From Web Search towards Agentic Deep Research: Incentivizing Search with Reasoning Agents

Weizhi Zhang[1*†], Yangning Li[2*], Wei-Chieh Huang[1*], Junyu Luo[3], Guancheng Wan[4], Liangwei Yang[5], Yuyao Yang[1], Chenxuan Xie[6], Chunyu Miao[1], Henry Peng Zou[1], Xiao Luo[4], Yusheng Zhao[3], Yankai Chen[1], Chunkit Chan[7], Peilin Zhou[8], Xinyang Zhang[9‡], Chenwei Zhang[9§], Jingbo Shang[10], Ming Zhang[3], Yangqiu Song[7], Irwin King[11], Philip S. Yu[1]

[1] UIC, [2] THU, [3] PKU, [4] UCLA, [5] Salesforce AI Research, [6] ZJUT, [7] HKUST, [8] HKUST (Guangzhou), [9] Amazon, [10] UCSD, [11] CUHK

Abstract

Information retrieval is a cornerstone of modern knowledge acquisition, enabling billions of queries each day across diverse domains. However, traditional keyword-based search engines are increasingly inadequate for handling complex, multi-step information needs. Our position is that Large Language Models (LLMs), endowed with reasoning and agentic capabilities, are ushering in a new foundational paradigm Agentic Deep Research. These systems transcend conventional information search techniques by tightly integrating autonomous reasoning, iterative retrieval, and information synthesis into a dynamic feedback loop. We trace the evolution from static web search to interactive, agent-based systems that plan, explore, and learn. We also introduce a test-time scaling law to formalize the impact of computational depth on reasoning and search. Supported by benchmark results and the rise of open-source trends, we show the position that Agentic Deep Research not only significantly outperforms existing approaches, but is also poised to become the foundation paradigm for future information search. All the related resources, including industry products, research papers, benchmark datasets, and open-source implementations, are collected for the community in https://github.com/DavidZWZ/Awesome-Deep-Research.

1 Introduction

"Introducing deep research: An agent that uses reasoning to synthesize large amounts of online information and complete multi-step research tasks for you."

– OpenAI

Every day, billions of people search for information online (Amendola et al., 2023), and rely heavily on these online resources to make decisions across personal, professional, and societal contexts (Zhang et al., 2024). For decades, traditional web search engines based on keyword matching have served as the primary gateway to digital information. While once revolutionary, these systems increasingly struggle with complex, multi-faceted queries that demand nuanced understanding and synthesis (Mo et al., 2024). The growing inadequacy highlights their inherent limitations in contextual comprehension and knowledge integration.

Against this backdrop, information seeking and synthesis are undergoing a profound transformation with large language models (LLMs). Rather than merely enhancing traditional search paradigms, LLMs are poised to fundamentally replace them in addressing complex information needs. Initially, we witnessed the rise of LLMs as standalone, knowledgeable chatbots, which challenged the dominance of web search by offering more direct answers and a degree of synthesis, thereby reducing the user's burden of sifting through numerous links (Liu et al., 2023). However, single LLMs are tethered to static, offline knowledge. The subsequent integration of search and retrieval-augmented generation (RAG) marked a step forward, grounding LLMs in external data and mitigating issues like hallucination (Ma et al., 2023; Yang et al., 2025). Nevertheless, these naive RAG methods still struggle with real-world questions that require sophisticated multi-hop reasoning and strategic search planning, as they often cannot plan correct search paths for complex problems (Yao et al., 2023b).

Recently, test-time scaling (TTS) has emerged as a potent paradigm for boosting the reasoning and agentic capabilities of LLMs (Snell et al., 2024).

*Authors contributed equally to this research.
†Corresponding author: wzhan42@uic.edu
‡Work does not relate to the author's position at Amazon

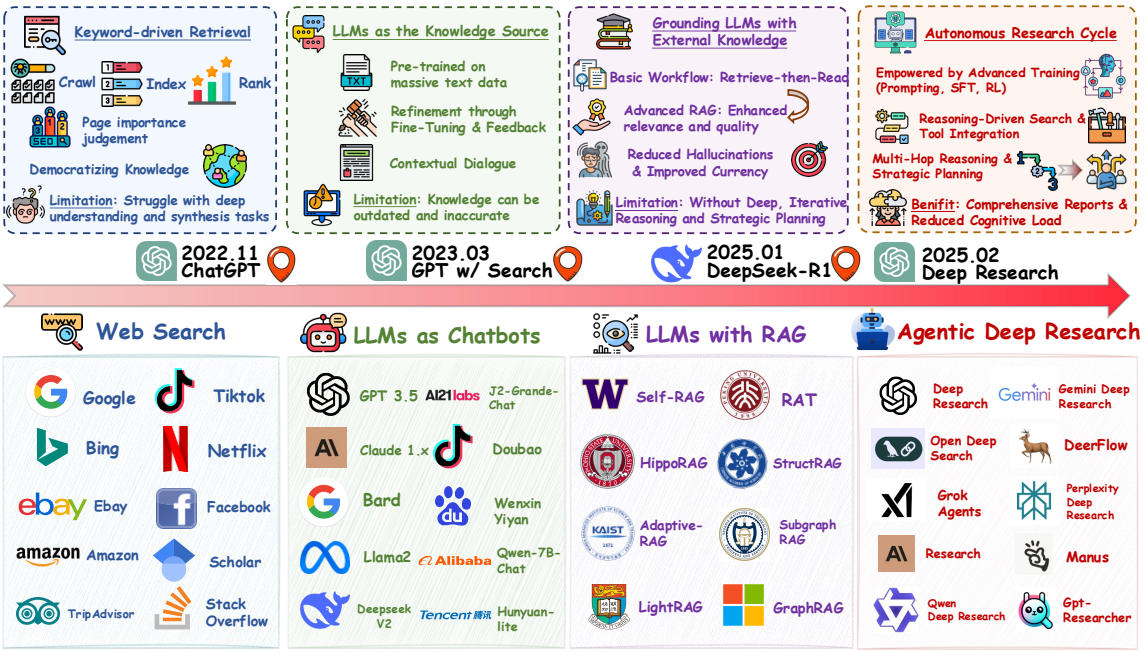

Figure 1: The evolution of information search paradigms.

It assigns additional computation during inference, enabling deeper problem-solving (Zou et al., 2025a; Gu et al., 2025). Equipped with TTS on reasoning and search, LLMs are set to drive a new search paradigm termed **Agentic Deep Research** systems, which are capable of autonomous **reasoning**, on-demand **searching**, and iterative information synthesis. Demonstrations from deep research products launched by OpenAI and Google highlight several key advantages of this paradigm: *(1) Comprehensive Understanding*: Ability to dissect and address complex, multifaceted queries that overwhelm traditional methods (Wei et al., 2022); *(2) Enhanced Synthesis*: Excels at synthesizing information from diverse, potentially conflicting sources into coherent and insightful narratives (Cheng et al., 2025); *(3) Reduced User Burden*: Significantly decreases the cognitive load and manual effort required from users by automating laborious search steps (Sami et al., 2024).

Our position. Agentic Deep Research frameworks with synergized reasoning and search capabilities will become the new foundation paradigm for future information search.

In this paper, we comprehensively investigate this paradigm shift and make four key contributions: (1) We systematically trace and analyze the evolutionary trajectory of information search paradigms, from traditional keyword-based search, through conversational LLM chatbots and naive retrieval-augmented LLMs, and ultimately to the Agentic Deep Research; (2) We discuss a preliminary scaling law for Deep Research, a novel hypothesis formalizing the relationship between inference-time computational resources allocation and the resulting improvements in LLMs reasoning capabilities and search depth; (3) We conduct extensive evaluations on existing LLMs, reasoning LLMs, and Agentic Deep Research models, coupled with analysis of open-source implementations to support our position; and (4) As the first position paper in the field of Deep Research, we offer alternative view discussions and critical future research directions, outlining both opportunities and challenges. Our discussion outlines a clear roadmap for this evolving field, underscoring how this evolution is fundamentally reshaping human interaction with real-world information and guiding further advancement on Agentic Deep Research.

2 Traditional Information Search Paradigms

The evolution of search paradigms, as in Figure 1, represents a fundamental transformation in how humans access and interact with information. This section examines three distinct frameworks that have shaped the landscape of information retrieval: traditional web search engines, Large Lan-

guage Models (LLMs) as chatbots, and LLMs with Retrieval Augmented Generation (RAG) systems. Each paradigm addresses prior information seeking challenges, from web ranking to interactive dialogue and knowledge-augmented generation.

2.1 Web Search

Web search has fundamentally transformed information access in modern society, enabling near-instantaneous retrieval of knowledge that previously required days or months to locate (Brin and Page, 1998). This has greatly accelerated knowledge access and scientific discovery. As the primary infrastructure for information retrieval in the Internet era, online search engines (e.g., Google, Bing) or content platforms (e.g., YouTube) have continuously shaped how humans obtain the expanding information landscape (Page et al., 1999).

Basically, web search operates through three fundamental processes: crawling (Khder, 2021), indexing (Hendriksen et al., 2024), and ranking (Robertson et al., 2009). Crawlers discover and collect web content across the Internet. The gathered data are then indexed into structured representations such as inverted indices, enabling efficient subsequent ranking. When a user submits a query, the ranking module employs algorithms, most notably PageRank (Page et al., 1999), to evaluate document relevance and authority based on link structures (Fuhr, 1992). Modern search systems have further evolved to incorporate semantic understanding or personalized user behavior (Wang et al., 2024b). However, the contexts presented to the user may still not be relevant and accurate, due to the limited context for user-specific complex queries (Leake and Scherle, 2001) and the influence of advertisement bidding (Linden et al., 2009).

2.2 LLMs as Chatbots

Benefiting from recent advances in natural language processing (NLP), large language models (LLMs) represent the latest evolution in information retrieval as user chatbots for tailored response generation. These approaches treat LLMs (e.g., ChatGPT (Achiam et al., 2023), Claude (Anthropic, 2023), and LLaMa (Touvron et al., 2023)) as knowledge bases that provide users with customized responses through interactive dialogue, generating integrated solutions that are not directly available online (Thirunavukarasu et al., 2023; Kasneci et al., 2023). Unlike traditional search engines that require users to repeatedly provide context and laboriously browse results, LLM-based chatbots can maintain contextual continuity throughout a conversation. In addition, they integrate vast amounts of web-derived knowledge within their parameters, serving as compact representations of extensive online information (Zeng et al., 2022).

However, despite these advantages, relying exclusively on internal LLM knowledge presents notable challenges: (1) *hallucinations*, where models generate plausible but inaccurate content (Tam et al., 2023; Yao et al., 2023a; Zhao et al., 2025); (2) *lack of awareness of recent events*, which compromises the timeliness of responses (Chen and Shu, 2023); and (3) *Limited context window*, hindering a comprehensive understanding of complex queries (Wang et al., 2024d). Therefore, integrating external information sources and employing advanced reasoning to verify retrieved data are crucial strategies for addressing these limitations, thus ensuring LLM chatbots deliver accurate, relevant, and up-to-date information (Peng et al., 2023).

2.3 LLMs with RAG

To address the inherent limitations of LLMs chatbots mentioned above, particularly their static knowledge and tendency to hallucinations, Retrieval Augmented Generation (RAG) has emerged as a promising paradigm (Prabhune and Berndt, 2024). RAG integrates the generative capabilities of LLMs with retrieval systems to dynamically access relevant external information. Early implementations of RAG primarily employed a straightforward "Retrieve-then-Read" workflow (Ma et al., 2023), typically involving a single-step retrieval from an external knowledge source. Although they improve upon purely parametric methods, such naive RAG systems can still struggle with inaccurate retrieval when faced with complex queries.

To solve this problem, multi-hop retrieval addresses the limitations of traditional single-hop retrieval by enabling iterative, sequential searches and reasoning steps across multiple data sources (Jiang et al., 2023). Multi-hop retrieval incorporates iterative refinement, where intermediate retrieval outcomes guide subsequent queries, progressively building comprehensive context (Zhang et al., 2025a). Although multi-hop retrieval has strong power generation capabilities, it also suffers from limitations due to the underlying techniques it employs. Early stage errors in reasoning paths can propagate throughout subsequent retrieval and reasoning steps, severely influencing the final output

integrity (Zhang et al., 2025a). Additionally, maintaining faithfulness to retrieved evidence poses ongoing difficulties, as LLMs frequently encounter conflicts between retrieved data and internal parametric knowledge (Zheng et al., 2025b).

3 Towards Agentic Deep Research

Many complex real-world problems, including open-domain question answering (Yang et al., 2015; Chen and Yih, 2020) and scientific discovery (Lu et al., 2024; Wang et al., 2023b; Baek et al., 2024; Schmidgall et al., 2025), inherently require an iterative interplay between information retrieval and reasoning. For example, in Table 1, standard LLMs with simple reasoning may still be confronted by the hallucination problem. A single search step often falls short of capturing comprehensive information, while isolated reasoning phases can fail to identify critical insights (Trivedi et al., 2023). By tightly integrating search and reasoning in a multi-step and interactive manner, Deep Research models can progressively enhance the relevance and depth of retrieved knowledge and simultaneously refine the reasoning process underlying query interpretation, ultimately producing more accurate and contextually nuanced responses. Here, reasoning actively influences search (e.g., refining search queries based on intermediate deductions) (Li et al., 2025d), while retrieved information recursively refines reasoning in a dynamic feedback loop. Unlike the previous LLMs with RAG framework in Section 2.3, where retrieval and reasoning occur in discrete and sequential stages, this approach treats them as interdependent, continuously co-evolving and benefiting each other. Based on this, we further summarize the common error types during information search in the Appendix A.

This evolution in search methodologies gives rise to a transformative paradigm we define as Agentic Deep Research. In this paradigm, language models take on the role of active information-seeking agents. Rather than a one-shot prompt + retrieve paradigm, an "agentic" LLM plans a series of steps: it can issue search queries, consult documents, browse the web, or even collaborate with other agents, all while refining its query understanding and response via iterative retrieval and reasoning. Inspired by the way human experts might research a question, we encapsulate this iterative synergy between *reasoning* and *search* in the term **Deep Research**, highlighting its dynamic and interactive essence. To substantiate our central position that LLM-driven Agentic Deep Research will become the predominant paradigm for future information-seeking, we ground our argument across three interlinked technical dimensions: reasoning capabilities as the foundation (Section 3.1), principled approaches to incentivize search (Section 3.2), and ecosystem-level momentum evidenced through benchmarks (Section 3.3).

3.1 Reasoning as A Preliminary Foundation

The evolution of reasoning capabilities in LLMs represents a crucial stepping stone toward truly agentic systems, particularly in the context of deep research tasks. While Chain-of-Thought (CoT) prompting (Wei et al., 2022) initially demonstrated the possibility of explicit reasoning processes, the real breakthrough lies in how reasoning mechanisms enable autonomous decision-making and strategic planning. The transformation from simple CoT to more sophisticated reasoning frameworks marks a fundamental shift of how to approach complex tasks. Rather than merely following predetermined patterns, modern reasoning frameworks enable systems to dynamically plan, execute, and adjust their actions based on intermediate outcomes. This capability is particularly evident in recent reinforcement learning approaches (Jaech et al., 2024; Guo et al., 2025), which have demonstrated unprecedented abilities in managing complex tasks. These systems can autonomously determine when to initiate search, formulate appropriate queries, and synthesize findings into a coherent understanding, forming the cornerstone of agentic behavior.

The DeepSeek-R1 (Guo et al., 2025) represents a significant milestone in this evolution, demonstrating how reinforcement learning can optimize reasoning processes for complex mathematical tasks. By learning from experience and feedback, these systems develop sophisticated strategies for information gathering and synthesis, moving beyond simple pattern matching to true strategic planning. This advancement in reasoning capabilities provides the essential foundation for agentic deep research by enabling systems to autonomously evaluate information needs, strategically decompose complex queries, synthesize information across multiple sources while maintaining logical consistency, and adapt search strategies based on intermediate results and feedback (Jin et al., 2025). These capabilities, rooted in advanced reasoning mechanisms, establish the preliminary foundation necessary for conducting deep research tasks that

Model Paradigm	Outcome & Analysis
Standard LLM (GPT-5)	Answer: **Tenali Raman (Tenali Rama)**. **Error Type:** Hallucinated Association. The model latches onto "humor" and "breaking the fourth wall" to guess Tenali Raman. It hallucinates that a 1980s series fits the "fewer than 50 episodes" constraint to force a fit, ignoring the specific "selfless ascetics" backstory constraint.
LLM + Search/RAG (GPT-5 with Search)	Answer: **Ellery Queen**. **Error Type:** Shallow Retrieval Steps. The model performs a search for "character breaking fourth wall TV show." It retrieves Ellery Queen (who breaks the fourth wall). However, it fails to verify the "selfless ascetics" constraint or the humor aspect, accepting a partial match from the first page of results.
Agentic Deep Research (OpenAI Deep Research)	Answer: **Plastic Man**. **Success Factors:** 1. Decomposition: Break down query (4th wall, ascetics, 60s–80s TV, $<$ 50 episodes). 2. Iterative Search: Finds "Plastic Man" via "ascetic backstory" queries. 3. Conflict Resolution: It encounters a conflicting data point (episode count $>$ 100) but reasons that this might include "segments" or "anthology" features. It performs a specific follow-up check to resolve this discrepancy before finalizing the answer.

Table 1: **Case study illustrating distinct failure modes and success factors for a information search question on BrowseComp (Wei et al., 2025a).** Question: "Please identify the fictional character who occasionally breaks the fourth wall with the audience, has a backstory involving help from selfless ascetics, is known for his humor, and had a TV show that aired between the 1960s and 1980s with fewer than 50 episodes." Ground truth: **Plastic Man**.

require strategic planning, iterative refinement, and complex decision-making. The integration of reinforcement learning with reasoning frameworks represents a crucial step toward truly agentic systems capable of conducting sophisticated research autonomously, marking significant advancements of traditional search paradigms (Chen et al., 2025b).

3.2 Incentivize Search with Reasoning Agents

Within this paradigm, reasoning is not merely an auxiliary component applied post-retrieval; rather, it constitutes the core mechanism that determines when, what, and how to search (Wu et al., 2025b). While prompting and supervised fine-tuning (SFT) serve as foundational techniques for instilling tool-use behaviors and basic query generation, they are inherently limited by their reliance on fixed instruction patterns and offline supervision (Wang et al., 2023a; Ghosh et al., 2024). In contrast, reinforcement learning (RL) provides a principled framework for cultivating truly agentic behavior, enabling models to explore, self-correct, and adaptively optimize their retrieval strategies in open-ended, interactive environments (Singh et al., 2025; Jin et al., 2025; Song et al., 2025). This shift toward RL-incentivized search marks a critical step toward developing autonomous agents capable of reasoning-driven information acquisition.

Prompting and In-Context Learning: Bridging Search and Reasoning. Prompting methods have laid important groundwork for coupling reasoning with external information retrieval. ReAct (Yao et al., 2023c) and its successors (Li et al., 2025b; Alzubi et al., 2025) introduced paradigms where LLMs alternate between reasoning steps and tool use, guiding models to break problems down and issue relevant search queries mid-process. This enables iterative refinement of reasoning with retrieved evidence, improving factuality and coherence. Extensions such as Search-o1 (Li et al., 2025b) and Open Deep Search (ODS) (Alzubi et al., 2025) prompt LLMs to actively consult web resources and integrate results into ongoing thought chains. In parallel, methods like Self-Ask (Press et al., 2023) and IRCoT (Trivedi et al., 2023) embed search directly within step-by-step reasoning, generating sub-questions and retrieving partial answers in a recursive loop. These prompting approaches offer flexible templates to scaffold retrieval-enhanced reasoning. However, they rely on fixed prompting logic and do not provide incentives for exploring better search or reasoning paths, limiting their scalability for open-ended or high-stakes tasks.

Supervised Fine-Tuning: Hard-Coding Search Patterns. Supervised fine-tuning (SFT) takes a more structured approach by directly training LLMs on datasets that combine reasoning and retrieval. Toolformer (Schick et al., 2023) and INTERS (Zhu et al., 2024) illustrate how models can be trained to learn when and how to query external

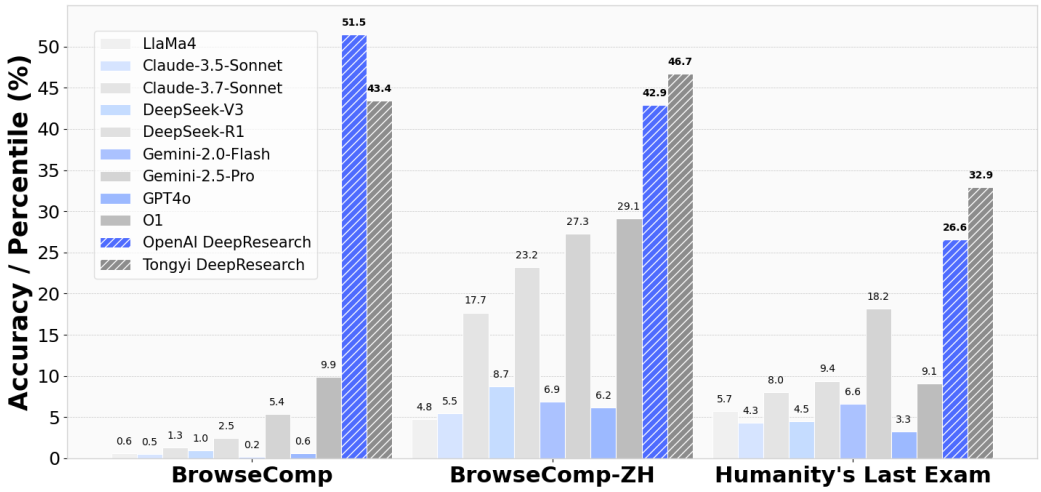

Figure 2: Benchmarks results of five standard LLMs, four reasoning LLMs, and two agentic deep research models on BrowseComp, BrowseComp-ZH, and Humanity's Last Exam.

tools, assess retrieved information, and integrate it logically into final outputs. SFT data typically comes from two sources: synthetic data generation (e.g., Toolformer (Schick et al., 2023), RAG-Studio (Mao et al., 2024)) or instructional reformulation of existing datasets (e.g., INTERS (Zhu et al., 2024), InstructRetro (Wang et al., 2024a)). These enable LLMs to follow structured retrieval-reasoning sequences. However, such methods primarily encode static behaviors learned from data, not dynamic, adaptive behaviors optimized for diverse environments. While prompting and SFT offer controlled environments for building retrieval-aware reasoning, they impose fixed search patterns and predefined goals. They do not equip agents with the ability to explore the open-ended, uncertain nature of real-world search tasks.

Reinforcement Learning: Synergizing Agentic Search with Reasoning in the Wild. Reinforcement learning (RL) fundamentally changes the search paradigm by letting agents learn through trial and error in interactive environments. Instead of being told how to search, RL-trained agents are incentivized (through feedback or reward functions) to discover, refine, and adapt their reasoning and search strategies for specific goals. Early systems like WebGPT (Nakano et al., 2021) and RAG-RL (Huang et al., 2025) demonstrated how reward signals (based on human feedback or factual correctness) can guide multi-step retrieval policies that improve response accuracy and trustworthiness. More modular designs like M-RAG (Wang et al., 2024e) separate reasoning and retrieval into specialized agents, each trained to collaborate via shared RL objectives. Recent RL-based systems such as Search-R1 (Jin et al., 2025), R1-Searcher (Song et al., 2025), DeepResearcher (Zheng et al., 2025a), ZeroSearch (Sun et al., 2025a), WebThinker (Li et al., 2025c), and WebAgent-R1 (Wei et al., 2025b) operate in various search environments from static local corpora and open search APIs to real-world web interfaces. These agents learn to decompose complex tasks, plan query sequences, verify evidence, and adjust their strategies based on environmental feedback. Importantly, ReSearch (Chen et al., 2025b) and ReARTeR (Sun et al., 2025b) go a step further by optimizing not just factual correctness, but also alignment with transparent, interpretable reasoning.

3.3 Benchmarks and Open-Source Trends

Deep Research Benchmarks To compare standard LLMs, reasoning-enhanced LLMs, and agentic deep research systems under realistic scenarios, we evaluate three representative benchmarks: BrowseComp (Wei et al., 2025a), BrowseComp-ZH (Zhou et al., 2025), and Humanity's Last Exam (HLE) (Phan et al., 2025). BrowseComp measures an agent's ability to perform multi-step, open-ended web search for non-trivial information, while BrowseComp-ZH extends this setting to the Chinese web. HLE, in contrast, emphasizes expert-level questions across diverse academic domains that cannot be solved by surface-level search alone. As shown in Figure 2, standard LLMs perform poorly on these benchmarks (typically below 10% on BrowseComp and below 20%

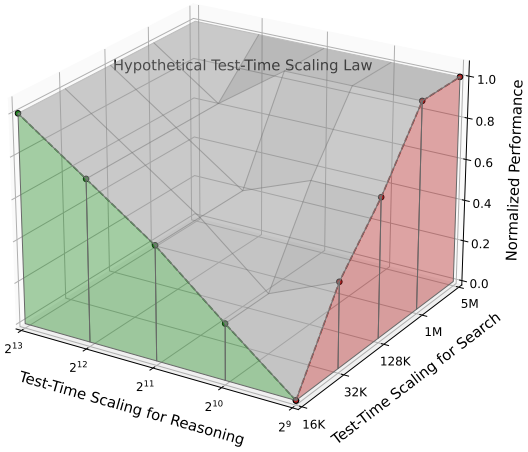

Figure 3: **Test-time scaling (TTS) law for agentic deep research** (measured in inference tokens). Data before performance normalization: reasoning TTS from S1 on AIME24 (Muennighoff et al., 2025) and search TTS from IterDRAG on MuSiQue (Yue et al., 2025).

on HLE), whereas deep research agents substantially outperform them, achieving over 40% on BrowseComp/BrowseComp-ZH and over 25% on HLE, highlighting the effectiveness of integrating reasoning with iterative search. More details for evaluation are explained in Appendix B.

Open-Source Trends Complementing benchmark results, open-source activity further reflects the rapid development of agentic deep research. Analysis of GitHub star trajectories (Appendix E), after filtering extreme outliers, highlights strong open-source momentum since early 2025. Notably, systems such as *deep-searcher* and *deer-flow* gained more than ten thousand stars within a few months, while smaller projects (e.g., *Deep-Researcher*, *R1-Searcher*) also collected hundreds of stars. This pattern suggests not only technical progress but a broader community shift toward reasoning-driven, agentic frameworks, reinforcing the view that agentic deep research is emerging as a dominant paradigm for future information-seeking systems.

4 Test-Time Scaling Law for Agentic Deep Research

Building upon the stated position, we introduce the Test-Time Scaling (TTS) law for Agentic Deep Research, a hypothesis predicting the performance improvements achievable through extended computational resources during inference. Figure 3 illustrates this hypothesis, normalized performance improves linearly through scaling of internal reasoning depth and external knowledge exploration. Data supporting this observation comes from two representative evaluations: S1 on the AIME24 dataset (Muennighoff et al., 2025), which tests reasoning-based scaling on advanced multi-step mathematical reasoning problems, and IterDRAG on MuSiQue (Yue et al., 2025), which focuses on search-based scaling via multi-hop retrieval tasks. As in Figure 3, the diagonal plane connecting empirical data points interpolated in the three-dimensional plot represents our hypothetical TTS for Agentic Deep Research. Tasks requiring deeper internal knowledge utilization (reasoning), such as solving complex math problems and logic puzzles, are better aligned with the green plane, whereas tasks demanding extensive external knowledge exploration (search), like medical QA, will align more closely with the red plane. More details for TTS on reasoning (internal knowledge), search (external knowledge), and their trade-offs along the scaling axes are elaborated in Appendix C

5 Alternative Views and Discussions

While this paper supports the Agentic Deep Research, it is essential to discuss alternative perspectives. Below, we provide a representative opposing view and our counterarguments. **More alternative view discussions can be found in Appendix D**.

Alternative view 1: Retaining human primacy in search. An alternative viewpoint contends that search should remain fundamentally a human-led activity. This counter-position emphasizes that humans with primary involvement in search is indispensable for ensuring trust, interpretability, and responsibility in open-ended inquiry tasks. Existing systems still lack robust modeling of user intent, contextual nuance, and domain-specific ethics. These elements are often essential in complex or high-stakes search scenarios such as scientific research and legal interpretation.

Response: Advocates of human-led search argue that delegating the full pipeline from information retrieval, reasoning, and synthesis to LLM agents risks introducing misaligned conclusions and trustworthy concerns. However, human labor could be prioritized in the management and decision-making process (Rong et al., 2023) to of-

fer greater transparency and traceability. Maintaining human guidance over search with human-in-the-loop (Lage et al., 2018) allows users to apply critical judgment, verify information provenance, and assume responsibility for downstream decisions, while largely improving productivity.

6 Open Problems and Opportunities

Human-in-Loop and Trustworthy This suggests that progress in AI search systems should prioritize augmenting human capabilities rather than replacing them. To build trustworthy Deep Research systems, several research questions need to be solved: *(1) Search Content Access Control.* Implementing fine-grained access control mechanisms ensures that users can only access information appropriate to their roles and permissions. This is particularly important in domains where role-based access to information is essential for privacy, regulatory confidentiality. *(2) Human Verification and Feedback Mechanisms.* Incorporating human oversight at critical stages of the AI search process can improve the accuracy and reliability of the system (Zou et al., 2025b,c). Designing systems that facilitate user feedback enables continuous learning and adaptation, aligning AI outputs more closely with user-specific expectations (Zhang et al., 2025c).

In-Domain Expert-Level Deep Research In highly specialized fields such as medicine, law, and biology, in-domain expert-level deep research is essential to ensure accuracy, relevance, and usability. Domain-specific research poses unique challenges that general-purpose agents cannot meet. We highlight some key research directions: *(1) Domain-Specific Database Construction.* In many fields, data is scattered across fragmented databases with inconsistent coverage and interfaces, e.g., in bioinformatics (Dyer et al., 2025) or medicine (Knox et al., 2024). This fragmentation hampers retrieval and reasoning. *(2) Domain-Grounded Reasoning.* Reasoning paradigms vary by field: legal reasoning relies on precedent, medical reasoning on diagnostic codes, and scientific reasoning on hypothesis testing. Agentic systems must align with these paradigms, adapting their planning and inference mechanisms to domain-specific workflows.

From Textual Space to Multi-Modality For Agentic Deep Research systems to truly emulate human research capabilities, they must transcend textual limitations and integrate diverse information modalities, including images, audio-visual content, and structured data (Bei et al., 2025). Critical research directions include: *(1) Cross-modal Semantic Alignment:* deep cross-modal semantic alignment within unified representational spaces that support context-aware reasoning across modalities (Li et al., 2025a); *(2) Information Fusion and Conflict Resolution:* robust information fusion with sophisticated conflict resolution mechanisms when modalities present inconsistent information (Bi et al., 2025); and *(3) Multi-modal Knowledge Acquisition:* efficient multi-modal knowledge retrieval and comprehension from heterogeneous, large-scale data repositories (Abootorabi et al., 2025).

Efficient Test-Time Scaling As Deep Research systems advance, operational efficiency and cost management become critical scaling determinants. Optimization must target both reasoning and search: *(1) Reasoning Efficiency*, balancing computational depth against resource utilization through techniques like capability transfer to smaller models or latent reasoning approaches (Sui et al., 2025). Such efficiency improvements potentially extend beyond mere compression of reasoning chains to manifest as higher-order intelligence through *resource self-management* (Li et al., 2025e). *(2) Search Efficiency*: Optimizing search beyond naive retrieval requires addressing scalable retrieval from massive heterogeneous sources, efficient long-context processing, and iterative search with intelligent query refinement and adaptive stopping criteria (Singh et al., 2025). Future systems should implement budget-aware strategies that dynamically adjust search workflows.

7 Conclusion

This paper traces a clear trajectory from traditional web search paradigms to the rise of Agentic Deep Research. By addressing the limitations of prior search systems and highlighting the transformative potential of iterative reasoning and reinforcement learning–driven search, we show that agentic approaches consistently surpass traditional models on complex benchmarks. By outlining key open challenges and opportunities, we position Agentic Deep Research as not only the emerging dominant paradigm, but also a fundamentally human-centered, LLM-driven advance in knowledge acquisition and synthesis.

Limitations

Despite presenting a foundational vision for the Agentic Deep Research paradigm, this paper primarily assumes reliable reasoning and search orchestration within LLMs in future research, but real-world deployment still faces challenges in trustworthiness, human-in-the-loop integration, and computational efficiency. Additionally, the current discussion focuses mainly on text-based modalities on general query settings and lacks exploration into multimodal or user-personalized search scenarios.

References

Mohammad Mahdi Abootorabi, Amirhosein Zobeiri, Mahdi Dehghani, Mohammadali Mohammadkhani, Bardia Mohammadi, Omid Ghahroodi, Mahdieh Soleymani Baghshah, and Ehsaneddin Asgari. 2025. Ask in any modality: A comprehensive survey on multimodal retrieval-augmented generation. *arXiv preprint arXiv:2502.08826*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Salaheddin Alzubi, Creston Brooks, Purva Chiniya, Edoardo Contente, Chiara von Gerlach, Lucas Irwin, Yihan Jiang, Arda Kaz, Windsor Nguyen, Sewoong Oh, et al. 2025. Open deep search: Democratizing search with open-source reasoning agents. *arXiv preprint arXiv:2503.20201*.

Maddalena Amendola, Andrea Passarella, and Raffaele Perego. 2023. Social search: Retrieving information in online social platforms–a survey. *Online Social Networks and Media*, 36:100254.

Anthropic. 2023. Claude. https://www.anthropic.com.

Anthropic. 2024. Model context protocol. https://modelcontextprotocol.io/introduction.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *arXiv preprint arXiv:2404.07738*.

Yuanchen Bei, Weizhi Zhang, Siwen Wang, Weizhi Chen, Sheng Zhou, Hao Chen, Yong Li, Jiajun Bu, Shirui Pan, Yizhou Yu, et al. 2025. Graphs meet ai agents: Taxonomy, progress, and future opportunities. *arXiv preprint arXiv:2506.18019*.

Jing Bi, Susan Liang, Xiaofei Zhou, Pinxin Liu, Junjia Guo, Yunlong Tang, Luchuan Song, Chao Huang, Guangyu Sun, Jinxi He, et al. 2025. Why reasoning matters? a survey of advancements in multimodal reasoning (v1). *arXiv preprint arXiv:2504.03151*.

Sergey Brin and Lawrence Page. 1998. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117.

Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? In *NeurIPS 2023 Workshop on Regulatable ML*.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th annual meeting of the association for computational linguistics: tutorial abstracts*, pages 34–37.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z Pan, Wen Zhang, Huajun Chen, Fan Yang, et al. 2025a. Research: Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Fan Yang, Zenan Zhou, Weipeng Chen, Haofen Wang, Jeff Z Pan, et al. 2025b. Learning to reason with search for llms via reinforcement learning. *arXiv preprint arXiv:2503.19470*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025c. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, et al. 2025. A survey on knowledge-oriented retrieval-augmented generation. *arXiv preprint arXiv:2503.10677*.

Sarah C Dyer, Olanrewaju Austine-Orimoloye, Andrey G Azov, Matthieu Barba, If Barnes, Vianey Paola Barrera-Enriquez, Arne Becker, Ruth Bennett, Martin Beracochea, Andrew Berry, et al. 2025. Ensembl 2025. *Nucleic Acids Research*, 53(D1):D948–D957.

Norbert Fuhr. 1992. Probabilistic models in information retrieval. *The computer journal*, 35(3):243–255.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Ramaneswaran S, Deepali Aneja, Zeyu Jin, Ramani Duraiswami, and Dinesh Manocha. 2024. A closer look at the limitations of instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, pages 15559–15589.

Zhengyao Gu, Henry Peng Zou, Yankai Chen, Aiwei Liu, Weizhi Zhang, and Philip S Yu. 2025. Scaling laws for many-shot in-context learning with self-generated annotations. *arXiv preprint arXiv:2503.03062*.

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

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. 2023. Reasoning with language model is planning with world model. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Gijs Hendriksen, Michael Dinzinger, Sheikh Mastura Farzana, Noor Afshan Fathima, Maik Fröbe, Sebastian Schmidt, Saber Zerhoudi, Michael Granitzer, Matthias Hagen, Djoerd Hiemstra, et al. 2024. The open web index: Crawling and indexing the web for public use. In *European Conference on Information Retrieval*, pages 130–143. Springer.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6609–6625.

Jerry Huang, Siddarth Madala, Risham Sidhu, Cheng Niu, Julia Hockenmaier, and Tong Zhang. 2025. Rag-rl: Advancing retrieval-augmented generation via rl and curriculum learning. *arXiv preprint arXiv:2503.12759*.

Xu Huang, Weiwen Liu, Xiaolong Chen, Xingmei Wang, Hao Wang, Defu Lian, Yasheng Wang, Ruiming Tang, and Enhong Chen. 2024. Understanding the planning of llm agents: A survey. *arXiv preprint arXiv:2402.02716*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7969–7992.

Zhouyu Jiang, Mengshu Sun, Lei Liang, and Zhiqiang Zhang. 2024. Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach. *arXiv preprint arXiv:2407.13101*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

Moaiad Ahmad Khder. 2021. Web scraping or web crawling: State of art, techniques, approaches and application. *International Journal of Advances in Soft Computing & Its Applications*, 13(3).

Craig Knox, Mike Wilson, Christen M Klinger, Mark Franklin, Eponine Oler, Alex Wilson, Allison Pon, Jordan Cox, Na Eun Chin, Seth A Strawbridge, et al. 2024. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275.

Isaac Lage, Andrew Ross, Samuel J Gershman, Been Kim, and Finale Doshi-Velez. 2018. Human-in-the-loop interpretability prior. *Advances in neural information processing systems*, 31.

David B Leake and Ryan Scherle. 2001. Towards context-based search engine selection. In *Proceedings of the 6th international conference on Intelligent user interfaces*, pages 109–112.

Mingxiao Li, Na Su, Fang Qu, Zhizhou Zhong, Ziyang Chen, Zhaopeng Tu, and Xiaolong Li. 2025a. Vista: Enhancing vision-text alignment in mllms via cross-modal mutual information maximization. *arXiv preprint arXiv:2505.10917*.

Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and Zhicheng Dou. 2025b. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint arXiv:2501.05366*.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025c. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.

Yangning Li, Weizhi Zhang, Yuyao Yang, Wei-Chieh Huang, Yaozu Wu, Junyu Luo, Yuanchen Bei, Henry Peng Zou, Xiao Luo, Yusheng Zhao, et al. 2025d. Towards agentic rag with deep reasoning: A survey of rag-reasoning systems in llms. *arXiv preprint arXiv:2507.09477*.

Zheng Li, Qingxiu Dong, Jingyuan Ma, Di Zhang, and Zhifang Sui. 2025e. Selfbudgeter: Adaptive token allocation for efficient llm reasoning. *arXiv preprint arXiv:2505.11274*.

Jintao Liang, Gang Su, Huifeng Lin, You Wu, Rui Zhao, and Ziyue Li. 2025. Reasoning rag via system 1 or system 2: A survey on reasoning agentic retrieval-augmented generation for industry challenges. *arXiv preprint arXiv:2506.10408*.

Greg Linden, Christopher Meek, Max Chickering, and Chris Meek. 2009. The pollution effect: Optimizing keyword auctions by favoring relevant advertising. In *Fifth workshop on Ad Auctions*.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, Zihao Wu, Lin Zhao, Dajiang Zhu, Xiang Li, Ning Qiang, Dingang Shen, Tianming Liu, and Bao Ge. 2023. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-Radiology*, 1(2):100017.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Xinbei Ma, Yeyun Gong, Pengcheng He, Hai Zhao, and Nan Duan. 2023. Query rewriting in retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5303–5315.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Kelong Mao, Zheng Liu, Hongjin Qian, Fengran Mo, Chenlong Deng, and Zhicheng Dou. 2024. Rag-studio: Towards in-domain adaptation of retrieval augmented generation through self-alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 725–735.

Fengran Mo, Kelong Mao, Ziliang Zhao, Hongjin Qian, Haonan Chen, Yiruo Cheng, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Jian-Yun Nie. 2024. A survey of conversational search. *Preprint*, arXiv:2410.15576.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford infolab.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, et al. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. 2025. Humanity's last exam. *arXiv preprint arXiv:2501.14249*.

Sonal Prabhune and Donald J Berndt. 2024. Deploying large language models with retrieval augmented generation. *arXiv preprint arXiv:2411.11895*.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Yao Rong, Tobias Leemann, Thai-Trang Nguyen, Lisa Fiedler, Peizhu Qian, Vaibhav Unhelkar, Tina Seidel, Gjergji Kasneci, and Enkelejda Kasneci. 2023. Towards human-centered explainable ai: A survey of user studies for model explanations. *IEEE transactions on pattern analysis and machine intelligence*, 46(4):2104–2122.

Abdul Malik Sami, Zeeshan Rasheed, Kai-Kristian Kemell, Muhammad Waseem, Terhi Kilamo, Mika Saari, Anh Nguyen Duc, Kari Systä, and Pekka Abrahamsson. 2024. System for systematic literature review using multiple ai agents: Concept and an empirical evaluation. *Preprint*, arXiv:2403.08399.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Zicheng Liu, and Emad Barsoum. 2025. Agent laboratory: Using llm agents as research assistants. *arXiv preprint arXiv:2501.04227*.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9248–9274.

Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. 2024. Scaling llm test-time compute optimally can be more effective than scaling model parameters. *arXiv preprint arXiv:2408.03314*.

Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang, and Ji-Rong Wen. 2025. R1-searcher: Incentivizing the search capability in llms via reinforcement learning. *arXiv preprint arXiv:2503.05592*.

Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Hanjie Chen, et al. 2025. Stop overthinking: A survey on efficient reasoning for large language models. *arXiv preprint arXiv:2503.16419*.

Hao Sun, Zile Qiao, Jiayan Guo, Xuanbo Fan, Yingyan Hou, Yong Jiang, Pengjun Xie, Fei Huang, and Yan Zhang. 2025a. Zerosearch: Incentivize the search capability of llms without searching. *arXiv preprint arXiv:2505.04588*.

Zhongxiang Sun, Qipeng Wang, Weijie Yu, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Song Yang, and Han Li. 2025b. Rearter: Retrieval-augmented reasoning with trustworthy process rewarding. *arXiv preprint arXiv:2501.07861*.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2023. Evaluating the factual consistency of large language models through news summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5220–5255.

Tongyi DeepResearch Team. 2025. Tongyi deepresearch: A new era of open-source ai researchers.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10014–10037.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023a. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Boxin Wang, Wei Ping, Lawrence Mcafee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. 2024a. Instructretro: Instruction tuning post retrieval-augmented pretraining. In *International Conference on Machine Learning*, pages 51255–51272. PMLR.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. 2023b. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Jiayi Wang, Tianyu Lu, Lin Li, Decheng Huang, et al. 2024b. Enhancing personalized search with ai: a hybrid approach integrating deep learning and cloud computing. *Journal of Advanced Computing Systems*, 4(10):1–13.

Mingzhu Wang, Yuzhe Zhang, Qihang Zhao, Junyi Yang, and Hong Zhang. 2024c. Redefining information retrieval of structured database via large language models. *arXiv preprint arXiv:2405.05508*.

Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. 2024d. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37:58118–58153.

Zheng Wang, Shu Xian Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024e. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. *arXiv preprint arXiv:2405.16420*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025a. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, et al. 2025b. Webagent-r1: Training web agents via end-to-end multi-turn reinforcement learning. *arXiv preprint arXiv:2505.16421*.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, et al. 2025a. Webwalker: Benchmarking llms in web traversal.

Junde Wu, Jiayuan Zhu, and Yuyuan Liu. 2025b. Agentic reasoning: Reasoning llms with tools for the deep research. *arXiv preprint arXiv:2502.04644*.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Minjia Zhang, Zhiyong Lu, and Aidong Zhang. 2024. Improving

retrieval-augmented generation in medicine with iterative follow-up questions. In *Biocomputing 2025: Proceedings of the Pacific Symposium*, pages 199–214. World Scientific.

Wooseong Yang, Weizhi Zhang, Yuqing Liu, Yuwei Han, Yu Wang, Junhyun Lee, and Philip S Yu. 2025. Cold-start recommendation with knowledge-guided retrieval-augmented generation. *arXiv preprint arXiv:2505.20773*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.

Jia-Yu Yao, Kun-Peng Ning, Zhen-Hui Liu, Mu-Nan Ning, Yu-Yang Liu, and Li Yuan. 2023a. Llm lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint arXiv:2310.01469*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023b. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023c. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Zhenrui Yue, Honglei Zhuang, Aijun Bai, Kai Hui, Rolf Jagerman, Hansi Zeng, Zhen Qin, Dong Wang, Xuanhui Wang, and Michael Bendersky. 2025. Inference scaling for long-context retrieval augmented generation. In *The Thirteenth International Conference on Learning Representations*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. 2022. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*.

Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Yankai Lin, and Jiaxin Mao. 2024. Usimagent: Large language models for simulating search users. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2687–2692.

Ningning Zhang, Chi Zhang, Zhizhong Tan, Xingxing Yang, Weiping Deng, and Wenyong Wang. 2025a. Credible plan-driven rag method for multi-hop question answering. *arXiv preprint arXiv:2504.16787*.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, and Xiao Huang. 2025b. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, et al. 2025c. Personaagent: When large language model agents meet personalization at test time. *arXiv preprint arXiv:2506.06254*.

Yusheng Zhao, Qixin Zhang, Xiao Luo, Weizhi Zhang, Zhiping Xiao, Wei Ju, Philip S Yu, and Ming Zhang. 2025. Dynamic text bundling supervision for zero-shot inference on text-attributed graphs. *arXiv preprint arXiv:2505.17599*.

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

Zheng Zheng, Xinyi Ni, and Pengyu Hong. 2025b. Multiple abstraction level retrieve augment generation. *arXiv preprint arXiv:2501.16952*.

Peilin Zhou, Bruce Leon, Xiang Ying, Can Zhang, Yifan Shao, Qichen Ye, Dading Chong, Zhiling Jin, Chenxuan Xie, Meng Cao, et al. 2025. Browsecomp-zh: Benchmarking web browsing ability of large language models in chinese. *arXiv preprint arXiv:2504.19314*.

Yutao Zhu, Peitian Zhang, Chenghao Zhang, Yifei Chen, Binyu Xie, Zheng Liu, Ji-Rong Wen, and Zhicheng Dou. 2024. Inters: Unlocking the power of large language models in search with instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2782–2809.

Henry Peng Zou, Zhengyao Gu, Yue Zhou, Yankai Chen, Weizhi Zhang, Liancheng Fang, Yibo Wang, Yangning Li, Kay Liu, and Philip S Yu. 2025a. Testnuc: Enhancing test-time computing approaches and scaling through neighboring unlabeled data consistency. *arXiv preprint arXiv:2502.19163*.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Yankai Chen, Chunyu Miao, Hoang Nguyen, Yue Zhou, Weizhi Zhang, Liancheng Fang, Langzhou He, et al. 2025b. A survey on large language model based human-agent systems. *arXiv preprint arXiv:2505.00753*.

Henry Peng Zou, Wei-Chieh Huang, Yaozu Wu, Chunyu Miao, Dongyuan Li, Aiwei Liu, Yue Zhou, Yankai Chen, Weizhi Zhang, Yangning Li, et al. 2025c. A call for collaborative intelligence: Why human-agent systems should precede ai autonomy. *arXiv preprint arXiv:2506.09420*.

A Taxonomy of Error Types

Here, we summarize a taxonomy of error types observed in information search problems in Table 2. Errors are grouped by the primary stage at which failure occurs, including retrieval, reasoning, and synthesis. Each category is further decomposed into representative error types with precise operational definitions. This taxonomy is used to standardize failure analysis across case studies, enabling consistent attribution of model errors to specific breakdowns.

B Deep Research Benchmarks

We followed the official benchmark settings (aligned with the intended use) of each dataset to ensure fair and consistent comparison across different LLMs and reasoning models. Both open-source and closed-source LLMs were evaluated via API calls in a single standardized run. The total computational budget was approximately USD 1.5k. For two deep research models, we report the results from official papers (Wei et al., 2025a; Zhou et al., 2025; Team, 2025) due to restricted API access to these models. Detailed descriptions of the individual benchmarks are provided below.

BrowseComp (Wei et al., 2025a) is a benchmark introduced by OpenAI to evaluate the capabilities of AI agents in conducting complex web searches. Comprising 1,266 challenging questions, it assesses an agent's ability to persistently navigate the internet to locate hard-to-find, entangled information. Unlike traditional benchmarks that focus on retrieving easily accessible facts, BrowseComp emphasizes tasks where answers are deliberately obscured, requiring agents to demonstrate advanced reasoning, strategic search planning, and adaptability. The benchmark's design ensures that while answers are difficult to discover, they are straightforward to verify, facilitating reliable evaluation of agent performance. BrowseComp serves as a critical tool for advancing research in developing AI systems capable of sophisticated information retrieval and reasoning across the web.

BrowseComp-ZH (Zhou et al., 2025) is a high-difficulty benchmark developed to evaluate the web browsing and reasoning capabilities of large language models (LLMs) within the Chinese information ecosystem. Recognizing that existing benchmarks like BrowseComp focus primarily on English-language contexts, BrowseComp-ZH addresses the unique challenges posed by the Chinese web, including linguistic complexity, fragmented information across diverse platforms, and varying search engine infrastructures. The benchmark comprises 289 multi-hop questions spanning 11 diverse domains such as film, art, medicine, geography, history, and technology. Each question is meticulously reverse-engineered from a concise, objective, and easily verifiable answer (e.g., a date, number, or proper noun). A two-stage quality control protocol ensures high question difficulty and answer uniqueness. Notably, the questions are designed so that answers are not readily retrievable via standard search engines, requiring models to engage in complex reasoning and information synthesis.

Humanity's Last Exam (HLE) (Phan et al., 2025) is a multi-modal benchmark designed to evaluate the reasoning and problem-solving capabilities of large language models (LLMs) across a broad spectrum of academic disciplines. Developed collaboratively by the Center for AI Safety and Scale AI, HLE comprises 3,000 expert-crafted questions spanning mathematics, humanities, natural sciences, and more. Unlike BrowseComp and BrowseComp-ZH, where agents can locate and answer questions by analyzing information retrieved from the web, HLE presents 'closed-book' academic challenges (the answers aren't directly available online) that demand deep reasoning and specialized domain expertise, going well beyond what surface-level online searches can support. Each question in HLE is designed to be unambiguous and verifiable, yet not readily answerable through internet search, thereby testing the intrinsic reasoning abilities of LLMs.

WebWalkerQA (Wu et al., 2025a) is a benchmark introduced at ACL 2025 to evaluate the end-to-end web navigation and multi-hop information synthesis capabilities of LLM-based agents. Unlike benchmarks such as BrowseComp that emphasize difficult query formulation or HLE that focuses on closed-book reasoning, WebWalkerQA requires models to iteratively traverse real web pages, follow hyperlinks, and aggregate evidence across multiple sources to answer a question. The tasks are grounded in dynamic, noisy web environments where relevant information is scattered and interleaved with distractors, making single-pass retrieval insufficient. This design explicitly tests long-horizon planning, adaptive browsing strate-

Error Category	Specific Error Type	Definition & Mechanism
Retrieval Failures	Shallow Retrieval (Premature Stop)	The agent executes a query and accepts surface-level results (e.g., snippets or summaries) without deeper verification, failing to uncover obscured or non-salient facts.
Retrieval Failures	Query Misalignment	The agent formulates search queries that are overly broad, underspecified, or semantically drifted from the user's precise intent, leading to the retrieval of irrelevant or noisy information.
Retrieval Failures	Fragmentation Miss	The agent fails to aggregate partial evidence distributed across multiple sources, incorrectly treating the absence of a single definitive source as a confirmed negative.
Reasoning Missteps	Hallucinated Association	The agent forces an association based on internal parametric knowledge or prior beliefs, overriding or ignoring contradictory evidence obtained from external retrieval.
Reasoning Missteps	Propagation Error (Cascading)	An error introduced in an early reasoning step (e.g., incorrect query decomposition or constraint interpretation) propagates downstream, rendering subsequent retrieval or verification steps ineffective.
Reasoning Missteps	Planning Failure	The agent fails to identify or execute the required multi-step verification strategy, often manifesting as repetitive search loops or missing necessary sequential checks for sub-constraints.
Synthesis Errors	Conflict Resolution Failure	When confronted with conflicting information across sources (e.g., inconsistent dates or statistics), the agent fails to investigate the discrepancy and instead makes an arbitrary or unverified selection.
Synthesis Errors	Context Overload / Loss	The agent retrieves excessive irrelevant information, causing dilution or loss of focus on the original detailed constraints specified in the user prompt.

Table 2: **Taxonomy of error types in agentic deep research.** Distinct light background colors indicate retrieval-, reasoning-, and synthesis-stage failures, mirroring the visual language used in prior case-study tables.

gies, and cross-page reasoning, while maintaining verifiable answers for reliable evaluation, thereby providing a more realistic and challenging setting for assessing agentic information-seeking systems.

We further evaluate different models on WebWalkerQA (Wu et al., 2025a), to compare them of multi-step web navigation and cross-page evidence synthesis in realistic, open-world settings. As shown in Table 3, standard LLMs and reasoning models perform poorly (all below 12% accuracy), highlighting their limitations in long-horizon planning and multi-hop retrieval. In contrast, Tongyi DeepResearch achieves 72.20%, yielding a dramatic performance gap (6–10× improvement). This result provides strong empirical evidence that agentic frameworks with iterative search and tool use are fundamentally more effective for complex information-seeking tasks, supporting our claim that Agentic Deep Research represents a dominant paradigm beyond pure LLM inference.

Model	Accuracy (%)
Claude-3.5-Sonnet	1.47
Claude-3.7-Sonnet	4.56
GPT4o	6.47
O1	6.32
Gemini-2.0-Flash	6.62
Gemini-2.5-Pro	11.03
Tongyi DeepResearch	**72.20**

Table 3: Performance comparison on the WebWalkerQA benchmark.

C TTS for Agentic Deep Research

C.1 Reasoning: TTS of Internal Knowledge Utilization

Recent studies have shown that LLMs do more than recall memorized facts or patterns. When given extra computational resources to "think", they can also perform deeper reasoning steps. This pattern, known as the test-time scaling law for reasoning, indicates that as a model takes more inference steps,

including building longer chains of thought (Chen et al., 2025c), iterative self-refinement (Madaan et al., 2023), or self-consistency decoding (Hao et al., 2023), its accuracy in complex tasks steadily improves. This phenomenon also suggests that the internal knowledge embedded in LLMs is not fully exposed in a single forward pass. Instead, a deeper understanding can be progressively uncovered through extended inference. In the context of Agentic Deep Research, this highlights a key shift from producing one-shot answers to engaging in active, multi-step reasoning. Instead of relying solely on external retrieval or prompting tricks, the model's internal knowledge becomes a reusable and expandable resource. This inference-time flexibility is crucial for handling complex, open-ended queries, and positions reasoning as a core axis of scalable capability in LLM-based research agents (Huang et al., 2024).

C.2 Search: TTS of External Knowledge Exploration

In addition to making full use of the internal knowledge of LLMs, exploring abundant external knowledge effectively is another key to achieving TTS in Agentic Deep Research systems. When performing external knowledge search, it is often difficult to obtain all the important information in a single-step retrieval (Shao et al., 2023; Jiang et al., 2024). To overcome this bottleneck, consistent with the agentic RAG insights, existing work mainly explores the scaling potential of the search phase through *iterative* search and *long-context* RAG. Representatively, many works have explored improving the test-time retrieval performance through iterative RAG/search, which introduces a dynamic and multi-step retrieval to knowledge search via task decomposition (Trivedi et al., 2023; Asai et al., 2023; Xiong et al., 2024). They generally show that iterative multi-step retrieval under *a proper number of iterations* can also enhance RAG's performance. Yue et al. (2025) recently further observed an important experimental phenomenon that under optimal inference parameters, the performance improves nearly linearly with increasing test-time computation. The gradual expansion of knowledge retrieval from split local text chunks to an increasingly precise retrieval executed within nearly global knowledge bases would become a more powerful development trend in the test-time search process.

C.3 TTS Trade-offs in Search and Reasoning

Agentic Deep Research systems integrate both search and reasoning, each consuming part of a limited token budget. Under such constraints, a natural trade-off emerges: allocating more tokens to search (e.g., issuing broader or more detailed queries) reduces the capacity available for reasoning (e.g., multi-hop inference or synthesis), and vice versa. This balance is task-dependent. Search-heavy tasks such as multi-hop RAG (Xiong et al., 2024; Shao et al., 2023) or literature surveys prioritize broader content access, while reasoning-heavy tasks like causal analysis or math verification (Snell et al., 2024) require deeper internal processing. Building systems that adaptively allocate token budgets between search and reasoning based on task characteristics is critical for maximizing effectiveness and efficiency in Agentic Deep Research. We anticipate the emergence of a test-time scaling law that governs the optimal balance between search and reasoning under different task conditions. Furthermore, training models that can dynamically allocate and manage budget across these two components is a defining capability of next-generation deep research systems.

D Alternative Views and Discussions

Alternative view 2: Search infrastructure is not suitable for agents. Some argue that the existing information search infrastructure that is built around the human query process, proprietary data indexing, and legacy UI is fundamentally mismatched for deep-research agents. Unlike human users, who can flexibly interpret context and click through interfaces, agents struggle to operate effectively when content is fragmented, unstandardized, or locked behind proprietary system protections.

Response: We acknowledge these challenges but stress that the infrastructure itself is rapidly evolving to meet agentic needs. New interoperability standards like the Model Context Protocol (MCP) (Anthropic, 2024) are explicitly designed to allow LLM agents to connect with tools, APIs, and secure data sources in a structured, machine-readable way. Meanwhile, researchers are actively reformatting knowledge domains into contexts more suitable for LLM consumption. For example, web retrieval APIs (Wang et al., 2024c) and more structured databases (tabular data and knowledge graph) explicitly designed for LLMs to access. These developments suggest that while the legacy

web infrastructure has limitations, the ecosystem is already shifting toward protocols and contexts that enable Agentic Deep Research at scale.

Alternative view 3: Computation costs are expensive. A further concern is that coupling deep reasoning with search significantly increases inference costs, both economically and environmentally. This is not economical for many users if it dramatically increases computational overhead in their daily usage. Additionally, the substantial token overhead enlarges computational expenditure in general, making it environmentally unsustainable.

Response: We would like to emphasize three key points to address the concerns: (i) The computational cost of agentic deep research dynamically adapts to the difficulty of the task at hand, ensuring that inference is extremely efficient by scaling resources only as needed and avoiding unnecessary overhead for simpler queries. (ii) The ultimate goal is to save user time by significantly boosting productivity and reducing human effort in complex information seeking. While each stage of technology evolution may increase energy consumption, this increase is justified by the substantial savings in human cognitive load and manual labor. (iii) The growing adoption of environmentally friendly energy sources, such as nuclear micro-reactors, wind power, and solar energy, helps sustainably support the energy demands of LLM agents, mitigating their environmental impact.

E Open-Source Deep Research Implementations Overview

Table 4 presents a comprehensive overview of recent open-source deep research implementations that integrate reasoning and search capabilities across different model families. The listed implementations employ diverse base models, including Qwen, Llama, DeepSeek, OpenAI series, and Gemini, and adopt varying optimization paradigms such as prompt-based and reinforcement learning (RL)-based strategies. Early frameworks like Agentic Reasoning (Wu et al., 2025b) and Search-o1 (Li et al., 2025b) rely primarily on prompting, whereas later works such as Search-R1 (Jin et al., 2025) and ZeroSearch (Sun et al., 2025a), incorporate RL training on datasets like HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020). Some systems, such as Webthinker (Li et al., 2025c), further expand evaluation to more challenging tasks (e.g., Humanity's Last Exam). Meanwhile, community projects like gpt-researcher, deep-searcher, and node-DeepResearch focus on prompt-based orchestration across multiple model families (DeepSeek, Claude, Gemini, etc.) without dedicated training data. Overall, the table illustrates a clear evolution from prompt-based prototypes to RL-driven search-reasoning agents operating over increasingly challenging and diverse benchmarks.

F Related Work

In this section we discuss our work within the broader literature on retrieval-augmented generation (RAG) and search domain, especially related agentic RAG survey. While a substantial body of prior survey paper (Singh et al., 2025; Liang et al., 2025) has explored RAG in LLM agents, most of these efforts assume access to restricted or static knowledge sources and focus on improving relevance, faithfulness, or reasoning depth within closed retrieval settings. Recent surveys on agentic RAG (Singh et al., 2025) and Reasoning RAG (Liang et al., 2025) systematically organize methods that enhance retrieval and reasoning. However, their scope is largely limited to retrieval from local corpora. In contrast, our paper focuses the Agentic Deep Research, which involves navigating the open web, observation-driven reasoning, and performing interleaved reasoning-search information synthesis. Zhang et al. (2025b) further categorize agents into recommendation- and search-oriented paradigms, without engaging with systems designed for deep, iterative research. In contrast, our work advances a forward-looking position on Agentic Deep Research as a distinct paradigm in which agents navigate the open web, interleave observation-driven reasoning with adaptive search, and synthesize evidence across multiple steps. Rather than offering an exhaustive retrospective survey, we aim to articulate a conceptual roadmap for this emerging direction, which is critical for future real-world information seeking tasks.

Name	Base Model	Optimization	Training Data	Evaluation Data	Link	Stars
Search-R1 (Jin et al., 2025)	Llama, Qwen	RL	NQ, HotpotQA	NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, Bamboogle	Link	3.7K
DeepResearcher (Zheng et al., 2025a)	Qwen	RL	NQ, TQ, HotpotQA, 2WikiMultiHopQA	MuSiQue, Bamboogle, PopQA, NQ, TQ, HotpotQA, 2WikiMultiHopQA	Link	0.7K
R1-Searcher (Song et al., 2025)	Llama, Qwen	RL	HotpotQA, 2WikiMultiHopQA	HotpotQA, 2WikiMultiHopQA, MuSiQue, Bamboogle	Link	0.7K
ReSearch (Chen et al., 2025a)	Qwen	RL	MuSiQue	HotpotQA, 2WikiMultiHopQA, MuSiQue, Bamboogle	Link	1.3K
ZeroSearch (Sun et al., 2025a)	Llama, Qwen	RL	NQ, HotpotQA	NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, Bamboogle	Link	1.2K
Webthinker (Li et al., 2025c)	Qwen	RL	SuperGPQA, WebWalkerQA, OpenThoughts, NaturalReasoning, NuminaMath	GPQA, GAIA, WebWalkerQA, Humanity's Last Exam	Link	1.4K
Agentic Reasoning (Wu et al., 2025b)	N/A	Prompt	N/A	GPQA	Link	0.7K
Search-o1 (Li et al., 2025b)	Qwen	Prompt	N/A	GPQA, MATH500, AMC2023, AIME2024, LiveCodeBench, Natural Questions, TriviaQA, HotpotQA, 2Wiki, MuSiQue, Bamboogle	Link	1.1K
Open Deep Search (Alzubi et al., 2025)	DeepSeek, Llama	Prompt	N/A	SimpleQA, FRAME	Link	3.8K
gpt-researcher	OpenAI Series	Prompt	N/A	N/A	Link	24.6K
deep-searcher	DeepSeek, OpenAI Series, Claude, Gemini, Grok, Qwen, Llama, GLM	Prompt	N/A	N/A	Link	7.3K
nanoDeepResearch	OpenAI Series, Claude	Prompt	N/A	N/A	Link	0.2K
DeerFlow	OpenAI Series, Qwen	Prompt	N/A	N/A	Link	18.9K
deep-research	DeepSeek, OpenAI Series	Prompt	N/A	N/A	Link	18.3K
open-deep-research	OpenAI Series, DeepSeek, Claude, Gemini	Prompt	N/A	N/A	Link	2.1K
r1-reasoning-rag	DeepSeek	Prompt	N/A	N/A	Link	0.3K
node-DeepResearch	Gemini, OpenAI Series	Prompt	N/A	N/A	Link	5.0K
deep-research	Gemini, OpenAI Series, DeepSeek, Claude, Grok	Prompt	N/A	N/A	Link	4.3K

Table 4: Overview of open-source deep research implementations. All open-source implementations are released in year of 2025, and star number collected by Dec 31, 2025.