.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025. Search-o1: Agentic Search-Enhanced Large Reasoning Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Meta AI. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Blog post.

OpenAI. 2025. GPT-5.1. https://openai.com/index/gpt-5-1/. GPT-5.1: A smarter, more conversational ChatGPT.

Hossein A Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. 2023. A survey on asking clarification questions datasets in conversational systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 6252–6278.

Leila Tavakoli, Johanne R Trippas, Hamed Zamani, Falk Scholer, and Mark Sanderson. 2022. Mimics-duo: Offline & online evaluation of search clarification. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3198–3208.

Xiaohua Wang, Zhenghua Wang, Xuan Gao, Feiran Zhang, Yixin Wu, Zhibo Xu, Tianyuan Shi, Zhengyuan Wang, Shizheng Li, Qi Qian, and 1 others. 2024. Searching for best practices in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17716–17736.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025. BrowseComp: A Simple Yet Challenging Benchmark for Browsing Agents. *arXiv preprint arXiv:2504.12516*.

Yuning Wu, Jiahao Mei, Ming Yan, Chenliang Li, Shaopeng Lai, Yuran Ren, Wang Zijia, Ji Zhang, Mengyue Wu, Qin Jin, and Fei Huang. 2025. Writingbench: A comprehensive benchmark for generative writing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

xAI. 2025. Grok 4.1 model card. Model card, xAI. Accessed: 2026-01-02.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik R Narasimhan. 2025. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *The Thirteenth International Conference on Learning Representations*.

Hyeonggeun Yun and Jinkyu Jang. 2025. Interaction-Driven Browsing: A Human-in-the-Loop Conceptual Framework Informed by Human Web Browsing for Browser-Using Agents. *arXiv preprint arXiv:2509.12049*.

Tong Zhang, Peixin Qin, Yang Deng, Chen Huang, Wenqiang Lei, Junhong Liu, Dingnan Jin, Hongru Liang, and Tat-Seng Chua. 2024a. CLAMBER: A Benchmark of Identifying and Clarifying Ambiguous Information Needs in Large Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 10746–10766.

Xuan Zhang, Yang Deng, Zifeng Ren, See Kiong Ng, and Tat-Seng Chua. 2024b. Ask-before-Plan: Proactive Language Agents for Real-World Planning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10836–10863.

Xuan Zhang, Yongliang Shen, Zhe Zheng, Linjuan Wu, Wenqi Zhang, Yuchen Yan, Qiuying Peng, Jun Wang, and Weiming Lu. 2025. AskToAct: Enhancing LLMs Tool Use via Self-Correcting Clarification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Penghao Zhao, Hailin Zhang, Qinhan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang, Wentao Zhang, Jie Jiang, and Bin Cui. 2024. Retrieval-Augmented Generation for AI-Generated Content: A Survey. *arXiv preprint arXiv:2402.19473*.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. DeepResearcher: Scaling Deep Research via Reinforcement Learning in Real-world Environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, and 1 others. 2024. OpenResearcher: Unleashing AI for Accelerated Scientific

Research. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, and 1 others. 2025. BrowseComp-ZH: Benchmarking Web Browsing Ability of Large Language Models in Chinese. *arXiv preprint arXiv:2504.19314*.

Yujia Zhou, Yan Liu, Xiaoxi Li, Jiajie Jin, Hongjin Qian, Zheng Liu, Chaozhuo Li, Zhicheng Dou, Tsung-Yi Ho, and Philip S Yu. 2024. Trustworthiness in Retrieval-Augmented Generation Systems: A Survey. *arXiv preprint arXiv:2409.10102*.

A Hyperparameter Configuration

To ensure reproducibility, we detail the specific hyperparameter configurations for both the execution of the framework and the calculation of evaluation metrics. These settings are summarized in Table 6.

Framework Execution Parameters. We impose specific constraints on the research process to prevent unbounded execution and maintain a realistic simulation environment.

- **Iteration Limits:** We set the *Max Supervisor Iterations* to **6** and *Max Researcher Tool Calls* to **5**. These values are aligned with the default configuration of the *Open Deep Research* architecture, serving as a standard baseline to control the depth of reasoning without incurring excessive latency.
- **Concurrency and Context:** To model the parallel nature of human research teams, we allow up to **3 concurrent research units**. Furthermore, we enforce a **50,000-character** limit on raw web content, balancing information retention with context management.

Evaluation Metrics Configuration. Table 6 further details the parameters used to compute our interaction-aware metrics, categorized by the specific metric they support:

- **Report Similarity:** We utilize **Qwen/Qwen3-0.6B** as the embedding backbone for calculating cosine similarity. Its 32k-token context window is essential for encoding full-length research reports, ensuring that the similarity score reflects global semantic consistency rather than truncated segments.
- **Multi-Granularity F1-Score:** To compute F1-scores at the chunk level, we adopt a sliding window approach with a **300-token chunk size** and **50-token overlap**. A strict hard match threshold of $\tau = \mathbf{0.8}$ is applied to filter out low-confidence matches, ensuring capturing genuine structural overlap.
- **LLM Aspect Coverage Score (LLM-ACS):** For evaluating intent fulfillment, we generate between **8 and 20 specific aspects** per query. This range provides sufficient granularity to evaluate intent coverage comprehensively while avoiding trivial details.

Description	Setting
Execution Constraints	
Max Supervisor Iterations	6
Max Researcher Tool Calls	5
Max Concurrent Research Units	3
Max Content Length	50,000 chars
Report Similarity & Multi-Granularity F1-Score	
Embedding Model	`Qwen/Qwen3-0.6B`
Multi-Granularity F1-Score	
Chunk Size	300 tokens
Chunk Overlap	50 tokens
Hard Match Threshold (τ)	0.8
LLM Aspect Coverage Score (LLM-ACS)	
Generated Aspects (M)	8–20

Table 6: Summary of hyperparameter configurations for the Interactive Deep Research framework and the IDRBench evaluation suite.

B Examples of Ambiguity Injection

Table 7 presents selected examples of ambiguity injection from our dataset. In these pairs, the **Original Query** represents a highly specified user request, characterized by explicit constraints, rich background context, and detailed output requirements (e.g., specific technical limitations, target demographics, or required data dimensions). The complete version of the dataset is available in our GitHub repository.

The **Ambiguity Injected Query** is derived from the original text. As illustrated in the table, while the **core user intent** (such as performing a comparative analysis, conducting a medical review, or summarizing a cultural topic) is strictly preserved, the specific **details and constraints** are intentionally omitted. For instance, in Example 68, the technical constraint regarding the "standard Cluster Autoscaler relying on pending pods" is removed, leaving a broader request for "approaches beyond the standard." This transformation results in prompts

ID	Original Query	Ambiguity Injected Query
54	In the field of FinTech, machine learning algorithms are now widely applied to asset allocation and investment decisions. Examples include classic models like Mean-Variance and Black-Litterman, as well as emerging deep learning models. While these models have shown certain advantages under different market conditions, each also has its limitations. For instance, the Mean-Variance model assumes asset returns follow a normal distribution, which often doesn't align with actual market conditions. The Black-Litterman model relies on subjective view inputs, introducing a degree of subjectivity. Although deep learning models can handle complex non-linear relationships, they suffer from poor interpretability. So, what are the core differences between these various models in terms of risk measurement, return prediction, and asset allocation? And is it possible to combine their strengths to build a more general-purpose and effective modeling framework?	What are the main differences between traditional and machine learning models in asset allocation regarding risk measurement and return prediction, and can their strengths be integrated into a more effective framework?
68	I need to dynamically adjust Kubernetes (K8S) cluster node counts based on fluctuating business request volumes, ensuring resources are scaled up proactively before peak loads and scaled down promptly during troughs. The standard Cluster Autoscaler (CA) isn't suitable as it relies on pending pods and might not fit non-elastic node group scenarios. What are effective implementation strategies, best practices, or existing projects that address predictive or scheduled autoscaling for K8S nodes?	What are effective approaches or tools for predictive or scheduled autoscaling of Kubernetes nodes beyond the standard Cluster Autoscaler, especially for handling varying request volumes?
76	The significance of the gut microbiota in maintaining normal intestinal function has emerged as a prominent focus in contemporary research, revealing both beneficial and detrimental impacts on the equilibrium of gut health. Disruption of microbial homeostasis can precipitate intestinal inflammation and has been implicated in the pathogenesis of colorectal cancer. Conversely, probiotics have demonstrated the capacity to mitigate inflammation and retard the progression of colorectal cancer. Within this domain, key questions arise: What are the predominant types of gut probiotics? What precisely constitutes prebiotics and their mechanistic role? Which pathogenic bacteria warrant concern, and what toxic metabolites do they produce? How might these findings inform and optimize our daily dietary choices?	What is the role of gut microbiota and its balance in intestinal health and disease, and how can insights into probiotics, prebiotics, and harmful bacteria guide dietary choices to support gut health?
91	I would like a detailed analysis of the Saint Seiya franchise (anime/manga). The analysis should be structured around the different classes of armor (Cloths, Scales, Surplices, God Robes, etc.), such as Bronze Saints, Silver Saints, Gold Saints, Marina Generals, Specters, God Warriors, etc. For each significant character within these categories, provide details on their power level, signature techniques, key appearances/story arcs, and final outcome/fate within the series.	Provide an overview of the Saint Seiya franchise focusing on the major armor classes and their representative characters, discussing their roles and significance within the series.
95	Create comprehensive, in-depth study notes for the Diamond Sutra (Vajracchedikā Prajñāpāramitā Sūtra). These notes should offer deep analysis and interpretation from various perspectives, exploring its teachings and relevance in contexts such as daily life, the workplace/career, business practices, marriage, parenting, emotional well-being, and interpersonal dynamics.	Create comprehensive study notes for the Diamond Sutra, including analysis of its teachings and relevance in various aspects of life.

Table 7: Examples of Ambiguity Injection

that are significantly shorter and inherently more ambiguous, effectively simulating the underspecified nature of real-world initial user queries.

C Core Agent Prompt Designs

We detail the prompt specifications for the three agents central to the interactive deep research framework.

Evaluator. This agent (Figure 4) functions as the interaction gatekeeper. It analyzes the current research context to determine whether the information gain from user clarification outweighs the interruption burden. Instead of indiscriminate questioning, it enforces a binary decision based on specific guidelines tailored to the different research stages.

Questioner. When interaction is triggered, the Questioner formulates targeted inquiries. The prompt (Figure 5) explicitly constrains the agent to focus on high-level scope, intent, and structural ambiguities rather than trivial technical details. It ensures that questions are concise and tonally adapted to the user's language to minimize cognitive load.

User Simulator. This agent (Figure 6) acts as a proxy for human feedback, enabling scalable and reproducible evaluation. It is strictly grounded in the Reference Document. The prompt instructs the simulator to provide natural, goal-oriented guidance that steers the research trajectory toward the target result without hallucinating requirements.

D Ethical Considerations

The dataset constructed in this work is derived from the publicly available dataset and is used in strict adherence to its original license and usage terms. We have rigorously reviewed the data samples to verify that they do not contain personally identifiable information (PII), offensive text, or sensitive content. Additionally, we utilized Large Language Models to assist in data construction, specifically for generating ambiguous queries through summarization, with human verification to ensure semantic consistency. As this work focuses on benchmarking and evaluating the capabilities of research agents rather than deploying a user-facing generative system, we do not foresee any significant ethical or societal risks associated with the release or use of this dataset.

Evaluator

You are a Research Evaluator in a deep research pipeline. The pipeline's goal is to generate a comprehensive research report (markdown format) based on the user's research topic. Your current task is to analyze the research progress and decide if we need to ask the user for clarification or more specific direction.

Today's date is {date}.

Context
<research topic>{research_topic}</research topic>
<research plan>{research_plan}</research plan>
<current research progress>{current_research_progress}</current research progress>

Current Phase: {current_phase}
- Research iteration: {iteration_count}
- Remaining question opportunities: {remaining_opportunities}

Guidelines for Decision
1. Pre-Research Phase (no findings yet):
   - Research Outline: How should the research outline be formulated? Does it align with user expectations?
   - Report Genre: What is the desired genre of the report (e.g., academic survey, practical guide, etc.)?
   - Key Concepts: Are there any key terms or concepts in the topic that need clarification or interpretation?
2. Mid-Research Phase (research in progress):
   - Scope and Priorities: What aspects should be included in the research? Which areas deserve deeper exploration?
   - Research Completeness: Are the current plan and findings complete and well-rounded? Any obvious gaps or missing elements?
3. Final Phase (ready to generate final report):
   - Report Structure: What is the expected structure of the final report? How should content be organized across sections?
   - Presentation Preferences: How should the content be presented (e.g., use of tables, citation format, etc.)?

Trade-off to Consider
- Benefit of asking: Asking for user clarification can help produce a report that better aligns with user expectations.
- Cost of asking: Each question may increase user interaction burden and slow down the research process.
- Consideration: A well-targeted question often saves more effort than it costs.

Constraints
- If you decide to ask, you will be able to ask up to two questions.
- Questions must address the core focus of the research topic - avoid trivial or peripheral matters.
- Each question should focus ONLY on one single aspect of the research process. Stay focused on high-level aspects rather than technical details.

Task
1. Based on the guidelines and constraints, evaluate whether the corresponding aspects are already clear.
2. Considering the trade-off and remaining opportunities, decide if the questions worth asking.

Output
(All the "thinking", "should_ask", and "reason" fields are required in the final response):
- In the "thinking" field: first think about whether to ask questions based on the Task steps.
- In the "should_ask" field: provide your decision (true or false).
- In the "reason" field: provide a brief explanation for your decision.

CRITICAL: All fields can be written in English.

Figure 4: Evaluator's prompt

Questioner

You are a Research Assistant in a deep research pipeline. The pipeline's goal is to generate a comprehensive research report (markdown format) based on the user's research topic. Your current task is to ask clarifying questions to guide the ongoing research.

Context
<research topic>{research_topic}</research topic>
<research plan>{research_plan}</research plan>
<current research progress>{current_research_progress}</current research progress>

Reason for Asking
{reason_for_asking}

Current Phase: {current_phase}
- Research iteration: {iteration_count}
- Remaining question opportunities: {remaining_opportunities}

Guidelines for Decision
1. Pre-Research Phase (no findings yet):
   - Research Outline: How should the research outline be formulated? Does it align with user expectations?
   - Report Genre: What is the desired genre of the report (e.g., academic survey, practical guide, etc.)?
   - Key Concepts: Are there any key terms or concepts in the topic that need clarification or interpretation?
2. Mid-Research Phase (research in progress):
   - Scope and Priorities: What aspects should be included in the research? Which areas deserve deeper exploration?
   - Research Completeness: Are the current plan and findings complete and well-rounded? Any obvious gaps or missing elements?
3. Final Phase (ready to generate final report):
   - Report Structure: What is the expected structure of the final report? How should content be organized across sections?
   - Presentation Preferences: How should the content be presented (e.g., use of tables, citation format, etc.)?

Constraints
- You can ask up to two questions in this round. Each question should be concise (max 3 sentences) and focus on one aspect only.
- Questions must address the core focus of the research topic - avoid trivial or peripheral matters.
- FOCUS ON THE BIG PICTURE: Ask about high-level aspects of the research process - NOT technical details or specifications.

Output
(All the "thinking", "summary", and "question" fields are required in the final response):
- In the "thinking" field: first think step by step about how to ask questions based on the guidelines.
- In the "summary" field: provide a brief summary of the current research progress (max 200 words).
- In the "question" field: provide question(s) in numbered format (e.g., 1) ... 2) ...).

CRITICAL: The content of the "summary" and "question" fields will be passed to the user, so they MUST be written in the SAME language type as the research topic. The "thinking" field can be written in English.

Figure 5: Questioner's prompt

User Simulator

You are a User Simulator, acting as a user who has given a research question to request a comprehensive markdown research report. The researcher is currently conducting research for the user and has encountered some questions that require user feedback. Your role is to simulate the user and provide answers based on the expected report, which represents what the user ideally wants from the research.

Your Goal
Answer the researcher's questions strategically to guide them toward producing a final report as similar as possible to the expected report.

Context
<question>{question}</question>
<expected report>{article}</expected report>

Guidelines
1. Natural User Tone: Simulate a real user expressing requirements (e.g., "I think..." or "I don't want..."). NEVER mention the given expected report itself (e.g., "The expected report says..." or "Not specified in the expected report").
2. Steer Research Focus: Answer the questions to guide the researcher to produce a report similar to the expected report. ALWAYS steer toward the major focus of the expected report, not trivial details. If the question focuses on minor or less important aspects, actively redirect attention to the major focus of the expected report (e.g., "Price doesn't matter much here - what's more important is comparing the performance and scalability of each approach.").
3. Strict Fidelity & Independent Thinking: Answer only based on the expected report. NEVER fabricate details or deliverables not explicitly mentioned in the expected report. If the question provides options, DO NOT feel obligated to pick one. Ignore misleading options and answer based on the expected report.
4. Be Comprehensive Yet Brief: Provide comprehensive answers that fully convey your requirements. Omitting key parts could impact the comprehensiveness of the research. Keep each individual answer under 50 words.
5. Answer Format: When providing multiple answers, use numbered format (e.g., 1) ... 2) ...) to separate different answers. Within each individual answer, use plain natural language sentences ONLY - avoid using lists, bullet points, or other structured formats. Speak naturally as a real user would.

Output
(Both the "thinking" and "answer" fields are required in the final response):
- In the "thinking" field: first search for relevant information from the Expected Report, then check and refine it against each Guideline step by step.
- In the "answer" field: provide your answer(s) in numbered format (e.g., 1) ... 2) ...).

CRITICAL: The "answer" field MUST be written in the SAME language as the expected report. The "thinking" field can be written in English.

Figure 6: User Simulator's prompt