

A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications

RENJUN XU* and JINGWEN PENG, Zhejiang University, China

This survey examines the rapidly evolving field of Deep Research systems—AI-powered applications that automate complex research workflows through the integration of large language models, advanced information retrieval, and autonomous reasoning capabilities. We analyze more than 80 commercial and non-commercial implementations that have emerged since 2023, including OpenAI/DeepResearch, Gemini/DeepResearch, Perplexity/DeepResearch, and numerous open-source alternatives. Through comprehensive examination, we propose a novel hierarchical taxonomy that categorizes systems according to four fundamental technical dimensions: foundation models and reasoning engines, tool utilization and environmental interaction, task planning and execution control, and knowledge synthesis and output generation. We explore the architectural patterns, implementation approaches, and domain-specific adaptations that characterize these systems across academic, scientific, business, and educational applications. Our analysis reveals both the significant capabilities of current implementations and the technical and ethical challenges they present regarding information accuracy, privacy, intellectual property, and accessibility. The survey concludes by identifying promising research directions in advanced reasoning architectures, multimodal integration, domain specialization, human-AI collaboration, and ecosystem standardization that will likely shape the future evolution of this transformative technology. By providing a comprehensive framework for understanding Deep Research systems, this survey contributes to both the theoretical understanding of AI-augmented knowledge work and the practical development of more capable, responsible, and accessible research technologies. The paper resources can be viewed at <https://github.com/scienceaix/deepresearch>.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence**; *Natural language processing*; • **Computer systems organization** → *Embedded and cyber-physical systems*; • **Information systems** → *Information retrieval*; • **Human-centered computing** → *Collaborative and social computing*.

Additional Key Words and Phrases: Deep Research, Large Language Models, Autonomous Agents, AI Systems, Research Automation, Information Retrieval, Knowledge Synthesis, Human-AI Collaboration, Multi-Agent Systems, Tool-Using Agents

*Corresponding author: rux@zju.edu.cn

Authors' Contact Information: Renjun Xu; Jingwen Peng, Zhejiang University, China.

CONTENTS

Abstract	1
Contents	2
1 Introduction	4
1.1 Definition and Scope of Deep Research	4
1.2 Historical Context and Technical Evolution	5
1.3 Significance and Practical Implications	6
1.4 Research Questions and Contribution of this Survey	7
2 The Evolution and Technical Framework of Deep Research	7
2.1 Foundation Models and Reasoning Engines: Evolution and Advances	7
2.2 Tool Utilization and Environmental Interaction: Evolution and Advances	10
2.3 Task Planning and Execution Control: Evolution and Advances	11
2.4 Knowledge Synthesis and Output Generation: Evolution and Advances	13
3 Comparative Analysis and Evaluation of Deep Research Systems	14
3.1 Cross-Dimensional Technical Comparison	14
3.2 Application-Based System Suitability Analysis	16
3.3 Performance Metrics and Benchmarking	18
4 Implementation Technologies and Challenges	21
4.1 Architectural Implementation Patterns	21
4.2 Infrastructure and Computational Optimization	27
4.3 System Integration and Interoperability	29
4.4 Technical Challenges and Solutions	33
5 Evaluation Methodologies and Benchmarks	35
5.1 Functional Evaluation Frameworks	35
5.2 Non-Functional Evaluation Metrics	37
5.3 Cross-Domain Evaluation Benchmarks	39
5.4 Emerging Evaluation Approaches	41
5.5 Comparative Evaluation Methodology	42
6 Applications and Use Cases	44
6.1 Academic Research Applications	44
6.2 Scientific Discovery Applications	46
6.3 Business Intelligence Applications	49
6.4 Financial Analysis Applications	51
6.5 Educational Applications	52
6.6 Personal Knowledge Management Applications	54
7 Ethical Considerations and Limitations	56
7.1 Information Accuracy and Hallucination Concerns	56
7.2 Privacy and Data Security	59
7.3 Source Attribution and Intellectual Property	61
7.4 Accessibility and Digital Divide	63

A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications	3
8 Future Research Directions	64
8.1 Advanced Reasoning Architectures	64
8.2 Multi-Modal Deep Research	68
8.3 Domain-Specific Optimization	70
8.4 Human-AI Collaboration and Standardization	72
9 Conclusion	76
9.1 Key Findings and Contributions	76
9.2 Limitations and Outlook	78
9.3 Broader Implications	79
9.4 Final Thoughts	80
References	81

1 Introduction

Rapid advancement of artificial intelligence has precipitated a paradigm shift in how knowledge is discovered, validated, and utilized across academic and industrial domains. Traditional research methodologies, reliant on manual literature reviews, experimental design, and data analysis, are increasingly supplemented—and in some cases supplanted—by intelligent systems capable of automating end-to-end research workflows. This evolution has given rise to a novel domain we term “Deep Research”, which signifies the convergence of large language models (LLMs), advanced information retrieval systems, and automated reasoning frameworks to redefine the boundaries of scholarly inquiry and practical problem-solving.

1.1 Definition and Scope of Deep Research

Deep Research refers to the systematic application of AI technologies to automate and enhance research processes through three core dimensions:

- (1) **Intelligent Knowledge Discovery:** Automating literature search, hypothesis generation, and pattern recognition across heterogeneous data sources
- (2) **End-to-End Workflow Automation:** Integrating experimental design, data collection, analysis, and result interpretation into unified AI-driven pipelines
- (3) **Collaborative Intelligence Enhancement:** Facilitating human-AI collaboration through natural language interfaces, visualizations, and dynamic knowledge representation

To clearly delineate the boundaries of Deep Research, we distinguish it from adjacent AI systems as follows:

自主工作流程能力、专业研究工具以及端到端的研究协调能力

- **Differentiating from General AI Assistants:** While general AI assistants like ChatGPT can answer research questions, they lack the autonomous workflow capabilities, specialized research tools, and end-to-end research orchestration that define Deep Research systems. Recent surveys have highlighted this crucial distinction between specialized research systems and general AI capabilities [73, 76], with particular emphasis on how domain-specific tools fundamentally transform research workflows compared to general-purpose assistants [213, 318]. 专业工具集成与协调
- **Differentiating from Single-Function Research Tools:** Specialized tools like citation managers, literature search engines, or statistical analysis packages address isolated research functions but lack the integrated reasoning and cross-functional orchestration of Deep Research systems. Tools like scispace [242] and You.com [313] represent earlier attempts at research assistance but lack the end-to-end capabilities that define true Deep Research systems.
- **Differentiating from Pure LLM Applications:** Applications that simply wrap LLMs with research-oriented prompts lack the environmental interaction, tool integration, and workflow automation capabilities that characterize true Deep Research systems. 环境交互、工具集成和工作流自动化功能

This survey specifically examines systems that exhibit at least two of the three core dimensions, with a focus on those incorporating large language models as their foundational reasoning engine. Our scope encompasses commercial offerings such as OpenAI/DeepResearch [197], Google’s Gemini/DeepResearch [89], and Perplexity/DeepResearch [209], alongside open-source implementations including dzhng/deep-research [321], HKUDS/Auto-Deep-Research [112], and numerous others detailed in subsequent sections. We exclude purely bibliometric tools or single-stage automation systems lacking integrated cognitive capabilities,

such as research assistance tools like `Elicit` [74], `ResearchRabbit` [228], `Consensus` [63], or citation tools like `Scite` [243]. Additional specialized tools like `STORM` [278], which focuses on scientific text retrieval and organization, are valuable but lack the end-to-end deep research capabilities central to our survey scope.

1.2 Historical Context and Technical Evolution

The trajectory of Deep Research can be mapped through three evolutionary stages that reflect both technological advancements and implementation approaches:

1.2.1 Origin and Early Exploration (2023 - February 2025). It should be noted that workflow automation frameworks like `n8n` [183], `QwenLM/Qwen-Agent` [224], etc. had already been in existence long before the boom of deep research. Their early establishment demonstrated the pre-existing groundwork in related technological domains, highlighting that the development landscape was not solely shaped by the emergence of deep research, but had a more diverse and earlier-rooted origin. The concept of Deep Research emerged from the shift of AI assistants towards intelligent agents. In December 2024, Google Gemini pioneered this functionality with its initial Deep Research implementation, focusing on basic multi-step reasoning and knowledge integration [60]. This phase laid the groundwork for subsequent advancements, setting the stage for more sophisticated AI-driven research tools. Many of these advances built upon earlier workflow automation tools like `n8n` [183] and agent frameworks such as `AutoGPT` [250] and `BabyAGI` [311] that had already established foundations for autonomous task execution. Other early contributions to this ecosystem include `cline2024` [61], which pioneered integrated research workflows, and `open_operator` [36], which developed foundational browser automation capabilities essential for web-based research.

多步推理
知识集成

1.2.2 Technological Breakthrough and Competitive Rivalry (February - March 2025). The rise of DeepSeek's open-source models [68] revolutionized the market with efficient reasoning and cost-effective solutions. In February 2025, OpenAI's release of Deep Research, marked a significant leap forward [197]. Powered by the o3 model, it demonstrated advanced capabilities such as autonomous research planning, cross-domain analysis, and high-quality report generation, achieving accuracy rates exceeding previous benchmarks in complex tasks. Concurrently, Perplexity launched its free-to-use Deep Research in February 2025 [209], emphasizing rapid response and accessibility to capture the mass market. Open-source projects such as `nickscamara/open-deep-research` [42], `mshumer/OpenDeepResearcher` [249], `btahir_open_deep_research` [37], and `GPT-researcher` [16] emerged as community-driven alternatives to commercial platforms. The ecosystem continued to expand with lightweight implementations like `Automated-AI-Web-Researcher-Ollama` [267], designed for local execution with limited resources, and modular frameworks such as `Langchain-AI/Open_deep_research` [131] that provided composable components for custom research workflows.

1.2.3 Ecosystem Expansion and Multi-modal Integration (March 2025 - Present). The third stage is characterized by the maturation of a diverse ecosystem. Open-source projects like `Jina-AI/node-DeepResearch` [121] enable localized deployment and customization, while commercial closed-source versions from OpenAI and Google continue to push boundaries with multi-modal support and multi-agent collaboration capabilities. The integration of advanced search technologies and report generation frameworks further enhances the tool's utility across academic research, financial analysis, and other fields. Meanwhile, platforms like `Manus` [164] and `AutoGLM-Research` [330], `MGX` [171], and `Devin` [62] are incorporating advanced AI research capabilities

某些平台整合先进的人工智能研究功能增强其服务

to enhance their services. Concurrently, Anthropic launched **Claude/Research** [13] in April 2025, introducing agentic search capabilities that systematically explore multiple angles of queries and deliver comprehensive answers with verifiable citations. Agent frameworks such as **OpenManus** [193], **Camel-AI/OWL** [43], and **TARS** [39] further expand the ecosystem with specialized capabilities and domain-specific optimizations.

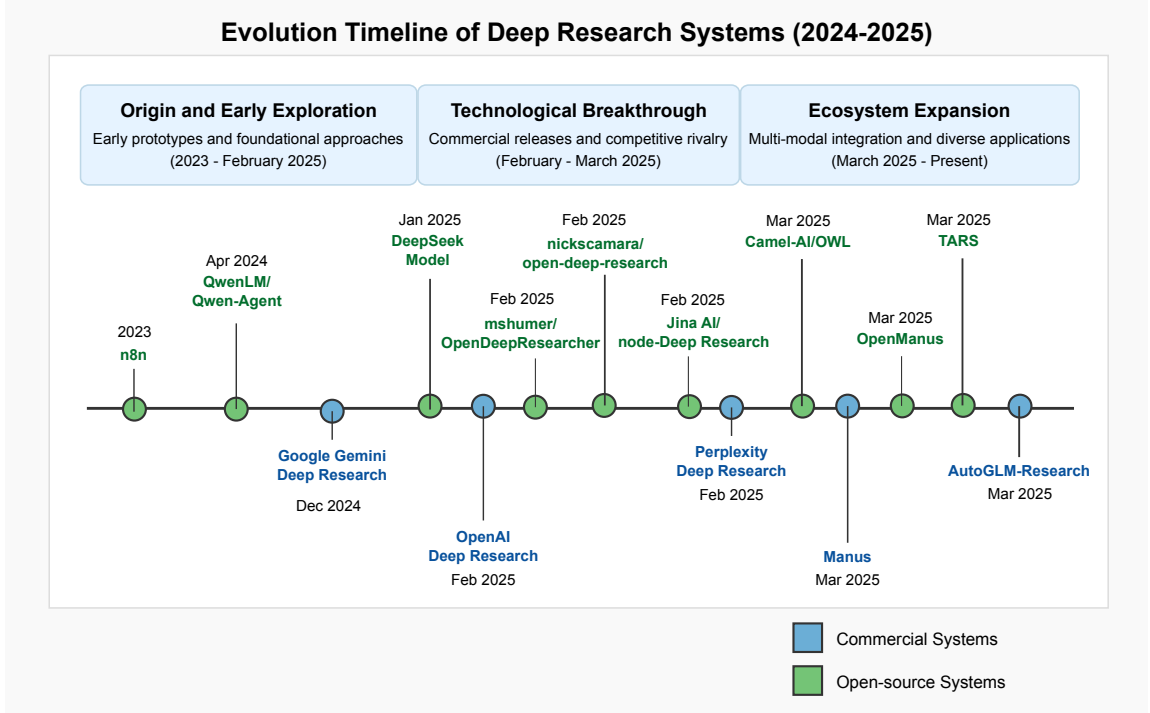


Fig. 1. Evolution Timeline of Deep Research Systems

1.3 Significance and Practical Implications

Deep Research demonstrates transformative potential across multiple domains:

- (1) **Academic Innovation:** Accelerating hypothesis validation through automated literature synthesis (e.g., HotpotQA [307] performance benchmarks) and enabling researchers to explore broader interdisciplinary connections that might otherwise remain undiscovered. The transformative potential of Deep Research extends beyond individual applications to fundamentally reshape scientific discovery processes. As Sourati and Evans [256] argue, human-aware artificial intelligence can significantly accelerate science by augmenting researchers' capabilities while adapting to their conceptual frameworks and methodological approaches. This human-AI synergy represents a fundamental shift from traditional automation toward collaborative intelligence that respects and enhances human scientific intuition. Complementary work by Khalili and Bouchachia [128] further demonstrates how systematic approaches to building science discovery machines can transform hypothesis generation, experimental design, and theory refinement through integrated AI-driven research workflows.

- (2) **Enterprise Transformation:** Enabling data-driven decision-making at scale through systems like `Agent-RL/ReSearch` [2] and `smolagents/open_deep_research` [115] that can analyze market trends, competitive landscapes, and strategic opportunities with unprecedented depth and efficiency.
- (3) **Democratization of Knowledge:** Reducing barriers to entry through open-source implementations like `grapeot/deep_research_agent` [263] and `OpenManus` [193], making sophisticated research capabilities accessible to individuals and organizations regardless of technical expertise or resource constraints.

1.4 Research Questions and Contribution of this Survey

This survey addresses three fundamental questions:

框架影响

- (1) How do **architectural choices** (system architecture, implementation approach, functional capabilities) impact Deep Research effectiveness? 出现的创新
- (2) What technical **innovations** have emerged in LLM fine-tuning, retrieval mechanisms, and workflow orchestration across the spectrum of Deep Research implementations? 现有系统
- (3) How do existing systems **balance performance, usability, and ethical considerations**, and what patterns emerge from comparing **approaches** like those of `n8n` [183] and `OpenAI/AgentsSDK` [199]?

Our contributions manifest in three dimensions:

- (1) **Methodological:** Proposing a novel taxonomy categorizing systems by their technical architecture, from foundation models to knowledge synthesis capabilities
- (2) **Analytical:** Conducting comparative analysis of representative systems across evaluation metrics, highlighting the strengths and limitations of different approaches
- (3) **Practical:** Identifying key challenges and formulating a roadmap for future development, with specific attention to emerging architectures and integration opportunities

The remainder of this paper follows a structured exploration beginning with conceptual frameworks (Section 2), technical innovations and comparative analysis (Sections 3-4), implementation technologies (Section 5), evaluation methodologies (Section 6), applications and use cases (Section 7), ethical considerations (Section 8), and future directions (Section 9).

2 The Evolution and Technical Framework of Deep Research

This section presents a comprehensive technical taxonomy for understanding Deep Research systems, organized around four fundamental technological capabilities that define these systems. For each capability, we examine the evolutionary **trajectory** and **technical innovations** while highlighting representative **implementations** that exemplify each approach. 轨迹, 创新, 实现

2.1 Foundation Models and Reasoning Engines: Evolution and Advances

The foundation of Deep Research systems lies in their underlying AI models and reasoning capabilities, which have evolved from general-purpose language models to specialized research-oriented architectures.

研究专业模型

2.1.1 From General-Purpose LLMs to Specialized Research Models. The progression from general LLMs to research-specialized models represents a fundamental shift in deep research capabilities:

Hierarchical Technical Framework of Deep Research Systems

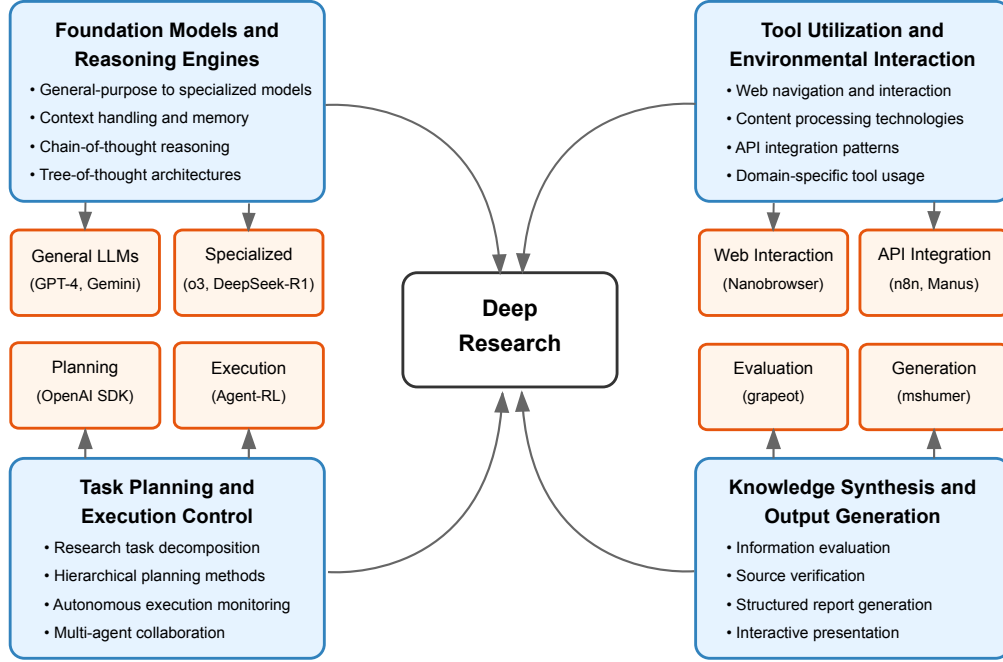


Fig. 2. Hierarchical Technical Framework of Deep Research Systems

Technical Evolution Trajectory. Early implementations relied on general-purpose LLMs with minimal task-specific optimization. Current systems feature models specifically enhanced for research tasks through architectural modifications, specialized training corpora, and fine-tuning regimes focused on analytical and reasoning capabilities. The transition from models like GPT-4 to OpenAI's o3 demonstrates significant improvements in abstraction, multi-step reasoning, and knowledge integration capabilities essential for complex research tasks [198, 200].

Representative Systems. OpenAI/DeepResearch [197] exemplifies this evolution with its o3-based model optimized specifically for web browsing and data analysis. The system leverages chain-of-thought and tree-of-thought reasoning techniques to navigate complex information landscapes. Google's Gemini/DeepResearch [60] similarly employs Gemini 2.5 Pro with enhanced reasoning capabilities and a million-token context window to process extensive information. These approaches build upon foundational work in reasoning enhancement techniques like chain-of-thought prompting [291], self-consistency [287], and human preference alignment [205] that have been adapted specifically for research-intensive tasks. In the open-source domain, AutoGLM-Research [330] demonstrates how specialized training regimes can optimize existing models like

ChatGLM for research-intensive tasks, achieving significant performance gains through targeted enhancements to reasoning components.

上下文和内存机制

2.1.2 Context Understanding and Memory Mechanisms. The ability to process, retain, and utilize extensive contextual information represents a crucial advancement in Deep Research systems:

Technical Evolution Trajectory. Early systems struggled with limited context windows, hampering their ability to synthesize information from multiple sources. Contemporary implementations employ sophisticated memory management techniques including episodic buffers, hierarchical compression, and attention-based retrieval mechanisms that extend effective context far beyond model limitations. The million-token context windows of models like Grok 3 [299] and Gemini 2.5 Pro [60], along with the context optimization in OpenAI's o3 model [195], have dramatically expanded the information processing capabilities of these systems. Advanced systems now distinguish between working memory (active reasoning context) and long-term memory (knowledge repository), allowing for more human-like research processes.

Representative Systems. Perplexity/DeepResearch [209] has pioneered efficient context processing by leveraging DeepSeek-R1's capabilities while implementing proprietary mechanisms for structured information management. The system can analyze hundreds of sources while maintaining coherent reasoning threads. Similarly, Camel-AI/OWL [43] employs an innovative open-weight approach to memory management, allowing for dynamic allocation of attention resources based on information relevance and task requirements. Both systems demonstrate how effective memory architectures can significantly enhance research performance even with comparable base model capabilities.

内存架构也能提升性能

2.1.3 Enhancements in Reasoning Capabilities. Advanced reasoning mechanisms distinguish modern Deep Research systems from conventional LLM applications:

Technical Evolution Trajectory. Early implementations relied primarily on zero-shot or few-shot prompting for reasoning tasks. Current systems integrate explicit reasoning frameworks including chain-of-thought, tree-of-thought, and graph-based reasoning architectures. Recent work by Lang et al. [132] demonstrates how debate-driven reasoning can facilitate weak-to-strong generalization, enabling more robust performance on complex research tasks through structured argumentative processes. These approaches implement reasoning patterns that more closely mirror human scientific discourse, with explicit representation of alternative viewpoints and structured evaluation of competing hypotheses. Advanced implementations like OpenAI's o3 incorporate self-critique, uncertainty estimation, and recursive reasoning refinement [198, 200]. This evolution enables increasingly sophisticated forms of evidence evaluation, hypothesis testing, and knowledge synthesis essential for high-quality research outputs.

辩论驱动等 更接近人类科学

专用工具包集成和模块化推理框架

Representative Systems. QwenLM/Qwen-Agent [224] exemplifies advanced reasoning capabilities through its specialized toolkit integration and modular reasoning framework. The system employs a multi-stage reasoning process with explicit planning, information gathering, analysis, and synthesis phases optimized for research workflows. Similar capabilities are evident in smolagents/open_deep_research [115], which implements a flexible reasoning architecture that can adapt to different research domains and methodologies. Systems like CycleResearcher [294] demonstrate how integrating automated review processes into research workflows can enhance accuracy through structured feedback loops. These approaches implement explicit verification steps

多阶段推理

自动化审核

that identify potential errors and inconsistencies before generating final research outputs. The application of AI to complex domains like mathematics further illustrates this progress, where models are increasingly viewed from a cognitive science perspective to enhance their reasoning abilities [320], achieving notable milestones such as silver-medal standards in solving International Mathematical Olympiad problems [7]. These systems highlight how reasoning enhancements can dramatically improve research quality even without requiring the largest or most computationally intensive base models.

2.2 Tool Utilization and Environmental Interaction: Evolution and Advances

Deep Research systems must effectively interact with external environments to gather and process information, representing a fundamental capability beyond core language model functions[144].

能力超越基础模型

2.2.1 Web Interaction Technology Development. The ability to navigate and extract information from the web represents a foundational capability for deep research:

之前的只是简单搜索，无多少交互

Technical Evolution Trajectory. Initial implementations relied on simple API-based search queries with limited interaction capabilities. Current systems employ sophisticated web navigation including dynamic content handling, authentication management, and interactive element manipulation. Advanced implementations feature semantic understanding of web structures, allowing for adaptive information extraction and multi-page navigation flows. This evolution has dramatically expanded access to web-based information sources and the ability to extract insights from complex web environments.

Representative Systems. Nanobrowser [184] represents a purpose-built browser environment designed specifically for AI agent use, offering optimized rendering and interaction capabilities for research tasks. It enables fine-grained control of web navigation while maintaining security and performance. Similarly, AutoGLM [330] demonstrates sophisticated GUI interaction capabilities across both web and mobile interfaces, allowing it to access information through interfaces designed for human use. These systems showcase how specialized web interaction technologies can significantly expand the information gathering capabilities of Deep Research systems.

2.2.2 Content Processing Technology Advancements. Beyond basic navigation, the ability to process diverse content formats is crucial for comprehensive research:

之前只能提取html'文本，现在可根据格式提取关键信息

Technical Evolution Trajectory. Early systems were limited primarily to text extraction from HTML sources. Modern implementations support multi-modal content processing including structured data tables, embedded visualizations, PDF documents, and interactive applications. Advanced systems like those built on OpenAI's o3 can extract semantic structure from unstructured content, identify key information from diverse formats, and integrate insights across modalities [201]. This evolution has dramatically expanded the range of information sources that can be incorporated into research processes.

不同文档不同处理

Representative Systems. The dzhng/deep-research [321] project exemplifies advanced content processing through its specialized modules for different document types and formats. It implements custom extraction logic for academic papers, technical documentation, and structured data sources. Similarly, nickscamara/open-deep-research [42] features sophisticated content normalization pipelines that transform diverse formats into consistent knowledge representations suitable for analysis. Both systems demonstrate how

将内容规范化，提高一致认知性

specialized content processing can significantly enhance the quality and comprehensiveness of research outputs.

2.2.3 Specialized Tool Integration Progress. Integration with domain-specific tools extends Deep Research capabilities beyond general information processing:

Technical Evolution Trajectory. Initial systems relied on general-purpose web search and basic API integrations. The integration of diverse tools has been dramatically advanced by frameworks like ToolLLM [222], which enables large language models to master over 16,000 real-world APIs, significantly expanding the interaction capabilities of research systems. Similarly, AssistGPT [82] demonstrates how general multi-modal assistants can plan, execute, inspect, and learn across diverse environments, creating unified research experiences that seamlessly incorporate varied information sources and interaction modalities. LLaVA-Plus [152] further extends these capabilities through explicit tool learning mechanisms, enabling research assistants to adaptively incorporate specialized tools within multimodal workflows. Current implementations feature complex toolchains including specialized databases, analytical frameworks, and domain-specific services. Advanced systems dynamically select and orchestrate tools based on research requirements, effectively composing custom research workflows from available capabilities. Some implementations like those leveraging OpenAI's Codex [194] can even generate custom code to process research data or implement analytical models on demand, further extending analytical capabilities. This evolution has enabled increasingly sophisticated analysis and domain-specific research applications.

工具集合

工具流自动化, web间交互

Representative Systems. Manus [164] exemplifies sophisticated tool orchestration through its extensive API integration framework and tool selection mechanisms. The system can incorporate domain-specific research tools and services into unified workflows, significantly expanding its analytical capabilities. Similarly, n8n [183] provides a flexible workflow automation platform that can be configured for research tasks, allowing for integration with specialized data sources and analytical services. Steward extends web interaction capabilities by implementing natural language-driven navigation and operation across websites, overcoming scalability limitations of traditional automation frameworks while maintaining low operational costs [261]. These systems highlight how tool integration can extend Deep Research capabilities into specialized domains and complex analytical workflows.

2.3 Task Planning and Execution Control: Evolution and Advances

Effective research requires sophisticated planning and execution mechanisms to coordinate complex, multi-stage workflows.

规划与执行

2.3.1 Research Task Planning Development. The ability to decompose research objectives into manageable tasks represents a fundamental advancement:

早期简单的任务分解与线性执行

Technical Evolution Trajectory. Early approaches employed simple task decomposition with linear execution flows, similar to those found in early agent frameworks like MetaGPT [111] and AgentGPT [230]. Modern systems implement hierarchical planning with dynamic refinement based on intermediate results and discoveries. Advanced planning approaches increasingly incorporate structured exploration methodologies to navigate complex solution spaces efficiently. AIDE [120] demonstrates how tree search algorithms can

树形：计算资源换性能

effectively explore the space of potential code solutions for machine learning engineering, trading computational resources for enhanced performance through strategic reuse and refinement of promising pathways. Advanced implementations incorporate resource-aware planning, considering time constraints, computational limitations, and information availability. However, incorporating AI tools for tasks like automated code review has been observed to increase pull request closure durations despite benefits, as evidenced in studies such as Cihan et al. [59], highlighting the critical need to account for temporal impacts in such resource-aware systems. This evolution has enabled increasingly sophisticated research strategies adaptive to both task requirements and available resources.

目标分解、执行跟踪和自适应细化

Representative Systems. The OpenAI/AgentsSDK [199] provides a comprehensive framework for research task planning, with explicit support for goal decomposition, execution tracking, and adaptive refinement. It enables the development of applications with sophisticated planning capabilities for research workflows. Similarly, Flowith/OracleMode [77] implements specialized planning mechanisms optimized for research tasks, with particular emphasis on information quality assessment and source prioritization. These systems demonstrate how advanced planning capabilities can significantly improve research efficiency and effectiveness.

注重信息质量评估和信息源优先级排序

2.3.2 Autonomous Execution and Monitoring Advances. Reliable execution of research plans requires sophisticated control and monitoring mechanisms: 控制与监控

Technical Evolution Trajectory. Initial systems employed basic sequential execution with limited error handling. Current implementations feature concurrent execution paths, comprehensive monitoring, and dynamic response to execution challenges. Advanced systems implement self-supervision with explicit success criteria, failure detection, and autonomous recovery strategies. This evolution has dramatically improved the reliability and autonomy of Deep Research systems across complex tasks.

学习更好的执行策略

Representative Systems. Agent-RL/ReSearch [2] exemplifies advanced execution control through its reinforcement learning-based approach to research execution. The system learns effective execution strategies from experience, continuously improving its ability to navigate complex research workflows. Its adaptive execution mechanisms can recover from failures and adjust strategies based on intermediate results, highlighting how sophisticated control mechanisms can enhance research reliability and effectiveness.

2.3.3 Multi-Agent Collaboration Framework Development. Complex research often benefits from specialized agent roles and collaborative approaches:

Technical Evolution Trajectory. Early systems relied on monolithic agents with undifferentiated capabilities. Modern implementations employ specialized agent roles with explicit coordination mechanisms and information sharing protocols. Advanced systems feature dynamic role allocation, consensus-building mechanisms, and sophisticated conflict resolution strategies. This evolution has enabled increasingly complex collaborative research workflows and improved performance on challenging tasks[49]. For instance, frameworks employing multi-agent debate have been shown to improve evaluation consistency [48], while research into generative AI voting demonstrates resilience to model biases in collective decision-making [162].

多 Agent 辩论，生成式人工智能投票

Representative Systems. The smolagents/open_deep_research [115] framework demonstrates effective multi-agent collaboration through its modular agent architecture and explicit coordination mechanisms. It

enables the composition of specialized research teams with complementary capabilities and shared objectives. Similarly, TARS [39] implements a sophisticated agent collaboration framework within its desktop environment, allowing multiple specialized agents to contribute to unified research workflows. These systems highlight how multi-agent approaches can enhance research capabilities through specialization and collaboration.

多agent像一个团队一样协同

2.4 Knowledge Synthesis and Output Generation: Evolution and Advances

信息混合并生成

The ultimate value of Deep Research systems lies in their ability to synthesize disparate information into coherent, actionable insights.

2.4.1 Information Evaluation Technology Development. Critical assessment of information quality represents a crucial capability for reliable research:

信誉启发法

Technical Evolution Trajectory. Early systems relied primarily on source reputation heuristics with limited content-based assessment. Modern implementations employ sophisticated evaluation frameworks considering source characteristics, content features, and consistency with established knowledge. Advanced systems implement explicit uncertainty modeling, contradiction detection, and evidential reasoning approaches. This evolution has dramatically improved the reliability and trustworthiness of research outputs. Advances in knowledge retrieval based on generative AI enhance the ability to source and verify information [306].



Representative Systems. The grapeot/deep_research_agent [263] implements sophisticated information evaluation mechanisms with explicit quality scoring for diverse source types. It can assess information reliability based on both intrinsic content features and extrinsic source characteristics, enabling more discerning information utilization. These capabilities highlight how advanced evaluation mechanisms can significantly enhance research quality and reliability.

2.4.2 Report Generation Technology Advances. Effective communication of research findings requires sophisticated content organization and presentation:

Technical Evolution Trajectory. Initial systems produced simple text summaries with limited structure or coherence. Current implementations generate comprehensive reports with hierarchical organization, evidence integration, and coherent argumentation. Advanced systems produce adaptive outputs tailored to audience expertise, information needs, and presentation contexts. This evolution has dramatically improved the usability and impact of Deep Research outputs.

自适应输出

Representative Systems. The mshumer/OpenDeepResearcher [249] project exemplifies a dvanced report generation through its structured output framework and evidence integration mechanisms. It produces comprehensive research reports with explicit attribution, structured arguments, and integrated supporting evidence. These capabilities demonstrate how sophisticated report generation can enhance the utility and trustworthiness of Deep Research outputs. Additionally, the MegaWika dataset [22] offers a large-scale multilingual resource consisting of millions of articles and referenced sources, enabling collaborative AI report generation.

交互生成

协作式 AI 报告生成

2.4.3 Interactive Presentation Technology Development. Beyond static reports, interactive result exploration enhances insight discovery and utilization:

Technical Evolution Trajectory. Early systems produced fixed textual outputs with minimal user interaction. Modern implementations support dynamic exploration including drill-down capabilities, source verification, and alternative viewpoint examination. Advanced systems enable collaborative refinement through iterative feedback incorporation and adaptive response to user queries. This evolution has dramatically enhanced the utility and flexibility of Deep Research interfaces.

交互式演示功能，可视化流程

Representative Systems. HKUDS/Auto-Deep-Research [112] implements sophisticated interactive presentation capabilities, allowing users to explore research findings through dynamic interfaces, examine supporting evidence, and refine analysis through iterative interaction. These features highlight how interactive presentation technologies can enhance the utility and accessibility of Deep Research outputs, facilitating more effective knowledge transfer and utilization.

This technical framework provides a comprehensive foundation for understanding the capabilities and evolution of Deep Research systems. The subsequent sections will build on this framework to analyze implementation approaches, evaluate system performance, and explore applications across diverse domains.

3 Comparative Analysis and Evaluation of Deep Research Systems

全面的比较分析

Building upon the technical framework established in Section 2, this section provides a comprehensive comparative analysis of existing Deep Research systems across multiple dimensions. We examine how different implementations balance technical capabilities, application suitability, and performance characteristics to address diverse research needs.

3.1 Cross-Dimensional Technical Comparison

Deep Research systems demonstrate varying strengths across the four key technical dimensions identified in our framework. This section analyzes how different implementations balance these capabilities and the resulting performance implications.

3.1.1 Foundation Model and Reasoning Efficiency Comparison. The underlying reasoning capabilities of Deep Research systems significantly impact their overall effectiveness:

Table 1. Comparison of Foundation Model Characteristics

System	Base Model	Context Length	Reasoning Approach
OpenAI/DeepResearch [197]	o3	may up to 200k tokens [195]	Multi-step reasoning
Gemini/DeepResearch [60]	Gemini 2.5 Pro	1M tokens [167]	Chain-of-thought
Perplexity/DeepResearch [209]	DeepSeek-R1	128K tokens [210]	Iterative reasoning
Grok3Beta [299]	Grok 3	1M tokens [299]	Chain-of-thought
AutoGLM-Research [330]	ChatGLM	DOM	Step-by-step planning

DOM: Depends On the Model

Commercial systems from OpenAI and Google leverage proprietary models with extensive context windows and sophisticated reasoning mechanisms, enabling them to process larger volumes of information with greater coherence. OpenAI’s o3 model demonstrates particular strength in complex reasoning tasks, while Gemini 2.5 Pro excels in information integration across diverse sources. In contrast, Perplexity/DeepResearch achieves

competitive performance with the open-source DeepSeek-R1 model through optimized implementation and focused use cases.

Open-source implementations like **Camel-AI/OWL** [43] and **QwenLM/Qwen-Agent** [224] demonstrate that effective deep research capabilities can be achieved with more accessible models through specialized optimization. The open-weight approach of **Camel-AI/OWL** [43] enables flexible deployment across computing environments, while **QwenLM/Qwen-Agent** [224] leverages modular reasoning to compensate for more limited base model capabilities.

3.1.2 Tool Integration and Environmental Adaptability Comparison. The ability to interact with diverse information environments varies significantly across implementations:

Table 2. Environmental Interaction Capabilities of Deep Research Systems

System	Web Interaction	API Integration	Document Processing	GUI Navigation
Nanobrowser [184]	Headless browsing, JavaScript execution, dynamic content rendering	REST API connectors	Basic HTML parsing	Not implemented
AutoGLM [330]	Full browser automation, form interaction	RESTful and GraphQL support	PDF, Office formats, JSON	Element identification, click/input automation
dzhng/deep-research [321]	Multi-page navigation, cookie handling	OAuth authentication support	Academic paper extraction, table parsing	Not implemented
Manus [164]	JavaScript rendering, session management	150+ service integrations, webhook support	PDF with layout preservation, CSV processing	Basic element interaction
n8n [183]	Limited, via HTTP requests	200+ integration nodes, custom webhook endpoints	CSV/XML processing	Not implemented
TARS [39]	Viewport management, scroll handling	REST/SOAP support	Standard formats processing	Desktop application control, UI element recognition

Note: Capabilities documented based on system repositories, technical documentation, and published demonstrations as of April 2025.

Specialized tools like **Nanobrowser** [184] excel in web interaction capabilities, providing sophisticated navigation and content extraction optimized for research workflows. Systems like **dzhng/deep-research** [321] and **nickscamara/open-deep-research** [42] complement these capabilities with advanced document processing features that can extract structured information from diverse formats.

Comprehensive platforms like **Manus** [164] and **AutoGLM** [330] offer broader environmental interaction capabilities, balancing web browsing, API integration, and document processing. These systems can adapt to diverse research scenarios but may not match the specialized performance of more focused tools in specific domains. The workflow automation capabilities of **n8n** [183] provide exceptional flexibility for API integration but offer more limited direct interaction with web and document environments.

3.1.3 Task Planning and Execution Stability Comparison. Effective research requires reliable task planning and execution capabilities:

Table 3. Planning and Execution Capabilities of Deep Research Systems

System	Task Planning Mechanisms	Error Handling Features	Collaboration Infrastructure
OpenAI/AgentsSDK [199]	Hierarchical task decomposition, goal-oriented planning	Automated retry logic, exception handling	Supervisor-worker architecture
Flowith/OracleMode [77]	Constraint-based planning, information quality prioritization	Checkpoint-based recovery	Limited role-based workflow
Agent-RL/ReSearch [2]	Reinforcement learning planning, adaptive task ordering	Progressive fallback strategies, state restoration	Standard agent messaging protocol
smolagents/open_deep_research [115]	Task queue management, priority-based scheduling	Basic retry mechanisms	Multi-agent configuration, specialized role definitions
TARS [39]	Process template architecture, event-driven coordination	State persistence, interruption handling	Team-based agent organization, shared memory
grapeot/deep_research_agent [263]	Linear task execution, sequential processing	Timeout handling	Single-agent architecture

Note: Capabilities documented based on system repositories, technical documentation, and published implementations as of April 2025.

The **OpenAI/AgentsSDK** [199] demonstrates sophisticated planning capabilities with hierarchical task decomposition and adaptive execution, enabling complex research workflows with reliable completion rates. Similarly, **Flowith/OracleMode** [77] offers advanced planning mechanisms optimized for research tasks, though with more limited error recovery capabilities.

Agent-RL/ReSearch [2] employs reinforcement learning techniques to develop robust execution strategies, enabling exceptional error recovery capabilities that can adapt to unexpected challenges during research workflows. In contrast, **smolagents/open_deep_research** [115] and **TARS** [39] focus on multi-agent collaboration, distributing complex tasks across specialized agents to enhance overall research effectiveness.

Simpler implementations like **grapeot/deep_research_agent** [263] offer more limited planning and execution capabilities but may provide sufficient reliability for less complex research tasks, demonstrating the range of complexity available across the ecosystem.

3.1.4 Knowledge Synthesis and Output Quality Comparison. The ability to synthesize findings into coherent, reliable outputs varies significantly:

Table 4. Knowledge Synthesis Capabilities of Deep Research Systems

System	Source Evaluation Mechanisms	Output Structuring	User Interaction Features
OpenAI/DeepResearch [197]	Source corroboration, authority ranking algorithms	Hierarchical report generation, section organization	Query clarification dialogue, result expansion
Perplexity/DeepResearch [209]	Source diversity metrics, publication date filtering	Citation-based organization, inline attribution	Source exploration interface, follow-up questioning
mshumer/OpenDeepResearcher [249]	Publication venue filtering, citation count tracking	Template-based document generation, section templating	Minimal interaction, batch processing focus
HKUDS/Auto-Deep-Research [112]	Basic source categorization, recency filtering	Standard academic format, heading organization	Interactive result exploration, citation navigation
grapeot/deep_research_agent [263]	Evidence classification algorithms, contradictory claim detection	Minimal formatting, raw data presentation	Command-line interface, non-interactive
OpenManus [193]	Source type categorization, basic metadata filtering	Markdown formatting, hierarchy-based organization	Basic query refinement, result browsing

Note: Capabilities documented based on system repositories, technical documentation, and published implementations as of April 2025.

Commercial platforms like **OpenAI/DeepResearch** [197] and **Perplexity/DeepResearch** [209] demonstrate sophisticated information evaluation capabilities, effectively assessing source credibility and content reliability to produce high-quality syntheses. OpenAI’s implementation excels in report structure and organization, while Perplexity offers particularly strong citation practices for source attribution and verification.

Open-source implementations like **mshumer/OpenDeepResearcher** [249] focus on report structure and organization, producing well-formatted outputs that effectively communicate research findings. **HKUDS/Auto-Deep-Research** [112] emphasizes interactive exploration, allowing users to examine evidence and refine analyses through iterative interaction. Specialized tools like **grapeot/deep_research_agent** [263] prioritize information evaluation over presentation, focusing on reliable content assessment rather than sophisticated output formatting.

3.2 Application-Based System Suitability Analysis

不同应用环境的适应性差异

Beyond technical capabilities, Deep Research systems demonstrate varying suitability for different application contexts. This section examines how system characteristics align with key application domains.

3.2.1 Academic Research Scenario Adaptability Assessment. Academic research requires particular emphasis on comprehensive literature review, methodological rigor, and citation quality. Systems like **OpenAI/DeepResearch** [197] excel in this domain through their ability to access academic databases, comprehensively analyze research methodologies, and generate properly formatted citations. Other specialized academic research tools like **PaperQA** [80] and **Scite** [243] offer complementary capabilities focused specifically on scientific literature processing, while Google’s **NotebookLm** [95] provides structured knowledge workspaces for academic exploration.

OpenAI/DeepResearch [197] demonstrates exceptional suitability for academic research through its comprehensive literature coverage, methodological rigor, and high-quality citation practices. The system can effectively navigate academic databases, understand research methodologies, and produce well-structured

Table 5. Academic Research Application Features of Deep Research Systems

System	Academic Database Integration	Methodology Analysis Features	Citation Management
OpenAI/DeepResearch [197]	ArXiv, IEEE Xplore, PubMed, Google Scholar	Statistical method identification, study design classification	IEEE, APA, MLA, Chicago format support
Perplexity/DeepResearch [209]	ArXiv, PubMed, JSTOR, ACM Digital Library	Experimental design analysis, sample size assessment	Automated citation generation, DOI resolution
dzhng/deep-research [321]	ArXiv, Semantic Scholar, limited database access	Basic methodology extraction	BibTeX export, standard format support
Camel-AI/OWL [43]	Custom corpus integration, specialized domain databases	Research design pattern recognition, methodology comparison	Domain-specific citation formatting
mshumer/OpenDeepResearcher [249]	Open access databases, PDF repository processing	Methodology summary extraction	Standard citation format generation
HKUDS/Auto-Deep-Research [112]	University library integration, institutional repository access	Research approach categorization	Reference management, bibliography generation

Note: Features documented based on system repositories, technical documentation, and published use cases as of April 2025.

literature reviews with appropriate attribution. Perplexity/DeepResearch [209] offers similarly strong performance for literature coverage and citation quality, though with somewhat less methodological sophistication.

Open-source alternatives like Camel-AI/OWL [43] provide competitive capabilities for specific academic domains, particular strength in methodological understanding for specific domains. Systems like dzhng/deep-research [321], mshumer/OpenDeepResearcher [249], and HKUDS/Auto-Deep-Research [112] offer moderate capabilities across all dimensions, making them suitable for less demanding academic research applications or preliminary literature exploration.

企业决策

3.2.2 Enterprise Decision-Making Scenario Adaptability Assessment. Business intelligence and strategic decision-making emphasize information currency, analytical depth, and actionable insights:

Table 6. Enterprise Decision-Making Application Features of Deep Research Systems

System	Market Information Sources	Analytical Frameworks	Decision Support Features
Gemini/DeepResearch [60]	News API integration, SEC filings access, market data feeds	Competitor analysis templates, trend detection algorithms	Executive summary generation, recommendation formatting
Manus [164]	Financial data integrations, news aggregation, industry reports	Market sizing frameworks, SWOT analysis templates	Strategic options presentation, decision matrix generation
n8n [183]	CRM integration, marketing platform connectivity, custom data sources	Custom analytics workflow creation, data pipeline automation	Dashboard generation, notification systems
Agent-RL/ReSearch [2]	Configurable information source adapters, custom data inputs	Pattern recognition algorithms, causal analysis frameworks	Scenario planning tools, impact assessment matrices
Flowith/OracleMode [77]	Real-time data feeds, specialized industry sources	Industry-specific analytical templates, framework application	Strategic briefing generation, insight prioritization
TABS [39]	Enterprise system integration, desktop application data access	Basic analytical template application	Standardized reporting, data visualization

Note: Features documented based on system repositories, technical documentation, and published use cases as of April 2025.

Gemini/DeepResearch [60] demonstrates exceptional suitability for enterprise decision-making through its strong information currency, analytical capabilities, and actionable output formats. The system effectively navigates business information sources, analyzes market trends, and produces insights directly relevant to decision processes. Manus [164] offers similarly strong performance for information acquisition and analysis, though with somewhat less emphasis on actionable recommendation formatting. Microsoft Copilot [173] empowers organizations with powerful generative AI, enterprise-grade security and privacy, and is trusted by companies around the world. Similarly, the Adobe Experience Platform AI Assistant [181] employs knowledge graph-enhanced retrieval-augmented generation to accurately respond over private enterprise documents, significantly enhancing response relevance while maintaining provenance tracking.

Workflow automation platforms like n8n [183] provide particular strengths in information currency and actionability through their integration with enterprise data sources and business intelligence tools. Research-focused systems like Agent-RL/ReSearch [2] and Flowith/OracleMode [77] offer competitive analytical capabilities but may require additional processing to translate findings into actionable business recommendations.

3.2.3 Personal Knowledge Management Adaptability Assessment. Individual knowledge management emphasizes accessibility, personalization, and integration with existing workflows:

Table 7. Personal Knowledge Management Features of Deep Research Systems

System	User Interface Design	Customization Options	Existing Tool Integration
Perplexity/DeepResearch [209]	Web-based interface, mobile application support	Topic preference settings, information filtering options	Browser extension, sharing functionality
nickscamara/open-deep-research [42]	Command-line interface, web interface option	Modular configuration, source priority adjustment	Local file system integration, note-taking exports
OpenManus [193]	Desktop application, local web interface	Template customization, workflow configuration	Note application exports, knowledge base connections
Nanobrowser [184]	Programmatic interface, developer-focused API	Full configuration access, component-level customization	Browser automation framework compatibility
smolagents/open_deep_research [115]	Technical interface, Python library integration	Architecture-level customization, agent behavior configuration	Python ecosystem integration, custom adapter support
Jina-AI/node-DeepResearch [121]	Node.js integration, API-driven interface	Component-level configuration, pipeline customization	Node.js application ecosystem, JavaScript framework support

Note: Features documented based on system repositories, technical documentation, and published implementations as of April 2025.

Perplexity/DeepResearch [209] offers strong accessibility for personal knowledge management through its consumer-friendly interface and free access tier, though with more limited personalization capabilities. Open-source implementations like nickscamara/open-deep-research [42] and OpenManus [193] provide greater personalization possibilities through local deployment and customization, enabling adaptation to individual information management preferences.

Infrastructure tools like Nanobrowser [184] and Jina-AI/node-DeepResearch [121] offer particular strengths in workflow integration, allowing seamless incorporation into existing personal knowledge management systems and processes. More complex frameworks like smolagents/open_deep_research [115] provide sophisticated capabilities but may present accessibility challenges for non-technical users.

3.3 Performance Metrics and Benchmarking

Beyond qualitative comparisons, quantitative performance metrics provide objective assessment of Deep Research capabilities across systems. **定量性能指标**

3.3.1 Quantitative Evaluation Metrics. Standard benchmarks enable comparative evaluation of core research capabilities:

Table 8. Performance on Standard Evaluation Benchmarks

System	HLE Score* [212]	MMLU** Score [33]	HotpotQA Score [307]	GAIA Score(pass@1)*** [172]
OpenAI/DeepResearch [197]	26.6%	-	-	67.36%
Gemini-2.5 [60, 293]	18.8%	-	-	-
Gemini-2.0-Flash [89, 93]	-	77.9%	-	-
Perplexity/DeepResearch [209]	21.1%	-	-	-
Grok3Beta [299]	-	79.9%	-	-
Manus [164]	-	-	-	86.5%
Agent-RL/ReSearch [2]	-	-	37.51%	-

*Humanity’s Last Exam: Tests frontier research capabilities

**Massive Multitask Language Understanding: Tests general knowledge

***GAIA Score(pass@1): Average score

OpenAI/DeepResearch [30, 123, 197] demonstrates leading performance across various benchmark categories, particularly excelling in Humanity’s Last Exam (HLE) [212] hich measures advanced research and reasoning capabilities. Gemini/DeepResearch [60] shows comparable performance. According to the introduction of Google Deep Research with Gemini 2.5 Pro Experimental [60, 126], the new model demonstrated superior user preference over OpenAI/DeepResearch across four key metrics: instruction following (60.6% vs. 39.4%), Comprehensiveness (76.9% vs. 23.1%), Completeness (73.3% vs. 26.7%), and Writing quality (58.2% vs. 41.8%). These results suggest Gemini 2.5 Pro’s enhanced capability in synthesizing structured, high-fidelity research outputs. This capability is further amplified in fullstack applications, where the integration of Gemini

Table 9. Documented Performance Metrics from Deep Research Systems

System	Benchmark	Reported Score	Evaluation Context	Source
OpenAI/DeepResearch	HLE	26.6%	Humanity’s Last Exam	[197]
OpenAI/DeepResearch	GAIA (pass@1)	67.36%	General AI assistant tasks	[197]
Perplexity/DeepResearch	HLE	21.1%	Humanity’s Last Exam	[209]
Perplexity/DeepResearch	SimpleQA	93.9%	Factual question answering	[209]
Grok3Beta	MMLU	92.7%	Multitask language understanding	[299]
Manus	GAIA (pass@1)	86.5%	General AI assistant tasks	[164]
Agent-RL/ReSearch	HotpotQA	37.51%	Multi-hop question answering	[2]
AutoGLM	WebArena-Lite	55.2% (59.1% retry)	Web navigation tasks	[330]
AutoGLM	OpenTable	96.2%	Restaurant booking tasks	[330]

Note: Scores reflect performance on specific benchmarks as reported in cited publications. Direct comparison requires consideration of evaluation methodologies and task specifications.

models with frameworks like LangGraph facilitates research-augmented conversational AI for comprehensive query handling, as demonstrated in **Google-Gemini/Gemini-Fullstack-Langgraph-Quickstart** [94]. **Perplexity/DeepResearch** [209] achieves competitive results despite utilizing the open-source DeepSeek-R1 model, highlighting the importance of implementation quality beyond raw model capabilities.

Open-source implementations show progressively lower benchmark scores, though many still achieve respectable performance suitable for practical applications. Systems like **AutoGLM-Research** [330], **HKUDS/Auto-Deep-Research** [112], and **Camel-AI/OWL** [43] demonstrate that effective research capabilities can be achieved with more accessible models and frameworks, though with some performance trade-offs compared to leading commercial implementations.

Recent benchmark development has expanded evaluation to more specialized aspects of research assistance. The **AAAR-1.0** benchmark [157] specifically evaluates AI’s potential to assist research through 150 multi-domain tasks designed to test both retrieval and reasoning capabilities. Domain-specific approaches include **DSBench** [122], which evaluates data science agent capabilities across 20 real-world tasks [182, 283], **SciCode** [268] for scientific code generation, **MASSW** [323] for scientific workflow assistance, and **MMSci** [147] for multimodal scientific understanding across graduate-level materials. **ScienceQA** [160] offers a comprehensive multimodal science benchmark with chain-of-thought explanations for evaluating reasoning capabilities. Domain-specific benchmarks like **TPBench** [58] for theoretical physics and **AAAR-1.0** [157] for research assistance capabilities offer additional targeted evaluation approaches for specialized research applications. Multi-domain code generation benchmark like **DomainCodeBench** [328] is designed to systematically assess large language models across 12 software application domains and 15 programming languages. Interactive evaluation frameworks like **LatEval** [114] specifically assess systems’ capabilities in handling incomplete information through lateral thinking puzzles, providing insight into research abilities under uncertainty and ambiguity. Complementary approaches like **Mask-DPO** [100] focus on generalizable fine-grained factuality alignment, addressing a critical requirement for reliable research outputs. Domain-specific benchmarks such as **GMAI-MMBench** [51] provide comprehensive multimodal evaluation frameworks specifically designed for medical AI applications, while **AutoBench** [52] offers automated evaluation of scientific discovery capabilities, providing standardized assessment of core research functions. Other broad evaluation frameworks including **HELM** [149], **BIG-bench** [88], and **AGIEval** [331], provide complementary assessment dimensions. Specialized

multimodal benchmarks like INQUIRE [279] extend this landscape to ecological challenges, rigorously evaluating expert-level text-to-image retrieval tasks critical for accelerating biodiversity research.

Table 10. Specialized Deep Research Benchmarks

Benchmark	Focus Area	Evaluation Approach	Key Metrics
AAAR-1.0 [157]	Research assistance	150 multi-domain tasks	Retrieval and reasoning capability
DSBench [122]	Data science	20 real-world tasks	End-to-end completion rate
SciCode [268]	Scientific coding	Curated by scientists	Code quality, task completion
MASSW [323]	Scientific workflows	Benchmarking tasks	Workflow orchestration quality
MMSci [147]	Multimodal science	Graduate-level questions	Cross-modal understanding
TPBench [58]	Theoretical physics	Physics reasoning	Problem-solving accuracy

Note: These benchmarks represent domain-specific evaluation frameworks for specialized research capabilities.

3.3.2 *Qualitative Assessment Frameworks.* Beyond numeric benchmarks, qualitative evaluation provides insight into practical effectiveness: 实际效果

Table 11. Documented Output Characteristics of Deep Research Systems

System	Content Organization	Information Diversity	Verification Features	Novel Connection Mechanisms
OpenAI/DeepResearch [197]	Hierarchical structure with 5+ sections, executive summaries	Cross-domain source integration (reported in [197])	Statement-level citation linking, contradiction flagging	Cross-domain connection identification
Gemini/DeepResearch [60]	Multi-level heading organization, standardized formatting	Multi-perspective source inclusion (documented in [60])	Source credibility metrics, confidence indicators	Thematic pattern identification
Perplexity/DeepResearch [209]	Progressive information disclosure, expandable sections	Real-time source aggregation across platforms	Direct quote attribution, inline source linking	Timeline-based relationship mapping
n8nuser/OpenDeepResearcher [249]	Template-based document structure, consistent formatting	Topic-based categorization of sources	Basic citation framework, reference listing	Topic cluster visualization
grapeot/deep_research_agent [263]	Minimal formatting, content-focused presentation	Source type categorization, domain tracking	Source credibility scoring system based on metadata	Not implemented per repository documentation
Agent-RL/ReSearch [2]	Adaptive content organization based on information types	Exploratory search patterns documented in repository	Contradiction detection algorithms	Pattern-based insight generation documented in [2]

Note: Characteristics documented based on system technical documentation, published demonstrations, repository analysis, and official descriptions as of April 2025. Specific feature implementations may vary across system versions.

Commercial systems generally demonstrate stronger qualitative performance, particularly in output coherence and factual accuracy. **OpenAI/DeepResearch** [197] produces exceptionally well-structured reports with reliable factual content, while also achieving moderate innovation in connecting disparate sources. **Gemini/DeepResearch** [60] shows similar strengths in coherence and accuracy, with slightly less emphasis on novel insights.

Some open-source implementations show particular strengths in specific dimensions. **Agent-RL/ReSearch** [2] achieves notable performance in insight novelty through its exploration-focused approach, while **grapeot/deep_research_agent** [263] demonstrates strong factual accuracy through its emphasis on information verification. These specialized capabilities highlight the diversity of approaches within the Deep Research ecosystem.

3.3.3 *Efficiency and Resource Utilization Metrics.* **Practical deployment considerations include computational requirements and operational efficiency:**

Commercial cloud-based services offer optimized performance with moderate response times, though with dependency on external infrastructure and associated costs. **Perplexity/DeepResearch** [209] achieves particularly strong efficiency metrics, with relatively quick response times and high token efficiency despite its competitive output quality.

Open-source implementations present greater variability in efficiency metrics. Systems like **AutoGLM-Research** [330] and **QwenLM/Qwen-Agent** [224] require substantial computational resources but can be deployed in local environments, offering greater control and potential cost savings for high-volume usage.

Table 12. Efficiency and Resource Utilization

System	Response Time*	Compute Requirements	Token Efficiency**
OpenAI/DeepResearch [197]	5-30 min	Cloud-only	High (detailed, citation-rich)
Perplexity/DeepResearch [209]	2m59s	Cloud-only	-
Grok3Beta [299]	-	Cloud-only	-
Nanobrowser [184]	-	User-defined via LLM API key	-
n8n [183]	-	Self-hosted or cloud-based; scalable	-

*Typical response time for moderately complex research tasks

**Efficiency of token utilization relative to output quality

Lighter-weight implementations like `nickscamara/open-deep-research` [42] can operate with more limited resources but typically demonstrate longer response times and lower token efficiency.

This comparative analysis highlights the diversity of approaches and capabilities across the Deep Research ecosystem. While commercial implementations currently demonstrate leading performance on standard benchmarks, open-source alternatives offer competitive capabilities in specific domains and use cases, with particular advantages in customization, control, and potential cost efficiency for specialized applications. The subsequent sections will build on this analysis to examine implementation technologies, evaluation methodologies, and application domains in greater detail.

4 Implementation Technologies and Challenges

The practical realization of Deep Research systems involves numerous technical challenges spanning infrastructure design, system integration, and safeguard implementation. This section examines the key implementation technologies that enable effective Deep Research capabilities and the challenges that must be addressed for reliable, efficient operation.

4.1 Architectural Implementation Patterns

The diverse systems analyzed in this survey reveal several distinct architectural patterns that represent different approaches to implementing Deep Research capabilities. This section examines four fundamental architectural patterns: monolithic, pipeline-based, multi-agent, and hybrid implementations. For each pattern, we analyze the underlying structural principles, component interactions, information flow mechanisms, and representative systems. **整体架构、基于 pipeline 的架构、多 Agent 架构、混合架构**

以一个基础模型为核心
4.1.1 Monolithic Architecture Pattern. Monolithic implementations integrate all Deep Research capabilities within a unified architectural framework centered around a core reasoning engine. As illustrated in Figure 4, these systems employ a centralized control mechanism with direct integration of specialized modules.

The defining characteristics of this architecture include:

- **Centralized Control Flow:** All operations route through a primary reasoning engine that maintains global state and execution context
- **Tightly Coupled Integration:** Specialized modules (web browsing, document processing, etc.) are directly integrated with the central controller
- **Shared Memory Architecture:** Information state is maintained in a centralized memory system accessible to all components

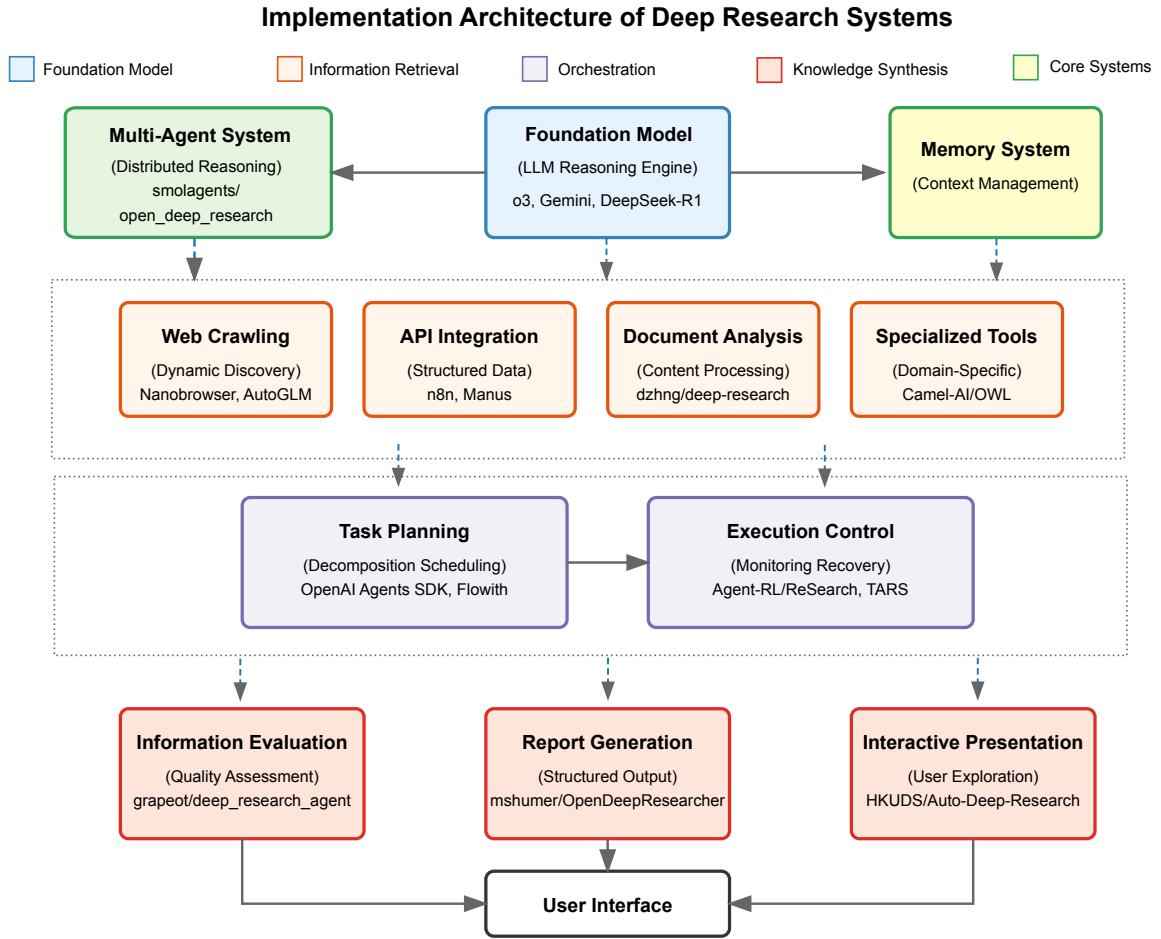


Fig. 3. Implementation Architecture of Deep Research Systems

- **Sequential Reasoning Processes:** Operations typically follow a structured sequence defined by the central controller

好的一致性，缺点不好扩展

This architectural pattern offers strong coherence and reasoning consistency through its unified control structure. However, it presents challenges for extensibility and can struggle with parallelization of complex operations. Representative implementations include OpenAI/DeepResearch [197] and grapeot/deep_research_agent [263], which demonstrate how this architecture enables coherent reasoning across diverse information sources while maintaining implementation simplicity.

workflow排序

4.1.2 Pipeline-Based Architecture Pattern. Pipeline architectures implement Deep Research capabilities through a sequence of specialized processing stages connected through well-defined interfaces. As shown in Figure 5, these systems decompose research workflows into discrete processing components with explicit data transformations between stages.

The key characteristics of pipeline implementations include:

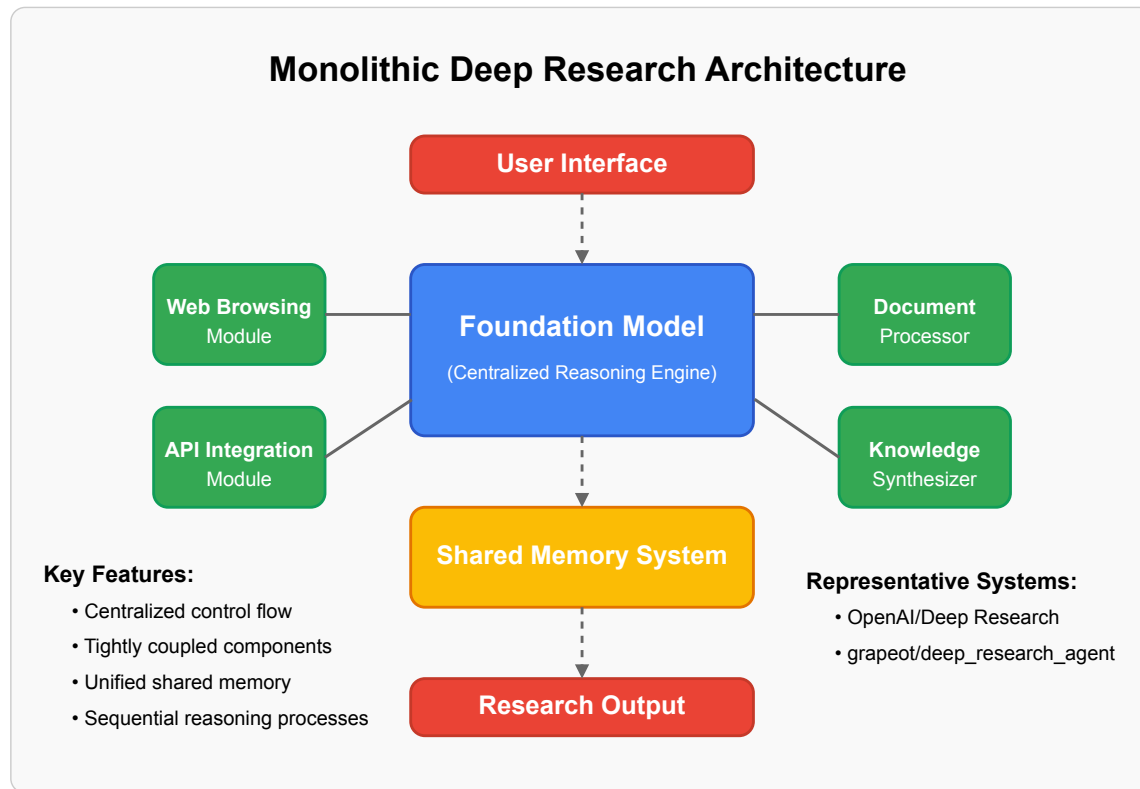


Fig. 4. Monolithic Deep Research Architecture

- **Sequential Component Organization:** Research tasks flow through a predefined sequence of specialized processing modules
- **Standardized Interfaces:** Clear data transformation specifications between pipeline stages enable modular component replacement
- **Staged Processing Logic:** Each component implements a specific transformation, with minimal dependence on global state 对全局状态的依赖小
- **Configurable Workflow Paths:** Advanced implementations enable conditional routing between alternative processing paths based on intermediary results workflow定制和组件可重用性方面好
跨组件迭代细化的复杂推理任务时可能会遇到困难。

Pipeline architectures excel in workflow customization and component reusability but may struggle with complex reasoning tasks requiring iterative refinement across components. Systems like `n8n` [183] and `dzhng/deep-research` [321] exemplify this approach, demonstrating how explicit workflow sequencing enables sophisticated research automation through composition of specialized components.

4.1.3 Multi-Agent Architecture Pattern. Multi-agent architectures implement Deep Research capabilities through ecosystems of specialized autonomous agents coordinated through explicit communication protocols. Figure 6 illustrates how these systems distribute research functionality across collaborating agents with differentiated roles and responsibilities.

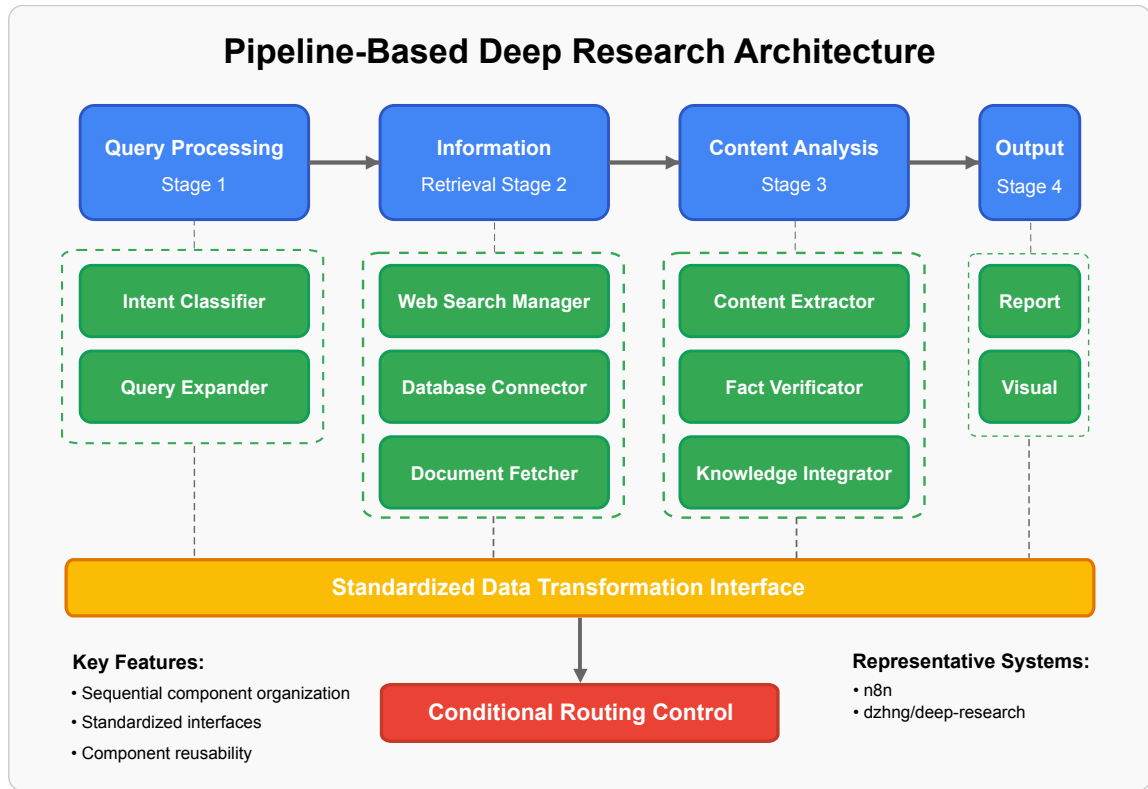


Fig. 5. Pipeline-Based Deep Research Architecture

The defining elements of multi-agent implementations include:

- **Distributed Functional Decomposition:** Research capabilities are distributed across specialized agents with defined roles (searcher, analyst, critic, etc.)
- **Explicit Coordination Mechanisms:** Standardized message passing and task delegation protocols enable inter-agent collaboration
- **Autonomous Decision Logic:** Individual agents maintain independent reasoning capabilities within their designated domains 个体 Agent在特定领域独立推理
- **Dynamic Task Allocation:** Advanced implementations employ flexible task assignment based on agent capabilities and current workload 任务分配
各专业能力并行处理优秀，易扩展；维护一致性难

Multi-agent architectures excel in complex research tasks requiring diverse specialized capabilities and parallel processing. Their distributed nature enables exceptional scaling for complex research workflows but introduces challenges in maintaining overall coherence and consistent reasoning across agents. Representative implementations include `smolagents/open_deep_research` [115] and `TARS` [39], which demonstrate how multi-agent coordination enables sophisticated research workflows through specialized agent collaboration.

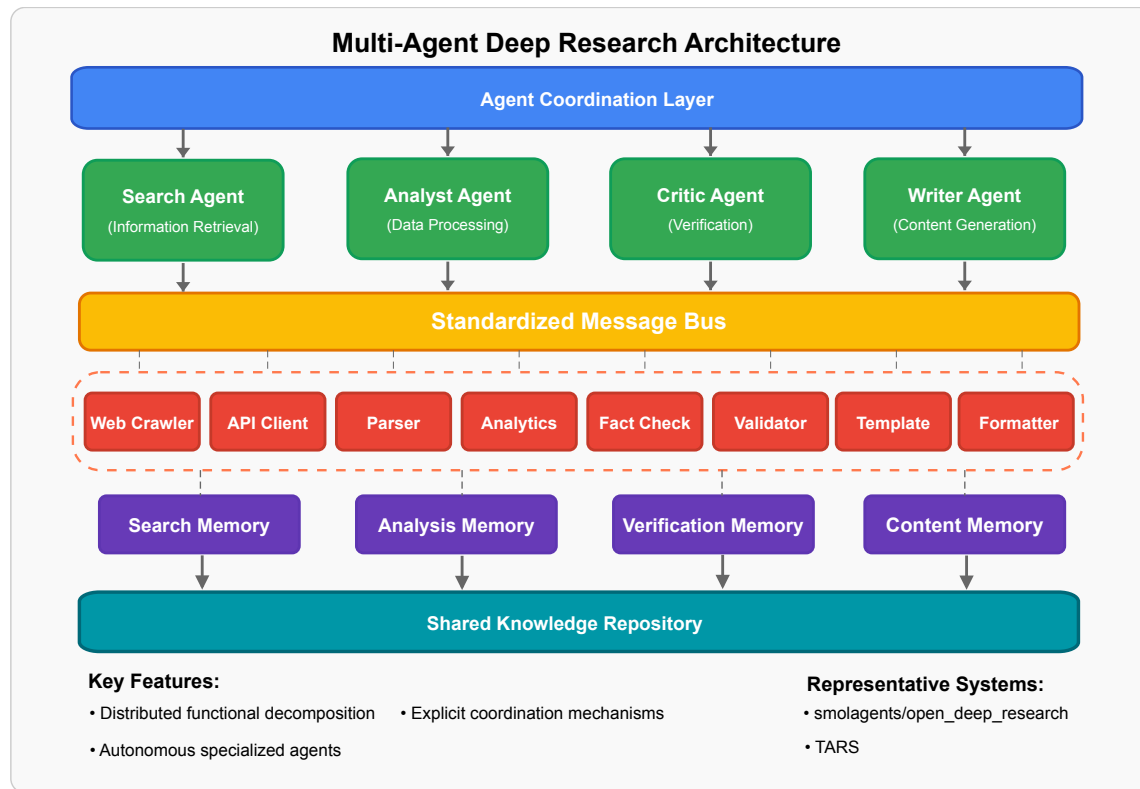


Fig. 6. Multi-Agent Deep Research Architecture

平衡各自优势

4.1.4 Hybrid Architecture Pattern. Hybrid architectures combine elements from multiple architectural patterns to balance their respective advantages within unified implementations. As shown in Figure 7, these systems employ strategic integration of architectural approaches to address specific research requirements.

Key characteristics of hybrid implementations include:

- **Tiered Architectural Organization:** Different architectural patterns are employed at different system levels based on functional requirements
- **Domain-Specific Optimization:** Architectural approaches are selected based on domain-specific processing requirements
特定领域的处理要求选择架构方法
- **Flexible Integration Mechanisms:** Standardized interfaces enable communication between components employing different architectural patterns
- **Adaptive Execution Frameworks:** Control mechanisms dynamically adjust processing approaches based on task characteristics

灵活；复杂，集成难

Hybrid architectures offer exceptional flexibility and optimization opportunities but introduce implementation complexity and potential integration challenges. Systems like Perplexity/DeepResearch [209] and Camel-AI/OWL [43] exemplify this approach, combining centralized reasoning with distributed information

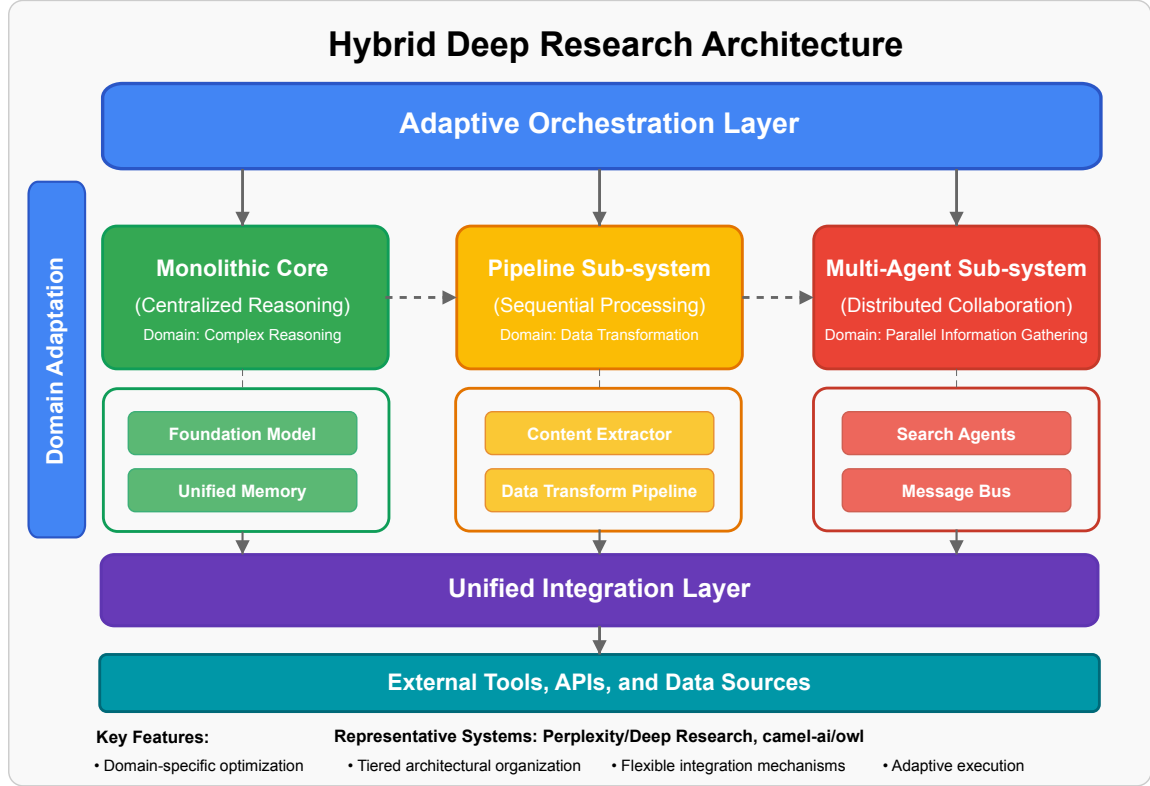


Fig. 7. Hybrid Deep Research Architecture

gathering and specialized processing pipelines to achieve sophisticated research capabilities with balanced performance characteristics.

agent框架增强deep research

4.1.5 Emerging Agent Framework Ecosystems. Beyond the core architectural patterns described above, the Deep Research ecosystem has been significantly enhanced by specialized agent frameworks that provide standardized components for agent development. Emerging systems incorporate specialized agent frameworks [54, 142, 301] that structure reasoning in ways particularly suited to complex research tasks requiring both depth and breadth of analysis. As detailed in comprehensive analyses of agent frameworks [133, 304], these systems offer varying approaches to agent orchestration, execution control, and reasoning orchestration.

Key frameworks include LangGraph [134], which provides graph-based control flow for language model applications, enabling complex reasoning patterns through explicit state management and transition logic. Google's Agent Development Kit (ADK) [91] offers a comprehensive framework for agent development with standardized interfaces for tool integration, planning, and execution monitoring. CrewAI [64] implements an agent collaboration framework designed specifically for multi-specialist workflows, enabling role-based task distribution with explicit coordination mechanisms. More experimental frameworks like Agno [3] explore agentic autonomy through self-improvement and meta-reasoning capabilities.

The TapeAgents framework [19] provides a particularly comprehensive approach to agent development and optimization, with explicit support for iterative refinement through systematic recording and analysis of agent behavior. These frameworks collectively demonstrate an ongoing shift toward standardized agent components that enhance development efficiency while enabling more complex reasoning and execution patterns.

4.1.6 Architectural Pattern Comparison. Table 13 provides a comparative analysis of these architectural patterns across key performance dimensions:

Table 13. Architectural Pattern Characteristics in Deep Research Systems

Characteristic	Monolithic	Pipeline	Multi-Agent	Hybrid
Control Structure	Centralized	Sequential	Distributed	Mixed
Component Coupling	Tight	Loose	Moderate	Variable
Failure Propagation	System-wide	Stage-limited	Agent-isolated	Component-dependent
Development Complexity	Minimal	Moderate	Substantial	Maximal
Deployment Flexibility	Limited	Moderate	Moderate	High
Representative Systems	grapeot/deep_research_agent	n8n, dzhng/deep-research	smolagents, TARS	Perplexity, Camel-AI/OWL

Note: Characteristics based on architectural analysis of surveyed systems. Quantitative performance comparison requires standardized benchmarking across identical tasks and environments.

Each architectural pattern presents distinct advantages and limitations that influence its suitability for specific Deep Research applications. Monolithic architectures excel in reasoning coherence and implementation simplicity, making them appropriate for focused research applications with well-defined workflows. Pipeline architectures offer exceptional extensibility and component reusability, enabling customized research workflows through modular composition. Multi-agent architectures provide superior parallelization and fault tolerance, supporting complex research tasks requiring diverse specialized capabilities. Hybrid architectures balance these characteristics through strategic integration, offering flexible optimization for diverse research requirements.

The architectural pattern selection significantly influences system capabilities, performance characteristics, and application suitability. As the Deep Research ecosystem continues to evolve, we anticipate further architectural innovation combining elements from these foundational patterns to address emerging application requirements and technical capabilities.

4.2 Infrastructure and Computational Optimization

Deep Research systems require sophisticated infrastructure to support their complex reasoning and information processing capabilities.

4.2.1 Distributed Reasoning Architectures. Effective reasoning across expansive information landscapes requires specialized architectural approaches. Frameworks like **AutoChain** [78] and **AutoGen** [298] have pioneered distributed agent paradigms that can be applied to research workflows. Advanced systems employ distributed reasoning architectures that decompose complex queries into parallel processing paths. **OpenAI/DeepResearch** [197] implements a hierarchical reasoning framework that distributes analytical tasks across multiple execution threads while maintaining coherent central coordination.

Implementation approaches increasingly leverage specialized frameworks for efficient LLM serving, including **LightLLM** [177], **Ollama** [192], **VLLM** [281], and **Web-LLM** [176] for browser-based deployment.

These frameworks enable more efficient utilization of computational resources, particularly important for resource-intensive research workflows requiring extensive model inference. Such optimizations are especially critical for open-source implementations operating with more constrained computational resources compared to commercial cloud-based alternatives.

Parallel Reasoning Pathways. Advanced systems employ distributed reasoning architectures that decompose complex queries into parallel processing paths. **OpenAI/DeepResearch** [197] implements a hierarchical reasoning framework that distributes analytical tasks across multiple execution threads while maintaining coherent central coordination. Similar approaches are evident in **Gemini/DeepResearch** [60], which leverages Google’s distributed computing infrastructure to parallelize information analysis while preserving reasoning consistency.

Open-source implementations like **HKUDS/Auto-Deep-Research** [112] and **Agent-RL/ReSearch** [2] demonstrate more accessible distributed reasoning approaches, utilizing task decomposition and asynchronous processing to enhance performance within more constrained computational environments. These systems show that effective parallelization can be achieved even without the extensive infrastructure of commercial platforms.

Memory and State Management. Distributed reasoning introduces significant challenges in memory coherence and state management. Commercial systems implement sophisticated state synchronization mechanisms that maintain consistent reasoning contexts across distributed components. OpenAI’s implementation utilizes a hierarchical memory architecture with explicit coordination protocols [200], while Google’s approach leverages its existing distributed computing frameworks adapted for reasoning workflows.

Open-source alternatives like **Camel-AI/OWL** [43] employ simplified but effective memory management approaches, including centralized knowledge repositories with controlled access patterns. These implementations demonstrate pragmatic solutions to state management challenges within more constrained technical environments.

4.2.2 Parallel Search and Information Retrieval. Information acquisition represents a primary bottleneck in Deep Research performance:

Concurrent Query Execution. Advanced systems implement sophisticated parallel search infrastructures to accelerate information gathering. **Perplexity/DeepResearch** [209] employs a multi-threaded search architecture that dispatches dozens of concurrent queries across different information sources, significantly accelerating the research process. Similar capabilities are evident in **dzhng/deep-research** [321], which implements a specialized scheduler for concurrent web queries with adaptive rate limiting to avoid service restrictions.

Infrastructure tools like **Nanobrowser** [184] provide optimized platforms for parallel browsing operations, enabling multiple concurrent page loads with shared resource management. These specialized components enhance the information gathering capabilities of integrated systems like **Manus** [164] and **Flowith/OracleMode** [77], which leverage concurrent browsing to accelerate their research workflows.

Query Coordination and Deduplication. Effective parallel search requires sophisticated coordination to avoid redundancy and ensure comprehensive coverage. Commercial systems implement advanced query

planning that dynamically adapts to intermediate results, adjusting search strategies based on discovered information. OpenAI’s implementation includes explicit deduplication mechanisms that identify and consolidate redundant sources, while Perplexity employs source diversification techniques to ensure broad coverage.

Open-source tools like `nickscamara/open-deep-research` [42] implement pragmatic approaches to query coordination, including simple but effective caching mechanisms and result fingerprinting to avoid redundant processing. These techniques demonstrate that effective coordination can be achieved with relatively straightforward implementation approaches.

4.2.3 Resource Allocation and Efficiency Optimization. Computational efficiency significantly impacts both performance and operational economics:

Adaptive Resource Allocation. Advanced systems implement dynamic resource allocation based on task characteristics and complexity. `Gemini/DeepResearch` [60] employs sophisticated workload prediction to provision computational resources adaptively, allocating additional capacity for more complex research tasks. Similar approaches are emerging in open-source implementations like `QwenLM/Qwen-Agent` [224], which incorporates task complexity estimation to guide resource allocation decisions.

Progressive Processing Strategies. Efficiency-focused implementations employ progressive processing approaches that incrementally refine results based on available information. `Perplexity/DeepResearch` [209] utilizes a staged analysis approach that provides preliminary findings quickly while continuing deeper analysis in the background. This strategy enhances perceived responsiveness while ensuring comprehensive results for complex queries.

Open-source alternatives like `mshumer/OpenDeepResearcher` [249] implement simpler but effective progressive strategies, including early result previews and incremental report generation. These approaches demonstrate pragmatic solutions to efficiency challenges without requiring sophisticated infrastructure.

4.3 System Integration and Interoperability

Deep Research systems must effectively coordinate diverse components and external services to deliver comprehensive capabilities.

4.3.1 API Design and Standardization. Consistent interfaces enable modular development and component interoperability:

Component Interface Standardization. Current Deep Research implementations employ largely incompatible architectures and interfaces. Future research could build upon emerging standardization efforts like Anthropic’s Model Context Protocol (MCP) [12] and Google’s Agent2Agent Protocol (A2A) [90, 92] to establish truly universal component interfaces. MCP provides a structured framework for model-tool interaction, enabling consistent integration patterns across diverse LLM applications, while A2A focuses on standardized agent-to-agent communication to facilitate multi-agent systems. These complementary approaches could form the foundation for comprehensive standardization enabling modular development

and interchangeable components across implementations. Early steps in this direction appear in frameworks like **OpenAI/AgentsSDK** [199], which provides standardized agent definitions, but more comprehensive standardization would require broader industry adoption of common protocols.

Workflow Automation. Several workflow automation platforms like **Dify** [259], **Coze** [38], and **Flowise** [5] have emerged as low-code environments for building LLM-powered applications, potentially offering standardized frameworks for Deep Research components. Advanced workflow orchestration platforms including **Temporal** [265], **Restate** [229], and **Orkes** [203] provide robust infrastructure for complex, stateful workflows with explicit support for long-running processes and reliability patterns crucial for sophisticated research applications. Implementation approaches might include defining standard message passing protocols between research components, establishing common data structures for research tasks and results, developing compatibility layers between competing standards, extending existing protocols with research-specific interaction patterns, and establishing common evaluation frameworks for component interoperability. These advances could accelerate ecosystem development by enabling specialized components from diverse developers to work seamlessly within unified frameworks, significantly enhancing the pace of innovation through componentization and reuse.

External Service Integration. Access to specialized external services significantly enhances research capabilities. Advanced retrieval frameworks like **LlamaIndex** [235] provide standardized interfaces for retrieval augmentation, enabling consistent integration patterns across diverse information sources and document formats. Systems like **n8n** [183] excel in external service integration through their comprehensive connector library and standardized authentication mechanisms. This capability enables access to specialized information sources and analytical services that extend beyond basic web search.

Open-source frameworks like **Jina-AI/node-DeepResearch** [121] implement simplified but effective API integration patterns, providing standardized wrappers for common services while maintaining extensibility for custom integrations. These approaches balance standardization with flexibility for diverse research requirements.

4.3.2 Tool Integration Frameworks. Effective orchestration of diverse tools enhances overall system capabilities:

Tool Selection and Composition. Advanced systems implement sophisticated tool selection based on task requirements and information context. **Manus** [164] features an adaptive tool selection framework that identifies appropriate tools for specific research subtasks, dynamically composing workflows based on available capabilities. Similar approaches are emerging in open-source implementations like **grapeot/deep_research_agent** [263], which includes basic tool selection heuristics based on task classification.

Tool Execution Monitoring. Reliable tool usage requires effective execution monitoring and error handling. Commercial systems implement sophisticated monitoring frameworks that track tool execution, detect failures, and implement recovery strategies. OpenAI’s implementation includes explicit success criteria verification and fallback mechanisms for tool failures, ensuring reliable operation even with unreliable external components.

Open implementations like **Agent-RL/ReSearch** [2] demonstrate more accessible monitoring approaches, including simplified execution tracking and basic retry mechanisms for common failure modes. These implementations show that effective monitoring can be achieved with relatively straightforward implementation strategies.

Recent advances in agent collaboration frameworks [145, 221] highlight significant challenges in agent coordination [46], particularly for complex research tasks requiring diverse, specialized capabilities working in concert toward unified research objectives.

4.3.3 Cross-Platform Compatibility. Deployment flexibility requires careful attention to environmental dependencies:

Platform Abstraction Layers. Cross-platform implementations employ abstraction layers to isolate core logic from environmental dependencies. **TARS** [39] implements a sophisticated abstraction architecture that separates its core reasoning framework from platform-specific integration components, enabling deployment across diverse environments. Similar approaches are evident in **Nanobrowser** [184], which provides consistent browsing capabilities across different operating systems.

Containerization and Deployment Standardization. Modern implementations leverage containerization to ensure consistent deployment across environments. **OpenManus** [193] provides explicit container configurations that encapsulate all dependencies, enabling reliable deployment across diverse infrastructures. Similar approaches are employed by **AutoGLM-Research** [330], which provides standardized deployment configurations for different environments. Alongside containerization, modern cloud platforms such as Vercel [280] offer streamlined, standardized deployment workflows for the web-based interfaces of many research applications.

4.3.4 Research-Oriented Coding Assistance Integration. The integration of AI-powered coding assistants represents an increasingly important dimension of Deep Research system capabilities, particularly for computational research workflows requiring custom analysis scripts, data processing pipelines[108], and research automation tools.

Coding Assistant Integration Patterns. Modern research workflows increasingly depend on custom code development for data analysis, visualization, and automation tasks. AI coding assistants have emerged as crucial tools for enhancing researcher productivity in these computational aspects. The landscape of coding assistance tools demonstrates varying approaches to integration with research workflows, from IDE-native completion systems to conversational code generation interfaces. Systems like GitHub Copilot [20, 86] provide seamless integration within development environments, enabling context-aware code completion for research scripts and analysis workflows. Complementary approaches like ChatGPT-based code generation [309] offer conversational interfaces that can translate research requirements into executable implementations. More specialized frameworks like **AutoDev** [275], **DSPy**[257], and **Pydantic-AI**[216] enable end-to-end automated development workflows particularly suited for research prototype generation and experimental tool creation. Additionally, tools like Bolt [32] allow researchers to create web applications directly from text descriptions, handling the coding process while they focus on their vision. Evolutionary coding agents like AlphaEvolve [190] further enhance capabilities by iteratively optimizing algorithms using autonomous pipelines of LLMs and evolutionary feedback mechanisms. Recent research explores the synergy between generative AI and software

engineering, leveraging techniques like zero-shot prompting to enhance coding assistants and streamline development processes [41]. However, research has revealed limitations in these assistants’ capabilities, such as ambiguous beliefs regarding research claims and a lack of credible evidence to support their responses [35]. A large-scale survey demonstrates that developers frequently decline initial suggestions, citing unmet functional or non-functional requirements and challenges in controlling the tool to generate desired outputs [148]. User resistance behaviors documented in such surveys highlight the need for comprehensive adoption strategies, including providing active support during initial use, clearly communicating system capabilities, and adhering to predefined collaboration rules to mitigate low acceptance rates[252]. This underscores the need for adaptive hint systems, which can provide personalized support for bug finding and fixing by tailoring to user understanding levels and program representations to improve accuracy in debugging tasks [226]. Pioneering studies employ physiological measurements such as EEG and eye tracking to quantify developers’ cognitive load during AI-assisted programming tasks, addressing critical gaps in understanding actual usage patterns and productivity impacts [106]. Furthermore, tools like CodeScribe address challenges in AI-driven code translation for scientific computing by combining prompt engineering with user supervision to automate conversion processes while ensuring correctness [69]. Similarly, CodeCompose’s multi-line suggestion feature deployed at Meta demonstrates substantial productivity improvements, saving 17% of keystrokes through optimized latency solutions despite initial usability challenges [72]. Moreover, for debugging tasks, ChatDBG [139] enhances debugging capabilities by enabling programmers to engage in collaborative dialogues for root cause analysis and bug resolution, leveraging LLMs to provide domain-specific reasoning. Intelligent QA assistants are also being developed to streamline bug resolution processes [308], and grey literature reviews indicate a growing trend in AI-assisted test automation [231]. Additionally, benchmarks like CodeMMLU [163] evaluate code understanding and reasoning across diverse tasks, revealing significant comprehension gaps in current models despite advanced generative capabilities. Empirical evaluations of ACATs through controlled development scenarios demonstrate nuanced variations in acceptance patterns, modification reasons, and effectiveness based on task characteristics and user expertise [260]. Generative AI tools significantly enhance developer productivity by accelerating learning processes and altering collaborative team workflows through reduced repetitive tasks, fundamentally transforming development paradigms [277]. To realize the vision of next-generation AI coding assistants, it is crucial to address integration gaps and establish robust design principles such as setting clear usage expectations and employing extendable backend architectures [186].

Table 14. Qualitative Assessment of AI Coding Assistants for Research Applications

System	Documented Capabilities	Integration Approach	Evaluation Evidence	Research-Specific Features
GitHub Copilot [86, 319]	Code completion, documentation	IDE-native integration	User study on practices [319]	Limited domain specialization
Amazon CodeWhisperer [175]	Security-focused suggestions	AWS ecosystem integration	Comparative evaluation [309]	Cloud research workflows
ChatGPT Code [309]	Conversational code generation	API-based interaction	Code quality assessment [309]	Natural language specification
Cursor [65]	Context-aware completion	Codebase integration	No published evaluation	Repository-level understanding
Codeium [206]	Multi-language support	Editor extensions	Comparative benchmark [206]	Analysis workflow support
AutoDev [275]	Automated development	Task automation pipeline	Empirical evaluation [275]	End-to-end implementation
GPT-Pilot [217]	Project scaffolding	Guided development process	Repository demonstrations	Research prototype generation

Note: Capabilities and evaluations based on published studies and documented features. Comparative performance requires standardized evaluation across identical tasks.

The diversity of coding assistance approaches highlights the importance of integration flexibility within Deep Research systems. While some implementations benefit from tightly integrated coding assistance that understands research context, others require more flexible interfaces that can accommodate diverse

development workflows and programming paradigms. This integration dimension becomes particularly crucial as research increasingly requires custom computational tools and analysis pipelines that extend beyond pre-existing software packages[75, 244, 295]. Recent work by Chen et al. [53] demonstrates that proactive programming assistants, which automatically provide suggestions to enhance productivity and user experience, represent a key advancement in this domain. Additionally, ChatDev [220] exemplifies how linguistic communication serves as a unifying bridge for multi-agent collaboration in software development, streamlining the entire lifecycle from design to testing. Moreover, research on integrating AI assistants in Agile meetings reveals critical links to team collaboration dynamics and provides roadmaps for facilitating their adoption in development contexts [40]. As demonstrated by Talissa Dreossi[70], this hybrid approach bridges the gap between the high performance of deep learning models and the transparency of symbolic reasoning, advancing AI by providing interpretable and trustworthy applications.

Research Workflow Code Generation. Advanced coding assistants specifically optimized for research contexts demonstrate particular value in translating research methodologies into executable implementations. Systems like GPT-Pilot [217] enable guided development of complete research applications, while domain-specific tools can generate analysis scripts aligned with particular research methodologies or data types. These capabilities enhance research efficiency by reducing the technical barriers between research design and computational implementation.

Implementation patterns typically involve integration with research data management systems, version control workflows, and collaborative development environments that support reproducible research practices. The effectiveness of such integration depends significantly on the coding assistant’s understanding of research-specific requirements including documentation standards, reproducibility considerations, and domain-specific libraries and frameworks commonly used in particular research fields[124].

4.4 Technical Challenges and Solutions

Deep Research systems face numerous technical challenges that must be addressed for reliable, trustworthy operation.

4.4.1 Hallucination Control and Factual Consistency. Maintaining factual accuracy represents a fundamental challenge for LLM-based research systems:

Source Grounding Techniques. Advanced implementations employ explicit source grounding to enhance factual reliability. **Perplexity/DeepResearch** [209] implements strict attribution requirements that link all generated content to specific sources, reducing unsupported assertions. Similar approaches are evident in **OpenAI/DeepResearch** [197], which maintains explicit provenance tracking throughout the reasoning process.

Open-source implementations like **grapeot/deep_research_agent** [263] demonstrate more accessible grounding approaches, including simple but effective citation tracking and verification mechanisms. These techniques show that meaningful improvements in factual reliability can be achieved with straightforward implementation strategies.

Contradiction Detection and Resolution. Effective research requires identification and resolution of contradictory information. Commercial systems implement sophisticated contradiction detection mechanisms that identify inconsistencies between sources and implement resolution strategies [296]. **Gemini/DeepResearch**

[60] includes explicit uncertainty modeling and conflicting evidence presentation, enhancing transparency when definitive conclusions cannot be reached.

Open implementations like HKUDS/Auto-Deep-Research [112] employ simpler but useful contradiction identification approaches, flagging potential inconsistencies for user review. These implementations demonstrate that even basic contradiction handling can significantly enhance research reliability.

4.4.2 Privacy Protection and Security Design. Research systems must safeguard sensitive information and protect against potential misuse:

Query and Result Isolation. Secure implementations employ strict isolation between user queries to prevent information leakage. Commercial platforms implement sophisticated tenant isolation that ensures complete separation between different users’ research activities. Similar concerns motivate open-source implementations like OpenManus [193], which enables local deployment for sensitive research applications.

Source Data Protection. Responsible implementation requires careful handling of source information. Systems like Flowith/OracleMode [77] implement controlled data access patterns that respect source restrictions including authentication requirements and access limitations. These approaches enhance compliance with source terms of service while ensuring comprehensive information access. Recent advancements include benchmarking frameworks such as CI-Bench [56], which evaluates how well systems adhere to contextual norms and privacy expectations.

4.4.3 Explainability and Transparency. The scientific context places particularly stringent requirements on explanation quality. Mengaldo [170] argues that transparent explanation is not merely a feature but a fundamental requirement for scientific applications, emphasizing that black-box approaches fundamentally contradict scientific methodology’s requirement for transparent reasoning and reproducible results. This perspective suggests that explanation capabilities may require different standards in scientific Deep Research applications compared to general AI systems. Trustworthy research systems must provide insight into their reasoning processes and sources:

Reasoning Trail Documentation. Advanced implementations maintain explicit documentation of the reasoning process. OpenAI/DeepResearch [197] includes comprehensive reasoning traces that expose the analytical steps leading to specific conclusions. Similar capabilities are emerging in open-source alternatives like mshumer/OpenDeepResearcher [249], which includes basic reasoning documentation to enhance result interpretability.

Source Attribution and Verification. Transparent systems provide clear attribution for all information and enable verification. Perplexity/DeepResearch [209] implements comprehensive citation practices with explicit links to original sources, enabling direct verification of all claims. Similar approaches are employed by dzhng/deep-research [321], which maintains rigorous source tracking throughout the research process.

These implementation technologies and challenges highlight the complex engineering considerations involved in creating effective Deep Research systems. While commercial platforms benefit from extensive infrastructure and specialized components, open-source implementations demonstrate that effective research capabilities can be achieved through pragmatic approaches to the same fundamental challenges. The diversity

of implementation strategies across the ecosystem reflects different priorities in balancing capability, efficiency, reliability, and accessibility.

5 Evaluation Methodologies and Benchmarks

Rigorous evaluation of Deep Research systems presents unique challenges due to their complex capabilities and diverse application contexts. This section examines established frameworks for assessment, identifies emerging evaluation standards, and analyzes the strengths and limitations of current approaches.

Multi-dimensional Evaluation Framework for Deep Research Systems

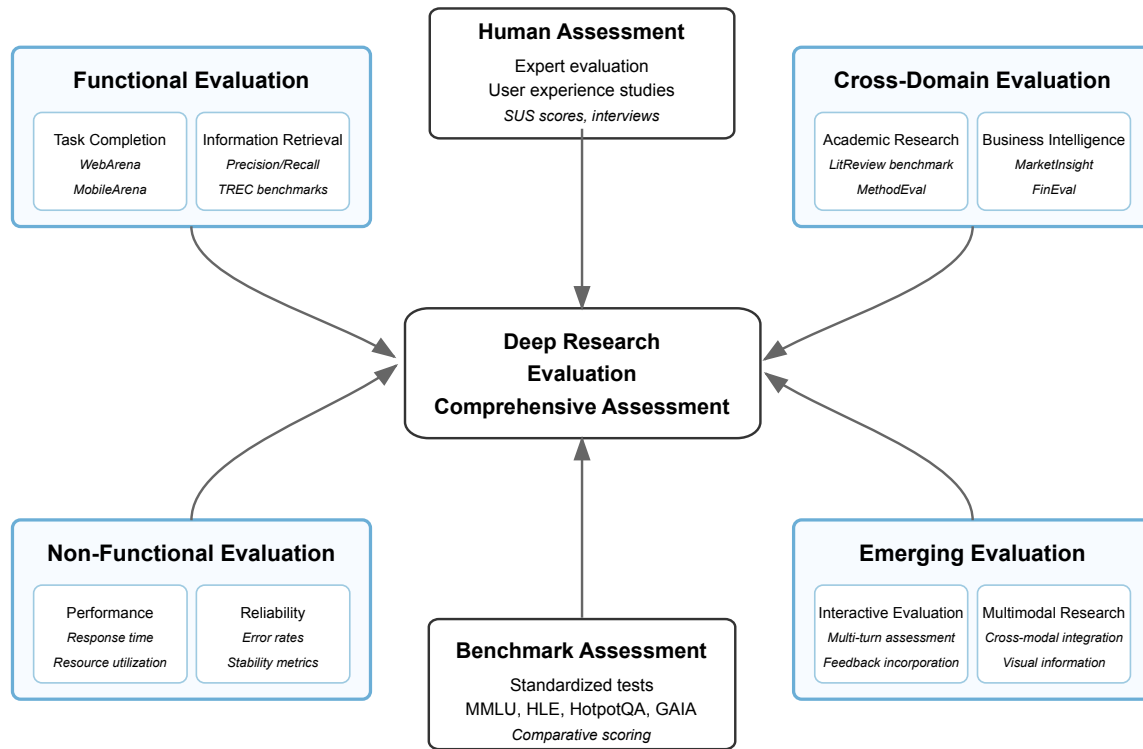


Fig. 8. Multi-dimensional Evaluation Framework for Deep Research Systems

5.1 Functional Evaluation Frameworks

Functional evaluation assesses core capabilities essential to effective research performance.

5.1.1 Task Completion Capability Assessment. The ability to successfully complete research tasks represents a fundamental evaluation dimension:

Task Success Rate Metrics. Quantitative assessment of task completion provides objective performance measures. Standardized evaluation suites like WebArena [332] measure successful completion of web-based

research tasks. For instance, **AutoGLM** [330] achieves a 55.2% success rate on VAB-WebArena-Lite (improving to 59.1% on a second attempt) and 96.2% on OpenTable evaluation tasks. Similarly, benchmarks like MobileArena evaluate successful completion of mobile interface tasks, where **AutoGLM** [330] demonstrates a 36.2% success rate on AndroidLab and 89.7% on common tasks in popular Chinese apps [153]. Domain-specific benchmarks, such as AutoPenBench for generative agents in penetration testing [85], provide further targeted assessments. These benchmarks provide meaningful comparative metrics, though with limitations in representing real-world research complexity.

These benchmarks provide meaningful comparative metrics, though with limitations in representing real-world research complexity. **Perplexity/DeepResearch** [209] explicitly highlights this distinction, noting that while benchmark performance provides comparative indicators, practical effectiveness depends significantly on task characteristics and domain specifics.

Multi-Attempt Resolution Rates. Effective research often involves iterative refinement with multiple attempts. Advanced evaluation frameworks incorporate multi-attempt metrics that assess system resilience and adaptability. **AutoGLM** [154] demonstrates significant performance improvement with second attempts (55.2% to 59.1% on WebArena-Lite), highlighting the importance of error recovery and adaptive strategies in practical research contexts.

Open-source frameworks like **Agent-RL/ReSearch** [2] explicitly emphasize iterative improvement through reinforcement learning approaches, demonstrating how evaluation methods that consider adaptability provide more comprehensive assessment than single-attempt metrics alone.

5.1.2 Information Retrieval Quality Evaluation. Effective information gathering forms the foundation of successful research:

Search Effectiveness Metrics. Information retrieval quality significantly impacts overall research performance. Evaluation frameworks employ metrics including precision (relevance of retrieved information), recall (comprehensiveness of coverage), and F1 scores (balanced measure of both). Systems like **Perplexity/DeepResearch** [209] demonstrate particular strength in recall metrics, effectively identifying comprehensive information across diverse sources.

Specialized information retrieval benchmarks like TREC [214] provide standardized assessment of search effectiveness. However, to the best of our knowledge, there is no specific evidence that the Deep Research systems from OpenAI, Google, Perplexity, or any of the open-source projects listed in this survey have been formally evaluated on TREC benchmarks [214]. This limitation motivates domain-specific evaluation approaches that better reflect particular research requirements.

Source Diversity Assessment. Comprehensive research requires balanced information from diverse perspectives and sources. Advanced evaluation frameworks incorporate explicit diversity metrics that assess the breadth of source utilization. Commercial systems like **Gemini/DeepResearch** [60] emphasize source diversity as a key performance indicator, while open implementations like **dzhng/deep-research** [321] incorporate specific mechanisms to ensure balanced source consideration.

Emerging evaluation approaches include explicit source spectra analysis that examines distribution across domains, perspectives, and publication types. These methods provide more nuanced assessment of

information gathering quality beyond simple relevance metrics, addressing concerns about potential bias in automated research processes.

5.1.3 Knowledge Synthesis Accuracy Assessment. Transforming information into accurate, coherent insights represents a crucial capability:

Factual Consistency Metrics. Reliable research requires accurate synthesis without introducing errors or misrepresentations. Evaluation frameworks employ fact verification techniques that compare generated content against source materials, identifying potential inaccuracies or unsupported claims. Systems like `grapeot/deep_research_agent` [263] emphasize factual verification through explicit source linking, enabling direct accuracy assessment. Benchmark suites like TruthfulQA [151] assess the truthfulness of language models under challenging conditions. While specific accuracy figures for `OpenAI/DeepResearch` [197] and `Perplexity/DeepResearch` [209] on TruthfulQA [151] are not publicly available, these systems have demonstrated notable performance on other rigorous benchmarks. For instance, `OpenAI/DeepResearch` [197] achieved a 26.6% accuracy [197] on Humanity’s Last Exam (HLE) [212]. Similarly, `Perplexity/DeepResearch` [209] attained a 21.1% accuracy [209] on the same benchmark. The development of unified, fine-grained, and multi-dimensional evaluation frameworks for summarization further advances the ability to assess the quality of synthesized content from LLMs [137]. These metrics provide standardized comparison points, though with recognized limitations in representing the complexity of real-world research synthesis.

Logical Coherence Assessment. Effective research requires logically sound integration of information into coherent analyses. Sophisticated evaluation approaches employ reasoning validity assessment that examines logical structures and inference patterns in research outputs. This dimension proves particularly challenging for automated assessment, often requiring expert human evaluation for reliable scoring.

Commercial systems like `OpenAI/DeepResearch` [197] and `Gemini/DeepResearch` [60] emphasize logical coherence in their evaluation frameworks, while open-source alternatives like `mshumer/OpenDeepResearcher` [249] incorporate simplified but useful logical consistency checks. These approaches highlight the importance of sound reasoning in effective research outputs beyond simple factual accuracy.

5.2 Non-Functional Evaluation Metrics

Beyond core functionality, practical effectiveness depends on operational characteristics that impact usability and deployment.

5.2.1 Performance and Efficiency Metrics. Operational efficiency significantly impacts practical utility:

Response Time Profiling. Timeliness represents a crucial dimension of research effectiveness. Evaluation frameworks incorporate response time metrics that measure completion duration across standardized tasks. Commercial systems demonstrate varying performance characteristics, with `Perplexity/DeepResearch` [209] achieving relatively quick response times (2-5 minutes for moderate tasks) while `OpenAI/DeepResearch` [197] typically requires longer processing (5-10 minutes) for similar complexity.

Open-source implementations generally demonstrate longer response times, though with significant variation based on implementation approaches and deployment environments. Systems like `nickscamara/`

open-deep-research [42] emphasize accessibility over performance optimization, while **QwenLM/Qwen-Agent** [224] incorporates specific optimizations to enhance response times within resource constraints.

Resource Utilization Assessment. Computational efficiency enables broader deployment and accessibility. Comprehensive evaluation includes resource profiling that measures memory consumption, computational requirements, and energy utilization across standardized workloads. Specialized benchmarks like Minerva assess programmable memory capabilities of language models, offering insights into their efficiency in handling long-context information [300]. Commercial cloud-based systems obscure some of these metrics due to their managed infrastructure, though with operational costs providing indirect resource indicators. Open implementations like **Camel-AI/OWL** [43] and **AutoGLM-Research** [330] provide more transparent resource profiles, enabling direct assessment of deployment requirements and operational economics. These metrics highlight significant variation in efficiency across the ecosystem, with implications for practical deployment scenarios and accessibility.

5.2.2 Reliability and Stability Metrics. Consistent performance under diverse conditions ensures practical usability:

Error Rate Analysis. Reliability under challenging conditions significantly impacts user trust and adoption. Robust evaluation frameworks incorporate error rate metrics that measure failure frequency across diverse scenarios. Commercial systems generally demonstrate lower error rates compared to open-source alternatives, though with remaining challenges in complex or novel research contexts.

Specialized reliability testing employs adversarial scenarios designed to trigger failure modes, providing insight into system robustness. Systems like **OpenAI/DeepResearch** [197] and **Agent-RL/ReSearch** [2] incorporate explicit error recovery mechanisms that enhance reliability under challenging conditions, highlighting the importance of resilience in practical research applications.

Long-Term Stability Assessment. Consistent performance over extended operation provides crucial deployment confidence. Comprehensive evaluation includes stability metrics that measure performance consistency across extended sessions and repeated executions. This dimension proves particularly relevant for open-source implementations that must operate in diverse deployment environments with varying infrastructure stability.

Systems like **Flowith/OracleMode** [77] and **TARS** [39] emphasize operational stability through robust error handling and recovery mechanisms, enabling reliable performance in production environments. These capabilities highlight the importance of engineering quality beyond core algorithmic performance in practical research applications.

5.2.3 User Experience and Usability Metrics. Effective interaction significantly impacts practical utility:

Interface Usability Assessment. Intuitive interfaces enhance accessibility and effective utilization. Usability evaluation frameworks employ standardized usability metrics including System Usability Scale (SUS) [140] scores and task completion time measurements. Commercial systems typically demonstrate stronger usability characteristics, with **Perplexity/DeepResearch** [209] particularly emphasizing intuitive interaction for non-technical users. Open-source alternatives show greater variability, with implementations like **HKUDS/Auto-Deep-Research** [112] incorporating specific interface enhancements to improve accessibility.

User studies provide more nuanced usability assessment beyond standardized metrics. Evaluations of systems like **Manus** [164] and **Flowith/OracleMode** [77] incorporate explicit user feedback to identify interaction challenges and improvement opportunities. These approaches highlight the importance of human-centered design in practical research applications beyond technical performance. Similarly, frameworks such as **AdaptoML-UX** [87] enable HCI researchers to employ automated ML pipelines without specialized expertise, facilitating robust model development and customization.

Learning Curve Assessment. Approachability for new users significantly impacts adoption and effective utilization. Comprehensive evaluation includes learning curve metrics that measure time-to-proficiency across user segments with varying technical backgrounds. Commercial systems generally demonstrate gentler learning curves, with **Perplexity/DeepResearch** [209] explicitly designed for accessibility to non-technical users.

Open implementations show greater variability, with systems like **n8n** [183] requiring more technical expertise for effective deployment and utilization. More accessible alternatives like **nickscamara/open-deep-research** [42] incorporate simplified interfaces designed for broader accessibility, highlighting diverse approaches to the accessibility-sophistication balance across the ecosystem.

5.3 Cross-Domain Evaluation Benchmarks

Standardized benchmarks enable objective comparison across systems and domains.

5.3.1 Academic Research Task Benchmarks. Specialized benchmarks assess capabilities relevant to scholarly research:

Literature Review Benchmarks. Comprehensive literature synthesis represents a fundamental academic research task requiring sophisticated information retrieval, critical analysis, and synthesis capabilities. To the best of our knowledge, no benchmark suite is specifically designed to evaluate systems’ ability to identify relevant literature, synthesize key findings, and highlight research gaps across scientific domains. We propose leveraging existing high-quality literature reviews published in *Nature Reviews* journals as gold standards. Citation networks from academic knowledge graphs—such as Microsoft Academic Graph, Semantic Scholar Academic Graph, and Open Academic Graph—could provide complementary evaluation data by measuring a system’s ability to traverse citation relationships and identify seminal works[1, 31].

While direct literature review benchmarks remain underdeveloped, several indirect benchmarks offer insight into related capabilities. **OpenAI/DeepResearch** [197] demonstrates leading performance, achieving 26.6% accuracy on Humanity’s Last Exam (HLE) [212] and averaging 72.57% on the GAIA benchmark [172], reflecting strong performance in complex reasoning tasks essential for literature synthesis. Similarly, **Perplexity/DeepResearch** [209] achieves 21.1% accuracy on HLE [212] and 93.9% on SimpleQA [290], indicating robust factual retrieval capabilities.

These benchmarks include challenging cases requiring integration across multiple disciplines, identification of methodological limitations, and disambiguation of conflicting findings—all crucial for effective literature review. Such tasks demonstrate the importance of sophisticated reasoning capabilities beyond simple information retrieval. While specific performance metrics for systems like **Camel-AI/OWL** [43] are not publicly

available, their specialized academic optimization suggests potential effectiveness in handling complex synthesis tasks.

Methodology Evaluation Benchmarks. Critical assessment of research methodology requires sophisticated analytical capabilities. To the best of our knowledge, no benchmark is specifically designed for quantitative methodology assessment of strengths and limitations. A comprehensive methodology evaluation benchmark would need to assess a system’s ability to identify flaws in research design, statistical approaches, sampling methods, and interpretive limitations across diverse disciplines. An effective benchmark might incorporate multi-layered evaluation criteria including: reproducibility assessment, identification of confounding variables, appropriate statistical power analysis, and proper handling of uncertainty. Future benchmarks could utilize expert-annotated corpora of research papers with methodological strengths and weaknesses clearly marked, creating a gold standard against which systems’ analytical capabilities can be measured while minimizing bias through diverse evaluation metrics that reflect methodological best practices across different fields of inquiry.

Beyond standard benchmarks, case study evaluations of complete AI scientist systems provide valuable insights into current capabilities. Beel et al. [24] conduct a detailed assessment of Sakana’s AI Scientist for autonomous research, examining whether current implementations represent genuine progress toward “Artificial Research Intelligence” or remain limited in fundamental ways, highlighting the gap between current benchmarks and comprehensive research capability evaluation.

5.3.2 Business Analysis Task Benchmarks. Standardized evaluation for business intelligence applications:

Market Analysis Benchmarks. Strategic decision support necessitates a comprehensive understanding of market dynamics. Advanced AI systems, such as OpenAI/DeepResearch [197], are designed to analyze competitive landscapes, identify market trends, and generate strategic recommendations based on diverse business information. OpenAI/DeepResearch has demonstrated significant capabilities in handling complex, multi-domain data analysis tasks, providing detailed insights and personalized recommendations. Similarly, Google’s Gemini/DeepResearch [60] offers robust performance in processing extensive datasets, delivering concise and factual reports efficiently.

These benchmarks include challenging scenarios requiring integration of quantitative financial data with qualitative market dynamics and regulatory considerations. Such tasks highlight the importance of both analytical depth and domain knowledge, with systems like Manus [164] demonstrating strong performance through specialized business intelligence capabilities.

Financial Analysis Benchmarks. Economic assessment requires sophisticated quantitative reasoning combined with contextual understanding of market dynamics. The FinEval benchmark [103] provides a standardized framework for measuring systems’ capabilities in analyzing financial statements, evaluating investment opportunities, and assessing economic risk factors across diverse scenarios. To our knowledge, no Deep Research projects have yet published official FinEval benchmark results, though several commercial demonstrations suggest strong performance in this domain. OpenAI/DeepResearch [197] has demonstrated particular strength in quantitative financial analysis through its ability to process complex numerical data while incorporating relevant market context. Meanwhile, open-source implementations show more variable performance, though specialized systems like n8n [183] achieve competitive results through strategic

integration with financial data sources and analytical tools. These patterns highlight the critical importance of domain-specific integrations and data accessibility in financial analysis applications, extending beyond core language model capabilities to create truly effective analytical systems.

5.3.3 General Knowledge Management Benchmarks. Broad applicability assessment across general research domains:

Factual Research Benchmarks. Accurate information gathering forms the foundation of effective research. The SimpleQA benchmark [290] evaluates language models’ ability to answer short, fact-seeking questions with a single, indisputable answer. **Perplexity/DeepResearch** [209] demonstrates exceptional performance on this benchmark, achieving an accuracy of 93.9% [209]. OpenAI’s Deep Research tool, integrated into ChatGPT, offers comprehensive research capabilities, though specific accuracy metrics on SimpleQA [290] are not publicly disclosed [197]. Similarly, Google’s **Gemini/DeepResearch** provides robust information synthesis features, but detailed performance data on SimpleQA [290] is not available.

These metrics provide useful baseline performance indicators, though with recognized limitations in representing more complex research workflows. Comparative evaluation highlights the importance of information quality beyond simple factual recall, with sophisticated systems demonstrating more nuanced performance profiles across complex tasks.

Humanities and Social Sciences Benchmarks. Comprehensive evaluation requires assessment beyond STEM domains. The MMLU benchmark [33] evaluates systems’ performance across humanities and social science research tasks, including historical analysis, ethical evaluation, and social trend identification. Performance shows greater variability compared to STEM-focused tasks, with generally lower accuracy across all systems while maintaining similar relative performance patterns. These benchmarks highlight remaining challenges in domains requiring nuanced contextual understanding and interpretive reasoning. Commercial systems maintain performance leads, though with open alternatives like **smolagents/open_deep_research** [115] demonstrating competitive capabilities in specific humanities domains through specialized component design.

5.4 Emerging Evaluation Approaches

Beyond established benchmarks, novel evaluation methods address unique aspects of Deep Research performance.

Interactive Evaluation Frameworks. Traditional static benchmarks often fail to capture the dynamic and interactive nature of real-world research workflows. To address this gap, interactive evaluation frameworks have been developed to assess AI systems’ abilities to iteratively refine research strategies through multiple interaction rounds. Notably, QuestBench [141] is a novel benchmark which specifically assesses an AI system’s ability to identify missing information and ask appropriate clarification questions, a crucial skill for real-world research scenarios where problems are often underspecified. To the best of our knowledge, no deep research system investigated in this survey has yet been publicly evaluated using QuestBench. Nonetheless, these systems have demonstrated strong performance in other interactive evaluations, highlighting their effectiveness in supporting iterative research processes.

Multimodal Research Evaluation. Comprehensive research increasingly involves diverse content modalities. Advanced evaluation frameworks incorporate multimodal assessment that measures systems’ ability to integrate information across text, images, data visualizations, and structured content. Commercial systems generally demonstrate stronger multimodal capabilities, with **Gemini/DeepResearch** [60] particularly excelling in image-inclusive research tasks.

Open implementations show emerging multimodal capabilities, with systems like **Jina-AI/node-DeepResearch** [121] incorporating specific components for multimodal content processing. These approaches highlight the growing importance of cross-modal integration in practical research applications beyond text-centric evaluation.

Ethical and Bias Assessment. Responsible research requires careful attention to ethical considerations and potential biases. Comprehensive evaluation increasingly incorporates explicit assessment of ethical awareness, bias detection, and fairness in information processing. Commercial systems implement sophisticated safeguards, with **OpenAI/DeepResearch** [197] incorporating explicit ethical guidelines and bias mitigation strategies. Open implementations show varied approaches to these considerations, with systems like **grapeot/deep_research_agent** [263] emphasizing transparency in source selection and attribution.

These evaluation dimensions highlight the importance of responsibility beyond technical performance, addressing growing concerns about potential amplification of existing information biases through automated research systems. Ongoing development of standardized ethical evaluation frameworks represents an active area of research with significant implications for system design and deployment.

The diverse evaluation approaches outlined in this section highlight both the complexity of comprehensive assessment and the ongoing evolution of evaluation methodologies alongside system capabilities. While standard benchmarks provide useful comparative metrics, practical effectiveness depends on alignment between system capabilities, evaluation criteria, and specific application requirements. This alignment represents a key consideration for both system developers and adopters seeking to integrate Deep Research capabilities into practical workflows.

5.5 Comparative Evaluation Methodology

To ensure systematic and consistent evaluation across diverse Deep Research systems, we have developed a comprehensive evaluation framework. This section outlines our methodological approach, evaluation criteria selection, and application consistency across systems.

5.5.1 Systems Selection Criteria. Our evaluation encompasses various Deep Research systems selected based on the following criteria:

- **Functional Completeness:** Systems must implement at least two of the three core dimensions of Deep Research as defined in Section 1.1
- **Public Documentation:** Sufficient technical documentation must be available to enable meaningful analysis
- **Active Development:** Systems must have demonstrated active development or usage within the past 12 months

- **Representational Balance:** Selection ensures balanced representation of commercial, open-source, general-purpose, and domain-specialized implementations

5.5.2 Evaluation Dimensions and Metrics Application. Our evaluation employs a consistent set of dimensions across all systems, though the specific benchmarks within each dimension vary based on system focus and available performance data. Table 15 presents the evaluation coverage across representative systems.

Table 15. Evaluation Metrics Application Across Systems

System	Functional Benchmarks	Performance Metrics	Efficiency Metrics	Domain-Specific Benchmarks	Usability Assessment
OpenAI/DeepResearch	HLE, GAIA	Factual accuracy	Response time	Academic citation	User interface
Gemini/DeepResearch	MMLU	Output coherence	Cloud compute	Market analysis	Mobile support
Perplexity/DeepResearch	HLE, SimpleQA	Source diversity	Response time	Legal search	Multi-device
Grok3Beta	MMLU	Source verification	Cloud efficiency	Financial analysis	Voice interface
Manus	GAIA	Cross-domain	API latency	Business analysis	Dashboard
Agent-RL/ReSearch	HotpotQA	Planning efficiency	Local compute	Scientific research	CLI interface
AutoGLM-Research	WebArena	GUI navigation	Mobile efficiency	Domain adaptation	Accessibility
n8n	Workflow	API integration	Self-hosted	Enterprise workflow	No-code design

5.5.3 Data Collection Methods. Our evaluation data comes from four primary sources:

- (1) **Published Benchmarks:** Performance metrics reported in peer-reviewed literature or official system documentation
- (2) **Technical Documentation Analysis:** Capabilities and limitations outlined in official documentation, APIs, and technical specifications
- (3) **Repository Examination:** Analysis of open-source code repositories for architectural patterns and implementation approaches
- (4) **Experimental Verification:** Where inconsistencies exist, we conducted direct testing of publicly available systems to verify capabilities

When benchmark results are unavailable for specific systems, we indicate this gap explicitly rather than extrapolating performance. This approach ensures transparency regarding the limits of our comparative analysis while maintaining the integrity of available evaluation data.

5.5.4 Cross-System Comparison Challenges. Several methodological challenges exist in comparing Deep Research systems:

- **Benchmark Diversity:** Different systems emphasize different benchmarks based on their focus areas
- **Implementation Transparency:** Commercial systems often provide limited details about internal architectures
- **Rapid Evolution:** Systems undergo frequent updates, potentially rendering specific benchmark results obsolete
- **Domain Specialization:** Domain-specific systems excel on targeted benchmarks but may perform poorly on general evaluations

We address these challenges through qualitative architectural analysis alongside quantitative benchmarks, enabling meaningful comparison despite data limitations. Section 3.3 presents the resulting comparative analysis, highlighting both performance differentials and the limitations of direct comparison across heterogeneous implementations.

6 Applications and Use Cases

The technical capabilities of Deep Research systems enable transformative applications across diverse domains. This section examines implementation patterns, domain-specific adaptations, and representative use cases that demonstrate the practical impact of these technologies.

Deep Research Application Domains and Use Cases

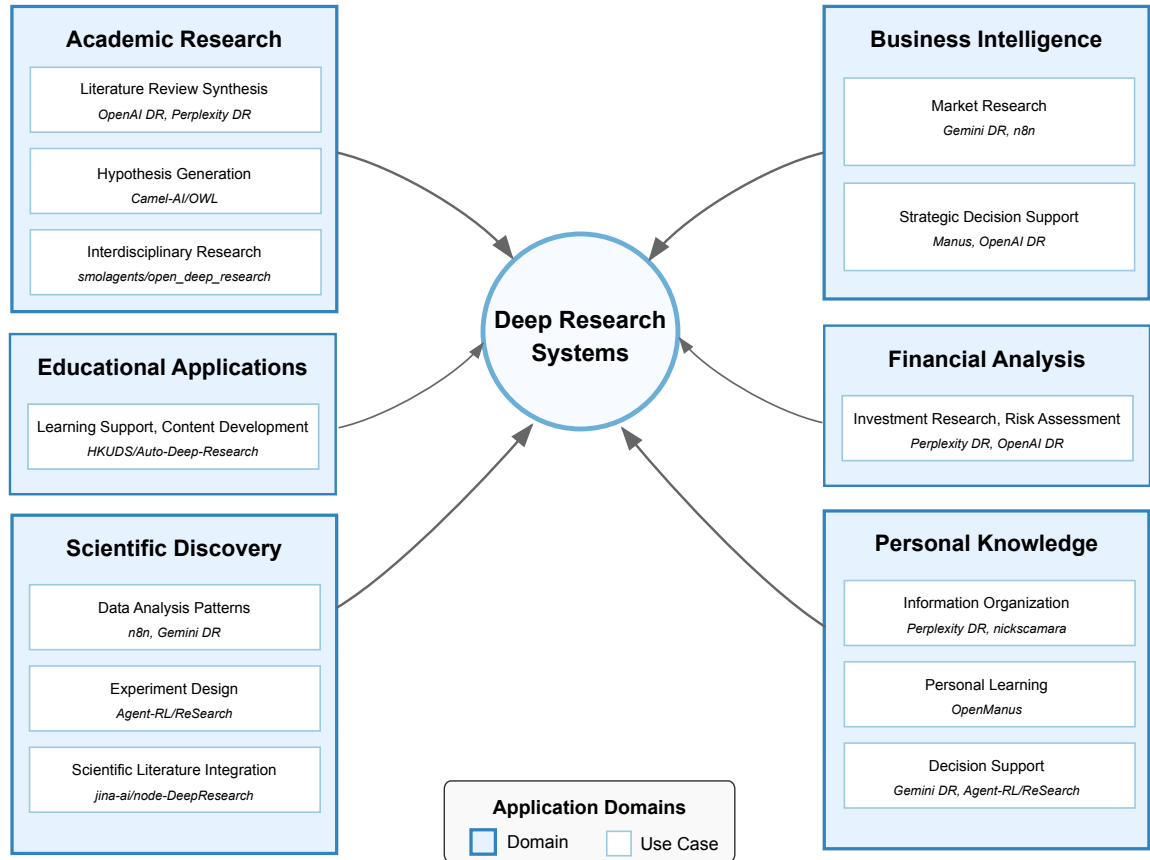


Fig. 9. Deep Research Application Domains and Use Cases

6.1 Academic Research Applications

Deep Research systems offer significant enhancements to scholarly research workflows.

6.1.1 Literature Review and Synthesis. Comprehensive literature analysis forms the foundation of effective research:

Systematic Review Automation. Deep Research systems demonstrate particular effectiveness for systematic literature reviews requiring exhaustive coverage of existing research. Systems like Google’s **Gemini/DeepResearch** [60] can efficiently analyze thousands of research papers, a capability that has significant implications for fields like biomedicine where the volume of literature makes comprehensive manual review increasingly challenging [289]. **OpenAI/DeepResearch** [197] has been successfully deployed for medical research reviews, analyzing thousands of publications to identify intervention efficacy patterns with significantly reduced human effort compared to traditional methods. Similar capabilities are evident in **Perplexity/DeepResearch** [209] and **Gemini/DeepResearch** [60], which enables rapid synthesis of research findings across disciplinary boundaries. Generative AI frameworks integrating retrieval-augmented generation further automate systematic reviews by expanding user queries to retrieve relevant scholarly articles and reduce time and resource burdens [234].

Open-source implementations like **dzhng/deep-research** [321] have found adoption in academic settings where local deployment and customization are prioritized. Specialized scientific implementations like **AI-Researcher** [109] extend these capabilities with domain-specific optimizations for academic literature processing and analysis. These systems enable literature review automation with greater control over search scope and synthesis methods, particularly valuable for specialized research domains with unique requirements. Implementation patterns typically involve customization of search strategies, source weightings, and output formats to align with disciplinary conventions.

Research Gap Identification. Beyond simple synthesis, advanced systems effectively identify unexplored areas and research opportunities. **Gemini/DeepResearch** [60] has demonstrated this capability in interdisciplinary contexts, identifying connection opportunities between distinct research domains that might otherwise remain undiscovered. This application leverages the system’s ability to process extensive literature across fields while identifying patterns and absences in existing research coverage.

Open implementations like **HKUDS/Auto-Deep-Research** [112] incorporate specific mechanisms for gap analysis, including explicit detection of methodological limitations and underexplored variables across research corpora. These capabilities highlight the potential for automated systems to not only synthesize existing knowledge but actively contribute to research direction through systematic gap identification.

6.1.2 Hypothesis Generation and Testing. AI-assisted hypothesis development enhances research creativity and validation:

Hypothesis Formulation Support. Deep Research systems effectively generate testable hypotheses based on existing literature and theoretical frameworks. **OpenAI/DeepResearch** [197] provides explicit hypothesis generation capabilities, identifying potential causal relationships and testable predictions derived from literature synthesis. These features enable researchers to explore broader possibility spaces than might be practical through manual review alone.

Specialized frameworks like **Camel-AI/OWL** [43] implement domain-specific hypothesis generation for scientific applications, incorporating field-specific constraints and validation criteria. These approaches highlight how domain adaptation enhances the practical utility of hypothesis generation capabilities beyond

generic formulation. Implementation patterns typically involve iterative refinement with researcher feedback to align generated hypotheses with specific research objectives.

Preliminary Validation Assessment. Advanced systems support hypothesis validation through evidence assessment and methodological planning. **Gemini/DeepResearch** [60] enables preliminary hypothesis testing through automated data source identification, statistical power analysis, and potential confound identification. These capabilities streamline the transition from hypothesis formulation to empirical testing, reducing manual effort in research design.

Open implementations like **Agent-RL/ReSearch** [2] incorporate specific validation planning components, guiding researchers through experimental design considerations based on hypothesis characteristics. These approaches demonstrate how Deep Research capabilities extend beyond information gathering to actively support the complete research workflow from conception through validation planning.

6.1.3 Interdisciplinary Research Support. Cross-domain integration represents a particular strength of automated research systems:

Cross-Domain Knowledge Translation. Deep Research systems effectively bridge terminological and conceptual gaps between disciplines. **Perplexity/DeepResearch** [209] demonstrates this capability through explicit concept mapping between fields, enabling researchers from diverse backgrounds to explore unfamiliar domains with reduced onboarding barriers. This application leverages the system’s broad knowledge base to identify conceptual parallels across disciplinary boundaries.

Open frameworks like **smolagents/open_deep_research** [115] implement specialized agents for disciplinary translation, with explicit focus on terminological mapping and concept alignment. These approaches highlight how multi-agent architectures can effectively address the challenges of interdisciplinary communication through specialized component design[117].

Methodology Transfer Facilitation. Advanced systems enable effective adaptation of research methods across domains. **OpenAI/DeepResearch** [197] supports methodology transfer through explicit identification of adaptation requirements and implementation guidance when applying techniques from one field to another. This capability accelerates methodological innovation by facilitating cross-pollination between research traditions. Implementation patterns typically involve specialized methodological components like those in **QwenLM/Qwen-Agent** [224], which incorporates explicit methodology modeling to identify transfer opportunities and adaptation requirements. This is particularly relevant in fields like engineering, where AI is beginning to impact established design procedures for complex dynamical systems [67]. These approaches demonstrate how Deep Research systems can actively contribute to methodological innovation beyond simple information retrieval and synthesis.

6.2 Scientific Discovery Applications

Deep Research technologies enable enhanced scientific investigation across disciplines.

6.2.1 Data Analysis and Pattern Recognition. Automated analysis enhances insight extraction from complex scientific data:

Large-Scale Data Synthesis. Deep Research systems effectively integrate findings across extensive datasets to identify broader patterns. **Gemini/DeepResearch** [60] has been applied to climate science research, synthesizing findings across hundreds of climate models and observational datasets to identify consistent patterns and outliers. This application leverages the system’s ability to process and integrate diverse data formats while maintaining analytical coherence. Open implementations like **n8n** [183] enable similar capabilities through workflow automation that coordinates specialized analytical tools across complex data processing pipelines. Furthermore, **SqlCompose** [161] enhances analytical workflows by automating SQL authoring to reduce syntax barriers and improve efficiency in large-scale data operations, as demonstrated through enterprise deployment and user feedback. Systems like **DataInquirer** quantitatively measure workflow patterns and task execution consistency, revealing significant variations across practitioners while also assessing AI tool impacts on aligning novice approaches with expert practices [325]. AI assistants specifically designed for data wrangling tasks can provide semi-automated support in transforming and cleaning data through interactive recommendations, thereby enhancing workflow efficiency [211]. Other systems assist domain experts in making sense of multi-modal personal tracking data through visualization and human-in-the-loop LLM agents [143]. Additionally, no-code machine-readable documentation frameworks support responsible dataset evaluation by facilitating quality assessment and accuracy verification during large-scale data synthesis [233]. These approaches demonstrate how tool integration capabilities extend analytical reach beyond the core language model’s native capabilities, particularly valuable for quantitative scientific applications.

Anomaly Detection and Investigation. Advanced systems effectively identify unexpected patterns and facilitate targeted investigation. **OpenAI/DeepResearch** [197] demonstrates this capability in pharmacological contexts, identifying unexpected drug interaction patterns across clinical literature and proposing mechanistic explanations for further investigation. This application combines pattern recognition with explanatory hypothesis generation to enhance scientific discovery.

Specialized tools like **grapeot/deep_research_agent** [263] implement focused anomaly detection capabilities, with particular emphasis on statistical outlier identification and contextual explanation. These approaches highlight how targeted optimization can enhance specific scientific workflows beyond general-purpose research capabilities[125].

6.2.2 Experiment Design and Simulation. AI assistance enhances experimental planning and virtual testing:

Experimental Protocol Optimization. Deep Research systems support experimental design through comprehensive protocol development and optimization. **Gemini/DeepResearch** [60] provides explicit protocol generation capabilities, incorporating existing methodological best practices while identifying potential confounds and control strategies. These features streamline experimental planning while enhancing methodological rigor.

Open implementations like **Agent-RL/ReSearch** [2] incorporate specialized experimental design components with particular emphasis on statistical power optimization and confound control. These approaches demonstrate how focused optimization can enhance specific scientific workflows through specialized component design targeting critical research phases.

Despite these capabilities, significant gaps remain between current systems and truly autonomous scientific discovery. Yu et al. [314] identify critical missing elements in current AI research systems, particularly highlighting limitations in open-ended exploration, creative hypothesis generation, and experimental design optimization that constrain their effectiveness in leading scientific discovery processes.

Theoretical Model Testing. Advanced systems enable accelerated testing of theoretical models through simulation and virtual experimentation. **OpenAI/DeepResearch** [197] supports this application through integration with computational modeling tools, enabling rapid assessment of theoretical predictions against existing evidence. This capability accelerates theory refinement by identifying empirical constraints and validation opportunities more efficiently than manual methods.

Implementation patterns typically involve specialized tool integration like that found in **Manus** [164], which provides sophisticated orchestration of computational modeling and simulation tools within research workflows. Systems like **AgentLaboratory** [237] further enhance these capabilities through specialized experimental design components that generate statistically rigorous protocols based on research objectives and methodological best practices. These approaches highlight how tool integration capabilities significantly enhance scientific applications beyond the language model’s native capabilities.

6.2.3 Scientific Literature Integration. Comprehensive knowledge integration enhances scientific understanding:

Cross-Modal Scientific Content Analysis. Deep Research systems effectively integrate information across text, data, and visualizations prevalent in scientific literature. **Gemini/DeepResearch** [60] demonstrates particular strength in this application, extracting and synthesizing information from scientific figures, tables, and text into cohesive analyses. This capability enables more comprehensive literature utilization than text-only approaches.

Open implementations like **Jina-AI/node-DeepResearch** [121] incorporate specialized components for multimodal scientific content processing, enabling similar capabilities in customizable frameworks. These approaches highlight the growing importance of multimodal processing in scientific applications, reflecting the diverse information formats prevalent in scientific communication.

Conflicting Evidence Resolution. Advanced systems help navigate contradictory findings common in scientific literature. **Perplexity/DeepResearch** [209] provides explicit conflict identification and resolution guidance, identifying methodological differences, contextual factors, and potential reconciliation approaches when faced with contradictory evidence. This capability enhances scientific understanding by providing structured approaches to evidence integration rather than simple aggregation.

Implementation patterns typically involve sophisticated evidence modeling like that found in **HKUDS/Auto-Deep-Research** [112], which implements explicit evidence weighting and confidence estimation mechanisms. These approaches demonstrate how specialized components for scientific evidence handling enhance the practical utility of Deep Research systems in complex scientific contexts.

6.2.4 Autonomous Scientific Discovery. Fully autonomous research systems represent an emerging direction that extends current Deep Research capabilities toward greater autonomy. Recent work in this area includes the AI Scientist system [159] that implements an automated discovery loop with hypothesis generation,

experimentation, and theory revision capacities. Similarly, the Dolphin system [316] demonstrates how closed-loop auto-research can integrate thinking, practice, and feedback mechanisms to implement systematic scientific discovery processes.

This evolution toward more autonomous operation represents a significant advancement beyond traditional tool-based approaches, enabling continuous research cycles with minimal human intervention while maintaining scientific rigor through structured validation processes. Systems like CycleResearcher [294] further enhance this approach by incorporating automated peer review mechanisms [150] that improve output quality through systematic feedback loops mimicking scientific review processes.

Practical implementation of these concepts appears in systems like AgentLaboratory [240], which demonstrates how LLM agents can function as effective research assistants within structured laboratory environments. Complementing these approaches, the concept of self-maintainability (SeM) addresses critical gaps in laboratory automation by enabling systems to autonomously adapt to disturbances and maintain operational readiness [191]. In addition, strategies such as BOLAA [156] orchestrate multiple specialized agents by employing a controller to manage communication among them, enhancing the resolution of complex tasks. Moreover, Automated Capability Discovery (ACD) [158] automates the evaluation of foundation models by designating one model as a scientist to propose open-ended tasks that systematically uncover unexpected capabilities and failures. Similarly, SeqMate [178] utilizes large language models to automate RNA sequencing data preparation and analysis, enabling user-friendly one-click analytics and report generation for biologists. The FutureHouse Platform [253] broadens accessibility by delivering the first publicly available superintelligent AI agents for scientific discovery through web interfaces and APIs. These implementations highlight both the significant potential and current limitations of autonomous scientific discovery systems, suggesting an evolutionary path toward increasingly capable research automation while maintaining appropriate human oversight and validation.

6.3 Business Intelligence Applications

Deep Research technologies enable enhanced strategic decision support in commercial contexts.

6.3.1 Market Research and Competitive Analysis. Comprehensive market understanding supports strategic planning:

Competitor Landscape Mapping. Deep Research systems effectively synthesize comprehensive competitive intelligence across diverse sources. **Gemini/DeepResearch** [60] enables detailed competitor analysis across financial disclosures, product announcements, market reception, and strategic positioning to identify competitive dynamics and market opportunities. This application leverages the system’s ability to integrate information across public and specialized business sources with current market context.

Open implementations like **n8n** [183] support similar capabilities through workflow automation that integrates specialized business intelligence data sources. These approaches demonstrate how effective tool integration can create sophisticated business intelligence applications by coordinating specialized components within consistent analytical frameworks.

Emerging Trend Identification. Advanced systems effectively identify early-stage market trends and potential disruptions. **OpenAI/DeepResearch** [197] demonstrates this capability through temporal pattern analysis

across industry publications, startup activity, and technology development indicators. This application combines historical pattern recognition with current signal detection to anticipate market evolution with greater lead time than manual methods alone.

Implementation patterns typically involve specialized analytical components like those in **Flowith/OracleMode** [77], which incorporates explicit trend modeling and weak signal amplification techniques. These approaches highlight how specialized optimization enhances business intelligence applications through components targeting specific analytical requirements.

6.3.2 *Strategic Decision Support.* AI-enhanced analysis informs high-stakes business decisions:

Investment Opportunity Assessment. Deep Research systems support investment analysis through comprehensive opportunity evaluation. **Perplexity/DeepResearch** [209] enables detailed investment analysis incorporating financial metrics, market positioning, competitive dynamics, and growth indicators within unified analytical frameworks. This application integrates quantitative financial assessment with qualitative market understanding to support more comprehensive investment evaluation.

Open frameworks like **mshumer/OpenDeepResearcher** [249] implement investment analysis components with particular emphasis on structured evaluation frameworks and comprehensive source integration. These approaches demonstrate how domain-specific optimization enhances practical utility for specialized business applications beyond generic research capabilities.

Risk Factor Identification. Advanced systems support risk management through comprehensive threat identification and assessment. **Gemini/DeepResearch** [60] provides explicit risk analysis capabilities, identifying potential threats across regulatory, competitive, technological, and market dimensions with associated impact and likelihood estimation. These features enable more comprehensive risk management than might be practical through manual analysis alone.

Implementation patterns typically involve specialized risk modeling components like those found in **Manus** [164], which incorporates explicit risk categorization and prioritization mechanisms. These approaches highlight how targeted optimization enhances specific business workflows through specialized components addressing critical decision support requirements.

6.3.3 *Business Process Optimization.* Research-driven insights enhance operational effectiveness:

Best Practice Identification. Deep Research systems effectively synthesize operational best practices across industries and applications. **OpenAI/DeepResearch** [197] enables comprehensive process benchmarking against industry standards and innovative approaches from adjacent sectors, identifying optimization opportunities that might otherwise remain undiscovered. This application leverages the system’s broad knowledge base to facilitate cross-industry learning and adaptation.

Open implementations like **TARS** [39] support similar capabilities through workflow analysis and recommendation components designed for business process optimization. These approaches demonstrate how domain adaptation enhances practical utility for specific business applications beyond general research capabilities.

Implementation Planning Support. Advanced systems support process change through comprehensive implementation guidance. **Gemini/DeepResearch** [60] provides detailed implementation planning incorporating change management considerations, resource requirements, and risk mitigation strategies derived from

similar initiatives across industries. This capability accelerates organizational learning by leveraging broader implementation experience than typically available within single organizations.

Implementation patterns typically involve specialized planning components like those in `QwenLM/Qwen-Agent` [224], `HuggingGPT`[246], `XAgent`[202], `Mastra`[168], `Letta`[138] and `SemanticKernel`[174] which incorporates explicit process modeling and change management frameworks. These approaches highlight how targeted optimization enhances specific business workflows through specialized components addressing critical implementation challenges.

6.4 Financial Analysis Applications

Deep Research technologies enable enhanced financial assessment and decision support.

6.4.1 Investment Research and Due Diligence. AI-enhanced analysis supports investment decisions across asset classes:

Comprehensive Asset Evaluation. Deep Research systems enable detailed asset analysis across financial and contextual dimensions. `Perplexity/DeepResearch` [209] supports investment research through integration of financial metrics, market positioning, competitive dynamics, and growth indicators within unified analytical frameworks. This application enhances investment decision quality through more comprehensive information integration than typically practical through manual methods alone.

Open implementations like `n8n` [183] enable similar capabilities through workflow automation that integrates specialized financial data sources and analytical tools. These approaches demonstrate how effective tool orchestration creates sophisticated financial applications by coordinating specialized components within consistent analytical frameworks.

Management Quality Assessment. Advanced systems support leadership evaluation through comprehensive background analysis. `OpenAI/DeepResearch` [197] enables detailed management assessment incorporating historical performance, leadership approach, strategic consistency, and reputation across diverse sources. This capability enhances investment evaluation by providing deeper leadership insights than typically available through standard financial analysis.

Implementation patterns typically involve specialized entity analysis components like those found in `Manus` [164], which incorporates explicit leadership evaluation frameworks. These approaches highlight how targeted optimization enhances specific financial workflows through specialized components addressing critical evaluation dimensions.

6.4.2 Financial Trend Analysis. Pattern recognition across financial data informs strategic positioning:

Multi-Factor Trend Identification. Deep Research systems effectively identify complex patterns across financial indicators and contextual factors. `Gemini/DeepResearch` [60] demonstrates this capability through integrated analysis of market metrics, macroeconomic indicators, sector-specific factors, and relevant external trends. This application enhances trend identification through more comprehensive factor integration than typically practical through manual analysis alone.

Open frameworks like `grapeot/deep_research_agent` [263] implement specialized trend analysis components with particular emphasis on statistical pattern detection and causal factor identification. However,

research indicates that the effectiveness of such AI systems may be limited in tasks requiring deep domain understanding, as their generated outputs can exhibit redundancy or inaccuracies [254]. These approaches demonstrate how domain-specific optimization enhances practical utility for specialized financial applications beyond generic analytical capabilities.

Scenario Development and Testing. Advanced systems support financial planning through structured scenario analysis. **OpenAI/DeepResearch** [197] enables detailed scenario development incorporating varied assumptions, historical precedents, and system dependencies with coherent projection across financial impacts. This capability enhances strategic planning by facilitating more comprehensive scenario exploration than typically practical through manual methods.

Implementation patterns typically involve specialized scenario modeling components like those in **Agent-RL/ReSearch** [2], which incorporates explicit dependency modeling and consistency verification mechanisms. These approaches highlight how targeted optimization enhances specific financial workflows through specialized components addressing critical planning requirements.

6.4.3 Risk Assessment and Modeling. Comprehensive risk analysis informs financial decisions:

Multi-Dimensional Risk Analysis. Deep Research systems enable integrated risk assessment across diverse risk categories. **Perplexity/DeepResearch** [209] supports comprehensive risk evaluation incorporating market, credit, operational, regulatory, and systemic risk factors within unified analytical frameworks. This application enhances risk management through more comprehensive factor integration than typically practical through compartmentalized analysis.

Open implementations like **nickscamara/open-deep-research** [42] implement risk analysis components with particular emphasis on integrated factor assessment and interaction modeling. These approaches demonstrate how domain adaptation enhances practical utility for specific financial applications beyond general analytical capabilities. Evaluations such as RedCode-Exec[101] show that agents are less likely to reject executing technically buggy code, indicating high risks, which highlights the need for stringent safety evaluations for diverse code agents.

Stress Testing and Resilience Assessment. Advanced systems support financial stability through sophisticated stress scenario analysis. **Gemini/DeepResearch** [60] provides detailed stress testing capabilities incorporating historical crisis patterns, theoretical risk models, and system dependency analysis to identify potential vulnerabilities. These features enable more comprehensive resilience assessment than might be practical through standardized stress testing alone.

Implementation patterns typically involve specialized stress modeling components like those found in **Flowith/OracleMode** [77], which incorporates explicit extreme scenario generation and impact propagation mechanisms. These approaches highlight how targeted optimization enhances specific financial workflows through specialized components addressing critical stability assessment requirements.

6.5 Educational Applications

Deep Research technologies enable enhanced learning and knowledge development. Educational approaches to research automation have shown particular promise in scientific education [236] and data science pedagogy [274], with systems like DS-Agent automating machine learning workflows through case-based reasoning

to reduce learners' technical barriers [102], highlighting the dual role of these systems in both conducting research and developing research capabilities in human learners. Smart AI reading assistants are also being developed to enhance reading comprehension through interactive support [266]. However, adoption challenges remain significant in educational contexts, where user resistance and ineffective system utilization can impede learning progress, requiring strategies such as active support during initial use and clear communication of system capabilities [252]. Specifically in data science education, learners encounter challenges similar to those faced by data scientists when interacting with conversational AI systems, such as difficulties in formulating prompts for complex tasks and adapting generated code to local environments [57]. Structured empirical evaluations of LLMs for data science tasks, such as the work by Nathalia Nascimento et al. [185], demonstrate their effectiveness in coding challenges and provide guidance for model selection in educational tools.

6.5.1 Personalized Learning Support. AI-enhanced research supports individualized educational experiences:

Adaptive Learning Path Development. Deep Research systems effectively generate customized learning pathways based on individual interests and knowledge gaps. **OpenAI/DeepResearch** [197] enables detailed learning plan development incorporating knowledge structure mapping, prerequisite relationships, and diverse learning resources tailored to individual learning styles and objectives. This application enhances educational effectiveness through more personalized learning journeys than typically available through standardized curricula.

Open implementations like **OpenManus** [193] implement personalized learning components with particular emphasis on interest-driven exploration and adaptive difficulty adjustment. These approaches demonstrate how educational adaptation enhances practical utility beyond general research capabilities.

Comprehensive Question Answering. Advanced systems provide detailed explanations tailored to learner context and prior knowledge. **Perplexity/DeepResearch** [209] demonstrates this capability through multi-level explanations that adjust detail and terminology based on learner background, providing conceptual scaffolding appropriate to individual knowledge levels. This capability enhances learning effectiveness by providing precisely targeted explanations rather than generic responses.

Implementation patterns typically involve specialized educational components like those in **HKUDS/Auto-Deep-Research** [112], which incorporates explicit knowledge modeling and explanation generation mechanisms. These approaches highlight how targeted optimization enhances educational applications through specialized components addressing critical learning support requirements.

6.5.2 Educational Content Development. Research-driven content creation enhances learning materials:

Curriculum Development Support. Deep Research systems effectively synthesize educational best practices and domain knowledge into coherent curricula. **Gemini/DeepResearch** [60] enables comprehensive curriculum development incorporating learning science principles, domain structure mapping, and diverse resource integration. This application enhances educational design through more comprehensive knowledge integration than typically practical for individual educators.

Open frameworks like **smolagents/open_deep_research** [115] implement curriculum development components with particular emphasis on learning progression modeling and resource alignment. These approaches

demonstrate how specialized adaptation enhances practical utility for educational applications beyond generic content generation.

Multi-Modal Learning Material Creation. Advanced systems generate diverse educational content formats tailored to learning objectives. **OpenAI/DeepResearch** [197] supports creation of integrated learning materials incorporating explanatory text, conceptual visualizations, practical examples, and assessment activities aligned with specific learning outcomes. This capability enhances educational effectiveness through more comprehensive content development than typically practical through manual methods alone.

Implementation patterns typically involve specialized content generation components like those in **QwenLM/Qwen-Agent** [224], which incorporates explicit learning objective modeling and multi-format content generation. These approaches highlight how targeted optimization enhances educational applications through specialized components addressing diverse learning modalities.

6.5.3 Academic Research Training. AI-assisted research skill development supports scholarly advancement:

Research Methodology Instruction. Deep Research systems effectively teach research methods through guided practice and feedback. **Perplexity/DeepResearch** [209] provides explicit methodology training, demonstrating effective research processes while explaining rationale and providing structured feedback on learner attempts. This application enhances research skill development through more interactive guidance than typically available through traditional instruction.

Open implementations like **Jina-AI/node-DeepResearch** [121] support similar capabilities through research practice environments with explicit guidance and feedback mechanisms. These approaches demonstrate how educational adaptation enhances practical utility for research training beyond simple information provision.

Critical Evaluation Skill Development. Maintaining critical thinking skills while leveraging AI research assistance presents unique educational challenges. Drosos et al. [71] demonstrate that carefully designed “provocations” can help restore critical thinking in AI-assisted knowledge work, suggesting important educational approaches for developing research skills that complement rather than rely entirely on AI capabilities. Advanced systems support critical thinking through guided source evaluation and analytical practice. **OpenAI/DeepResearch** [197] enables critical evaluation training, demonstrating source assessment, evidence weighing, and analytical reasoning while guiding learners through similar processes. This capability enhances critical thinking development through structured practice with sophisticated feedback.

Implementation patterns typically involve specialized educational components like those in **grapeot/deep_research_agent** [263], which incorporates explicit critical thinking modeling and guided practice mechanisms. These approaches highlight how targeted optimization enhances educational applications through specialized components addressing crucial scholarly skill development.

6.6 Personal Knowledge Management Applications

Deep Research technologies enable enhanced individual information organization and utilization.

6.6.1 Information Organization and Curation. AI-enhanced systems support personal knowledge development:

Personalized Knowledge Base Development. Deep Research systems effectively organize diverse information into coherent personal knowledge structures. **Perplexity/DeepResearch** [209] supports knowledge base development through automated information organization, connection identification, and gap highlighting tailored to individual interests and objectives. This application enhances personal knowledge management through more sophisticated organization than typically practical through manual methods alone.

Open implementations like **nickscamara/open-deep-research** [42] implement knowledge organization components with particular emphasis on personalized taxonomy development and relationship mapping. These approaches demonstrate how individual adaptation enhances practical utility for personal applications beyond generic information management.

Content Summarization and Abstraction. Advanced systems transform complex information into accessible personal knowledge. **OpenAI/DeepResearch** [197] provides multi-level content abstraction capabilities, generating overview summaries, detailed analyses, and conceptual maps from complex source materials tailored to individual comprehension preferences. This capability enhances information accessibility by providing precisely targeted representations rather than generic summaries.

Implementation patterns typically involve specialized content processing components like those in **Nanobrowser** [184], which incorporates explicit knowledge distillation and representation generation mechanisms. These approaches highlight how targeted optimization enhances personal knowledge applications through specialized components addressing individual information processing needs.

6.6.2 *Personal Learning and Development.* Research-driven insights support individual growth:

Interest-Driven Exploration. Deep Research systems effectively support curiosity-driven learning through guided exploration. **Gemini/DeepResearch** [60] enables interest-based knowledge discovery, identifying connections, extensions, and practical applications related to individual curiosities. This application enhances personal learning through more sophisticated guidance than typically available through standard search alone.

Open frameworks like **OpenManus** [193] implement exploration components with particular emphasis on interest mapping and discovery facilitation. These approaches demonstrate how personalization enhances practical utility for individual learning beyond generic information retrieval.

Skill Development Planning. Advanced systems support personal growth through comprehensive development guidance. **Perplexity/DeepResearch** [209] provides detailed skill development planning, incorporating learning resource identification, progression mapping, and practice guidance tailored to individual objectives and constraints. This capability enhances personal development through more comprehensive planning support than typically available through generic guidance.

Implementation patterns typically involve specialized planning components like those in **TARS** [39], which incorporates explicit skill modeling and development path generation. These approaches highlight how targeted optimization enhances personal growth applications through specialized components addressing individual development needs.

6.6.3 *Decision Support for Individual Users.* Research-enhanced decision making improves personal outcomes:

Complex Decision Analysis. Deep Research systems effectively support personal decisions through comprehensive option evaluation. **OpenAI/DeepResearch** [197] enables detailed decision analysis, incorporating multiple criteria, preference weighting, and consequence projection tailored to individual values and constraints. This application enhances decision quality through more sophisticated analysis than typically practical through manual methods alone.

Open implementations like **Agent-RL/ReSearch** [2] implement decision support components with particular emphasis on preference elicitation and consequence modeling. These approaches demonstrate how personalization enhances practical utility for individual decision making beyond generic information provision.

Life Planning and Optimization. Advanced systems support long-term planning through integrated life domain analysis. **Gemini/DeepResearch** [60] provides comprehensive life planning support, integrating career, financial, health, and personal considerations within coherent planning frameworks tailored to individual values and objectives. This capability enhances life optimization through more integrated planning than typically achievable through domain-specific approaches alone.

Implementation patterns typically involve specialized planning components like those in **Flowith/OracleMode** [77], which incorporates explicit value modeling and multi-domain integration mechanisms. These approaches highlight how targeted optimization enhances personal planning applications through specialized components addressing holistic life considerations.

The diverse applications outlined in this section demonstrate the broad practical impact of Deep Research technologies across domains. While specific implementation approaches vary across commercial and open-source ecosystems, common patterns emerge in domain adaptation, specialized component design, and integration with existing workflows. These patterns highlight how technical capabilities translate into practical value through thoughtful application design aligned with domain-specific requirements and user needs.

7 Ethical Considerations and Limitations

The integration of Deep Research systems into knowledge workflows introduces significant ethical considerations and technical limitations that must be addressed for responsible deployment. This section examines key challenges across four fundamental dimensions (see Figure 10): information integrity, privacy protection, source attribution and intellectual property, and accessibility.

7.1 Information Accuracy and Hallucination Concerns

Deep Research systems face fundamental challenges in maintaining factual reliability despite their sophisticated capabilities.

7.1.1 Factual Verification Mechanisms. Recent studies have highlighted significant challenges in reliable uncertainty communication [55], with particular concerns for research contexts where uncertainty boundaries may be unclear or contested. Some researchers have raised concerns about excessive reliance on AI-generated content in scholarly writing [27, 45, 104, 119, 146, 207, 282, 286, 324, 335], particularly when verification mechanisms are inadequate or bypassed. These limitations are further complicated by tendencies toward misleading responses in conversation [113], presenting particular challenges for interactive research workflows where iterative refinement may inadvertently amplify initial inaccuracies. AI support systems designed for

Ethical Dimensions of Deep Research Systems

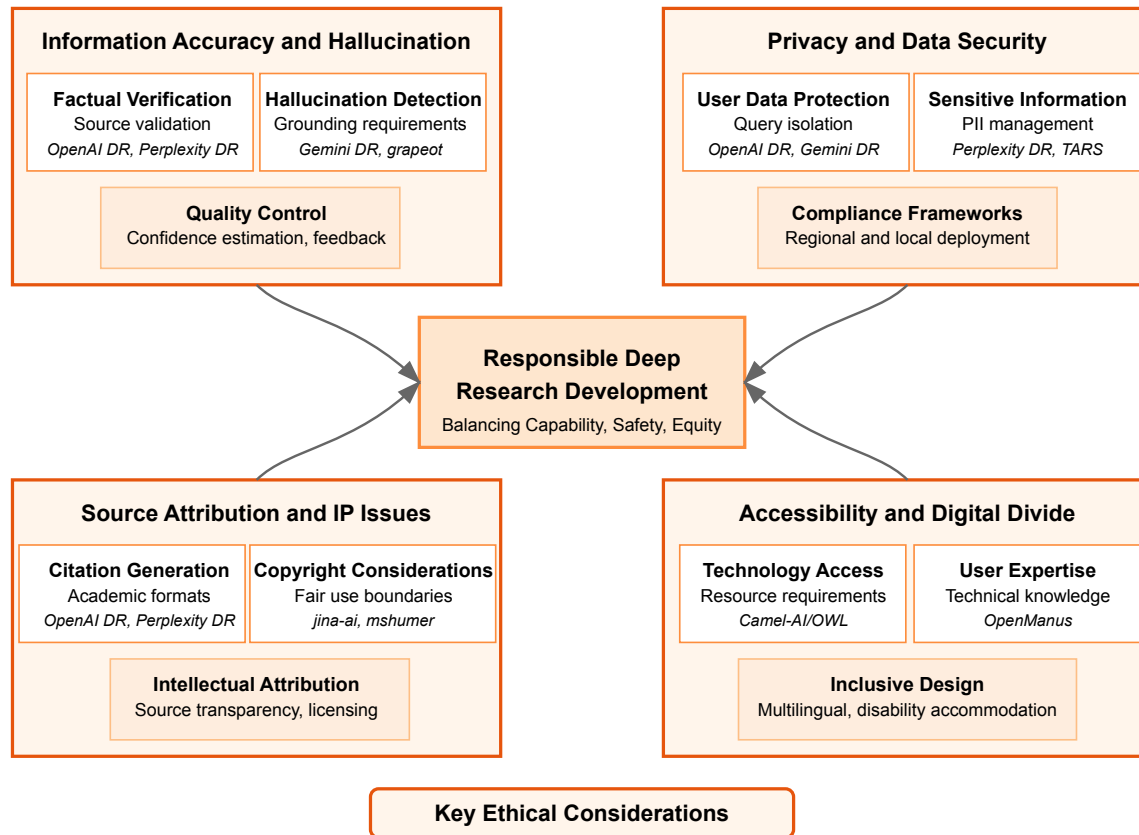


Fig. 10. Ethical Dimensions of Deep Research Systems

evidence-based expository writing tasks, such as literature reviews, offer frameworks to enhance verification through structured sensemaking over source documents [247]. Addressing these challenges requires technical advancements in uncertainty representation, improvements in decision workflow design [107] and interface design improvements that effectively communicate confidence boundaries to research users [270].

Ensuring information accuracy requires explicit verification strategies:

Source Verification Approaches. Leading implementations incorporate explicit source validation mechanisms to enhance factual reliability. **OpenAI/DeepResearch** [197] implements multi-level verification that confirms information across multiple independent sources before incorporation into research outputs, with detailed guidelines outlined in their system documentation [196]. Similarly, **Perplexity/DeepResearch** [209] implements automated fact-checking that independently verifies key claims against trusted reference sources before inclusion in final reports.

Open-source alternatives demonstrate varied approaches to verification. Systems like **grapeot/deep_research_agent** [263] emphasize explicit citation mechanisms that maintain direct links between claims

and sources, enabling straightforward verification. More sophisticated implementations like **HKUDS/Auto-Deep-Research** [112] incorporate specialized verification modules that assess source credibility and content consistency before information utilization.

Hallucination Detection and Prevention. Mitigating fabricated information represents a crucial challenge for LLM-based research systems. Commercial implementations employ advanced hallucination reduction techniques including strict grounding requirements and consistency verification. **Gemini/DeepResearch** [60] implements explicit uncertainty modeling that distinguishes between confirmed information and speculative extensions, enhancing transparency when definitive answers cannot be provided. Emerging paradigms like those proposed by Silver and Sutton [251] suggest a fundamental shift toward experience-driven learning, potentially transforming how research systems acquire and refine capabilities through interaction with information environments. Such approaches could enable more human-like research development through continuous improvement based on research experiences rather than static training alone, and could fundamentally mitigate hallucinations.

Open implementations demonstrate pragmatic approaches to hallucination reduction within more constrained technical environments. Systems like **Agent-RL/ReSearch** [2] employ preventative strategies including explicit sourcing requirements and conservative synthesis guidelines that prioritize factual reliability over comprehensive coverage. Complementary approaches like Mask-DPO [100] focus on generalizable fine-grained factuality alignment, addressing a critical requirement for reliable research outputs. Recent work from the GAIR NLP team on **DeepResearcher** [81] has advanced these capabilities through integrated neural verification and knowledge graph alignment techniques that significantly enhance factual reliability. These approaches highlight diverse strategies for addressing a fundamental challenge that impacts all LLM-based research systems.

7.1.2 Uncertainty Communication Approaches. Transparent uncertainty representation enhances result interpretation and appropriate utilization:

Confidence Estimation Methods. Advanced systems implement explicit confidence assessment for research findings and recommendations. **OpenAI/DeepResearch** [197] incorporates graduated confidence scoring that reflects evidence quality, consistency across sources, and reasoning reliability. This capability enhances result interpretation by clearly distinguishing between well-supported conclusions and more speculative findings.

Open-source implementations demonstrate simplified but effective confidence communication approaches. Systems like **mshumer/OpenDeepResearcher** [249] incorporate basic confidence indicators that signal information reliability through explicit markers in research outputs. These approaches highlight the importance of transparent uncertainty communication regardless of implementation sophistication.

Evidence Qualification Standards. Responsible systems clearly communicate limitations and contextual factors affecting result interpretation. Commercial implementations like **Perplexity/DeepResearch** [209] incorporate explicit evidence qualification that highlights contextual limitations, conflicting viewpoints, and temporal constraints affecting research findings. This practice enhances appropriate utilization by providing necessary context for result interpretation.

Open-source alternatives demonstrate varied approaches to evidence qualification. Systems like **dzhng/deep-research** [321] implement explicit limitation statements that identify key constraints affecting research

reliability. More sophisticated implementations like **Camel-AI/OWL** [43] incorporate structured evidence models that represent both supporting and contradicting information within unified frameworks.

7.1.3 Quality Control Frameworks. Systematic approaches to quality assurance enhance overall reliability:

Pre-Release Verification Standards. Leading implementations employ comprehensive validation processes before result delivery. Gemini Deep Research implements structured quality verification including automated consistency checking, source validation, and reasoning verification before providing research outputs. These practices enhance overall reliability through systematic error identification and correction.

Open-source implementations demonstrate more varied quality control approaches. Systems like **nicksamara/open-deep-research** [42] incorporate simplified validation processes focusing on critical reliability factors including source verification and logical consistency. These approaches highlight how even basic quality control mechanisms can significantly enhance research reliability.

Feedback Integration Systems. Continuous improvement requires effective incorporation of accuracy feedback. As Deep Research systems advance toward greater autonomy, broader safety considerations become increasingly important. Bengio et al. [26] highlight potential risks from superintelligent agents and propose approaches like “Scientist AI” that balance capability with safer development paths, emphasizing the importance of integrated safety mechanisms in advanced research systems. Commercial systems implement sophisticated feedback integration including explicit accuracy reporting channels and systematic error pattern analysis. **OpenAI/DeepResearch** [197] includes dedicated correction mechanisms that incorporate verified accuracy feedback into system improvements, creating virtuous improvement cycles.

Open implementations demonstrate more community-oriented feedback approaches. Systems like **smolagents/open_deep_research** [115] incorporate collaborative improvement frameworks that enable distributed error identification and correction through community contributions. These approaches highlight diverse strategies for enhancing reliability through user engagement across implementation contexts.

7.2 Privacy and Data Security

Research systems must carefully protect sensitive information throughout the research process.

7.2.1 User Data Protection Mechanisms. Safeguarding user information requires comprehensive protection strategies:

Query Isolation Practices. Leading implementations employ strict isolation between user research sessions. Commercial systems like **OpenAI/DeepResearch** [197] and **Gemini/DeepResearch** [60] implement comprehensive tenant isolation that prevents information leakage between distinct users or organizations. These practices are particularly crucial for sensitive research applications in corporate or governmental contexts.

Open-source implementations demonstrate varied isolation approaches depending on deployment models. Systems designed for local deployment like **OpenManus** [193] enable complete isolation within organizational boundaries, enhancing privacy for sensitive applications. Cloud-dependent implementations typically incorporate more limited isolation mechanisms, highlighting deployment considerations for privacy-sensitive applications.

Data Minimization Strategies. Responsible systems limit sensitive data collection and retention. Commercial implementations increasingly emphasize data minimization, collecting only information necessary for service

provision and applying appropriate retention limitations. These practices enhance privacy protection by reducing potential exposure of sensitive information through either security incidents or authorized access.

Open implementations demonstrate diverse approaches to data management. Systems like **Nanobrowser** [184] enable complete local control of browsing data, preventing external exposure of research activities. Infrastructure frameworks like **Jina-AI/node-DeepResearch** [121] provide flexible configuration options that enable deployment-specific privacy controls aligned with organizational requirements.

7.2.2 Sensitive Information Handling. Special safeguards are required for particularly sensitive content categories:

Personal Identifier Management. Advanced systems implement specific protections for personally identifiable information. Commercial implementations like **Perplexity/DeepResearch** [209] incorporate automatic detection and redaction of personal identifiers from research outputs unless specifically relevant to research objectives. These practices prevent inadvertent exposure of personal information through research activities.

Open implementations demonstrate more varied approaches to identifier management. Systems like **TARS** [39] incorporate basic identifier detection focused on common patterns like email addresses and phone numbers. More sophisticated implementations like **QwenLM/Qwen-Agent** [224] provide configurable sensitivity controls that enable context-appropriate protection aligned with specific deployment requirements.

Protected Category Safeguards. Responsible systems implement enhanced protections for specially regulated information categories. Commercial implementations increasingly incorporate specialized handling for information categories including health data, financial records, and other regulated content types. These practices enhance compliance with domain-specific regulatory requirements governing sensitive information.

Open-source alternatives demonstrate more varied regulatory alignment. Systems like **n8n** [183] provide specialized workflow components for handling regulated data categories, enabling compliance-oriented implementations in sensitive domains. These approaches highlight how specialized components can address domain-specific regulatory requirements within flexible implementation frameworks.

7.2.3 Compliance with Regulatory Frameworks. Adherence to applicable regulations ensures legally appropriate operation:

Jurisdictional Compliance Adaptation. Advanced systems implement regionally appropriate operational standards. Commercial implementations increasingly incorporate jurisdiction-specific adaptations that align with regional privacy regulations including GDPR, CCPA, and other frameworks. These practices enhance legal compliance across diverse deployment environments with varying regulatory requirements.

Open implementations demonstrate more deployment-dependent compliance approaches. Systems designed for flexible deployment like **Flowith/OracleMode** [77] provide configurable privacy controls that enable adaptation to specific regulatory environments. These approaches highlight the importance of adaptable privacy frameworks that can address diverse compliance requirements across implementation contexts.

Transparency and Control Mechanisms. Responsible systems provide appropriate visibility and user authority over information processing. Emerging regulatory frameworks are increasingly focusing on AI agents with autonomous capabilities. Osogami [204] proposes that regulation of autonomous AI systems should specifically consider action sequence patterns rather than individual actions in isolation, which has particular implications for Deep Research systems that execute complex multi-step research workflows. Commercial implementations increasingly emphasize transparency through explicit processing disclosures

and user control mechanisms aligned with regulatory requirements. These practices enhance both regulatory compliance and user trust through appropriate information governance.

Open-source alternatives demonstrate varied transparency approaches. Systems like `HKUDS/Auto-Deep-Research` [112] provide detailed logging of information access and processing activities, enabling appropriate oversight and verification. These approaches highlight how transparent operation can enhance both compliance and trust across implementation contexts.

7.3 Source Attribution and Intellectual Property

Proper acknowledgment of information sources and respect for intellectual property rights are essential for ethical information utilization.

7.3.1 Citation Generation and Verification. Accurate source attribution requires reliable citation mechanisms:

Automated Citation Systems. Advanced implementations incorporate sophisticated citation generation for research outputs. Commercial systems like `OpenAI/DeepResearch` [197] and `Perplexity/DeepResearch` [209] implement automatic citation generation in standard academic formats, enhancing attribution quality and consistency. These capabilities support appropriate source acknowledgment without manual effort.

Open implementations demonstrate varied citation approaches. Systems like `mshumer/OpenDeepResearcher` [249] incorporate basic citation generation focused on fundamental bibliographic information. More sophisticated alternatives like `dzhng/deep-research` [321] provide enhanced citation capabilities including format customization and citation verification against reference databases.

Citation Completeness Verification. Responsible systems ensure comprehensive attribution for all utilized information. Commercial implementations increasingly incorporate citation coverage verification that identifies unsupported claims requiring additional attribution. These practices enhance attribution reliability by ensuring all significant claims maintain appropriate source connections.

Open-source alternatives demonstrate pragmatic approaches to attribution verification. Systems like `grapeot/deep_research_agent` [263] implement explicit source-claim mapping that maintains clear relationships between information and origins. These approaches highlight the importance of systematic attribution regardless of implementation sophistication.

7.3.2 Intellectual Attribution Challenges. Special attribution considerations apply to complex intellectual contributions:

Idea Attribution Practices. Research systems must appropriately acknowledge conceptual contributions beyond factual information. Commercial implementations increasingly emphasize concept-level attribution that acknowledges intellectual frameworks and theoretical approaches beyond simple facts. These practices enhance ethical information utilization by appropriately recognizing intellectual contributions.

Open implementations demonstrate varied idea attribution approaches. Systems like `Camel-AI/OWL` [43] incorporate explicit concept attribution that identifies theoretical frameworks and analytical approaches utilized in research outputs. These approaches highlight the importance of comprehensive attribution beyond basic factual sources.

Synthesized Knowledge Attribution. Attribution becomes particularly challenging for insights synthesized across multiple sources. Advanced systems implement specialized attribution approaches for synthetic insights

that acknowledge multiple contributing sources while clearly identifying novel connections. These practices enhance attribution accuracy for the increasingly common scenario of cross-source synthesis.

Open-source alternatives demonstrate pragmatic approaches to synthesis attribution. Systems like **Agent-RL/ReSearch** [2] implement explicit synthesis markers that distinguish between directly sourced information and system-generated connections. These approaches highlight the importance of transparent derivation even when direct attribution becomes challenging.

7.3.3 Copyright and Fair Use Considerations. Research activities interact with copyright protections in multiple dimensions:

Fair Use Evaluation Mechanisms. Research systems must navigate appropriate utilization of copyrighted materials. Commercial implementations increasingly incorporate fair use evaluation that considers purpose, nature, amount, and market impact when utilizing copyrighted content. These practices enhance legal compliance while enabling appropriate information utilization for legitimate research purposes.

Open implementations demonstrate varied copyright approaches. Systems like **Jina-AI/node-DeepResearch** [121] incorporate basic copyright acknowledgment focusing on proper attribution, while more sophisticated alternatives like **Manus** [164] provide enhanced copyright handling including content transformation assessment and restricted access mechanisms for sensitive materials.

Content Licensing Compliance. Responsible systems respect diverse license terms applicable to utilized content. Advanced implementations increasingly incorporate license-aware processing that adapts information utilization based on specific terms governing particular sources. These practices enhance compliance with varied license requirements across the information ecosystem.

Open implementations demonstrate more standardized licensing approaches. Systems like **grapeot/deep_research_agent** [263] incorporate simplified license categorization focusing on common frameworks including creative commons and commercial restrictions. These approaches highlight pragmatic strategies for license navigation within resource constraints.

7.3.4 Output Intellectual Property Frameworks. Clear rights management for research outputs enhances downstream utilization:

Output License Assignment. Complex questions arise regarding intellectual property in research outputs. Commercial systems increasingly implement explicit license assignment for generated content, clarifying intellectual property status for downstream utilization. These practices enhance transparency regarding usage rights for research outputs created through automated systems.

Open-source alternatives demonstrate varied approaches to output rights. Systems like **OpenManus** [193] incorporate explicit license designation for research outputs aligned with organizational policies and source restrictions. These approaches highlight the importance of clear intellectual property frameworks regardless of implementation context.

Derivative Work Management. Research systems must address whether outputs constitute derivative works of source materials. Commercial systems increasingly implement derivative assessment frameworks that evaluate the nature and extent of source transformation in research outputs. These practices enhance appropriate categorization for downstream utilization aligned with source licenses.

Open-source alternatives demonstrate varied derivation approaches. Systems such as **QwenLM/Qwen-Agent** [224] incorporate a basic transformation assessment focusing on content reorganization and analytical

addition. These approaches highlight the importance of thoughtful derivative consideration regardless of implementation sophistication.

7.4 Accessibility and Digital Divide

Equitable access to research capabilities requires addressing systematic barriers.

7.4.1 Technology Access Disparities. Recent work has highlighted both adoption barriers and opportunities for making Deep Research systems more accessible. Bianchini et al. [29] and Tonghe Zhuang et al. [334] identify specific organizational and individual factors affecting AI adoption in scientific research contexts, with implications for Deep Research deployment. Accessibility-focused approaches like those presented by Mowar et al. [179] demonstrate how AI coding assistants can be specifically designed to support accessible development practices, suggesting parallel opportunities for accessibility-centered Deep Research systems. Extending this, systems such as ResearchAgent [18] showcase how AI can lower barriers to scientific innovation by enabling iterative refinement of research ideas through collaborative feedback mechanisms, thus democratizing access to complex ideation processes.

Resource requirements create potential exclusion for various user segments:

Computational Requirement Considerations. Resource-intensive systems may exclude users without substantial computing access. Commercial cloud-based implementations address this challenge through shared infrastructure that reduces local requirements, though with associated cost barriers. Open-source alternatives demonstrate varied resource profiles, with systems like `Camel-AI/OWL` [43] emphasizing efficiency to enable broader deployment on limited hardware.

Cost Barrier Mitigation. Financial requirements create systematic access disparities across socioeconomic dimensions. Commercial implementations demonstrate varied pricing approaches, with systems like `Perplexity/DeepResearch` [209] offering limited free access alongside premium tiers. Open-source alternatives like `HKUDS/Auto-Deep-Research` [112] and `nicksamara/open-deep-research` [42] eliminate direct cost barriers while potentially introducing technical hurdles.

7.4.2 User Expertise Requirements. Technical complexity creates additional access barriers beyond resource considerations:

Technical Expertise Dependencies. Complex system deployment and operation may exclude users without specialized knowledge. Commercial implementations address this challenge through managed services that eliminate deployment complexity, though with reduced customization flexibility. Open-source alternatives demonstrate varied usability profiles, with systems like `OpenManus` [193] emphasizing simplified deployment to enhance accessibility despite local operation.

Domain Knowledge Prerequisites. Effective research still requires contextual understanding for appropriate utilization. Both commercial and open-source implementations increasingly incorporate domain guidance that assists users with limited background knowledge in specific research areas. These capabilities enhance accessibility by reducing domain expertise barriers to effective research utilization.

7.4.3 Inclusivity and Universal Design Approaches. Deliberate inclusive design can address systematic access barriers:

Linguistic and Cultural Inclusivity. Language limitations create significant barriers for non-dominant language communities. Commercial implementations increasingly offer multilingual capabilities, though with persistent quality disparities across languages. Open-source alternatives demonstrate varied language support, with systems like `Flowith/OracleMode` [77] emphasizing extensible design that enables community-driven language expansion beyond dominant languages.

Disability Accommodation Approaches. Accessible design ensures appropriate access for users with diverse abilities. Commercial implementations increasingly incorporate accessibility features including screen reader compatibility, keyboard navigation, and alternative format generation. Open-source alternatives demonstrate more varied accessibility profiles, highlighting an area for continued community development to ensure equitable access across implementation contexts.

The ethical considerations explored in this section highlight the complex responsibilities associated with Deep Research technologies beyond technical performance. While current implementations demonstrate varying approaches to these challenges across commercial and open-source ecosystems, consistent patterns emerge in the importance of factual verification, attribution quality, privacy protection, intellectual property respect, and accessible design. Addressing these considerations represents a critical priority for responsible development and deployment of these increasingly influential research technologies.

8 Future Research Directions

The rapidly evolving field of Deep Research presents numerous opportunities for technical advancement and application expansion. Recent work by Zheng et al. [329] proposes scaling deep research capabilities via reinforcement learning in real-world environments, while Wu et al. [297] explore enhancing reasoning capabilities of LLMs with tools specifically for deep research applications. The comprehensive framework for building effective agents outlined by Anthropic [11] provides additional design principles that could inform future Deep Research systems. This section examines promising research directions (illustrated in Figure 11) that could significantly enhance capabilities, address current limitations, and expand practical impact across domains, focusing on four key areas: advanced reasoning architectures, multimodal integration, domain specialization, and human-AI collaboration with standardization.

8.1 Advanced Reasoning Architectures

Enhanced reasoning capabilities represent a fundamental advancement opportunity for next-generation systems.

8.1.1 Context Window Optimization and Management. The information-intensive nature of deep research tasks presents fundamental challenges for context window utilization:

Information Compression and Prioritization. Current systems struggle with context window exhaustion when processing extensive research materials. Future architectures could incorporate sophisticated compression mechanisms that maintain semantic content while reducing token consumption. Early steps in this direction appear in systems like `OpenAI/DeepResearch` [197], which implements basic summarization for lengthy sources. Recent work on academic paper review systems demonstrates how hierarchical processing of extended research content can maintain coherence while managing context limitations [333]. Semantic navigation techniques offer complementary approaches by enabling efficient exploration of problem-solution

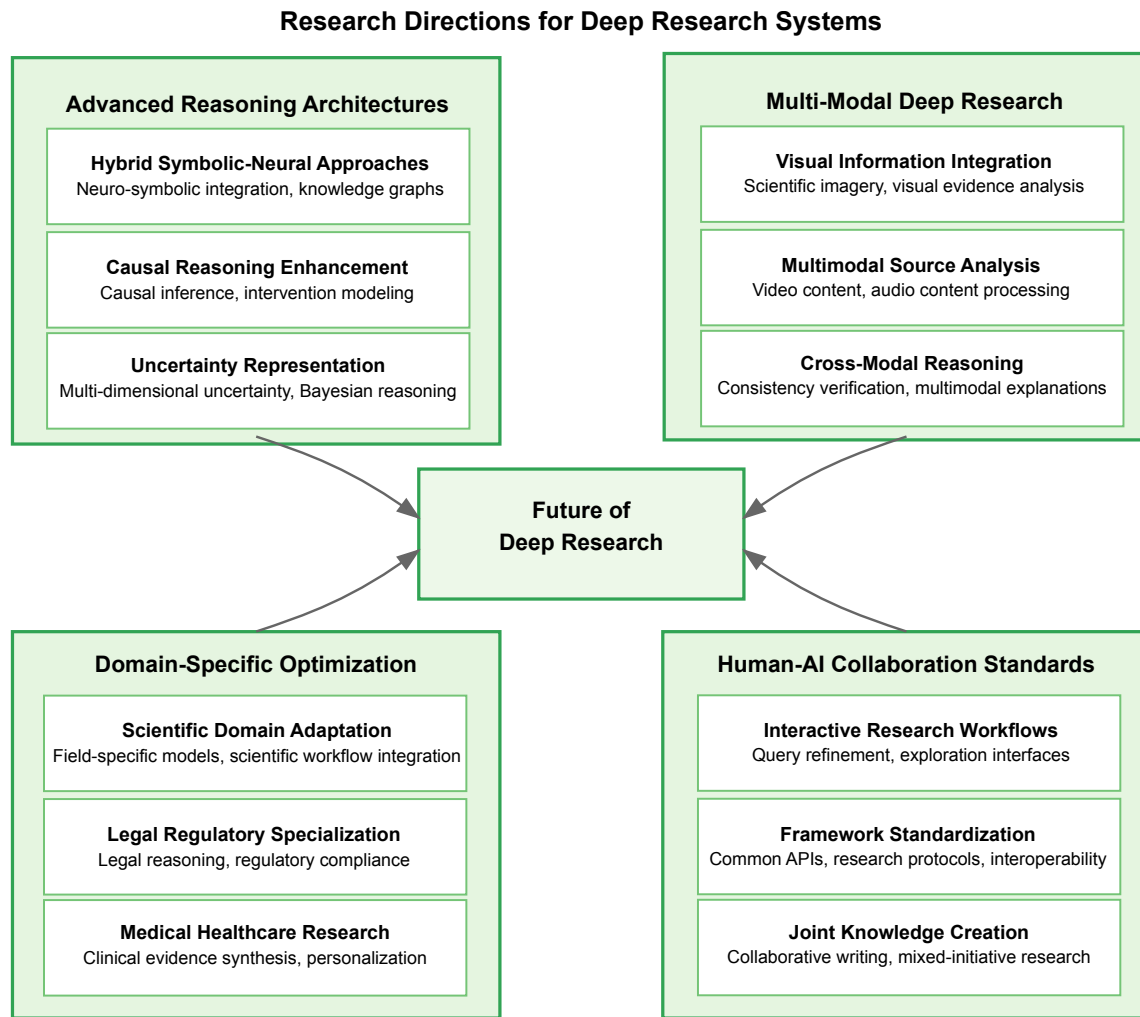


Fig. 11. Research Directions for Deep Research Systems

spaces within constrained domains, optimizing context usage through input filtering while enhancing generation quality [238]. More advanced approaches could develop adaptive compression that preserves crucial details while condensing secondary information based on query relevance.

Implementation opportunities include developing hierarchical summarization techniques that maintain multi-level representations of sources, implementing information relevance scoring that prioritizes context allocation to critical content, and designing dynamic context management that continuously optimizes window utilization throughout research workflows. These advances could significantly enhance information processing capabilities without requiring proportional increases in context length.

External Memory Architectures. Beyond compression, architectural innovations could fundamentally transform context window utilization. Future systems could implement sophisticated external memory

frameworks that maintain rich information representations outside the primary context window, accessing them through efficient retrieval mechanisms when needed. Systems like **Camel-AI/OWL** [43] demonstrate early steps with basic retrieval-augmented generation, but more comprehensive approaches could enable effectively unlimited knowledge integration.

Research directions include developing differentiable retrieval mechanisms that seamlessly integrate external knowledge within reasoning flows, implementing structured memory hierarchies that organize information for efficient access, and designing memory-aware reasoning processes that explicitly consider information availability when planning analytical approaches. These architectures could fundamentally address context limitations while enhancing reasoning transparency and reliability.

8.1.2 Hybrid Symbolic-Neural Approaches. Integration of complementary reasoning paradigms offers significant potential:

Neuro-Symbolic Integration. Current Deep Research systems rely primarily on neural approaches with limited explicit reasoning structures. Future systems could integrate symbolic reasoning components that provide formal logical capabilities alongside neural flexibility, enhancing both reliability and explainability. Early examples of this direction appear in systems like **Camel-AI/OWL** [43], which incorporates structured knowledge representation within primarily neural architectures. Future research could develop more sophisticated integration approaches that leverage the complementary strengths of both paradigms.

Implementation approaches might include explicit logical verification layers that validate neural-generated reasoning, hybrid architectures that select appropriate reasoning mechanisms based on task characteristics, or integrated systems that translate between symbolic and neural representations as needed throughout complex workflows. These approaches could address current challenges in reliability and consistency while maintaining the flexibility and generalization capabilities of neural foundations.

Advanced Knowledge Graph Integration. While current systems already incorporate basic knowledge graph capabilities, future approaches could implement more sophisticated integration with dynamic, contextually-aware knowledge structures. Beyond the entity relationship modeling seen in systems like **HKUDS/Auto-Deep-Research** [112], next-generation implementations could enable bidirectional updates where research findings automatically refine and expand knowledge graphs while simultaneously leveraging them for reasoning. Such approaches could incorporate uncertainty representation within graph structures, probabilistic reasoning across knowledge networks, and adaptive abstraction hierarchies that transform between detailed and high-level conceptual representations based on reasoning requirements. Research opportunities include developing dynamic knowledge graph construction techniques that automatically build and refine structured representations from unstructured sources, implementing graph-aware attention mechanisms that incorporate relationship structures into neural reasoning, and designing hybrid querying approaches that combine graph traversal with neural generation. These advances could enhance precision for complex reasoning tasks requiring structured relationship understanding.

8.1.3 Causal Reasoning Enhancement. Moving beyond correlation to causal understanding represents a crucial capability advancement:

Causal Inference Mechanisms. Current systems excel at identifying correlations but struggle with robust causal analysis. Future research could develop specialized causal reasoning components that systematically identify potential causal relationships, evaluate evidence quality, and assess alternative explanations. Recent

work in healthcare research by Schuemie et al. [241] demonstrates the challenges of establishing confident observational findings, highlighting the need for more sophisticated causal reasoning in research systems. Early steps in this direction appear in systems like **OpenAI/DeepResearch** [197], which incorporates basic causal language in relationship descriptions. Other research explores the use of AI to assist in mining causality, for instance, by searching for instrumental variables in economic analysis [105]. More sophisticated approaches could enable reliable causal analysis across domains. Implementation opportunities include developing causal graph construction techniques that explicitly model intervention effects and counterfactuals, implementing causal uncertainty quantification that represents confidence in causal assertions, and designing specialized prompt structures that guide causal reasoning through structured analytical patterns. These advances could enhance research quality for domains where causal understanding is particularly crucial, including medicine, social sciences, and policy analysis.

Intervention Modeling Techniques. Advanced causal understanding requires sophisticated intervention and counterfactual reasoning capabilities. Future systems could incorporate explicit intervention modeling that simulates potential actions and outcomes based on causal understanding, enhancing both explanatory and predictive capabilities. Early examples of this direction appear in systems like **Agent-RL/ReSearch** [2], which implements basic intervention simulation within reinforcement learning frameworks. More comprehensive approaches could enable sophisticated what-if analysis across domains.

Research directions include developing counterfactual generation techniques that systematically explore alternative scenarios based on causal models, implementing intervention optimization algorithms that identify high-leverage action opportunities, and designing domain-specific intervention templates that embed field-specific causal knowledge for common analysis patterns. These advances could enhance practical utility for decision support applications requiring sophisticated action planning and outcome prediction.

8.1.4 Uncertainty Representation and Reasoning. Sophisticated uncertainty handling enhances both accuracy and trustworthiness:

Multi-Dimensional Uncertainty Modeling. Current systems employ relatively simplistic uncertainty representations that inadequately capture different uncertainty types. Future research could develop multi-dimensional uncertainty frameworks that separately represent epistemic uncertainty (knowledge limitations), aleatoric uncertainty (inherent randomness), and model uncertainty (representation limitations). Early steps in this direction appear in systems like **Perplexity/DeepResearch** [209], which distinguishes between source uncertainty and integration uncertainty. More comprehensive approaches could enable more nuanced and reliable uncertainty communication.

Implementation opportunities include developing uncertainty propagation mechanisms that track distinct uncertainty types throughout reasoning chains, implementing uncertainty visualization techniques that effectively communicate multi-dimensional uncertainty to users, and designing uncertainty-aware planning algorithms that appropriately balance different uncertainty types in decision contexts. These advances could enhance both system reliability and appropriate user trust calibration.

Bayesian Reasoning Integration. Probabilistic reasoning frameworks offer principled approaches to uncertainty handling and knowledge integration. Future systems could incorporate explicit Bayesian reasoning components that systematically update beliefs based on evidence strength and prior knowledge, enhancing both accuracy and explainability. Early examples of this direction appear in systems like

`grapeot/deep_research_agent` [263], which implements basic evidence weighting within research workflows. More sophisticated integration could enable principled uncertainty handling across domains.

Research directions include developing scalable Bayesian inference techniques compatible with large-scale language models, implementing belief update explanation mechanisms that communicate reasoning in understandable terms, and designing domain-specific prior models that incorporate field-specific background knowledge for common analysis patterns. These advances could enhance reasoning quality for domains with inherent uncertainty or limited evidence.

8.2 Multi-Modal Deep Research

Expanding beyond text to incorporate diverse information modalities represents a significant advancement opportunity.

8.2.1 Visual Information Integration. Image understanding dramatically expands information access and analysis capabilities:

Scientific Image Analysis. Current systems demonstrate limited capabilities for extracting and interpreting visual scientific content. Future research could develop specialized visual understanding components for scientific images including graphs, diagrams, experimental images, and visualizations across domains. Early steps in this direction appear in systems like `Gemini/DeepResearch` [60], which incorporates basic chart extraction capabilities. Frameworks such as ChartCitor [96] provide fine-grained bounding box citations to enhance explainability for complex chart understanding, improving user trust and productivity. Specialized models like LHRs-Bot [180] demonstrate sophisticated reasoning capabilities for remote sensing imagery by leveraging geographic information and multimodal learning. The development of large-scale, domain-specific multimodal datasets for areas like entomology [272] and seafloor geology [188] is crucial for training more capable models. More comprehensive approaches could enable sophisticated analysis of visual scientific communication. Implementation opportunities include developing specialized scientific visualization parsers that extract quantitative data from diverse chart types, implementing diagram understanding systems that interpret complex scientific illustrations across domains, and designing domain-specific visual analysis components optimized for field-specific imagery like medical scans or astronomical observations. These advances could dramatically expand information access beyond text-centric sources.

Visual Evidence Integration. Effective research increasingly requires integration of visual evidence alongside textual sources. Future systems could implement sophisticated multimodal reasoning that incorporates visual evidence within comprehensive analytical frameworks, enabling true multimodal research synthesis. Recent analyses have identified multi-modal integration as a key missing capability in current AI research systems [315], highlighting the critical importance of cross-modal reasoning for scientific applications. Early examples of this direction appear in systems like `Gemini/DeepResearch` [60], which provides basic integration of image-derived information. More sophisticated approaches could enable balanced evidence integration across modalities.

Research directions include developing evidence alignment techniques that match textual and visual information addressing common questions, implementing cross-modal consistency verification that identifies conflicts between textual claims and visual evidence, and designing multimodal synthesis mechanisms that

generate integrated understanding across information types. These advances could enhance research quality for domains with significant visual information components.

8.2.2 Multimodal Source Analysis. Comprehensive understanding requires integrated analysis across diverse information formats:

Video Content Processing. Video represents an increasingly important but currently underutilized information source. Future research could develop specialized video understanding components that extract and interpret temporal visual information, including presentations, interviews, demonstrations, and dynamic processes. Initial steps in this direction are emerging in systems like OpenAI’s DALL-E 3, though not yet integrated into Deep Research workflows. Comprehensive integration could enable access to the extensive knowledge embedded in video content.

Implementation opportunities include developing lecture understanding systems that extract structured knowledge from educational videos, implementing process analysis components that interpret demonstrations and procedures, and designing integrated audio-visual analysis that combines visual information with spoken content for comprehensive understanding. These advances could expand information access to the rapidly growing corpus of video knowledge.

Audio Content Integration. Spoken information in podcasts, lectures, interviews, and discussions represents a valuable knowledge source. Future systems could incorporate sophisticated audio processing that extracts, interprets, and integrates spoken information within research workflows. Early examples of speech processing appear in transcription services, but comprehensive research integration remains limited. Advanced approaches could enable seamless incorporation of spoken knowledge alongside traditional text sources.

Research directions include developing speaker identification and attribution systems that maintain appropriate source tracking for spoken content, implementing domain-specific terminology extraction that accurately captures specialized vocabulary in varied acoustic conditions, and designing temporal alignment techniques that connect spoken information with related textual or visual content. These advances could expand information access while maintaining appropriate attribution and context.

8.2.3 Cross-Modal Reasoning Techniques. Effective multimodal research requires specialized reasoning approaches across information types:

Multi-Modal Chain of Thought Reasoning. Current reasoning processes typically operate primarily within single modalities despite handling diverse information types. Future systems could implement true multi-modal reasoning chains that explicitly incorporate diverse information types throughout the analytical process, not just in final outputs. Early steps appear in systems like Gemini/DeepResearch [60], which demonstrates basic visual incorporation in reasoning steps. More sophisticated approaches could enable reasoning flows that seamlessly transition between textual analysis, visual processing, numerical computation, and spatial reasoning based on task requirements.

Research opportunities include developing explicit multi-modal reasoning protocols that formalize information transfer between modalities, implementing cross-modal verification techniques that leverage complementary information types throughout reasoning chains, and designing unified representation frameworks that enable coherent reasoning across diverse information formats. These advances could significantly enhance reasoning quality for complex research tasks requiring integrated understanding across modalities,

moving beyond the current text-centric reasoning paradigms to more human-like analytical processes that naturally leverage the most appropriate modality for each reasoning component.

Cross-Modal Consistency Verification. Integrating diverse information modalities introduces new consistency challenges. Future research could develop specialized verification mechanisms that assess consistency across textual, visual, numerical, and temporal information, enhancing overall reliability. Early steps in this direction appear in systems like **Gemini/DeepResearch** [60], which implements basic cross-format validation. More sophisticated approaches could enable reliable integration of increasingly diverse information types.

Implementation opportunities include developing cross-modal contradiction detection algorithms that identify conflicts between information expressed in different formats, implementing uncertainty alignment techniques that reconcile confidence estimates across modalities, and designing multimodal fact verification systems that leverage complementary evidence types for enhanced reliability. These advances could address emerging challenges in multimodal information integration.

Multimodal Explanation Generation. Effective communication often requires coordinated explanation across modalities. Future systems could generate truly multimodal research outputs that combine textual, visual, and interactive components to enhance understanding and persuasiveness. Early examples of this direction appear in systems like **mshumer/OpenDeepResearcher** [249], which implements basic report visualization. More comprehensive approaches could enable sophisticated multimodal communication tailored to content requirements.

Research directions include developing coordinated generation architectures that produce aligned content across modalities, implementing adaptive format selection algorithms that identify optimal representation formats for different content types, and designing multimodal narrative structures that effectively combine diverse formats within coherent explanatory frameworks. These advances could enhance communication effectiveness across application domains.

8.3 Domain-Specific Optimization

Tailored enhancement for particular fields offers significant performance improvements for specialized applications.

8.3.1 Scientific Domain Adaptation. Scientific research presents unique requirements and opportunities for specialization:

Field-Specific Model Adaptation. Current systems employ relatively general architectures across scientific domains. Future research could develop specialized adaptation techniques that optimize performance for particular scientific fields including physics, chemistry, biology, and others with distinct knowledge structures and reasoning patterns. Early steps in this direction appear in systems like **AutoGLM-Research** [330], which implements domain-specific prompting. Domain-specialized research agents have demonstrated particular promise in physics [305], chemistry [6, 34, 50, 326], materials science [189], oceanography [28], geospatial analysis [165], patent research [227, 285], and broader scientific discovery workflows [84]. These specialized implementations highlight the value of domain adaptation beyond general research capabilities. More comprehensive adaptation could enable significant performance improvements for scientific applications.

Implementation approaches might include domain-specific fine-tuning regimes that emphasize field-relevant reasoning patterns, specialized architectural modifications that enhance performance for domain-characteristic

tasks, or hybrid systems that incorporate symbolic components for domain-specific formal reasoning. These approaches could address current limitations in scientific reasoning while maintaining general capabilities for cross-domain research.

Scientific Workflow Integration. Effective scientific application requires integration with existing research methodologies and tools. Future systems could implement specialized interfaces for scientific workflows including experimental design, data analysis, literature integration, and theory development. Early examples of this direction appear in systems like *n8n* [183], which provides workflow automation for data processing. Platforms designed to support machine learning development in fundamental science also illustrate this trend, enabling research in federated cloud environments [9]. More comprehensive integration could enable seamless incorporation within scientific research processes. Research assistant tools employing prompt-based templates demonstrate domain-agnostic support for tasks such as enhanced literature search queries and preliminary peer review, facilitating standardized assistance across diverse scientific fields [245]. User studies highlight varying automation needs across DS/ML workflows, suggesting targeted rather than complete end-to-end automation aligns with researcher preferences [284]. Research opportunities include developing experimental design assistants that generate and refine research protocols based on literature and objectives, implementing integrated analysis pipelines that combine automated and human analytical components, and designing theory development frameworks that link empirical findings with formal theoretical structures. These advances could enhance practical scientific impact beyond general information access [44, 288].

8.3.2 Legal and Regulatory Domain Specialization. Legal applications present distinct challenges requiring specialized adaptation:

Legal Reasoning Enhancement. Current systems struggle with the precision and structure of legal analysis. Future research could develop specialized legal reasoning components that incorporate case-based reasoning, statutory interpretation, and doctrinal analysis within coherent legal frameworks. Early steps in this direction appear in systems like *OpenAI/DeepResearch* [197], which incorporates basic legal language handling. More comprehensive specialization could enable sophisticated legal applications across practice areas.

Implementation opportunities include developing case analysis systems that extract and apply relevant precedent principles, implementing statutory interpretation frameworks that apply established analytical methodologies to legislative text, and designing multi-jurisdictional reasoning approaches that navigate conflicts of law across legal boundaries. These advances could enhance practical utility for legal research and analysis applications.

Regulatory Compliance Specialization. Compliance applications require comprehensive coverage with exceptional precision. Future systems could implement specialized compliance components that ensure complete regulatory coverage, systematic obligation identification, and reliable guidance across complex regulatory landscapes. Early examples of this direction appear in general information retrieval, but true compliance optimization remains limited. Advanced approaches could enable reliable automation of currently labor-intensive compliance processes.

Research directions include developing regulatory change tracking systems that monitor and interpret evolving requirements, implementing obligation extraction techniques that identify and classify compliance requirements across regulatory texts, and designing responsibility mapping approaches that connect regulatory

obligations with organizational functions and processes. These advances could enhance practical utility for compliance-intensive industries facing complex regulatory environments.

8.3.3 Medical and Healthcare Research Support. Healthcare applications present unique requirements and ethical considerations:

Clinical Evidence Synthesis. Medical applications require exceptional precision and comprehensive evidence integration. Future research could develop specialized medical components that synthesize clinical evidence across studies, guidelines, and practice observations while maintaining rigorous evaluation standards. Recent efforts such as Google’s co-scientist project [97] demonstrate the potential for AI to assist in scientific research including medical domains. Early steps in this direction appear in systems like *Perplexity/DeepResearch* [209], which implements enhanced citation for medical claims. More comprehensive specialization could enable reliable clinical decision support.

Implementation approaches might include evidence grading systems that apply established frameworks like GRADE [21] to clinical research, meta-analysis components that systematically integrate quantitative findings across studies, and guideline alignment techniques that map evidence to established clinical recommendations. These advances could enhance practical utility for evidence-based medicine while maintaining appropriate caution for this high-stakes domain.

Patient-Specific Research Adaptation. Personalized medicine requires adapting general knowledge to individual patient contexts. Future systems could implement specialized personalization components that adapt research findings based on patient characteristics, comorbidities, preferences, and other individual factors. Early examples of this direction appear in basic filtering of contraindications, but comprehensive personalization remains limited. Advanced approaches could enable truly personalized evidence synthesis for clinical applications.

Research opportunities include developing comorbidity reasoning systems that adjust recommendations based on condition interactions, implementing preference integration frameworks that incorporate patient values in evidence synthesis, and designing personalized risk-benefit analysis approaches that quantify individual trade-offs for treatment options. These advances could enhance clinical utility while respecting the complexity of individual patient contexts.

8.4 Human-AI Collaboration and Standardization

Enhancing human-AI partnership and establishing common standards represent crucial directions for practical research impact and ecosystem development.

8.4.1 Interactive Research Workflows. Effective collaboration requires sophisticated interaction throughout the research process:

Adaptive Query Refinement. Current systems offer limited interaction during query formulation and refinement. Future research could develop sophisticated refinement interfaces that collaboratively develop research questions through iterative clarification, expansion, and focusing based on initial results and user feedback. Early steps in this direction appear in systems like *HKUDS/Auto-Deep-Research* [112], which implements basic clarification dialogues, and benchmarks such as *QuestBench* [141], which evaluates AI systems’ ability to identify missing information and formulate appropriate clarification questions in underspecified reasoning tasks. More comprehensive approaches could enable truly collaborative question

development. Frameworks like **AutoAgent** [262] demonstrate how zero-code interfaces can enable non-technical users to effectively guide deep research processes through intuitive interaction patterns, while other systems are exploring methods that go beyond standard retrieval-augmented generation to better handle question identification in real-time conversations [4]. Implementation opportunities include developing intent clarification systems that identify potential ambiguities and alternatives in research questions, implementing scope adjustment interfaces that dynamically expand or narrow research focus based on initial findings, and designing perspective diversification tools that suggest alternative viewpoints relevant to research objectives. These advances could enhance research quality by improving question formulation through human-AI collaboration.

Interactive Exploration Interfaces. Current systems typically present relatively static research outputs. Future research could develop sophisticated exploration interfaces that enable dynamic navigation, drilling down, and expansion across research findings based on evolving interests. Early examples of this direction appear in systems like **OpenManus** [193], which provides basic exploration capabilities. Advanced approaches could enable truly interactive research experiences tailored to discovery patterns.

Research directions include developing information visualization techniques specifically designed for research navigation, implementing adaptive detail management that expands or collapses content areas based on user interest signals, and designing seamless source transition mechanisms that enable smooth movement between synthesis and original sources. These advances could enhance discovery by enabling more exploratory and serendipitous research experiences.

8.4.2 Expertise Augmentation Models. Effective augmentation requires adaptation to user expertise and objectives:

Expertise-Adaptive Interaction. Current systems offer limited adaptation to user knowledge levels and expertise. Future research could develop sophisticated adaptation mechanisms that tailor research approaches, explanations, and outputs based on user domain knowledge and research sophistication. Early steps in this direction appear in systems like **Perplexity/DeepResearch** [209], which implements basic terminology adjustment. More comprehensive adaptation could enable truly personalized research assistance aligned with individual expertise.

Implementation approaches might include expertise inference systems that dynamically assess user knowledge through interaction patterns, explanation adaptation mechanisms that adjust detail and terminology based on expertise models, and knowledge gap identification tools that highlight potentially unfamiliar concepts within research contexts. Furthermore, mechanisms that learn to strategically request expert assistance when encountering gaps exceeding autonomous capability - as formalized in the Learning to Yield and Request Control (YRC) coordination problem [66] - are crucial for optimizing intervention timing and resolution effectiveness. These advances could enhance research effectiveness across diverse user populations with varying domain familiarity.

Complementary Capability Design. Optimal augmentation leverages complementary human and AI strengths. Future systems could implement specialized interfaces designed around capability complementarity, emphasizing AI contributions in information processing while prioritizing human judgment for subjective evaluation and contextual understanding. Early examples of this direction appear in systems like **Agent-**

RL/ReSearch [2], which implements basic division of analytical responsibilities. More sophisticated approaches could enable truly synergistic human-AI research partnerships.

Research opportunities include developing explanation components specifically designed to facilitate human judgment rather than replace it, implementing confidence signaling mechanisms that highlight areas particularly requiring human evaluation, and designing interactive critique frameworks that enable efficient human feedback on system reasoning. Feng Xiong et al. [303] redefine the collaborative dynamics between human researchers and AI systems. These advances could enhance collaborative effectiveness by optimizing around natural capability distributions.

8.4.3 Framework Standardization Efforts. Common architectures enable modular development and component interoperability:

Component Interface Standardization. Advanced implementations employ standardized interfaces between major system components. The `OpenAI/AgentsSDK` [199] defines explicit interface standards for agent components, enabling modular development and component substitution. Emerging industry standards like Anthropic’s Model Context Protocol (MCP) [12] provide standardized interaction frameworks for large language models and tools, enabling consistent integration patterns across implementations. Similarly, Google’s Agent2Agent Protocol (A2A) [90, 92] establishes standardized communication patterns between autonomous agents, facilitating reliable multi-agent coordination. Open-source alternatives like `smolagents/open_deep_research` [115] implement comparable messaging protocols between agent components, highlighting industry convergence toward standardized interaction patterns. Projects like `Open_deep_search` [8] further demonstrate how standardized protocols enable effective collaboration between specialized research agents. Integration of diverse API interactions, as explored in `Tool11m` [223], provides additional standardization opportunities for managing external tool usage within research workflows.

Evaluation Metric Standardization. Current evaluation practices vary widely across implementations. Future research could establish standardized evaluation frameworks that enable consistent assessment and comparison across systems and components. Early examples of this direction appear in benchmarks like HLE [212] and MMLU [33], but comprehensive standardization remains limited. Advanced standardization could enable more efficient development through reliable quality signals and clear improvement metrics.

Research opportunities include developing standardized benchmark suites targeting specific research capabilities, implementing common evaluation methodologies across research domains and applications, and designing multi-dimensional assessment frameworks that provide nuanced performance profiles beyond simple accuracy metrics. These advances could enhance ecosystem quality by establishing clear standards and highlighting genuine improvements.

8.4.4 Cross-Platform Research Protocols. Interoperability across diverse systems enhances collective capabilities:

Research Result Exchange Formats. Current systems typically produce outputs in incompatible formats. Future research could develop standardized exchange formats that enable seamless sharing of research results across platforms and systems, enhancing collective capabilities. Early steps in this direction appear in basic document formats, but true research-specific standardization remains limited. Comprehensive standardization could enable research workflows spanning multiple specialized systems.

Implementation opportunities include defining standard structures for research findings with appropriate attribution and confidence metadata, establishing common formats for evidence representation across systems, and developing shared schemas for research questions and objectives to enable distributed processing. These advances could enhance capability through specialization and complementary system utilization.

Distributed Research Coordination. Advanced interoperability enables coordinated research across systems with complementary capabilities. Future research could develop sophisticated coordination frameworks that enable multi-system research workflows with appropriate task allocation, result integration, and process management. Early examples of this direction appear in workflows like those enabled by `n8n` [183], but comprehensive research-specific coordination remains limited. Advanced approaches could enable truly distributed research ecosystems with specialized components addressing distinct process elements.

Research directions include developing distributed search coordination protocols that efficiently leverage specialized search capabilities, implementing cross-system result verification techniques that ensure consistency across distributed findings, and designing efficient coordination protocols that minimize communication overhead in distributed research workflows. These advances could enhance collective capability through specialization and parallelization across the ecosystem.

8.4.5 Joint Human-AI Knowledge Creation. Moving beyond information retrieval to collaborative insight generation:

Collaborative Creation Environments. Advanced collaboration requires sophisticated content co-creation capabilities. Future research could develop specialized collaborative environments that enable fluid transition between human and AI contributions within unified document development. Early steps in this direction appear in systems like `mshumer/OpenDeepResearcher`, which implements basic collaborative document generation. Advanced interfaces like those explored in Self-Explanation in Social AI Agents [23] demonstrate how explanation capabilities can enhance collaborative research through more transparent reasoning processes. Similarly, innovative interaction paradigms like AI-Instruments [232] show how prompts can be embodied as instruments to abstract and reflect commands as general-purpose tools, suggesting novel approaches to research interface design that enhance collaborative capabilities through intuitive interaction patterns. Approaches where AI agents learn to assist other agents by observing them also show promise for developing more effective collaborative behaviors [127]. Effidit demonstrates comprehensive writing support through multifunctional capabilities including text polishing and context-aware phrase refinement, extending collaborative editing beyond basic generation [248]. More comprehensive approaches could enable truly integrated co-creation experiences.

Implementation opportunities include developing section suggestion systems that propose potential content expansions based on document context, implementing stylistic adaptation mechanisms that align AI-generated content with established document voice and approach, and incorporating implicit feedback mechanisms that interpret rejected suggestions as negative signals to refine outputs while preserving original intent [271], and designing seamless revision interfaces that enable efficient editing across human and AI contributions, like iterative human-AI co-editing as demonstrated by REVISE [302] – a framework allowing writers to dynamically modify summary segments through fill-in-the-middle generation. These advances could enhance collaborative productivity by reducing friction in joint content development [116].

Mixed-Initiative Research Design. Sophisticated collaboration includes shared determination of research direction and approach. Future systems could implement mixed-initiative frameworks that dynamically balance direction setting between human preferences and AI-identified opportunities throughout the research process. Early examples of this direction appear in systems like `smolagents/open_deep_research` [115], which implements basic suggestion mechanisms. Advanced approaches could enable truly collaborative research planning with balanced initiative distribution.

Research directions include developing opportunity identification systems that highlight promising but unexplored research directions, implementing trade-off visualization techniques that communicate potential research path alternatives and implications, and designing preference elicitation frameworks that efficiently capture evolving research priorities throughout the process, and integrating explainable reward function mechanisms to enhance human understanding of AI’s decision logic, thereby improving collaborative efficiency in value alignment contexts [239]. These advances could enhance discovery by combining human insight with AI-identified opportunities in balanced partnerships.

The future research directions outlined in this section highlight both the significant potential for advancement and the multi-faceted nature of Deep Research development. Progress will likely emerge through complementary advances across reasoning architectures, multimodal capabilities, domain specialization, human-AI collaboration, and ecosystem standardization. While commercial implementations like `OpenAI/DeepResearch` [197], `Gemini/DeepResearch` [60], and `Perplexity/DeepResearch` [209] will undoubtedly drive significant innovation, open-source alternatives and academic research will play crucial roles in expanding the boundaries of what’s possible and ensuring broad participation in this rapidly evolving field.

9 Conclusion

This survey has examined the rapidly evolving domain of Deep Research systems, tracing their development from initial implementations in 2023 through the sophisticated ecosystem emerging in 2025. Through comprehensive analysis of commercial offerings like `OpenAI/DeepResearch` [197], `Gemini/DeepResearch` [60], and `Perplexity/DeepResearch` [209], alongside open-source alternatives including `HKUDS/Auto-DeepResearch` [112], `dzhng/deep-research` [321], and numerous others, we have identified key technical patterns, implementation approaches, and application opportunities that characterize this transformative technology domain.

9.1 Key Findings and Contributions

Our analysis reveals several fundamental insights about the current state and trajectory of Deep Research systems:

Technical Architecture Patterns. Effective Deep Research implementations demonstrate consistent architectural patterns across foundation models, environmental interaction, task planning, and knowledge synthesis dimensions. Commercial implementations like `OpenAI/DeepResearch` [197] and `Gemini/DeepResearch` [60] typically leverage proprietary foundation models with extensive context lengths and sophisticated reasoning capabilities, while open-source alternatives like `Camel-AI/OWL` [43] and `QwenLM/Qwen-Agent` [224] demonstrate how effective research capabilities can be achieved with more accessible models through specialized optimization.

Environmental interaction capabilities show greater diversity, with specialized tools like **Nanobrowser** [184] and **dzhng/deep-research** [321] demonstrating exceptional effectiveness in web navigation and content extraction, while comprehensive platforms like **Manus** [164] and **AutoGLM-Search** [330] offer broader interaction capabilities across multiple environments. These patterns highlight both the value of specialization and the importance of comprehensive environmental access for effective research.

Task planning and execution approaches reveal similar diversity, with frameworks like **OpenAI/AgentsSDK** [199] and **Flowith/OracleMode** [77] providing sophisticated planning capabilities, while systems like **Agent-RL/ReSearch** [2] and **smolagents/open_deep_research** [115] emphasize execution reliability and collaborative approaches respectively. Knowledge synthesis capabilities demonstrate consistent emphasis on information evaluation, though with varied approaches to presentation and interactivity across implementations like **HKUDS/Auto-Deep-Research** [112] and **mshumer/OpenDeepResearcher** [249].

Implementation Approach Distinctions. Our analysis highlights meaningful distinctions between commercial and open-source implementation approaches. Commercial platforms typically offer optimized performance, sophisticated interfaces, and comprehensive capabilities, though with associated costs and customization limitations. Systems like **OpenAI/DeepResearch** [197] and **Perplexity/DeepResearch** [209] demonstrate exceptional performance on standard benchmarks, though with significant variation in application focus and interaction models.

Open-source implementations demonstrate greater architectural diversity and customization flexibility, though typically with increased deployment complexity and more limited performance on standard benchmarks. Projects like **dzhng/deep-research** [321], **nickscamara/open-deep-research** [42], and **HKUDS/Auto-Deep-Research** [112] offer complete research pipelines with varied architectural approaches, while specialized components like **Jina-AI/node-DeepResearch** [121] and **Nanobrowser** [184] enable customized workflows addressing specific requirements. Frameworks such as **AutoChain** [78] provide lightweight tools to simplify the creation and evaluation of custom generative agents, enabling rapid iteration for specialized applications.

These distinctions highlight complementary roles within the ecosystem, with commercial implementations offering accessibility and performance for general users, while open-source alternatives enable customization, control, and potentially lower operational costs for specialized applications and high-volume usage. This diversity enhances overall ecosystem health through competition, specialization, and diverse innovation paths.

Application Domain Adaptations. Our examination of application patterns reveals meaningful adaptations across domains including academic research [118, 273, 276], scientific discovery [6, 10, 25, 47, 79, 83, 98, 99, 110, 129, 130, 135, 155, 166, 169, 218, 255, 258, 264, 269, 310, 312, 322, 327], business intelligence [187], financial analysis, education [14, 215, 219, 317], and personal knowledge management [136, 336]. Academic applications exemplified by systems like **OpenAI/DeepResearch** [197] and **Camel-AI/OWL** [43] demonstrate particular emphasis on comprehensive literature coverage, methodological understanding, and citation quality. Scientific implementations like **Gemini/DeepResearch** [60] and **Agent-RL/ReSearch** [2] emphasize experimental design, data analysis, and theory development capabilities.

Business applications leveraging systems like **Manus** [164] and **n8n** [183] show stronger focus on information currency, competitive analysis, and actionable insight generation. Educational implementations demonstrate

adaptations for learning support, content development, and research skill training across systems like **Perplexity/DeepResearch** [209] and **OpenManus** [193]. These patterns highlight how general deep research capabilities translate into domain value through specialized adaptation addressing field-specific requirements and workflows.

Ethical Consideration Approaches. Our analysis reveals both common patterns and implementation diversity in addressing crucial ethical dimensions including information accuracy, privacy protection, intellectual property respect, and accessibility. Commercial implementations typically demonstrate sophisticated approaches to factual verification, with systems like **OpenAI/DeepResearch** [197] and **Perplexity/DeepResearch** [209] implementing multi-level verification and explicit attribution, while open-source alternatives like **grapeot/deep_research_agent** [263] and **HKUDS/Auto-Deep-Research** [112] demonstrate pragmatic approaches within more constrained technical environments.

Privacy protection shows similar patterns, with commercial systems implementing comprehensive safeguards appropriate to their cloud-based operation, while open-source alternatives like **OpenManus** [193] emphasize local deployment for sensitive applications. Attribution and intellectual property approaches demonstrate consistent emphasis on source transparency and appropriate utilization boundaries, though with varied implementation sophistication across the ecosystem.

These patterns highlight both shared ethical priorities across the ecosystem and implementation diversity reflecting different technical constraints, deployment models, and user requirements. This diversity represents a strength in addressing multi-faceted ethical challenges through complementary approaches and continuous innovation.

9.2 Limitations and Outlook

While this survey provides comprehensive analysis of current Deep Research systems and emerging trends, several limitations warrant acknowledgment:

Rapidly Evolving Landscape. The accelerating pace of development in this domain presents inherent challenges for comprehensive analysis. New systems and capabilities continue to emerge, with commercial offerings like **OpenAI/DeepResearch** [197], **Gemini/DeepResearch** [60], and **Perplexity/DeepResearch** [209] receiving frequent updates, while the open-source ecosystem continuously expands through new projects and enhancements to existing frameworks like **dzhng/deep-research** [321] and **HKUDS/Auto-Deep-Research** [112].

This survey captures the state of the art as of early 2025, but both technical capabilities and implementation approaches will continue to evolve rapidly. The classification framework and analysis methodology provided here offer a structural foundation for continued assessment as the field progresses through subsequent development phases.

Implementation Detail Limitations. Comprehensive technical analysis faces challenges due to limited implementation transparency, particularly for commercial systems. While open-source implementations like **nicksamara/open-deep-research** [42] and **Agent-RL/ReSearch** [2] enable detailed architectural examination, commercial systems like **OpenAI/DeepResearch** [197] and **Gemini/DeepResearch** [60] reveal limited internal details, restricting comprehensive comparative analysis of certain technical dimensions.

Our approach addresses this limitation through behavioral analysis, publicly available documentation examination, and consistent evaluation across standardized benchmarks and qualitative assessment frameworks. These methods enable meaningful comparison despite transparency variations, though complete architectural analysis remains challenging for proprietary implementations.

Application Impact Assessment. Evaluating real-world impact presents persistent challenges given the early deployment stage of many Deep Research systems. While initial applications demonstrate promising capabilities across domains including academic research[17, 208, 225, 292], business intelligence, and education[14, 215, 317], a comprehensive long-term impact assessment requires extended observation beyond the scope of this survey. Potential transformative effects on research methodologies, knowledge work, and information access patterns remain partially speculative despite encouraging early indications.

Future research should incorporate longitudinal analysis of deployment patterns, usage evolution, and organizational integration to assess realized impact beyond technical capabilities and early applications. Such analysis would complement the technical and architectural focus of the current survey with valuable perspectives on practical significance and societal implications.

9.3 Broader Implications

Beyond specific findings, this survey highlights several broader implications for the future of knowledge work and information access:

Research Methodology Transformation. Deep Research systems demonstrate potential to fundamentally transform research methodologies across domains. The comprehensive information access, advanced reasoning capabilities, and efficient knowledge synthesis demonstrated by systems like **OpenAI/DeepResearch** [197], **Gemini/DeepResearch** [60], and their open-source alternatives suggest significant opportunities to accelerate discovery, enhance comprehensiveness, and enable novel cross-domain connections beyond traditional research approaches.

Rather than simply automating existing processes, these systems enable fundamentally new research approaches leveraging capabilities exceeding human information processing in scale while complementing human insight, creativity, and contextual understanding. This complementarity suggests evolution toward collaborative research models rather than replacement of human researchers, with significant potential for productivity enhancement and discovery acceleration. However, Ashktorab et al. [15] highlight that in human-AI collaboration, users may exhibit overreliance behaviors, appending AI-generated responses even when conflicting, which can compromise data quality.

Knowledge Access Democratization. The emergence of accessible Deep Research implementations across commercial and open-source ecosystems demonstrates potential for broader knowledge democratization. Systems like **Perplexity/DeepResearch** [209] with free access tiers and open-source alternatives like **nicksamara/open-deep-research** [42] and **HKUDS/Auto-Deep-Research** [112] enable sophisticated research capabilities previously requiring specialized expertise and substantial resources, potentially reducing barriers to high-quality information access and analysis.

This democratization carries significant implications for education, entrepreneurship, civic participation, and individual knowledge development. While accessibility challenges remain, particularly regarding technical

expertise requirements and computational resources, the overall trajectory suggests broadening access to advanced research capabilities with potential positive impacts on knowledge equity across society.

Collective Intelligence Enhancement. Beyond individual applications, Deep Research systems demonstrate potential for collective intelligence enhancement through improved knowledge integration, insight sharing, and collaborative discovery. The capabilities demonstrated by systems like *Manus* [164], *Flowith/OracleMode* [77], and *smolagents/open_deep_research* [115] suggest opportunities for enhanced knowledge synthesis across organizational and disciplinary boundaries, potentially addressing fragmentation challenges in increasingly complex knowledge domains.

Rather than viewing these systems as isolated tools, their integration into collaborative knowledge ecosystems highlights potential for systemic enhancement of collective sense-making, evidence-based decision making, and shared understanding development. This perspective emphasizes the social and organizational dimensions of Deep Research impact beyond technical capabilities and individual productivity enhancement.

9.4 Final Thoughts

The rapid emergence and evolution of Deep Research systems represent a significant advancement in the application of artificial intelligence to knowledge discovery and utilization. While technical implementations will continue to evolve and specific systems will emerge and recede, the fundamental capability shift enabled by these technologies appears likely to persist and expand.

The diverse ecosystem spanning commercial platforms like *OpenAI/DeepResearch* [197], *Gemini/DeepResearch* [60], and *Perplexity/DeepResearch* [209], alongside open-source alternatives like *dzhng/deep-research* [321], *HKUDS/Auto-Deep-Research* [112], and numerous specialized components, demonstrates innovation across multiple technical dimensions, implementation approaches, and application domains. This diversity enhances overall ecosystem health through competition, specialization, and complementary development trajectories.

As research continues across advanced reasoning architectures, multimodal capabilities, domain specialization, human-AI collaboration, and ecosystem standardization, we anticipate continued rapid advancement building on the foundation established by current implementations. This evolution will likely yield increasingly sophisticated research capabilities with significant implications for knowledge work across domains, potentially transforming how information is discovered, validated, synthesized, and utilized throughout society.

The responsible development of these powerful capabilities requires continued attention to ethical considerations including information accuracy, privacy protection, intellectual property respect, and accessibility. By addressing these considerations alongside technical advancement, the Deep Research ecosystem can fulfill its potential for positive impact on knowledge discovery and utilization while minimizing potential harms or misuse.

In conclusion, Deep Research represents both a fascinating technical domain for continued research and a potentially transformative capability for practical knowledge work across society. The frameworks, analysis, and directions presented in this survey provide a foundation for continued examination of this rapidly evolving field with significant implications for the future of information access, knowledge synthesis, and discovery processes.

References

- [1] Adilzhan Adilkhanov, Amir Yelenov, Assylkhan Seitzhanov, Ayan Mazhitov, Azamat Abdikarimov, Danissa Sandykbayeva, Daryn Kenzhebek, Dinmukhammed Mukashev, Ilyas Umurbekov, Jabrail Chumakov, Kamila Spanova, Karina Burunchina, Madina Yergibay, Margulan Issa, Moldir Zabirowa, Nurdaulet Zhuzbay, Nurlan Kabydyshov, Nurlan Zhaniyar, Rasul Yermagambet, Rustam Chibar, Saltanat Seitzhan, Soibkhon Khajikhanov, Tasbolat Taunyazov, Temirlan Galimzhanov, Temirlan Kaiyrbay, Tleukhan Mussin, Togzhan Syrymova, Valeriya Kostyukova, Yerkebulan Massalim, Yermakhan Kassym, Zerde Nurbayeva, and Zhanat Kappassov. 2025. Survey on Vision-Language-Action Models. arXiv:2502.06851 [cs.CL] <https://arxiv.org/abs/2502.06851>
- [2] Agent-RL. 2024. ReSearch. <https://github.com/Agent-RL/ReSearch>.
- [3] Agno-AGI. 2025. Agno. <https://github.com/agno-agi/agno>.
- [4] Garima Agrawal, Sashank Gummuluri, and Cosimo Spera. 2024. Beyond-RAG: Question Identification and Answer Generation in Real-Time Conversations. arXiv:2410.10136 [cs.CL] <https://arxiv.org/abs/2410.10136>
- [5] Flowise AI. 2023. Flowise: Low-code LLM Application Building Tool. <https://flowiseai.com/>.
- [6] Nawaf Alampara, Mara Schilling-Wilhelmi, Martiño Ríos-García, Indrajeet Mandal, Pranav Khetarpal, Hargun Singh Grover, N. M. Anoop Krishnan, and Kevin Maik Jablonka. 2025. Probing the limitations of multimodal language models for chemistry and materials research. arXiv:2411.16955 [cs.LG] <https://arxiv.org/abs/2411.16955>
- [7] AlphaProof and AlphaGeometry teams. 2024. AI achieves silver-medal standard solving International Mathematical Olympiad problems. <https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/>.
- [8] 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:2503.20201 [cs.LG] <https://arxiv.org/abs/2503.20201>
- [9] Lucio Anderlini, Matteo Barbetti, Giulio Bianchini, Diego Ciangottini, Stefano Dal Pra, Diego Michelotto, Carmelo Pellegrino, Rosa Petrini, Alessandro Pascolini, and Daniele Spiga. 2025. Supporting the development of Machine Learning for fundamental science in a federated Cloud with the AI-INFN platform. arXiv:2502.21266 [cs.DC] <https://arxiv.org/abs/2502.21266>
- [10] Mehrad Ansari and Seyed Mohamad Moosavi. 2023. Agent-based Learning of Materials Datasets from Scientific Literature. <https://github.com/AI4ChemS/Eunomia>. arXiv:2312.11690 [cs.AI] <https://arxiv.org/abs/2312.11690>
- [11] Anthropic. 2024. Building effective agents. <https://www.anthropic.com/engineering/building-effective-agents>.
- [12] Anthropic. 2024. Model Context Protocol (MCP). <https://docs.anthropic.com/en/docs/agents-and-tools/mcp>.
- [13] Anthropic. 2025. Claude takes research to new places. <https://www.anthropic.com/news/research>.
- [14] Prakash Aryan. 2024. LLMs as Debate Partners: Utilizing Genetic Algorithms and Adversarial Search for Adaptive Arguments. arXiv:2412.06229 [cs.AI] <https://arxiv.org/abs/2412.06229>
- [15] Zahra Ashktorab, Qian Pan, Werner Geyer, Michael Desmond, Marina Danilevsky, James M. Johnson, Casey Dugan, and Michelle Bachman. 2024. Emerging Reliance Behaviors in Human-AI Text Generation: Hallucinations, Data Quality Assessment, and Cognitive Forcing Functions. arXiv:2409.08937 [cs.HC] <https://arxiv.org/abs/2409.08937>
- [16] assafelevic. 2023. GPT-Researcher. <https://github.com/assafelevic/gpt-researcher/>.
- [17] Ahmet Yasin Aytar, Kemal Kilic, and Kamer Kaya. 2024. A Retrieval-Augmented Generation Framework for Academic Literature Navigation in Data Science. arXiv:2412.15404 [cs.IR] <https://arxiv.org/abs/2412.15404>
- [18] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. arXiv:2404.07738 [cs.CL] <https://arxiv.org/abs/2404.07738>
- [19] Dzmitry Bahdanau, Nicolas Gontier, Gabriel Huang, Ehsan Kamalloo, Rafael Pardini, Alex Piché, Torsten Scholak, Oleh Shliazhko, Jordan Prince Tremblay, Karam Ghanem, Soham Parikh, Mitul Tiwari, and Quaizar Vohra. 2024. TapeAgents: a Holistic Framework for Agent Development and Optimization. arXiv:2412.08445 [cs.AI] <https://arxiv.org/abs/2412.08445>
- [20] Gal Bakal, Ali Dasdan, Yaniv Katz, Michael Kaufman, and Guy Levin. 2025. Experience with GitHub Copilot for Developer Productivity at Zoominfo. arXiv:2501.13282 [cs.SE] <https://arxiv.org/abs/2501.13282>
- [21] Howard Balshem, Mark Helfand, Holger J Schünemann, Andrew D Oxman, Regina Kunz and Jan Brozek, Gunn E Vist, Yngve Falck-Ytter, Joerg Meerpohl, Susan Norris, and Gordon H Guyatt. 2011. GRADE guidelines: 3. Rating the quality of evidence. <https://pubmed.ncbi.nlm.nih.gov/21208779/>.
- [22] Samuel Barham, Orion Weller, Michelle Yuan, Kenton Murray, Mahsa Yarmohammadi, Zhengping Jiang, Siddharth Vashishtha, Alexander Martin, Anqi Liu, Aaron Steven White, Jordan Boyd-Graber, and Benjamin Van Durme. 2023. MegaWika: Millions of reports and their sources across 50 diverse languages. arXiv:2307.07049 [cs.CL] <https://arxiv.org/abs/2307.07049>

- [23] Rhea Basappa, Mustafa Tekman, Hong Lu, Benjamin Faught, Sandeep Kakar, and Ashok K. Goel. 2024. *Social AI Agents Too Need to Explain Themselves*. Springer Nature Switzerland, 351–360. doi:10.1007/978-3-031-63028-6_29
- [24] Joeran Beel, Min-Yen Kan, and Moritz Baumgart. 2025. Evaluating Sakana’s AI Scientist for Autonomous Research: Wishful Thinking or an Emerging Reality Towards ‘Artificial Research Intelligence’ (ARI)? arXiv:2502.14297 [cs.IR] <https://arxiv.org/abs/2502.14297>
- [25] Morad Behandish, John Maxwell III, and Johan de Kleer. 2022. AI Research Associate for Early-Stage Scientific Discovery. arXiv:2202.03199 [cs.AI] <https://arxiv.org/abs/2202.03199>
- [26] Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. 2025. Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path? arXiv:2502.15657 [cs.AI] <https://arxiv.org/abs/2502.15657>
- [27] Karim Benharrak, Tim Zindulka, and Daniel Buschek. 2024. Deceptive Patterns of Intelligent and Interactive Writing Assistants. arXiv:2404.09375 [cs.HC] <https://arxiv.org/abs/2404.09375>
- [28] Zhen Bi, Ningyu Zhang, Yida Xue, Yixin Ou, Daxiong Ji, Guozhou Zheng, and Huajun Chen. 2024. OceanGPT: A Large Language Model for Ocean Science Tasks. <http://oceanppt.zjukg.cn/>. arXiv:2310.02031 [cs.CL] <https://arxiv.org/abs/2310.02031>
- [29] Stefano Bianchini, Moritz Müller, and Pierre Pelletier. 2024. Drivers and Barriers of AI Adoption and Use in Scientific Research. arXiv:2312.09843 [cs.CY] <https://arxiv.org/abs/2312.09843>
- [30] bindAI. 2025. ChatGPT Deep Research vs Perplexity – Which One Is Better? <https://blog.getbind.co/2025/02/03/chatgpt-deep-research-is-it-better-than-perplexity/>.
- [31] Francisco Bolanos, Angelo Salatino, Francesco Osborne, and Enrico Motta. 2024. Artificial Intelligence for Literature Reviews: Opportunities and Challenges. arXiv:2402.08565 [cs.AI] <https://arxiv.org/abs/2402.08565>
- [32] Bolt. 2024. Bolt. <https://bolt.new/>.
- [33] braca. 2025. MMLU benchmark: Testing LLMs multi-task capabilities. <https://www.braca.eu/post/mmlu-benchmark>.
- [34] Andres M Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. 2023. ChemCrow: Augmenting large-language models with chemistry tools. arXiv:2304.05376 [physics.chem-ph] <https://arxiv.org/abs/2304.05376>
- [35] Chris Brown and Jason Cusati. 2024. Exploring the Evidence-Based Beliefs and Behaviors of LLM-Based Programming Assistants. arXiv:2407.13900 [cs.SE] <https://arxiv.org/abs/2407.13900>
- [36] browserbase. 2025. Open-operator. <https://github.com/browserbase/open-operator>.
- [37] btahir. 2024. open_deep_research. <https://github.com/btahir/open-deep-research>.
- [38] ByteDance. 2024. Coze Space. <https://www.coze.cn/space-preview>.
- [39] ByteDance. 2025. agent-tars. <https://github.com/bytedance/UI-TARS-desktop/tree/main/apps/agent-tars>.
- [40] Beatriz Cabrero-Daniel, Tomas Herda, Victoria Pichler, and Martin Eder. 2024. Exploring Human-AI Collaboration in Agile: Customised LLM Meeting Assistants. arXiv:2404.14871 [cs.SE] <https://arxiv.org/abs/2404.14871>
- [41] Filipe Calegario, Vanilson Burégio, Francisco Erivaldo, Daniel Moraes Costa Andrade, Kailane Felix, Nathalia Barbosa, Pedro Lucas da Silva Lucena, and César França. 2023. Exploring the intersection of Generative AI and Software Development. arXiv:2312.14262 [cs.SE] <https://arxiv.org/abs/2312.14262>
- [42] Nicholas Camara. 2025. open-deep-research. <https://github.com/nickscamara/open-deep-research>.
- [43] Camel AI. 2025. OWL. <https://github.com/camel-ai/owl>.
- [44] Franck Cappello, Sandeep Madireddy, Robert Underwood, Neil Getty, Nicholas Lee-Ping Chia, Nesar Ramachandra, Josh Nguyen, Murat Keceli, Tanwi Mallick, Zilinghan Li, Marieme Ngom, Chenhui Zhang, Angel Yanguas-Gil, Evan Antoniuk, Bhavya Kailkhura, Minyang Tian, Yufeng Du, Yuan-Sen Ting, Azton Wells, Bogdan Nicolae, Avinash Maurya, M. Mustafa Rafique, Eliu Huerta, Bo Li, Ian Foster, and Rick Stevens. 2025. EAIRA: Establishing a Methodology for Evaluating AI Models as Scientific Research Assistants. arXiv:2502.20309 [cs.AI] <https://arxiv.org/abs/2502.20309>
- [45] Peter Cardon, Carolin Fleischmann, Jolanta Aritz, Minna Logemann, and Jeanette Heidewald. 2023. The Challenges and Opportunities of AI-Assisted Writing: Developing AI Literacy for the AI Age. <https://journals.sagepub.com/doi/abs/10.1177/23294906231176517>.
- [46] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. 2025. Why Do Multi-Agent LLM Systems Fail? arXiv:2503.13657 [cs.AI] <https://arxiv.org/abs/2503.13657>
- [47] Eric Chamoun, Michael Schlichtrull, and Andreas Vlachos. 2024. Automated Focused Feedback Generation for Scientific Writing Assistance. arXiv:2405.20477 [cs.CL] <https://arxiv.org/abs/2405.20477>
- [48] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. ChatEval: Towards Better LLM-based Evaluators through Multi-Agent Debate. <https://github.com/thunlp/ChatEval>. arXiv:2308.07201 [cs.CL] <https://arxiv.org/abs/2308.07201>

- [49] Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, Aili Chen, Nianqi Li, Lida Chen, Caiyu Hu, Siye Wu, Scott Ren, Ziquan Fu, and Yanghua Xiao. 2024. From Persona to Personalization: A Survey on Role-Playing Language Agents. arXiv:2404.18231 [cs.CL] <https://arxiv.org/abs/2404.18231>
- [50] Kexin Chen, Hanqun Cao, Junyou Li, Yuyang Du, Menghao Guo, Xin Zeng, Lanqing Li, Jiezhong Qiu, Pheng Ann Heng, and Guangyong Chen. 2024. An Autonomous Large Language Model Agent for Chemical Literature Data Mining. arXiv:2402.12993 [cs.IR] <https://arxiv.org/abs/2402.12993>
- [51] Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyang Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Junjun He, and Yu Qiao. 2024. GMAI-MMBench: A Comprehensive Multimodal Evaluation Benchmark Towards General Medical AI. <https://uni-medical.github.io/GMAI-MMBench.github.io/>. arXiv:2408.03361 [eess.IV] <https://arxiv.org/abs/2408.03361>
- [52] Tingting Chen, Srinivas Anumasa, Beibei Lin, Vedant Shah, Anirudh Goyal, and Dianbo Liu. 2025. AutoBench: An Automated Benchmark for Scientific Discovery in LLMs. <https://github.com/AutoBench/AutoBench>. arXiv:2502.15224 [cs.LG] <https://arxiv.org/abs/2502.15224>
- [53] Valerie Chen, Alan Zhu, Sebastian Zhao, Hussein Mozannar, David Sontag, and Ameet Talwalkar. 2025. Need Help? Designing Proactive AI Assistants for Programming. arXiv:2410.04596 [cs.HC] <https://arxiv.org/abs/2410.04596>
- [54] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2023. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. arXiv:2308.10848 [cs.CL] <https://arxiv.org/abs/2308.10848>
- [55] Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can AI Assistants Know What They Don't Know? arXiv:2401.13275 [cs.CL] <https://arxiv.org/abs/2401.13275>
- [56] Zhao Cheng, Diane Wan, Matthew Abueg, Sahra Ghalebikesabi, Ren Yi, Eugene Bagdasarian, Borja Balle, Stefan Mellem, and Shawn O'Banion. 2024. CI-Bench: Benchmarking Contextual Integrity of AI Assistants on Synthetic Data. <https://www.aimodels.fyi/papers/arxiv/ci-bench-benchmarking-contextual-integrity-ai-assistants>. arXiv:2409.13903 [cs.AI] <https://arxiv.org/abs/2409.13903>
- [57] Bhavya Chopra, Ananya Singha, Anna Fariha, Sumit Gulwani, Chris Parnin, Ashish Tiwari, and Austin Z. Henley. 2023. Conversational Challenges in AI-Powered Data Science: Obstacles, Needs, and Design Opportunities. arXiv:2310.16164 [cs.HC] <https://arxiv.org/abs/2310.16164>
- [58] Daniel J. H. Chung, Zhiqi Gao, Yurii Kvasiuk, Tianyi Li, Moritz Münchmeyer, Maja Rudolph, Frederic Sala, and Sai Chaitanya Tadepalli. 2025. Theoretical Physics Benchmark (TPBench) – a Dataset and Study of AI Reasoning Capabilities in Theoretical Physics. <https://tpbench.org/>. arXiv:2502.15815 [cs.LG] <https://arxiv.org/abs/2502.15815>
- [59] Umut Cihan, Vahid Haratian, Arda İçöz, Mert Kaan Gül, Ömercan Devran, Emircan Furkan Bayendur, Baykal Mehmet Uçar, and Eray Tüzün. 2024. Automated Code Review In Practice. arXiv:2412.18531 [cs.SE] <https://arxiv.org/abs/2412.18531>
- [60] Dave Citron. 2025. Deep Research is now available on Gemini 2.5 Pro Experimental. <https://blog.google/products/gemini/deep-research-gemini-2-5-pro-experimental/>.
- [61] Cline. 2024. Cline. <https://github.com/cline/cline>.
- [62] Cognition Labs. 2025. *Devin.ai*. <https://devin.ai>
- [63] Consensus. 2025. Consensus. <https://consensus.app/>.
- [64] crewAIInc. 2023. CrewAI. <https://github.com/crewAIInc/crewAI>.
- [65] Cursor. 2023. Cursor. <https://www.cursor.com/>.
- [66] Mohamad H. Danesh, Tu Trinh, Benjamin Plaut, and Nguyen X. Khanh. 2025. Learning to Coordinate with Experts. <https://github.com/modanesh/YRC-Bench>. arXiv:2502.09583 [cs.LG] <https://arxiv.org/abs/2502.09583>
- [67] Kristin M. de Payrebrune, Kathrin Flaßkamp, Tom Ströhla, Thomas Sattel, Dieter Bestle, Benedict Röder, Peter Eberhard, Sebastian Peitz, Marcus Stoffel, Gulakala Rutwik, Borse Aditya, Meike Wohlleben, Walter Sextro, Maximilian Raff, C. David Remy, Manish Yadav, Merten Stender, Jan van Delden, Timo Lüddecke, Sabine C. Langer, Julius Schultz, and Christopher Blech. 2024. The impact of AI on engineering design procedures for dynamical systems. arXiv:2412.12230 [eess.SY] <https://arxiv.org/abs/2412.12230>
- [68] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, et al. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] <https://arxiv.org/abs/2501.12948>
- [69] Akash Dhruv and Anshu Dubey. 2025. Leveraging Large Language Models for Code Translation and Software Development in Scientific Computing. arXiv:2410.24119 [cs.SE] <https://arxiv.org/abs/2410.24119>

- [70] Talissa Dreossi. 2025. Bridging Logic Programming and Deep Learning for Explainability through ILASP. *Electronic Proceedings in Theoretical Computer Science* 416 (Feb. 2025), 314–323. doi:10.4204/eptcs.416.31
- [71] Ian Drosos, Advait Sarkar, Xiaotong Xu, and Neil Toronto. 2025. "It makes you think": Provocations Help Restore Critical Thinking to AI-Assisted Knowledge Work. arXiv:2501.17247 [cs.HC] <https://arxiv.org/abs/2501.17247>
- [72] Omer Dunay, Daniel Cheng, Adam Tait, Parth Thakkar, Peter C Rigby, Andy Chiu, Imad Ahmad, Arun Ganesan, Chandra Maddila, Vijayaraghavan Murali, Ali Tayyebi, and Nachiappan Nagappan. 2024. Multi-line AI-assisted Code Authoring. arXiv:2402.04141 [cs.SE] <https://arxiv.org/abs/2402.04141>
- [73] Steffen Eger, Yong Cao, Jennifer D'Souza, Andreas Geiger, Christian Greisinger, Stephanie Gross, Yufang Hou, Brigitte Krenn, Anne Lauscher, Yizhi Li, Chenghua Lin, Nafise Sadat Moosavi, Wei Zhao, and Tristan Miller. 2025. Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation. arXiv:2502.05151 [cs.CL] <https://arxiv.org/abs/2502.05151>
- [74] Elicit. 2025. Elicit. <https://elicit.com/?redirected=true>.
- [75] Michael D. Ernst. 2017. Natural Language is a Programming Language: Applying Natural Language Processing to Software Development. <https://drops.dagstuhl.de/storage/00lipics/lipics-vol071-snapl2017/LIPICS.SNAPL.2017.4/LIPICS.SNAPL.2017.4.pdf>.
- [76] Wenqi Fan, Yajuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented Large Language Models. arXiv:2405.06211 [cs.CL] <https://arxiv.org/abs/2405.06211>
- [77] Flowith. 2024. Flowith Oracle Mode. <https://flowith.net/>.
- [78] Forethought-Technologies. 2023. AutoChain. <https://github.com/Forethought-Technologies/AutoChain>.
- [79] César França. 2023. AI empowering research: 10 ways how science can benefit from AI. arXiv:2307.10265 [cs.GL] <https://arxiv.org/abs/2307.10265>
- [80] Future-House. 2023. PaperQA. <https://github.com/Future-House/paper-qa>.
- [81] GAIR-NLP. 2025. DeepResearcher. <https://github.com/GAIR-NLP/DeepResearcher>.
- [82] Difei Gao, Lei Ji, Luwei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. 2023. AssistGPT: A General Multi-modal Assistant that can Plan, Execute, Inspect, and Learn. arXiv:2306.08640 [cs.CV] <https://arxiv.org/abs/2306.08640>
- [83] Shanghua Gao, Ada Fang, Yepeng Huang, Valentina Giunchiglia, Ayush Noori, Jonathan Richard Schwarz, Yasha Ektefaie, Jovana Kondic, and Marinka Zitnik. 2024. Empowering Biomedical Discovery with AI Agents. arXiv:2404.02831 [cs.AI] <https://arxiv.org/abs/2404.02831>
- [84] Alireza Ghafarollahi and Markus J. Buehler. 2024. SciAgents: Automating scientific discovery through multi-agent intelligent graph reasoning. arXiv:2409.05556 [cs.AI] <https://arxiv.org/abs/2409.05556>
- [85] Luca Gioacchini, Marco Mellia, Idilio Drago, Alexander Delsanto, Giuseppe Siracusano, and Roberto Bifulco. 2024. AutoPenBench: Benchmarking Generative Agents for Penetration Testing. <https://github.com/lucagioacchini/autopen-bench>. arXiv:2410.03225 [cs.CR] <https://arxiv.org/abs/2410.03225>
- [86] Github. 2021. Github Copilot. <https://github.com/features/copilot?ref=nav.poetries.top>.
- [87] Amr Goma, Michael Sargious, and Antonio Krüger. 2024. AdaptoML-UX: An Adaptive User-centered GUI-based AutoML Toolkit for Non-AI Experts and HCI Researchers. https://github.com/MichaelSargious/AdaptoML_UX. arXiv:2410.17469 [cs.HC] <https://arxiv.org/abs/2410.17469>
- [88] Google. 2021. BIG-bench. <https://github.com/google/BIG-bench>.
- [89] Google. 2024. Try Deep Research and our new experimental model in Gemini, your AI assistant. <https://blog.google/products/gemini/google-gemini-deep-research/>.
- [90] Google. 2025. A2A. <https://github.com/google/A2A>.
- [91] Google. 2025. Agent Development Kit. <https://google.github.io/adk-docs/>.
- [92] Google. 2025. Announcing the Agent2Agent Protocol (A2A). <https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/>.
- [93] Google. 2025. Gemini 2.0 Flash (Feb '25): Intelligence, Performance and Price Analysis. <https://artificialanalysis.ai/models/gemini-2-0-flash>.
- [94] Google. 2025. gemini-fullstack-langgraph-quickstart. <https://github.com/google-gemini/gemini-fullstack-langgraph-quickstart>.
- [95] Google. 2025. NotebookLm. <https://notebooklm.google/>.
- [96] Kanika Goswami, Puneet Mathur, Ryan Rossi, and Franck Dernoncourt. 2025. ChartCitor: Multi-Agent Framework for Fine-Grained Chart Visual Attribution. arXiv:2502.00989 [cs.CL] <https://arxiv.org/abs/2502.00989>
- [97] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni,

- Nenad Tomasev, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. 2025. Towards an AI co-scientist. https://storage.googleapis.com/coscientist_paper/ai_coscientist.pdf.
- [98] Alexander H. Gower, Konstantin Korovin, Daniel Brunnsåker, Filip Kronström, Gabriel K. Reder, Ievgeniia A. Tiukova, Ronald S. Reiserer, John P. Wikswo, and Ross D. King. 2024. The Use of AI-Robotic Systems for Scientific Discovery. arXiv:2406.17835 [cs.LG] <https://arxiv.org/abs/2406.17835>
- [99] Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. 2025. LLMs can Realize Combinatorial Creativity: Generating Creative Ideas via LLMs for Scientific Research. arXiv:2412.14141 [cs.AI] <https://arxiv.org/abs/2412.14141>
- [100] Yuzhe Gu, Wenwei Zhang, Chengqi Lyu, Dahua Lin, and Kai Chen. 2025. Mask-DPO: Generalizable Fine-grained Factuality Alignment of LLMs. arXiv:2503.02846 [cs.CL] <https://arxiv.org/abs/2503.02846>
- [101] Chengquan Guo, Xun Liu, Chulin Xie, Andy Zhou, Yi Zeng, Zinan Lin, Dawn Song, and Bo Li. 2024. Red-Code: Risky Code Execution and Generation Benchmark for Code Agents. <https://github.com/AI-secure/RedCode>. arXiv:2411.07781 [cs.SE] <https://arxiv.org/abs/2411.07781>
- [102] Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. 2024. DS-Agent: Automated Data Science by Empowering Large Language Models with Case-Based Reasoning. <https://github.com/guosyjl/DS-Agent>. arXiv:2402.17453 [cs.LG] <https://arxiv.org/abs/2402.17453>
- [103] Xin Guo, Haotian Xia, ZhaoWei Liu, Hanyang Cao, Zhi Yang, Zhiqiang Liu, Sizhe Wang, Jinyi Niu, Chuqi Wang, Yanhui Wang, Xiaolong Liang, Xiaoming Huang, Bing Zhu, Zhongyu Wei, Yun Chen, Weining Shen, and Liwen Zhang. 2024. FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models. arXiv:2308.09975 [cs.CL] <https://arxiv.org/abs/2308.09975>
- [104] Hilda Hadan, Derrick Wang, Reza Hadi Mogavi, Joseph Tu, Leah Zhang-Kennedy, and Lennart E. Nacke. 2024. The Great AI Witch Hunt: Reviewers Perception and (Mis)Conception of Generative AI in Research Writing. <https://arxiv.org/abs/2407.12015>.
- [105] Sukjin Han. 2024. Mining Causality: AI-Assisted Search for Instrumental Variables. arXiv:2409.14202 [econ.EM] <https://arxiv.org/abs/2409.14202>
- [106] Ebtesam Al Haque, Chris Brown, Thomas D. LaToza, and Brittany Johnson. 2025. Towards Decoding Developer Cognition in the Age of AI Assistants. arXiv:2501.02684 [cs.HC] <https://arxiv.org/abs/2501.02684>
- [107] Gaole He, Patrick Hemmer, Michael Vössing, Max Schemmer, and Ujwal Gadiraju. 2025. Fine-Grained Appropriate Reliance: Human-AI Collaboration with a Multi-Step Transparent Decision Workflow for Complex Task Decomposition. arXiv:2501.10909 [cs.AI] <https://arxiv.org/abs/2501.10909>
- [108] Kaveen Hiniduma, Suren Byna, Jean Luca Bez, and Ravi Madduri. 2024. AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI. In *Proceedings of the 36th International Conference on Scientific and Statistical Database Management (SSDBM 2024)*. ACM, 1–12. doi:10.1145/3676288.3676296
- [109] HKUDS. 2025. AI-Researcher. <https://github.com/HKUDS/AI-Researcher>.
- [110] Brendan Hogan, Anmol Kabra, Felipe Siqueira Pacheco, Laura Greenstreet, Joshua Fan, Aaron Ferber, Marta Ummus, Alecsander Brito, Olivia Graham, Lillian Aoki, Drew Harvell, Alex Flecker, and Carla Gomes. 2024. AiSciVision: A Framework for Specializing Large Multimodal Models in Scientific Image Classification. arXiv:2410.21480 [cs.LG] <https://arxiv.org/abs/2410.21480>
- [111] Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiaowu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. <https://github.com/geekan/MetaGPT>. arXiv:2308.00352 [cs.AI] <https://arxiv.org/abs/2308.00352>
- [112] Hong Kong University Data Science Lab. 2024. Auto-Deep-Research. <https://github.com/HKUDS/Auto-Deep-Research>.
- [113] Betty Li Hou, Kejian Shi, Jason Phang, James Aung, Steven Adler, and Rosie Campbell. 2024. Large Language Models as Misleading Assistants in Conversation. arXiv:2407.11789 [cs.CL] <https://arxiv.org/abs/2407.11789>
- [114] Shulin Huang, Shirong Ma, Yinghui Li, Mengzuo Huang, Wuhe Zou, Weidong Zhang, and Hai-Tao Zheng. 2024. LatEval: An Interactive LLMs Evaluation Benchmark with Incomplete Information from Lateral Thinking Puzzles. <https://github.com/THUKElab/LatEval>. arXiv:2308.10855 [cs.CL] <https://arxiv.org/abs/2308.10855>
- [115] HuggingFace. 2025. smolagents: open_deep_research. https://github.com/huggingface/smolagents/tree/main/examples/open_deep_research.
- [116] Faria Huq, Abdus Samee, David Chuan-En Lin, Alice Xiaodi Tang, and Jeffrey P Bigham. 2025. NoTeeline: Supporting Real-Time, Personalized Notetaking with LLM-Enhanced Micronotes. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. ACM, 1064–1081. doi:10.1145/3708359.3712086
- [117] Kurando IIDA and Kenjiro MIMURA. 2024. CATER: Leveraging LLM to Pioneer a Multidimensional, Reference-Independent Paradigm in Translation Quality Evaluation. arXiv:2412.11261 [cs.CL] <https://arxiv.org/abs/2412.11261>

- [118] Seyed Mohammad Ali Jafari. 2024. Streamlining the Selection Phase of Systematic Literature Reviews (SLRs) Using AI-Enabled GPT-4 Assistant API. arXiv:2402.18582 [cs.DL] <https://arxiv.org/abs/2402.18582>
- [119] Rishab Jain and Aditya Jain. 2023. Generative AI in Writing Research Papers: A New Type of Algorithmic Bias and Uncertainty in Scholarly Work. arXiv:2312.10057 [cs.CY] <https://arxiv.org/abs/2312.10057>
- [120] Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. 2025. AIDE: AI-Driven Exploration in the Space of Code. arXiv:2502.13138 [cs.AI] <https://arxiv.org/abs/2502.13138>
- [121] Jina AI. 2025. node-DeepResearch. <https://github.com/jina-ai/node-DeepResearch>.
- [122] Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. 2025. DSbench: How Far Are Data Science Agents from Becoming Data Science Experts? <https://github.com/LiqiangJing/DSBench>. arXiv:2409.07703 [cs.AI] <https://arxiv.org/abs/2409.07703>
- [123] Nicola Jones. 2025. OpenAI's 'deep research' tool: is it useful for scientists? <https://www.nature.com/articles/d41586-025-00377-9>.
- [124] Vijay Joshi and Iver Band. 2024. Disrupting Test Development with AI Assistants: Building the Base of the Test Pyramid with Three AI Coding Assistants. (Oct. 2024). doi:10.36227/techrxiv.173014488.82191966/v1
- [125] Majeed Kazemitabaar, Jack Williams, Ian Drosos, Tovi Grossman, Austin Zachary Henley, Carina Negreanu, and Advait Sarkar. 2024. Improving Steering and Verification in AI-Assisted Data Analysis with Interactive Task Decomposition. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (UIST '24)*. ACM, 1–19. doi:10.1145/3654777.3676345
- [126] CTOL Editors Ken. 2025. Gemini Launches Deep Research on 2.5 Pro Aiming to Redefine AI-Powered Analysis with Strong Lead Over OpenAI. <https://www.ctol.digital/news/gemini-deep-research-launch-2-5-pro-vs-openai/>.
- [127] Antti Keurulainen, Isak Westerlund, Samuel Kaski, and Alexander Ilin. 2021. Learning to Assist Agents by Observing Them. arXiv:2110.01311 [cs.AI] <https://arxiv.org/abs/2110.01311>
- [128] Abdullah Khalili and Abdelhamid Bouchachia. 2022. Toward Building Science Discovery Machines. arXiv:2103.15551 [cs.AI] <https://arxiv.org/abs/2103.15551>
- [129] Stefan Kramer, Mattia Cerrato, Sašo Džeroski, and Ross King. 2023. Automated Scientific Discovery: From Equation Discovery to Autonomous Discovery Systems. arXiv:2305.02251 [cs.AI] <https://arxiv.org/abs/2305.02251>
- [130] Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névoul, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, and Iryna Gurevych. 2024. What Can Natural Language Processing Do for Peer Review? arXiv:2405.06563 [cs.CL] <https://arxiv.org/abs/2405.06563>
- [131] Martin Lance. 2024. open_deep_research. https://github.com/langchain-ai/open_deep_research.
- [132] Hao Lang, Fei Huang, and Yongbin Li. 2025. Debate Helps Weak-to-Strong Generalization. arXiv:2501.13124 [cs.CL] <https://arxiv.org/abs/2501.13124>
- [133] LangChain. 2025. How to think about agent frameworks. <https://blog.langchain.dev/how-to-think-about-agent-frameworks/>. <https://docs.google.com/spreadsheets/d/1B37VxTBuGLEtSPVWtz7UMsCdtXrqV5hCjWkbHN8tfAo/>
- [134] langChain AI. 2024. LangGraph. <https://github.com/langchain-ai/langgraph>.
- [135] Andrew Laverick, Kristen Surrao, Inigo Zubeldia, Boris Bolliet, Miles Cranmer, Antony Lewis, Blake Sherwin, and Julien Lesgourgues. 2024. Multi-Agent System for Cosmological Parameter Analysis. arXiv:2412.00431 [astro-ph.IM] <https://arxiv.org/abs/2412.00431>
- [136] Eunhae Lee. 2024. Towards Ethical Personal AI Applications: Practical Considerations for AI Assistants with Long-Term Memory. arXiv:2409.11192 [cs.CY] <https://arxiv.org/abs/2409.11192>
- [137] Yuho Lee, Taewon Yun, Jason Cai, Hang Su, and Hwanjun Song. 2024. UniSumEval: Towards Unified, Fine-Grained, Multi-Dimensional Summarization Evaluation for LLMs. <https://github.com/DISL-Lab/UniSumEval-v1.0>. arXiv:2409.19898 [cs.CL] <https://arxiv.org/abs/2409.19898>
- [138] Letta-AI. 2023. Letta. <https://github.com/letta-ai/letta>.
- [139] Kyla Levin, Nicolas van Kempen, Emery D. Berger, and Stephen N. Freund. 2025. ChatDBG: An AI-Powered Debugging Assistant. arXiv:2403.16354 [cs.SE] <https://arxiv.org/abs/2403.16354>
- [140] James R. Lewis. 2018. The System Usability Scale: Past, Present, and Future. *International Journal of Human-Computer Interaction* 34, 7 (2018), 577–590. doi:10.1080/10447318.2018.1455307 arXiv:https://doi.org/10.1080/10447318.2018.1455307
- [141] Belinda Z. Li, Been Kim, and Zi Wang. 2025. QuestBench: Can LLMs ask the right question to acquire information in reasoning tasks? arXiv:2503.22674 [cs.AI] <https://arxiv.org/abs/2503.22674>
- [142] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. arXiv:2303.17760 [cs.AI] <https://arxiv.org/abs/2303.17760>

- [143] Jiachen Li, Xiwen Li, Justin Steinberg, Akshat Choubey, Bingsheng Yao, Xuhai Xu, Dakuo Wang, Elizabeth Mynatt, and Varun Mishra. 2025. Vital Insight: Assisting Experts' Context-Driven Sensemaking of Multi-modal Personal Tracking Data Using Visualization and Human-In-The-Loop LLM Agents. arXiv:2410.14879 [cs.HC] <https://arxiv.org/abs/2410.14879>
- [144] Xuefeng Li, Haoyang Zou, and Pengfei Liu. 2025. TORL: Scaling Tool-Integrated RL. <https://github.com/GAIR-NLP/ToRL>. <https://arxiv.org/pdf/2503.23383>
- [145] Yuan Li, Yixuan Zhang, and Lichao Sun. 2023. MetaAgents: Simulating Interactions of Human Behaviors for LLM-based Task-oriented Coordination via Collaborative Generative Agents. arXiv:2310.06500 [cs.AI] <https://arxiv.org/abs/2310.06500>
- [146] Zhuoyan Li, Chen Liang, Jing Peng, and Ming Yin. 2024. How Does the Disclosure of AI Assistance Affect the Perceptions of Writing? arXiv:2410.04545 [cs.CL] <https://arxiv.org/abs/2410.04545>
- [147] Zekun Li, Xianjun Yang, Kyuri Choi, Wanrong Zhu, Ryan Hsieh, HyeonJung Kim, Jin Hyuk Lim, Sungyoung Ji, Byungju Lee, Xifeng Yan, Linda Ruth Petzold, Stephen D. Wilson, Woosang Lim, and William Yang Wang. 2025. MMSci: A Dataset for Graduate-Level Multi-Discipline Multimodal Scientific Understanding. <https://github.com/Leezekun/MMSci>. arXiv:2407.04903 [cs.CL] <https://arxiv.org/abs/2407.04903>
- [148] Jenny T. Liang, Chenyang Yang, and Brad A. Myers. 2023. A Large-Scale Survey on the Usability of AI Programming Assistants: Successes and Challenges. arXiv:2303.17125 [cs.SE] <https://arxiv.org/abs/2303.17125>
- [149] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. arXiv:2211.09110 [cs.CL] <https://arxiv.org/abs/2211.09110>
- [150] Jialiang Lin, Jiaxin Song, Zhangping Zhou, Yidong Chen, and Xiaodong Shi. 2023. Automated scholarly paper review: Concepts, technologies, and challenges. *Information Fusion* 98 (Oct. 2023), 101830. doi:10.1016/j.inffus.2023.101830
- [151] Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958 [cs.CL] <https://arxiv.org/abs/2109.07958>
- [152] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, Lei Zhang, Jianfeng Gao, and Chunyuan Li. 2023. LLaVA-Plus: Learning to Use Tools for Creating Multimodal Agents. arXiv:2311.05437 [cs.CV] <https://arxiv.org/abs/2311.05437>
- [153] Xiao Liu, Bo Qin, Dongzhu Liang, Guang Dong, Hanyu Lai, Hanchen Zhang, Hanlin Zhao, Iat Long Iong, Jiadai Sun, Jiaqi Wang, Junjie Gao, Junjun Shan, Kangning Liu, Shudan Zhang, Shuntian Yao, Siyi Cheng, Wentao Yao, Wenyi Zhao, Xinghan Liu, Xinyi Liu, Xinying Chen, Xinyue Yang, Yang Yang, Yifan Xu, Yu Yang, Yujia Wang, Yulin Xu, Zehan Qi, Yuxiao Dong, and Jie Tang. 2024. AutoGLM: Autonomous Foundation Agents for GUIs. arXiv:2411.00820 [cs.HC] <https://arxiv.org/abs/2411.00820>
- [154] Xiao Liu, Bo Qin, Dongzhu Liang, Guang Dong, Hanyu Lai, Hanchen Zhang, Hanlin Zhao, Iat Long Iong, Jiadai Sun, Jiaqi Wang, Junjie Gao, Junjun Shan, Kangning Liu, Shudan Zhang, Shuntian Yao, Siyi Cheng, Wentao Yao, Wenyi Zhao, Xinghan Liu, Xinyi Liu, Xinying Chen, Xinyue Yang, Yang Yang, Yifan Xu, Yu Yang, Yujia Wang, Yulin Xu, Zehan Qi, Yuxiao Dong, and Jie Tang. 2024. AutoGLM: Autonomous Foundation Agents for GUIs. arXiv:2411.00820 [cs.HC] <https://arxiv.org/abs/2411.00820>
- [155] Zijun Liu, Kaiming Liu, Yiqi Zhu, Xuanyu Lei, Zonghan Yang, Zhenhe Zhang, Peng Li, and Yang Liu. 2024. AIGS: Generating Science from AI-Powered Automated Falsification. arXiv:2411.11910 [cs.LG] <https://arxiv.org/abs/2411.11910>
- [156] Zhiwei Liu, Weiran Yao, Jianguo Zhang, Le Xue, Shelby Heinecke, Rithesh Murthy, Yihao Feng, Zeyuan Chen, Juan Carlos Niebles, Devansh Arpit, Ran Xu, Phil Mui, Huan Wang, Caiming Xiong, and Silvio Savarese. 2023. BOLAA: Benchmarking and Orchestrating LLM-augmented Autonomous Agents. <https://github.com/salesforce/BOLAA>. arXiv:2308.05960 [cs.AI] <https://arxiv.org/abs/2308.05960>
- [157] Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. 2025. AAAR-1.0: Assessing AI's Potential to Assist Research. <https://renzelou.github.io/AAAR-1.0/>. arXiv:2410.22394 [cs.CL] <https://arxiv.org/abs/2410.22394>
- [158] Cong Lu, Shengran Hu, and Jeff Clune. 2025. Automated Capability Discovery via Model Self-Exploration. arXiv:2502.07577 [cs.LG] <https://arxiv.org/abs/2502.07577>

- [159] 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:2408.06292 [cs.AI] <https://arxiv.org/abs/2408.06292>
- [160] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. <https://scienceqa.github.io/>. arXiv:2209.09513 [cs.CL] <https://arxiv.org/abs/2209.09513>
- [161] Chandra Maddila, Negar Ghorbani, Kosay Jabre, Vijayaraghavan Murali, Edwin Kim, Parth Thakkar, Nikolay Pavlovich Laptev, Olivia Harman, Diana Hsu, Rui Abreu, and Peter C. Rigby. 2024. AI-Assisted SQL Authoring at Industry Scale. arXiv:2407.13280 [cs.SE] <https://arxiv.org/abs/2407.13280>
- [162] Srijoni Majumdar, Edith Elkind, and Evangelos Pournaras. 2025. Generative AI Voting: Fair Collective Choice is Resilient to LLM Biases and Inconsistencies. arXiv:2406.11871 [cs.AI] <https://arxiv.org/abs/2406.11871>
- [163] Dung Nguyen Manh, Thang Phan Chau, Nam Le Hai, Thong T. Doan, Nam V. Nguyen, Quang Pham, and Nghi D. Q. Bui. 2025. CodeMMLU: A Multi-Task Benchmark for Assessing Code Understanding & Reasoning Capabilities of CodeLLMs. <https://github.com/FSoft-AI4Code/CodeMMLU>. arXiv:2410.01999 [cs.SE] <https://arxiv.org/abs/2410.01999>
- [164] Manus. 2025. Manus. <https://manus.im/>.
- [165] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2024. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. arXiv:2310.06213 [cs.CL] <https://arxiv.org/abs/2310.06213>
- [166] David M. Markowitz. 2024. From Complexity to Clarity: How AI Enhances Perceptions of Scientists and the Public's Understanding of Science. arXiv:2405.00706 [cs.CL] <https://arxiv.org/abs/2405.00706>
- [167] Jonathan Mast. 2025. ChatGPT's Deep Research vs. Google's Gemini 1.5 Pro with Deep Research: A Detailed Comparison. <https://whitebeardstrategies.com/ai-prompt-engineering/chatgpts-deep-research-vs-googles-gemini-1-5-pro-with-deep-research-a-detailed-comparison/>.
- [168] Mastra-AI. 2025. Mastra. <https://github.com/mastra-ai/mastra>.
- [169] Shray Mathur, Noah van der Vleuten, Kevin Yager, and Esther Tsai. 2024. VISION: A Modular AI Assistant for Natural Human-Instrument Interaction at Scientific User Facilities. arXiv:2412.18161 [cs.AI] <https://arxiv.org/abs/2412.18161>
- [170] Gianmarco Mengaldo. 2025. Explain the Black Box for the Sake of Science: the Scientific Method in the Era of Generative Artificial Intelligence. arXiv:2406.10557 [cs.AI] <https://arxiv.org/abs/2406.10557>
- [171] MGX Technologies. 2025. *MGX.dev*. <https://mgx.dev>
- [172] Gregoire Mialon, Clementine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. GAIA: A Benchmark for General AI Assistants. <https://huggingface.co/gaia-benchmark>. <https://arxiv.org/pdf/2311.12983>
- [173] Microsoft. 2023. Microsoft Copilot. <https://www.microsoft.com/en-us/microsoft-copilot/organizations>.
- [174] Microsoft. 2023. Semantic-kernel. <https://github.com/microsoft/semantic-kernel>.
- [175] mirayayerdem. 2022. Github-Copilot-Amazon-Whisperer-ChatGPT. <https://github.com/mirayayerdem/Github-Copilot-Amazon-Whisperer-ChatGPT>.
- [176] Mlc-ai. 2023. web-llm. <https://github.com/mlc-ai/web-llm>.
- [177] ModelTC. 2025. lightllm. <https://github.com/ModelTC/lightllm>.
- [178] Devam Mondal and Atharva Inamdar. 2024. SeqMate: A Novel Large Language Model Pipeline for Automating RNA Sequencing. arXiv:2407.03381 [q-bio.GN] <https://arxiv.org/abs/2407.03381>
- [179] Peya Mowar, Yi-Hao Peng, Jason Wu, Aaron Steinfeld, and Jeffrey P. Bigham. 2025. CodeA11y: Making AI Coding Assistants Useful for Accessible Web Development. arXiv:2502.10884 [cs.HC] <https://arxiv.org/abs/2502.10884>
- [180] Dilxat Muhtar, Zhenshi Li, Feng Gu, Xueliang Zhang, and Pengfeng Xiao. 2024. LHRS-Bot: Empowering Remote Sensing with VGI-Enhanced Large Multimodal Language Model. <https://github.com/NJU-LHRS/LHRS-Bot>. arXiv:2402.02544 [cs.CV] <https://arxiv.org/abs/2402.02544>
- [181] Manisha Mukherjee, Sungchul Kim, Xiang Chen, Dan Luo, Tong Yu, and Tung Mai. 2025. From Documents to Dialogue: Building KG-RAG Enhanced AI Assistants. arXiv:2502.15237 [cs.IR] <https://arxiv.org/abs/2502.15237>
- [182] Sheshera Mysore, Mahmood Jasim, Haoru Song, Sarah Akbar, Andre Kenneth Chase Randall, and Narges Mahyar. 2023. How Data Scientists Review the Scholarly Literature. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval (CHIIR '23)*. ACM, 137–152. doi:10.1145/3576840.3578309
- [183] n8n. 2023. n8n. <https://github.com/n8n-io/n8n>.
- [184] Nanobrowser Team. 2024. Nanobrowser. <https://github.com/nanobrowser/nanobrowser>.
- [185] Nathalia Nascimento, Everton Guimaraes, Sai Sanjna Chintakunta, and Santhosh Anitha Boominathan. 2024. LLM4DS: Evaluating Large Language Models for Data Science Code Generation. <https://github.com/DataForScience/LLM4DS>. arXiv:2411.11908 [cs.SE] <https://arxiv.org/abs/2411.11908>
- [186] Khanh Nghiem, Anh Minh Nguyen, and Nghi D. Q. Bui. 2024. Envisioning the Next-Generation AI Coding Assistants: Insights & Proposals. arXiv:2403.14592 [cs.SE] <https://arxiv.org/abs/2403.14592>

- [187] Alex Nguyen, Zilong Wang, Jingbo Shang, and Dheeraj Mekala. 2024. DOCMASTER: A Unified Platform for Annotation, Training, & Inference in Document Question-Answering. arXiv:2404.00439 [cs.CL] <https://arxiv.org/abs/2404.00439>
- [188] Kien X. Nguyen, Fengchun Qiao, Arthur Trembanis, and Xi Peng. 2024. SeafloorAI: A Large-scale Vision-Language Dataset for Seafloor Geological Survey. <https://github.com/deep-real/SeafloorAI>. arXiv:2411.00172 [cs.CV] <https://arxiv.org/abs/2411.00172>
- [189] Ziqi Ni, Yahao Li, Kaijia Hu, Kunyuan Han, Ming Xu, Xingyu Chen, Fengqi Liu, Yicong Ye, and Shuxin Bai. 2024. MatPilot: an LLM-enabled AI Materials Scientist under the Framework of Human-Machine Collaboration. arXiv:2411.08063 [physics.soc-ph] <https://arxiv.org/abs/2411.08063>
- [190] Alexander Novikov, Ngăn Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. 2025. AlphaEvolve: A coding agent for scientific and algorithmic discovery. <https://deepmind.google/discover/blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/>. <https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/alphaevolve-a-gemini-powered-coding-agent-for-designing-advanced-algorithms/AlphaEvolve.pdf>
- [191] Koji Ochiai, Yuya Tahara-Arai, Akari Kato, Kazunari Kaizu, Hirokazu Kariyazaki, Makoto Umeno, Koichi Takahashi, Genki N. Kanda, and Haruka Ozaki. 2025. Automating Care by Self-maintainability for Full Laboratory Automation. arXiv:2501.05789 [q-bio.QM] <https://arxiv.org/abs/2501.05789>
- [192] Ollama. 2023. Ollama. <https://github.com/ollama/ollama>.
- [193] Open Manus Team. 2025. OpenManus. <https://github.com/mannaandpoem/OpenManus>.
- [194] OpenAI. 2025. codex. <https://github.com/openai/codex>.
- [195] OpenAI. 2025. Compare models - OpenAI API. <https://platform.openai.com/docs/models/compare?model=o3>.
- [196] OpenAI. 2025. Deep Research System Card. <https://cdn.openai.com/deep-research-system-card.pdf>.
- [197] OpenAI. 2025. Introducing Deep Research. <https://openai.com/index/introducing-deep-research/>.
- [198] OpenAI. 2025. Introducing OpenAI o3 and o4-mini. <https://openai.com/index/introducing-o3-and-o4-mini/>.
- [199] OpenAI. 2025. OpenAI Agents SDK. <https://github.com/openai/openai-agents-python>.
- [200] OpenAI. 2025. OpenAI o3 and o4-mini System Card. <https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf>.
- [201] OpenAI. 2025. Thinking with images. <https://openai.com/index/thinking-with-images/>.
- [202] OpenBMB. 2023. XAgent. <https://github.com/OpenBMB/XAgent>.
- [203] Orkes. 2022. Orkes. <https://orkes.io/use-cases/agentic-workflows>.
- [204] Takauki Osogami. 2025. Position: AI agents should be regulated based on autonomous action sequences. arXiv:2503.04750 [cs.CY] <https://arxiv.org/abs/2503.04750>
- [205] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] <https://arxiv.org/abs/2203.02155>
- [206] Md Sultanul Islam Ovi, Nafisa Anjum, Tasmina Haque Bithe, Md. Mahabubur Rahman, and Mst. Shahnaj Akter Smrity. 2024. Benchmarking ChatGPT, Codeium, and GitHub Copilot: A Comparative Study of AI-Driven Programming and Debugging Assistants. arXiv:2409.19922 [cs.SE] <https://arxiv.org/abs/2409.19922>
- [207] Carlos Alves Pereira, Tanay Komarlu, and Wael Mobeirek. 2023. The Future of AI-Assisted Writing. arXiv:2306.16641 [cs.HC] <https://arxiv.org/abs/2306.16641>
- [208] Mike Perkins and Jasper Roe. 2024. Generative AI Tools in Academic Research: Applications and Implications for Qualitative and Quantitative Research Methodologies. arXiv:2408.06872 [cs.HC] <https://arxiv.org/abs/2408.06872>
- [209] Perplexity. 2025. Introducing Perplexity Deep Research. <https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research>.
- [210] Perplexity. 2025. Sonar by Perplexity. <https://docs.perplexity.ai/guides/model-cards#research-models>.
- [211] Tomas Petricek, Gerrit J. J. van den Burg, Alfredo Nazábal, Taha Ceritli, Ernesto Jiménez-Ruiz, and Christopher K. I. Williams. 2022. AI Assistants: A Framework for Semi-Automated Data Wrangling. arXiv:2211.00192 [cs.DB] <https://arxiv.org/abs/2211.00192>
- [212] 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:2501.14249 [cs.LG] <https://arxiv.org/abs/2501.14249>
- [213] Evangelos Pournaras. 2023. Science in the Era of ChatGPT, Large Language Models and Generative AI: Challenges for Research Ethics and How to Respond. arXiv:2305.15299 [cs.CY] <https://arxiv.org/abs/2305.15299>
- [214] Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghaddam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Ragnarök: A Reusable RAG Framework and Baselines for TREC 2024 Retrieval-Augmented

- Generation Track. arXiv:2406.16828 [cs.IR] <https://arxiv.org/abs/2406.16828>
- [215] James Prather, Juho Leinonen, Natalie Kiesler, Jamie Gorson Benario, Sam Lau, Stephen MacNeil, Narges Norouzi, Simone Opel, Vee Pettit, Leo Porter, Brent N. Reeves, Jaromir Savelka, David H. Smith IV, Sven Strickroth, and Daniel Zingaro. 2024. Beyond the Hype: A Comprehensive Review of Current Trends in Generative AI Research, Teaching Practices, and Tools. arXiv:2412.14732 [cs.CY] <https://arxiv.org/abs/2412.14732>
 - [216] Pydantic. 2024. Pydantic-AI. <https://github.com/pydantic/pydantic-ai>.
 - [217] Pythagora-io. 2024. gpt-pilot. <https://github.com/Pythagora-io/gpt-pilot>.
 - [218] Jingyuan Qi, Zian Jia, Minqian Liu, Wangzhi Zhan, Junkai Zhang, Xiaofei Wen, Jingru Gan, Jianpeng Chen, Qin Liu, Mingyu Derek Ma, Bangzheng Li, Haohui Wang, Adithya Kulkarni, Muhao Chen, Dawei Zhou, Ling Li, Wei Wang, and Lifu Huang. 2024. MetaScientist: A Human-AI Synergistic Framework for Automated Mechanical Metamaterial Design. arXiv:2412.16270 [cs.AI] <https://arxiv.org/abs/2412.16270>
 - [219] Laryn Qi, J. D. Zamfirescu-Pereira, Taehan Kim, Björn Hartmann, John DeNero, and Narges Norouzi. 2024. A Knowledge-Component-Based Methodology for Evaluating AI Assistants. arXiv:2406.05603 [cs.CY] <https://arxiv.org/abs/2406.05603>
 - [220] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative Agents for Software Development. <https://github.com/OpenBMB/ChatDev>. <https://aclanthology.org/2024.acl-long.810.pdf>
 - [221] Shuofei Qiao, Runnan Fang, Zhisong Qiu, Xiaobin Wang, Ningyu Zhang, Yong Jiang, Pengjun Xie, Fei Huang, and Huajun Chen. 2025. Benchmarking Agentic Workflow Generation. <https://github.com/zjunlp/WorBench>. arXiv:2410.07869 [cs.CL] <https://arxiv.org/abs/2410.07869>
 - [222] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. arXiv:2307.16789 [cs.AI] <https://arxiv.org/abs/2307.16789>
 - [223] Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. arXiv:2307.16789 [cs.AI] <https://arxiv.org/abs/2307.16789>
 - [224] Qwen LM. 2024. Qwen-Agent. <https://github.com/QwenLM/Qwen-Agent>.
 - [225] Joaquin Ramirez-Medina, Mohammadmehdi Ataei, and Alidad Amirfazli. 2025. Accelerating Scientific Research Through a Multi-LLM Framework. arXiv:2502.07960 [physics.app-ph] <https://arxiv.org/abs/2502.07960>
 - [226] Ruchit Rawal, Victor-Alexandru Pădurean, Sven Apel, Adish Singla, and Mariya Toneva. 2024. Hints Help Finding and Fixing Bugs Differently in Python and Text-based Program Representations. arXiv:2412.12471 [cs.SE] <https://arxiv.org/abs/2412.12471>
 - [227] Runtao Ren, Jian Ma, and Jianxi Luo. 2025. Large language model for patent concept generation. *Advanced Engineering Informatics* 65 (May 2025), 103301. doi:10.1016/j.aei.2025.103301
 - [228] ResearchRabbit. 2025. ResearchRabbit. <https://www.researchrabbit.ai/>.
 - [229] Restate. 2024. Restate. <https://restate.dev/>.
 - [230] reworkd. 2023. AgentGPT. <https://github.com/reworkd/AgentGPT>.
 - [231] Filippo Ricca, Alessandro Marchetto, and Andrea Stocco. 2025. A Multi-Year Grey Literature Review on AI-assisted Test Automation. <https://arxiv.org/pdf/2408.06224>.
 - [232] Nathalie Riche, Anna Offenwanger, Frederic Gmeiner, David Brown, Hugo Romat, Michel Pahud, Nicolai Marquardt, Kori Inkpen, and Ken Hinckley. 2025. AI-Instruments: Embodying Prompts as Instruments to Abstract & Reflect Graphical Interface Commands as General-Purpose Tools. <https://arxiv.org/abs/2502.18736>.
 - [233] Anthony Cintron Roman, Jennifer Wortman Vaughan, Valerie See, Steph Ballard, Jehu Torres, Caleb Robinson, and Juan M. Lavista Ferres. 2024. Open Datasheets: Machine-readable Documentation for Open Datasets and Responsible AI Assessments. arXiv:2312.06153 [cs.LG] <https://arxiv.org/abs/2312.06153>
 - [234] Kaushik Roy, Vedant Khandelwal, Harshul Surana, Valerie Vera, Amit Sheth, and Heather Heckman. 2023. GEAR-Up: Generative AI and External Knowledge-based Retrieval Upgrading Scholarly Article Searches for Systematic Reviews. arXiv:2312.09948 [cs.IR] <https://arxiv.org/abs/2312.09948>
 - [235] Run-llama. 2023. LlamaIndex. https://github.com/run-llama/llama_index.
 - [236] Sergey V Samsonau, Aziza Kurbonova, Lu Jiang, Hazem Lashen, Jiamu Bai, Theresa Merchant, Ruoxi Wang, Laiba Mehnaz, Zecheng Wang, and Ishita Patil. 2024. Artificial Intelligence for Scientific Research: Authentic Research Education Framework. arXiv:2210.08966 [cs.CY] <https://arxiv.org/abs/2210.08966>
 - [237] SamuelSchmidgall. 2025. AgentLaboratory. <https://github.com/SamuelSchmidgall/AgentLaboratory>.

- [238] Thomas Sandholm, Sarah Dong, Sayandev Mukherjee, John Feland, and Bernardo A. Huberman. 2024. Semantic Navigation for AI-assisted Ideation. arXiv:2411.03575 [cs.HC] <https://arxiv.org/abs/2411.03575>
- [239] Lindsay Sanneman and Julie Shah. 2021. Explaining Reward Functions to Humans for Better Human-Robot Collaboration. arXiv:2110.04192 [cs.RO] <https://arxiv.org/abs/2110.04192>
- [240] 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:2501.04227 [cs.HC] <https://arxiv.org/abs/2501.04227>
- [241] Martijn J. Schuemie, M. Soledad Cepeda, Marc A. Suchard, Jianxiao Yang, Yuxi Tian, Alejandro Schuler, Patrick B. Ryan, David Madigan, and George Hripcsak. 2020. How Confident Are We About Observational Findings in Healthcare: A Benchmark Study. *Harvard Data Science Review* 2, 1 (2020). doi:10.1162/99608f92.147cc28e
- [242] Scispace. 2024. Scispace. <https://scispace.com/>.
- [243] Scite. 2025. Scite. <https://scite.ai/>.
- [244] Agnia Sergeyuk, Yaroslav Golubev, Timofey Bryksin, and Iftekhhar Ahmed. 2025. Using AI-based coding assistants in practice: State of affairs, perceptions, and ways forward. *Information and Software Technology* 178 (Feb. 2025), 107610. doi:10.1016/j.infsof.2024.107610
- [245] Mahsa Shamsabadi and Jennifer D'Souza. 2024. A FAIR and Free Prompt-based Research Assistant. arXiv:2405.14601 [cs.CL] <https://arxiv.org/abs/2405.14601>
- [246] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. <https://github.com/microsoft/JARVIS>. <https://arxiv.org/pdf/2303.17580>
- [247] Zejiang Shen, Tal August, Pao Siangliulue, Kyle Lo, Jonathan Bragg, Jeff Hammerbacher, Doug Downey, Joseph Chee Chang, and David Sontag. 2023. Beyond Summarization: Designing AI Support for Real-World Expository Writing Tasks. arXiv:2304.02623 [cs.CL] <https://arxiv.org/abs/2304.02623>
- [248] Shuming Shi, Enbo Zhao, Duyu Tang, Yan Wang, Piji Li, Wei Bi, Haiyun Jiang, Guoping Huang, Leyang Cui, Xinting Huang, Cong Zhou, Yong Dai, and Dongyang Ma. 2022. Effidit: Your AI Writing Assistant. arXiv:2208.01815 [cs.CL] <https://arxiv.org/abs/2208.01815>
- [249] Michael Shumer. 2025. OpenDeepResearcher. <https://github.com/mshumer/OpenDeepResearcher>.
- [250] Significant-Gravitas. 2023. AutoGPT. <https://github.com/Significant-Gravitas/AutoGPT>.
- [251] David Silver and Richard Sutton. 2025. Welcome to the Era of Experience. <https://storage.googleapis.com/deepmind-media/Era-of-Experience%20/The%20Era%20of%20Experience%20Paper.pdf>.
- [252] Auste Simkute, Ewa Luger, Michael Evans, and Rhianne Jones. 2024. "It is there, and you need it, so why do you not use it?" Achieving better adoption of AI systems by domain experts, in the case study of natural science research. arXiv:2403.16895 [cs.HC] <https://arxiv.org/abs/2403.16895>
- [253] Michael Skarlinski, Tyler Nadolski, James Braza, Remo Storni, Mayk Caldas, Ludovico Mitchener, Michaela Hinks, Andrew White, and Sam Rodrigues. 2025. FutureHouse Platform: Superintelligent AI Agents for Scientific Discovery. <https://www.futurehouse.org/research-announcements/launching-futurehouse-platform-ai-agents>.
- [254] Xinyi Song, Kexin Xie, Lina Lee, Ruizhe Chen, Jared M. Clark, Hao He, Haoran He, Jie Min, Xinlei Zhang, Simin Zheng, Zhiyang Zhang, Xinwei Deng, and Yili Hong. 2025. Performance Evaluation of Large Language Models in Statistical Programming. arXiv:2502.13117 [stat.AP] <https://arxiv.org/abs/2502.13117>
- [255] Jamshid Sourati and James Evans. 2021. Accelerating science with human versus alien artificial intelligences. arXiv:2104.05188 [cs.AI] <https://arxiv.org/abs/2104.05188>
- [256] Jamshid Sourati and James Evans. 2023. Accelerating science with human-aware artificial intelligence. arXiv:2306.01495 [cs.AI] <https://arxiv.org/abs/2306.01495>
- [257] StanfordNLP. 2024. DSPy. <https://github.com/stanfordnlp/dspy>.
- [258] Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. 2025. Many Heads Are Better Than One: Improved Scientific Idea Generation by A LLM-Based Multi-Agent System. arXiv:2410.09403 [cs.AI] <https://arxiv.org/abs/2410.09403>
- [259] Ltd. Suzhou Yuling Artificial Intelligence Technology Co. 2023. Dify: Open-source LLM Application Development Platform. <https://dify.ai/>.
- [260] Xin Tan, Xiao Long, Xianjun Ni, Yinghao Zhu, Jing Jiang, and Li Zhang. 2024. How far are AI-powered programming assistants from meeting developers' needs? arXiv:2404.12000 [cs.SE] <https://arxiv.org/abs/2404.12000>
- [261] Brian Tang and Kang G. Shin. 2024. Steward: Natural Language Web Automation. arXiv:2409.15441 [cs.AI] <https://arxiv.org/abs/2409.15441>
- [262] Jiabin Tang, Tianyu Fan, and Chao Huang. 2025. AutoAgent: A Fully-Automated and Zero-Code Framework for LLM Agents. arXiv:2502.05957 [cs.AI] <https://arxiv.org/abs/2502.05957>
- [263] Yan Tang. 2025. deep_research_agent. https://github.com/grapeot/deep_research_agent.

- [264] Tadahiro Taniguchi, Shiro Takagi, Jun Otsuka, Yusuke Hayashi, and Hiro Taiyo Hamada. 2024. Collective Predictive Coding as Model of Science: Formalizing Scientific Activities Towards Generative Science. arXiv:2409.00102 [physics.soc-ph] <https://arxiv.org/abs/2409.00102>
- [265] Temporalio. 2020. Temporal. <https://github.com/temporalio/temporal>.
- [266] Enkeleda Thaqi, Mohamed Omar Mantawy, and Enkelejda Kasneci. 2024. SARA: Smart AI Reading Assistant for Reading Comprehension. In *Proceedings of the 2024 Symposium on Eye Tracking Research and Applications (ETRA '24)*. ACM, 1–3. doi:10.1145/3649902.3655661
- [267] TheBlewish. 2024. Automated-AI-Web-Researcher-Ollama. <https://github.com/TheBlewish/Automated-AI-Web-Researcher-Ollama>.
- [268] Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu Huerta, and Hao Peng. 2024. SciCode: A Research Coding Benchmark Curated by Scientists. <https://scicode-bench.github.io/>. arXiv:2407.13168 [cs.AI] <https://arxiv.org/abs/2407.13168>
- [269] Ievgeniia A. Tiukova, Daniel Brunnsåker, Erik Y. Bjurström, Alexander H. Gower, Filip Kronström, Gabriel K. Reder, Ronald S. Reiserer, Konstantin Korovin, Larisa B. Soldatova, John P. Wikswo, and Ross D. King. 2024. Genesis: Towards the Automation of Systems Biology Research. arXiv:2408.10689 [cs.AI] <https://arxiv.org/abs/2408.10689>
- [270] Irina Tolstykh, Aleksandra Tsybina, Sergey Yakubson, Aleksandr Gordeev, Vladimir Dokholyan, and Maksim Kuprashevich. 2024. GigaCheck: Detecting LLM-generated Content. arXiv:2410.23728 [cs.CL] <https://arxiv.org/abs/2410.23728>
- [271] Benjamin Towle and Ke Zhou. 2024. Enhancing AI Assisted Writing with One-Shot Implicit Negative Feedback. arXiv:2410.11009 [cs.CL] <https://arxiv.org/abs/2410.11009>
- [272] Thanh-Dat Truong, Hoang-Quan Nguyen, Xuan-Bac Nguyen, Ashley Dowling, Xin Li, and Khoa Luu. 2025. Insect-Foundation: A Foundation Model and Large Multimodal Dataset for Vision-Language Insect Understanding. <https://uark-cviu.github.io/projects/insect-foundation/>. arXiv:2502.09906 [cs.CV] <https://arxiv.org/abs/2502.09906>
- [273] Joseph Tu, Hilda Hadan, Derrick M. Wang, Sabrina A Sgandurra, Reza Hadi Mogavi, and Lennart E. Nacke. 2024. Augmenting the Author: Exploring the Potential of AI Collaboration in Academic Writing. arXiv:2404.16071 [cs.HC] <https://arxiv.org/abs/2404.16071>
- [274] Xinming Tu, James Zou, Weijie J. Su, and Linjun Zhang. 2023. What Should Data Science Education Do with Large Language Models? arXiv:2307.02792 [cs.CY] <https://arxiv.org/abs/2307.02792>
- [275] Michele Tufano, Anisha Agarwal, Jinu Jang, Roshanak Zilouchian Moghaddam, and Neel Sundaresan. 2024. AutoDev: Automated AI-Driven Development. arXiv:2403.08299 [cs.SE] <https://arxiv.org/abs/2403.08299>
- [276] Aleksei Turobov, Diane Coyle, and Verity Harding. 2024. Using ChatGPT for Thematic Analysis. arXiv:2405.08828 [cs.HC] <https://arxiv.org/abs/2405.08828>
- [277] Rasmus Ulfesnes, Nils Brede Moe, Viktoria Stray, and Marianne Skarpen. 2024. Transforming Software Development with Generative AI: Empirical Insights on Collaboration and Workflow. arXiv:2405.01543 [cs.SE] <https://arxiv.org/abs/2405.01543>
- [278] Stanford University. 2025. STORM. <https://storm.genie.stanford.edu/>.
- [279] Edward Vendrow, Omiros Pantazis, Alexander Shepard, Gabriel Brostow, Kate E. Jones, Oisín Mac Aodha, Sara Beery, and Grant Van Horn. 2024. INQUIRE: A Natural World Text-to-Image Retrieval Benchmark. <https://inquire-benchmark.github.io/>. arXiv:2411.02537 [cs.CV] <https://arxiv.org/abs/2411.02537>
- [280] Vercel. 2020. Vercel. <https://vercel.com/>.
- [281] Vllm-project. 2023. vllm. <https://github.com/vllm-project/vllm>.
- [282] Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Parsa Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling Downstream Gender Bias from Large Language Models: A Study on AI Educational Writing Assistance. arXiv:2311.03311 [cs.CL] <https://arxiv.org/abs/2311.03311>
- [283] April Yi Wang, Dakuo Wang, Jaimie Drozdal, Michael Muller, Soya Park, Justin D. Weisz, Xuye Liu, Lingfei Wu, and Casey Dugan. 2022. Documentation Matters: Human-Centered AI System to Assist Data Science Code Documentation in Computational Notebooks. *ACM Transactions on Computer-Human Interaction* 29, 2 (Jan. 2022), 1–33. doi:10.1145/3489465
- [284] Dakuo Wang, Q. Vera Liao, Yunfeng Zhang, Udayan Khurana, Horst Samulowitz, Soya Park, Michael Muller, and Lisa Amini. 2021. How Much Automation Does a Data Scientist Want? arXiv:2101.03970 [cs.LG] <https://arxiv.org/abs/2101.03970>
- [285] Suyuan Wang, Xueqian Yin, Menghao Wang, Ruofeng Guo, and Kai Nan. 2024. EvoPat: A Multi-LLM-based Patents Summarization and Analysis Agent. arXiv:2412.18100 [cs.DL] <https://arxiv.org/abs/2412.18100>
- [286] Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, Yibin Liu, Jialong Wu, Shengwei Ding, Long Li, Zhiwei Huang, Xinle Deng, Teng Yu, Gangan Ma, Han

- Xiao, Zixin Chen, Danjun Xiang, Yunxia Wang, Yuanyuan Zhu, Yi Xiao, Jing Wang, Yiru Wang, Siran Ding, Jiayang Huang, Jiayi Xu, Yilihamu Tayier, Zhenyu Hu, Yuan Gao, Chengfeng Zheng, Yueshu Ye, Yihang Li, Lei Wan, Xinyue Jiang, Yujie Wang, Siyu Cheng, Zhule Song, Xiangru Tang, Xiaohua Xu, Ningyu Zhang, Huaqun Chen, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Weaver: Foundation Models for Creative Writing. arXiv:2401.17268 [cs.CL] <https://arxiv.org/abs/2401.17268>
- [287] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. arXiv:2203.11171 [cs.CL] <https://arxiv.org/abs/2203.11171>
- [288] Yao Wang, Mingxuan Cui, and Arthur Jiang. 2025. Enabling AI Scientists to Recognize Innovation: A Domain-Agnostic Algorithm for Assessing Novelty. arXiv:2503.01508 [cs.AI] <https://arxiv.org/abs/2503.01508>
- [289] Ying-Mei Wang and Tzeng-J Chen. 2025. AI's deep research revolution: Transforming biomedical literature analysis. https://journals.lww.com/jcma/citation/9900/ai_s_deep_research_revolution_transforming.508.aspx.
- [290] Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. <https://cdn.openai.com/papers/simpleqa.pdf>.
- [291] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903 [cs.CL] <https://arxiv.org/abs/2201.11903>
- [292] Shufa Wei, Xiaolong Xu, Xianbiao Qi, Xi Yin, Jun Xia, Jingyi Ren, Peijun Tang, Yuxiang Zhong, Yihao Chen, Xiaoqin Ren, Yuxin Liang, Liankai Huang, Kai Xie, Weikang Gui, Wei Tan, Shuanglong Sun, Yongquan Hu, Qinxian Liu, Nanjin Li, Chihao Dai, Lihua Wang, Xiaohui Liu, Lei Zhang, and Yutao Xie. 2023. AcademicGPT: Empowering Academic Research. arXiv:2311.12315 [cs.CL] <https://arxiv.org/abs/2311.12315>
- [293] Sarah Welsh. 2025. AI Benchmark Deep Dive: Gemini 2.5 and Humanity's Last Exam. <https://arize.com/blog/ai-benchmark-deep-dive-gemini-humanitys-last-exam/>.
- [294] Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. CycleResearcher: Improving Automated Research via Automated Review. arXiv:2411.00816 [cs.CL] <https://arxiv.org/abs/2411.00816>
- [295] Man Fai Wong, Shangxin Guo, Ching Nam Hang, Siu Wai Ho, and Chee Wei Tan. 2023. Natural Language Generation and Understanding of Big Code for AI-Assisted Programming: A Review. *Entropy* 25, 6 (June 2023), 888. doi:10.3390/e25060888
- [296] Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Derek F. Wong, and Lidia S. Chao. 2024. A Survey on LLM-Generated Text Detection: Necessity, Methods, and Future Directions. arXiv:2310.14724 [cs.CL] <https://arxiv.org/abs/2310.14724>
- [297] Junde Wu, Jiayuan Zhu, and Yuyuan Liu. 2025. Agentic Reasoning: Reasoning LLMs with Tools for the Deep Research. arXiv:2502.04644 [cs.AI] <https://arxiv.org/abs/2502.04644>
- [298] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. <https://github.com/microsoft/autogen>. arXiv:2308.08155 [cs.AI] <https://arxiv.org/abs/2308.08155>
- [299] xAI. 2025. Grok 3 Beta - The age of reasoning agents. <https://x.ai/news/grok-3>.
- [300] Menglin Xia, Victor Ruehle, Saravan Rajmohan, and Reza Shokri. 2025. Minerva: A Programmable Memory Test Benchmark for Language Models. https://github.com/gkamradt/LLMTest_NeedleInAHaystack. arXiv:2502.03358 [cs.CL] <https://arxiv.org/abs/2502.03358>
- [301] Tianbao Xie, Fan Zhou, Zhoujun Cheng, Peng Shi, Luoxuan Weng, Yitao Liu, Toh Jing Hua, Junning Zhao, Qian Liu, Che Liu, Leo Z. Liu, Yiheng Xu, Hongjin Su, Dongchan Shin, Caiming Xiong, and Tao Yu. 2023. OpenAgents: An Open Platform for Language Agents in the Wild. arXiv:2310.10634 [cs.CL] <https://arxiv.org/abs/2310.10634>
- [302] Yujia Xie, Xun Wang, Si-Qing Chen, Wayne Xiong, and Pengcheng He. 2023. Interactive Editing for Text Summarization. arXiv:2306.03067 [cs.CL] <https://arxiv.org/abs/2306.03067>
- [303] Feng Xiong, Xinguo Yu, and Hon Wai Leong. 2024. AI-Empowered Human Research Integrating Brain Science and Social Sciences Insights. arXiv:2411.12761 [cs.HC] <https://arxiv.org/abs/2411.12761>
- [304] Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. 2024. TheAgentCompany: Benchmarking LLM Agents on Consequential Real World Tasks. arXiv:2412.14161 [cs.CL] <https://arxiv.org/abs/2412.14161>
- [305] Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. 2025. UGPhysics: A Comprehensive Benchmark for Undergraduate Physics Reasoning with Large Language Models. <https://github.com/YangLabHKUST/UGPhysics>. arXiv:2502.00334 [cs.CL] <https://arxiv.org/abs/2502.00334>

- [306] Te-Lun Yang, Jyi-Shane Liu, Yuen-Hsien Tseng, and Jyh-Shing Roger Jang. 2025. Knowledge Retrieval Based on Generative AI. arXiv:2501.04635 [cs.IR] <https://arxiv.org/abs/2501.04635>
- [307] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. arXiv:1809.09600 [cs.CL] <https://arxiv.org/abs/1809.09600>
- [308] Yi Yao, Jun Wang, Yabai Hu, Lifeng Wang, Yi Zhou, Jack Chen, Xuming Gai, Zhenming Wang, and Wenjun Liu. 2024. BugBlitz-AI: An Intelligent QA Assistant. arXiv:2406.04356 [cs.SE] <https://arxiv.org/abs/2406.04356>
- [309] Burak Yetiştiren, Işık Özsoy, Miray Ayerdem, and Eray Tüzün. 2023. Evaluating the Code Quality of AI-Assisted Code Generation Tools: An Empirical Study on GitHub Copilot, Amazon CodeWhisperer, and ChatGPT. arXiv:2304.10778 [cs.SE] <https://arxiv.org/abs/2304.10778>
- [310] Xiaoxin Yin. 2024. "Turing Tests" For An AI Scientist. <https://github.com/MatthewFilipovich/pycharge>. arXiv:2405.13352 [cs.AI] <https://arxiv.org/abs/2405.13352>
- [311] yoheinakajima. 2024. BabyAGI. <https://github.com/yoheinakajima/babyagi>.
- [312] Seungri Yoon, Woosang Jeon, Sanghyeok Choi, Taehyeong Kim, and Tae In Ahn. 2025. Knowledge Synthesis of Photosynthesis Research Using a Large Language Model. arXiv:2502.01059 [cs.CL] <https://arxiv.org/abs/2502.01059>
- [313] You.com. 2023. You.com. <https://you.com/about>.
- [314] Hengjie Yu and Yaochu Jin. 2025. Unlocking the Potential of AI Researchers in Scientific Discovery: What Is Missing? arXiv:2503.05822 [cs.CY] <https://arxiv.org/abs/2503.05822>
- [315] Hengjie Yu and Yaochu Jin. 2025. Unlocking the Potential of AI Researchers in Scientific Discovery: What Is Missing? arXiv:2503.05822 [cs.CY] <https://arxiv.org/abs/2503.05822>
- [316] Jiakang Yuan, Xiangchao Yan, Shiyang Feng, Bo Zhang, Tao Chen, Botian Shi, Wanli Ouyang, Yu Qiao, Lei Bai, and Bowen Zhou. 2025. Dolphin: Moving Towards Closed-loop Auto-research through Thinking, Practice, and Feedback. arXiv:2501.03916 [cs.AI] <https://arxiv.org/abs/2501.03916>
- [317] Siyu Yuan, Cheng Jiayang, Lin Qiu, and Deqing Yang. 2024. Boosting Scientific Concepts Understanding: Can Analogy from Teacher Models Empower Student Models? arXiv:2406.11375 [cs.CL] <https://arxiv.org/abs/2406.11375>
- [318] Hector Zenil, Jesper Tegnér, Felipe S. Abrahão, Alexander Lavin, Vipin Kumar, Jeremy G. Frey, Adrian Weller, Larisa Soldatova, Alan R. Bundy, Nicholas R. Jennings, Koichi Takahashi, Lawrence Hunter, Saso Dzeroski, Andrew Briggs, Frederick D. Gregory, Carla P. Gomes, Jon Rowe, James Evans, Hiroaki Kitano, and Ross King. 2023. The Future of Fundamental Science Led by Generative Closed-Loop Artificial Intelligence. arXiv:2307.07522 [cs.AI] <https://arxiv.org/abs/2307.07522>
- [319] Beiq Zhang, Peng Liang, Xiyu Zhou, Aakash Ahmad, and Muhammad Waseem. 2023. Practices and Challenges of Using GitHub Copilot: An Empirical Study. In *Proceedings of the 35th International Conference on Software Engineering and Knowledge Engineering (SEKE2023, Vol. 2023)*. KSI Research Inc., 124–129. doi:10.18293/seke2023-077
- [320] Cedegao E. Zhang, Katherine M. Collins, Adrian Weller, and Joshua B. Tenenbaum. 2023. AI for Mathematics: A Cognitive Science Perspective. arXiv:2310.13021 [q-bio.NC] <https://arxiv.org/abs/2310.13021>
- [321] David Zhang. 2025. deep-research. <https://github.com/dzhng/deep-research>.
- [322] Kevin Zhang and Hod Lipson. 2024. Aligning AI-driven discovery with human intuition. arXiv:2410.07397 [cs.LG] <https://arxiv.org/abs/2410.07397>
- [323] Xingjian Zhang, Yutong Xie, Jin Huang, Jinge Ma, Zhaoying Pan, Qijia Liu, Ziyang Xiong, Tolga Ergen, Dongsub Shim, Honglak Lee, and Qiaozhu Mei. 2024. MASSW: A New Dataset and Benchmark Tasks for AI-Assisted Scientific Workflows. <https://github.com/xingjian-zhang/massw>. arXiv:2406.06357 [cs.CL] <https://arxiv.org/abs/2406.06357>
- [324] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. ACM, 1–30. doi:10.1145/3586183.3606800
- [325] Jinjin Zhao, Avidgor Gal, and Sanjay Krishnan. 2024. A System for Quantifying Data Science Workflows with Fine-Grained Procedural Logging and a Pilot Study. arXiv:2405.17845 [cs.HC] <https://arxiv.org/abs/2405.17845>
- [326] Zihan Zhao, Bo Chen, Jingpiao Li, Lu Chen, Liyang Wen, Pengyu Wang, Zichen Zhu, Danyang Zhang, Yansi Li, Zhongyang Dai, Xin Chen, and Kai Yu. 2024. ChemDFM-X: towards large multimodal model for chemistry. <https://github.com/OpenDFM/ChemDFM-X>. *Science China Information Sciences* 67, 12 (Dec. 2024). doi:10.1007/s11432-024-4243-0
- [327] Raigul Zheldibayeva. 2025. The impact of AI and peer feedback on research writing skills: a study using the CGScholar platform among Kazakhstani scholars. arXiv:2503.05820 [cs.CY] <https://arxiv.org/abs/2503.05820>
- [328] Dewu Zheng, Yanlin Wang, Ensheng Shi, Xilin Liu, Yuchi Ma, Hongyu Zhang, and Zibin Zheng. 2025. Top General Performance = Top Domain Performance? DomainCodeBench: A Multi-domain Code Generation Benchmark. <https://github.com/DeepSoftwareAnalytics/MultiCodeBench>. arXiv:2412.18573 [cs.SE] <https://arxiv.org/abs/2412.18573>

- [329] 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. arXiv:2504.03160 [cs.AI] <https://arxiv.org/abs/2504.03160>
- [330] Zhipu AI. 2025. AutoGLM-Research. <https://autoglm-research.zhipuai.cn/>.
- [331] Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. AGIEval: A Human-Centric Benchmark for Evaluating Foundation Models. arXiv:2304.06364 [cs.CL] <https://arxiv.org/abs/2304.06364>
- [332] Shuyan Zhou, Chengrun Chi, Ceyao Zheng, Bailin Zhang, Yonatan Bisk, Daniel Fried, Ishan Misra, Karthik Raghunathan, Tongshuang Zhao, Baian Zhou, et al. 2024. WebArena: A Benchmark for Web Agents. <https://github.com/web-arena-x/webarena>.
- [333] Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process. arXiv:2503.08569 [cs.CL] <https://arxiv.org/abs/2503.08569>
- [334] Tonghe Zhuang and Zhicheng Lin. 2024. The why, what, and how of AI-based coding in scientific research. arXiv:2410.02156 [cs.CY] <https://arxiv.org/abs/2410.02156>
- [335] Dennis Zyska, Nils Dycke, Jan Buchmann, Ilia Kuznetsov, and Iryna Gurevych. 2023. CARE: Collaborative AI-Assisted Reading Environment. arXiv:2302.12611 [cs.CL] <https://arxiv.org/abs/2302.12611>
- [336] Tolga Çöplü, Arto Bendiken, Andrii Skomorokhov, Eduard Bateiko, Stephen Cobb, and Joshua J. Bouw. 2024. Prompt-Time Symbolic Knowledge Capture with Large Language Models. <https://github.com/HaltiaAI/paper-PTSKC>. arXiv:2402.00414 [cs.CL] <https://arxiv.org/abs/2402.00414>