Towards Agentic RAG with Deep Reasoning: A Survey of RAG-Reasoning Systems in LLMs

Yangning Li[1*], **Weizhi Zhang**[2*], **Yuyao Yang**[2], **Wei-Chieh Huang**[2], **Yaozu Wu**[3]
Junyu Luo[4], **Yuanchen Bei**[5], **Henry Peng Zou**[2], **Xiao Luo**[6], **Yusheng Zhao**[4]
Chunkit Chan[7], **Yankai Chen**[2,‡], **Zhongfen Deng**[2], **Yinghui Li**[1], **Hai-Tao Zheng**[1,‡],
Dongyuan Li[3], **Renhe Jiang**[3], **Ming Zhang**[4], **Yangqiu Song**[7], **Philip S. Yu**[2]
[1]Shenzhen International Graduate School, Tsinghua University
[2]University of Illinois Chicago [3]The University of Tokyo [4]Peking University
[5]University of Illinois Urbana-Champaign [6]University of Wisconsin–Madison [7]HKUST
yn-li23@mails.tsinghua.edu.cn, wzhan42@uic.edu

Abstract

Retrieval-Augmented Generation (RAG) lifts the factuality of Large Language Models (LLMs) by injecting external knowledge, yet it falls short on problems that demand multi-step inference; conversely, purely reasoning-oriented approaches often hallucinate or mis-ground facts. This survey synthesizes both strands under a unified reasoning-retrieval perspective. We first map how advanced reasoning optimizes each stage of RAG (**Reasoning-Enhanced RAG**). Then, we show how retrieved knowledge of different type supply missing premises and expand context for complex inference (**RAG-Enhanced Reasoning**). Finally, we spotlight emerging **Synergized RAG-Reasoning** frameworks, where (agentic) LLMs iteratively interleave search and reasoning to achieve state-of-the-art performance across knowledge-intensive benchmarks. We categorize methods, datasets, and open challenges, and outline research avenues toward deeper RAG-Reasoning systems that are more effective, multimodally-adaptive, trustworthy, and human-centric. The collection is available at https://github.com/DavidZWZ/Awesome-RAG-Reasoning.

1 Introduction

The remarkable progress in Large Language Models (LLMs) has transformed a wide array of fields, showcasing unprecedented capabilities across diverse tasks (Zhao et al., 2023). Despite these advancements, the effectiveness of LLMs remains hindered by two fundamental limitations: knowledge hallucinations, due to the static and parametric manner of their knowledge storage (Huang et al., 2025b); and struggles with complex reasoning, especially when tackling real-world problems (Chang et al., 2024). These limitations have driven the development of two major directions: Retrieval-Augmented Generation (RAG) (Fan et al., 2024a), which provides LLMs with external knowledge; and various methods aimed at enhancing their inherent reasoning abilities (Chen et al., 2025c).

The two limitations are inherently intertwined: missing knowledge can impede reasoning, and flawed reasoning hinders knowledge utilization (Tonmoy et al., 2024). Naturally, researchers have increasingly explored combining retrieval with reasoning, though early work followed *two separate, one-way enhancements*. The first, **Reasoning-enhanced RAG** (Gao et al., 2023b) (Reasoning → RAG), leverages reasoning to improve specific stages of the RAG pipeline. The second path, **RAG-enhanced Reasoning** (Fan et al., 2024a) (RAG → Reasoning), supplies external factual grounding or contextual cues to bolster LLM reasoning.

While beneficial, the above methods remain bound to a static Retrieval-Then-Reasoning (RTR) framework, offering only localized improvements to individual components. Several inherent limitations persist: *(1) Retrieval Adequacy and Accuracy* cannot be guaranteed; Pre-retrieved knowledge may fail to align with the actual knowledge needs that emerge during reasoning, especially in complex tasks (Zheng et al., 2025; Li et al., 2025d). *(2) Reasoning Depth* remains constrained. When retrieved knowledge contains errors or conflicts, it can adversely interfere with the model's inherent reasoning capabilities (Li et al., 2025b; Chen et al., 2025a). *(3) System Adaptability* proves insufficient. The RTR framework lacks mechanisms for iterative feedback or dynamic retrieval during reasoning. This rigidity limits its effectiveness in scenarios that require adaptive reasoning, such as open-domain QA or scientific discovery (Xiong et al., 2025; Alzubi et al., 2025).

As shown in Figure 1, these shortcomings have catalyzed a paradigm shift toward **Synergized Re-**

[*] Equal Contribution. ‡ Corresponding author. Zhongfeng contributed to this paper prior to Amazon.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12120–12145
November 4-9, 2025 ©2025 Association for Computational Linguistics

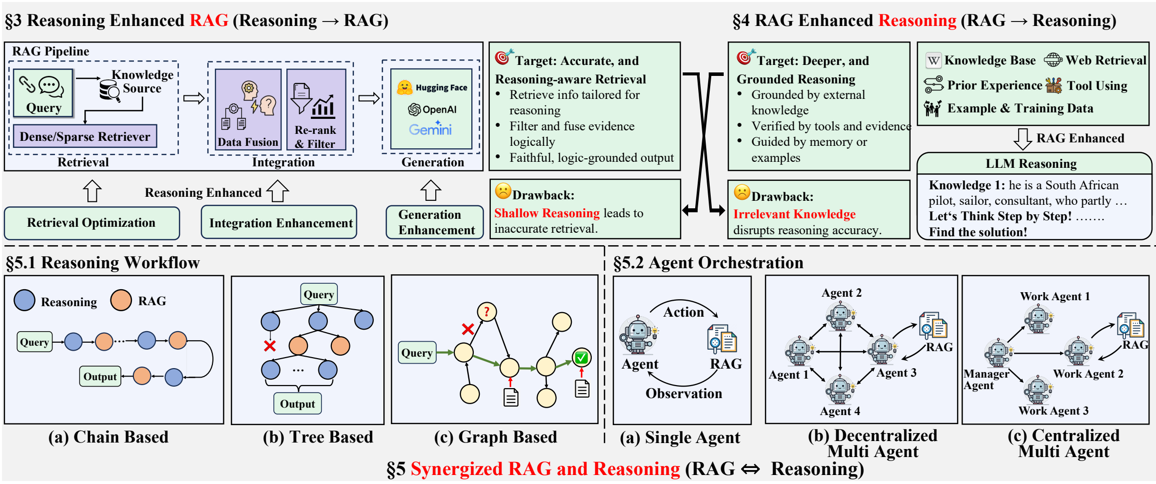

Figure 1: Overview of the RAG-Reasoning System. The *Reasoning-Enhanced RAG* methods and *RAG-Enhanced Reasoning* methods represent **one-way** enhancements. In contrast, the *Synergized RAG-Reasoning System* performs reasoning and retrieval **iteratively**, enabling mutual enhancements.

trieval and Reasoning within LLMs (RAG ⇔ Reasoning). These methods support a dynamic, iterative interplay where reasoning actively guides retrieval, and newly retrieved knowledge, in turn, continuously refines the reasoning process. This trend is further exemplified by recent "Deep Research" products from OpenAI[1], Gemini[2], Perplexity[3], and others, which emphasize tightly coupled retrieval and reasoning (Zhang et al., 2025f). These systems employ agentic capabilities to orchestrate multi-step web search and leverage reasoning to comprehensively interpret retrieved content, solving problems demanding in-depth investigation.

This survey charts the shift from isolated enhancements to cutting-edge synergized frameworks where retrieval and reasoning are deeply interwoven and co-evolve. While surveys on RAG (Fan et al., 2024a; Gao et al., 2023b) and LLM Reasoning (Chen et al., 2025c; Li et al., 2025e) exist, a dedicated synthesis focusing on their integration remains lacking. Our goal is to provide a comprehensive overview of how the symbiosis between retrieval and reasoning is advancing LLM capabilities, with particular emphasis on the move towards a synergized RAG and Reasoning framework.

The survey is structured as follows: Section 2 introduces the background; Section 3 and 4 review two one-way enhancements, respectively. Section 5 unifies both lines into synergized RAG–Reasoning frameworks. Section 6 lists benchmarks, and Section 7 outlines open challenges.

2 Background and Preliminary

RAG mitigates knowledge cut-off of LLMs through three sequential stages: *(i) Retrieval*, fetching task-relevant content from external knowledge stores; *(ii) Integration*, deduplicating, resolving conflicts, and re-ranking the retrieved content; and *(iii) Generation*, reasoning over the curated context to produce the final answer. Concurrently, Chain-of-Thought technique has significantly enhanced the reasoning capabilities of modern LLMs by encouraging them to "*think step by step*" before answering. The synergy between the structured RAG pipeline and these multi-step reasoning capacities grounds the emerging RAG-Reasoning paradigm explored in this survey.

3 Reasoning-Enhanced RAG

Traditional RAG methods first retrieve relevant documents, then concatenate the retrieved knowledge with the original query to generate the final answer. These methods often fail to capture the deeper context or intricate relationships necessary for complex reasoning tasks. By integrating reasoning capabilities across **Retrieval**, **Integration**, and **Generation** stages of the RAG pipeline, the system can identify and fetch the most relevant information, reducing hallucinations and improving response accuracy.[4]

[1]https://openai.com/index/introducing-deep-research/

[2]https://gemini.google/overview/deep-research/

[3]https://www.perplexity.ai/hub/blog/introducing-perplexity-deep-research

[4]If reasoning only serves to better leverage **fixed** retrieved knowledge in a unidirectional manner, it is considered within

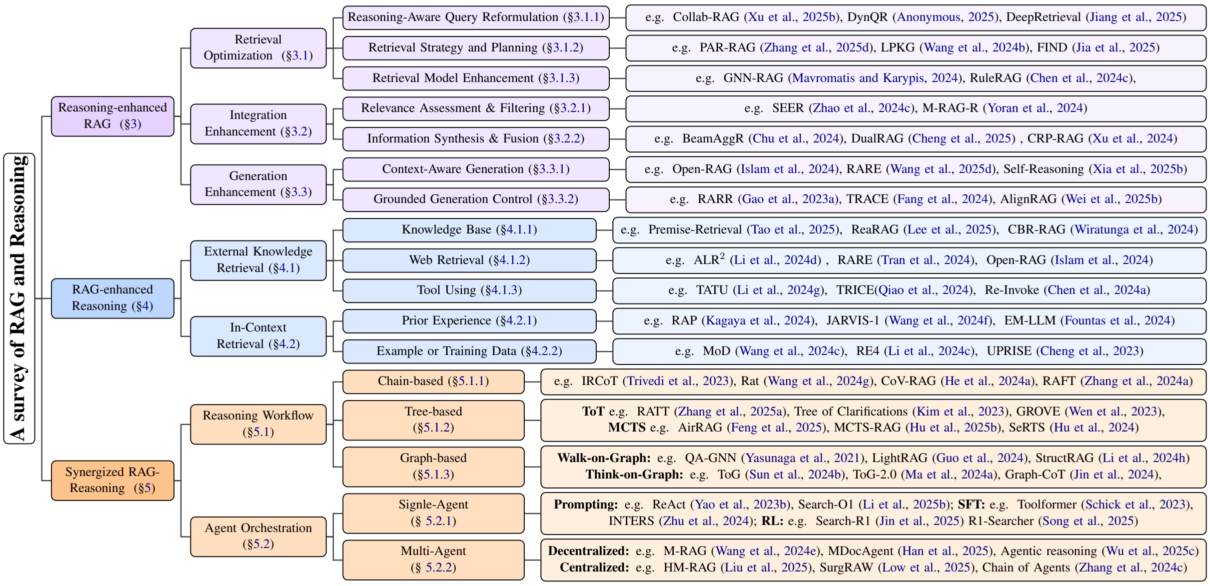

Figure 2: Taxonomy of Recent Advances in RAG-Reasoning System.

3.1 Retrieval Optimization

Retrieval optimization leverages reasoning to improve result relevance and quality. Existing methods are broadly categorized (1) Reasoning-Aware Query Reformulation, (2) Retrieval Strategy and Planning, and (3) Retrieval Model Enhancement.

3.1.1 Reasoning-Aware Query Reformulation

It reformulates the original query to better retrieve reasoning-relevant context. First, query decomposition breaks down complex queries into simpler sub-queries (Xu et al., 2025b). Second, query reformulation recasts ambiguous queries into more clear ones. To align with reasoning needs of generator, certain works train rewrites with RL signals (Anonymous, 2025; Wang et al., 2025c). Third, query expansion enrich the semantic richness of the query via CoT reasoning (Dhuliawala et al., 2024; Li et al., 2024e; Lee et al., 2024).

3.1.2 Retrieval Strategy and Planning

This section covers global retrieval guidance. Advance planning uses a reasoning model to generate a complete retrieval blueprint prior to execution. PAR-RAG (Zhang et al., 2025d) applies CoT for multi-step planning, mitigating local optima. LPKG (Wang et al., 2024b) fine-tunes LLMs on knowledge graphs to encode relational structure. In contrast, adaptive retrieval decision methods make a one-step prediction on whether and how to retrieve. FIND (Jia et al., 2025) and adaptive RAG (Jeong et al., 2024) use classifiers to assess query complexity and select retrieval strategies, reducing unnecessary calls. Marina et al. (2025) further adds features like entity popularity and question type.

3.1.3 Retrieval Model Enhancement

A line of work enhances retrievers with reasoning via two strategies. The first one leverages structured knowledge: GNN-RAG (Mavromatis and Karypis, 2024) encodes knowledge graphs with GNNs for implicit multi-hop reasoning, while RuleRAG (Chen et al., 2024c) appends symbolic rules to guide retrieval toward logical consistency. Another strategy integrates explicit reasoning: Ji et al. (2024) combines CoT with the query to improve intermediate knowledge recall in multi-hop QA.

3.2 Integration Enhancement

Integration enhancement uses reasoning to assess relevance and merge heterogeneous evidence, preventing irrelevant content from disrupting generation. Methods fall into two categories: (1) relevance assessment and (2) information synthesis.

3.2.1 Relevance Assessment & Filtering

These methods assess the relevance of each retrieved fragment to the user query through deeper reasoning. SEER (Zhao et al., 2024c) employs assessor experts to select faithful, helpful, and concise evidence while discarding irrelevant content. Yoran et al. (2024) improves robustness by filtering non-entailing passages using an NLI model, then

§3.3. In contrast, if reasoning **dynamically triggers new retrieval**, it is discussed in §5.

fine-tuning the LLM on mixed relevant/irrelevant contexts to help it ignore residual noise.

3.2.2 Information Synthesis & Fusion

Once relevant snippets are identified, the challenge is to fuse them into a coherent evidence set. BeamAggR (Chu et al., 2024) enumerates sub-question answer combinations and aggregates them via probabilistic reasoning. DualRAG (Cheng et al., 2025) combines reasoning-augmented querying with progressive knowledge aggregation to filter and organize retrieved information into an evolving outline. CRP-RAG (Xu et al., 2024) builds a reasoning graph to retrieve, evaluate, and aggregate knowledge at each node, dynamically selecting knowledge-sufficiency paths before generation.

3.3 Generation Enhancement

Even with retrieved context, traditional RAG may still generate unfaithful content without reasoning. Reasoning during generation addresses this issue through two main approaches: (1) context-aware synthesis and (2) grounded generation control.

3.3.1 Context-Aware Synthesis Strategies

Context-aware generation ensures outputs remain relevance while reducing noise. Selective–context utilization prunes or re-weights content based on task relevance. Open-RAG (Islam et al., 2024) uses a sparse expert mixture to dynamically select knowledge modules, while RARE (Wang et al., 2025d) adds domain knowledge to prompts to promote reliance on external context over memorization. Reasoning path generation builds explicit logical chains to enhance transparency, e.g., Ranaldi et al. (2024) generate contrasting explanations by comparing paragraph relevance step-by-step, guiding the model toward accurate conclusions. Self-Reasoning (Xia et al., 2025b) constructs structured reasoning chains through sequential evidence selection and verification.

3.3.2 Grounded Generation Control

Grounded generation control introduces verification mechanisms to ensure outputs remain anchored to retrieved evidence through reasoning. Fact verification methods use reasoning to assess factual consistency between generated content and retrieved evidence, e.g., Self-RAG (Asai et al., 2023) introduces reflection markers during decoding to trigger critical review and correction. Citation generation links generated content to source materials to enhance traceability and credibility, as in RARR (Gao et al., 2023a), which inserts citations while preserving stylistic coherence. Faithful reasoning ensures that each reasoning step adheres to retrieved evidence without introducing unverified content. TRACE (Fang et al., 2024) builds knowledge graphs to form coherent evidence chains, while AlignRAG (Wei et al., 2025b) applies criticism alignment to refine reasoning paths.

4 RAG-Enhanced Reasoning

Integrating external knowledge or in-context knowledge during reasoning can help LLMs reduce hallucinations and bridge logical gaps. External retrieval leverages structured sources like databases or web content, providing factual grounding, like IAG (Zhang et al., 2023). In-context retrieval utilizes internal contexts like prior interactions or training examples, enhancing contextual coherence, like RA-DT (Schmied et al., 2024). Both strategies collectively improve factual accuracy, interpretability, and logical consistency of reasoning processes.

4.1 External Knowledge Retrieval

External knowledge retrieval incorporates web content, database information, or external tools into reasoning, effectively filling knowledge gaps. Targeted retrieval improves factual accuracy, enabling language models to reliably address complex queries by grounding reasoning steps in verified external evidence.

4.1.1 Knowledge Base

Knowledge base (KB) typically stores arithmetic, commonsense, or logical knowledge in databases, books, or documents, with retrieval approaches varying by task. For question answering (QA) reasoning, AlignRAG (Wei et al., 2025b), MultiHop-RAG (Tang and Yang, 2024), and CRP-RAG (Xu et al., 2025a) retrieve interconnected factual entries from general KBs to enhance sequential reasoning. In specialized reasoning tasks, mathematical approaches like Premise-Retrieval (Tao et al., 2025) and ReaRAG (Lee et al., 2025) utilize formal lemmas from theorem libraries for structured deduction; legal approaches like CASEGPT (Yang, 2024) and CBR-RAG (Wiratunga et al., 2024) extract judicial precedents for analogical reasoning. For code generation tasks, CodeRAG (Li et al., 2025a) and Koziolek et al. (2024) access code snippets from repositories, ensuring syntactic correctness.

4.1.2 Web Retrieval

Web retrieval accesses dynamic online content like web pages, news or social media. Specifically, in fact-checking tasks, approaches such as VeraCT Scan (Niu et al., 2024), Ragar (Khaliq et al., 2024), PACAR (Zhao et al., 2024b), and STEEL (Li et al., 2024b) verify claims step-by-step using evidence from news or social media, enhancing logical reasoning. Meanwhile, QA-based reasoning like RARE (Tran et al., 2024), RAG-Star (Jiang et al., 2024), MindSearch (Chen et al., 2024b), and OPEN-RAG (Islam et al., 2024) iteratively refine reasoning with broad web content, aligning with current trends in agentic search, which involve synthesizing complex online materials to enhance context-aware and robust reasoning. Conversely, in specialized areas like medical domain, FRVA (Fan et al., 2024b) and ALR^2 (Li et al., 2024d) retrieve literature for accurate diagnostics.

4.1.3 Tool Using

Tool-using approaches leverage external resources like calculators, libraries, or APIs to enhance reasoning interactively. In QA-based reasoning, Re-Invoke (Chen et al., 2024a), AVATAR (Wu et al., 2024), ToolkenGPT (Hao et al., 2023), and Tool-LLM (Qin et al., 2023) invoke calculators or APIs (e.g., Yahoo Finance, Wikidata), improving numerical accuracy and factual precision. Within the context of scientific modeling, SCIAGENT (Ma et al., 2024b) and TRICE (Qiao et al., 2024) integrate symbolic computation tools (e.g., WolframAlpha), strengthening computational robustness. Similarly, in mathematical computation, llm-tool-use (Luo et al., 2025b) autonomously employs calculators for accurate numerical reasoning. Distinctively in code generation tasks, RAR (Dutta et al., 2024) retrieves code documentation via OSCAT libraries, ensuring syntactic accuracy and executable logic.

4.2 In-context Retrieval

In-context retrieval leverages a model's internal experiences or retrieved examples from demonstrations and training data to guide reasoning. This retrieval provides relevant exemplars, guiding models to emulate reasoning patterns and enhancing accuracy and logical coherence in novel questions.

4.2.1 Prior Experience

Prior experience refers to past interactions or successful strategies stored in a model's internal memory, with retrieval varying by task. In tasks involving planning and decision-making tasks such as robot path finding, RAHL (Sun et al., 2024a) and RA-DT (Schmied et al., 2024) leverage past decisions and reinforcement signals for sequential reasoning. For interactive reasoning tasks, JARVIS-1 (Wang et al., 2024f), RAP (Kagaya et al., 2024), and EM-LLM (Fountas et al., 2024) dynamically recall multimodal interactions and conversational histories, facilitating adaptive reasoning for personalized interactions. In the domain for **logical reasoning**, CoPS (Yang et al., 2024a) retrieves structured prior cases for robust logical reasoning in medical and legal scenarios.

4.2.2 Example or Training Data

Unlike approaches relying on prior experiences, example-based reasoning retrieves external examples from demonstrations or training data. For example, In complex text-understanding, RE4 (Li et al., 2024c) and Fei et al. (2024) utilize annotated sentence pairs to enhance relation recognition. Addressing QA-based reasoning, OpenRAG (Zhou and Chen, 2025), UPRISE (Cheng et al., 2023), MoD (Wang et al., 2024c), and Dr.ICL (Luo et al., 2023) select demonstrations closely matching queries, improving generalization. Additionally, in code generation tasks, PERC (Yoo et al., 2025) retrieves pseudocode by semantic or structural similarity from datasets like HumanEval, ensuring alignment with target code.

5 Synergized RAG-Reasoning

Many real-world problems, such as open-domain question answering (Yang et al., 2015; Chen and Yih, 2020) and scientific discovery (Lu et al., 2024; Wang et al., 2023; Baek et al., 2024; Schmidgall et al., 2025), require an iterative approach where new evidence continuously informs better reasoning and vice versa. A single retrieval step may not provide sufficient information, and a single round of reasoning may overlook key insights (Trivedi et al., 2023). By tightly integrating retrieval and reasoning in a multi-step, interactive manner, these systems can progressively refine both the search relevance of retrieved information and the reasoning-based understanding of the original query. We focus on two complementary perspectives within existing approaches: reasoning workflows, which emphasize structured, often pre-defined inference formats for multi-step reasoning; and agent orchestration, which focus on how agents interact with environment and coordinate with each others.

5.1 Reasoning Workflow

Broadly, the reasoning workflows can be categorized as chain-based, tree-based, or graph-based, reflecting an evolution from linear reasoning chains to branching and expressive reasoning structures.

5.1.1 Chain-based

Chain-of-Thought (CoT) (Wei et al., 2022) structures the reasoning process as a linear sequence of intermediate steps. However, relying solely on the parametric knowledge of LLMs can lead to error propagation. To solve this, IRCoT (Trivedi et al., 2023) and Rat (Wang et al., 2024g) interleave retrieval operations between reasoning steps. Several recent methods further improve the robustness and rigor of this chain-based paradigm via verification and filtering. CoV-RAG (He et al., 2024a) introduces a chain-of-verification that checks and corrects each reasoning step against retrieved references. To combat noisy or irrelevant context, approaches like RAFT (Zhang et al., 2024a) fine-tune LLMs to ignore distractor documents, while Chain-of-Note (Yu et al., 2024) prompts the model to take sequential "reading notes" on retrieved documents to filter out unhelpful information.

5.1.2 Tree-based

Tree-based reasoning methods typically adopt either Tree-of-Thought (ToT) (Yao et al., 2023a) or Monte Carlo Tree Search (MCTS) (Browne et al., 2012) approaches. **ToT** extends the CoT to explicitly construct a deterministic reasoning tree and branch multiple logical pathways. Examples include RATT (Zhang et al., 2025a), which construct retrieval-augmented thought trees to simultaneously evaluate multiple reasoning trajectories. Such ToT principles avoid LLM being trapped by an early mistaken assumption and have been applied to address ambiguous questions (Kim et al., 2023), to cover different diagnostic possibilities (Yang and Huang, 2025), and to create complex stories (Wen et al., 2023). Conversely, **MCTS**-based approaches like AirRAG (Feng et al., 2025), ARise (Zhang et al., 2025h), MCTS-RAG (Hu et al., 2025b), and SeRTS (Hu et al., 2024) employ probabilistic tree search, dynamically prioritizing exploration based on heuristic probabilities. To ensure retrieval and reasoning quality, AirRAG (Feng et al., 2025) incorporates self-consistency checks, and MCTS-RAG (Hu et al., 2025b) integrates adaptive MCTS retrieval to refine evidence and reduce hallucinations.

5.1.3 Graph-based

Walk-on-Graph methods mainly rely on graph learning techniques for the retrieval and reasoning. For example, PullNet (Sun et al., 2019), QA-GNN (Yasunaga et al., 2021), and GreaseLM (Zhang et al., 2022b) directly integrate graph neural networks (GNNs) to iteratively aggregate information from neighbor nodes, excelling at modeling the intricate relationships inherent in graph-structured data. Methods such as SR (Zhang et al., 2022a), LightRAG (Guo et al., 2024), and StructRAG (Li et al., 2024h) employ lightweight graph techniques such as vector indexing and PageRank to efficiently retrieve and reason in multi-hop context, providing the LLM with high-quality, structured content tailored for the queries. In contrast, **Think-on-Graph** methods integrate graph structures directly into the LLM reasoning loop, enabling dynamic and iterative retrieval and reasoning processes guided by the LLMs themselves. In the Think-on-Graph (ToG) framework (Sun et al., 2024b; Ma et al., 2024a), the LLM uses the KG as a "reasoning playground": at each step, it decides which connected entity or relation to explore next, gradually building a path that leads to the answer. While Graph-CoT (Jin et al., 2024) introduces a three-stage iterative loop (reasoning, graph interaction, and execution), KGP (Wang et al., 2024d) prioritize first constructing a document-level KG, both enabling LLM-driven graph traversal agent to navigate passages in each step with globally coherent context. GraphReader (Li et al., 2024f) and GIVE (He et al.) further refines this paradigm by coupling LLM reasoning with explicit external subgraph evidence and memories at each step.

5.2 Agent Orchestration

According to agent architectures (Luo et al., 2025a), we organize existing work into single-agent and multi-agent. Particularly, we have attached recent advances in agentic deep research and implementations in Appendix B.

5.2.1 Single-Agent

Single agentic system interweaves knowledge retrieval (search) into an LLM's reasoning loop, enabling dynamic information lookup at each step of problem solving and incentivizing it to actively seek out relevant evidence when needed.

The ReAct (Yao et al., 2023b) paradigm and its derivatives (Li et al., 2025b; Alzubi et al., 2025) have pioneered this **prompting** strategy by guiding LLMs to explicitly alternate between reasoning steps and external tool interactions, such as database searches. Different from ReAct that separates reasoning and action, with explicit commands like "search" triggering external retrieval, methods such as Self-Ask (Press et al., 2023) and IRCoT (Trivedi et al., 2023) prompt the model to recursively formulate and answer sub-questions, enabling interleaved retrieval within the Chain-of-Thought (step-by-step retrieval and reasoning). Involving self-reflection strategies, DeepRAG (Guan et al., 2025) and Self-RAG (Asai et al., 2024) empower LLMs to introspectively assess their knowledge limitations and retrieve only when necessary.

Rather than relying solely on prompting or static retrievers, Toolformer (Schick et al., 2023) and INTERS (Zhu et al., 2024) represent a complementary approach via **supervised fine-tuning** (SFT) LLMs on instruction-based or synthetic datasets that interleave search and reasoning. Synthetic data generation (Schick et al., 2023; Mao et al., 2024; Zhang et al., 2024a) aims to create large-scale, diverse, and task-specific datasets for search without the need for extensive human annotation. In contrast, instruction-based data reformulation (Zhu et al., 2024; Wang et al., 2024a; Lin et al., 2023; Nguyen et al., 2024) repurposes existing datasets into instructional formats to fine-tune models for improved generalization and alignment with human-like reasoning. INTERS (Zhu et al., 2024) exemplifies this approach by introducing a SFT dataset encompassing 20 tasks, derived from 43 distinct datasets with manually written templates.

Reinforcement learning (RL)-incentivized approaches provides a mechanism to optimize answer quality via reward signals on incentivizing agents' behaviors – what to search, how to integrate retrieved evidence, and when to stop, aiming at complex knowledge-intensive tasks (or "deep research" questions). Notable efforts like WebGPT (Nakano et al., 2021) and RAG-RL (Huang et al., 2025a) focus on improving reasoning fidelity by rewarding outputs based on factual correctness or human preference. More recent contributions operate directly in dynamic environments (e.g., live web search, local search tools), training agents to explore, reflect, and self-correct in noisy real-world conditions. For example, Search-R1 (Jin et al., 2025) learns to generate <search> token during reasoning and concurrently R1-Searcher (Song et al., 2025) builds on RL-driven search demonstrating strong generalization across domains. Deep-Researche (Zheng et al., 2025) make step further by introducing the first end-to-end RL-trained research agent that interacts with the open web. These settings showcase emergent capabilities, like decomposition, iterative verification, and retrieval planning, that supervised methods often hard to instill. Moreover, ReSearch (Chen et al., 2025b) and ReARTeR (Sun et al., 2025c) tackle a deeper challenge: not just producing correct answers, but aligning reasoning steps with both factuality and interpretability.

5.2.2 Multi-Agent

The exploration of multi-agent collaboration within RAG and reasoning has led to diverse orchestrations: centralized architectures (harness collective intelligence from workers-manager paradigm) and decentralized architectures (leverage complementary capabilities from role-specialized agents).

Decentralized architectures deploy multiple agents to collaboratively perform retrieval, reasoning, and knowledge integration, aiming to broaden coverage of relevant information and fully exploit the heterogeneous strengths of specialized agents. Wang et al. (2024e) and Salve et al. (2024) introduce multi-agent systems where each agent retrieves from a partitioned database or a specific data source (relational databases, NoSQL document stores, etc.). Beyond retrieval, Collab-RAG (Xu et al., 2025b) and RAG-KG-IL (Yu and McQuade, 2025) integrate different model capacities and assign them different roles in reasoning and knowledge integration. This philosophy extends to multimodal settings as in MDocAgent (Han et al., 2025), which employs a team of text and image agents to process and reason the document-based QA. A general formulation is seen in Agentic reasoning (Wu et al., 2025c), which unites tool-using agents for search, computation, and structured reasoning, orchestrated to solve complex analytical tasks.

Centralized architectures structure agents in hierarchical centralized patterns, supporting efficient task decomposition and progressive refinement. HM-RAG (Liu et al., 2025) and SurgRAW (Low et al., 2025) both employ decomposer-retriever-decider architectures, where different agent roles isolate subproblems such as multimodal processing or surgical decision-making. Wu et al. (2025a) and Iannelli et al. (2024) emphasize dynamic routing and system reconfiguration, respectively—enabling intelligent agent selection based on task relevance or resource constraints. Chain of Agents (Zhang et al., 2024c) and the cooperative multi-agent control framework for on-ramp merging (Zhang et al., 2025c) illustrate hierarchical agent designs where layered processing enables long-context summarization or policy refinement. Collectively, these works demonstrate how centralized control and hierarchical pipelining foster efficiency and adaptability in multi-agent RAG-reasoning systems.

6 Benchmarks and Datasets

Benchmarks and datasets for simultaneously evaluating knowledge (RAG) and reasoning capability cover a wide range of complexities, from basic fact retrieval to intricate multi-step reasoning in general or specific domains. We categorize notable benchmarks in several tasks and list them in Table 1 and highlight their details and properties. These representative tasks include Web browsing, such as BrowseComp (Wei et al., 2025a), single-hop QA, such as TriviaQA (Joshi et al., 2017), multi-hop QA, such as HotpotQA (Yang et al., 2018), multiple-choice QA, such as MMLU-Pro (Wang et al., 2025b), mathematics, such as MATH (Hendrycks et al., 2021), and code-centric evaluations from LiveCodeBench (Jain et al., 2024). More tasks can refer to Appendix A and Table 2.

7 Future Work

Future research directions for Synergized RAG-Reasoning systems center around enhancing both reasoning and retrieval capabilities to meet real-world demands for accuracy, efficiency, trust, and user alignment. We outline several key challenges and opportunities below.

• **Reasoning Efficiency.** Despite their advantages in complex reasoning, Synergized RAG-Reasoning systems can suffer significant latency due to iterative retrieval and multi-step reasoning loops (Sui et al., 2025). For instance, executing a single deep research query can take over 10 minutes in practical settings. This issue is especially pronounced in chain-based workflows discussed in Section 5. Future research should explore reasoning efficiency through latent reasoning approaches and strategic control over reasoning depth via thought distillation and length-penalty (Xia et al., 2025a; Zhang et al., 2025b). Beyond reasoning itself, emerging directions in models compression like quantization, pruning, and knowledge distillation is worth to explore for efficient small RAG-reasoning systems.

• **Retrieval Efficiency.** On the retrieval side, efficiency demands budget-aware query planning and memory-aware mechanisms that cache prior evidence or belief states to reduce redundant access (Zhao et al., 2024a). Additionally, adaptive retrieval control, learning when and how much to retrieve based on uncertainty signals can reduce wasteful operations. These technical paths push the system beyond static RAG, toward dynamic self-regulation of efficient retreival behaviors under real-world constraints.

• **Human-Agent Collaboration.** Many applications of RAG-Reasoning, such as literature reviews or interactive programming, are inherently personalized and cannot assume users know precisely what to ask or how to process retrieved results (Sun et al., 2025b). Corresponding to Section 5.2, humans can act as advanced agents, providing nuanced feedback to steer reasoning processes. Future systems should develop methods for modeling user intent under uncertainty (Zhang et al., 2025e; Yang et al., 2025), building interactive interfaces for iterative clarification, and designing agents that adapt reasoning strategies based on user expertise and preferences (Zhang et al., 2025g). This human-in-the-loop approach (Zou et al., 2025) is essential for creating robust and user-aligned RAG-Reasoning systems in open-ended domains.

• **Agentic Structures and Capabilities.** A key feature of Synergized RAG-Reasoning is its agentic architecture, where the system autonomously decides the roles of different agents and which tools or retrieval strategies to invoke during inference stages (Luo et al., 2025a; Bei et al., 2025). To fully exploit this potential, future research should focus on developing agent frameworks capable of dynamic tool selection, retrieval planning, and adaptive orchestration across reasoning workflows. Such capabilities enable flexible, context-aware problem solving and are critical for handling diverse, complex tasks (Schneider, 2025).

• **Multimodal Retrieval.** As also shown in our benchmark analysis, most existing Synergized RAG-Reasoning systems remain confined to text-only tasks. However, real-world applications increasingly require the ability to retrieve and integrate multimodal content (Liang et al., 2024; Hu et al., 2025a). Future research should move beyond the traditional vision-text paradigm to achieve genuine multimodality. This advancement necessitates strengthening foundational abilities of MLLMs, including grounding and cross-modal reasoning

Task	*Dataset*	Domain	Knowledge Source	Knowledge Type	Reasoning	Size	Input	Output
Web Browsing	*BrowseComp (Wei et al., 2025a)*	General	Human, Internet	Commonsense, Logical	Deductive	1,266	Question/Text	Natural Language
	GAIA (Mialon et al., 2023)	General	Internet, TooL	Commonsense, Logical	Deductive	466	Question/Text, Image/File/Code	Natural Language
	WebWalkerQA (Wu et al., 2025b)	General	Human, LLM	Commonsense, Logical	Deductive	680	Question/Text	Natural Language
Single-hop QA	*TriviaQA (Joshi et al., 2017)*	General	Internet	Commonsense, Logical	Deductive	650,000+	Question/Text	Natural Language
	NQ (Kwiatkowski et al., 2019)	General	Internet	Commonsense, Logical	Deductive	307,373	Question/Text	Natural Language
Multi-hop QA	*2WikiMultiHopQA (Ho et al., 2020)*	General	Internet	Commonsense, Logical	Deductive	192,606	Question/Text	Natural Language
	HotpotQA (Yang et al., 2018)	General	Internet	Commonsense	Deductive	113,000	Question/Text	Natural Language
	MuSiQue (Trivedi et al., 2022)	General	Previous Resource, Internet	Commonsense, Logical	Deductive	25,000	Question/Text	Natural Language
Multi-choice QA	*QuALITY (Pang et al., 2022)*	Narrative	Books	Commonsense, Logical	Deductive, Abductive	6,737	Question/Text, Options	Options
	MMLU-Pro (Wang et al., 2025b)	Science	Previous Resource, Internet	Arithmetic, Commonsense, Logical	Deductive, Inductive	12,032	Question/Text, Options	Natural Langue, Number, Options
Math	*MATH (Hendrycks et al., 2021)*	Math	Exam	Arithmetic, Logic	Deductive	12,500	Question/Text, Figure, Equation	Natural Langue, Number
	AQuA (Ling et al., 2017)	Math	Exam, Internet, Previous Resource	Arithmetic, Logic	Deductive	100,000	Question/Text, Options, Equation	Natural Langue, Options
Code	*Refactoring Oracle (Tsantalis et al., 2020)*	Software	Internet, Human	Logical	Deductive	7,226	Code, Instruction	Code
	LiveCodeBench (Jain et al., 2024)	Contest	Internet	Logical	Deductive, Abductive	500+	Question/Text, Code, Instruction	Code, Test Output

Table 1: Overview of representative knowledge and reasoning intensive benchmarks by task category.

(Liang et al., 2024). Additionally, enhancing the agentic capabilities of these models through hybrid-modal chain-of-thought reasoning is crucial, enabling interaction with the real world via multi-modal search tools (Wang et al., 2025a). Concurrently, developing unified multimodal retrievers that can jointly embed images, tables, text, and heterogeneous documents is essential.

• **Retrieval Trustworthiness.** Synergized RAG-Reasoning systems remain vulnerable to adversarial attacks through poisoned or misleading external knowledge sources. Ensuring the trustworthiness of retrieved content is therefore crucial for maintaining fully reliable downstream reasoning (Huang et al., 2024). Techniques like watermarking and digital fingerprinting have been employed to enhance system traceability. However, there's a pressing need to develop more dynamic and adaptive methods that can keep pace with the evolving landscape of LLMs, emerging attack techniques, and shifting model contexts (Liu et al., 2024). Existing studies have also individually explored uncertainty quantification and robust generation to bolster system reliability (Shorinwa et al., 2025). Future research should aim to integrate these approaches, as their combination can mutually reinforce system robustness and trustworthiness. Moreover, future efforts should also focus on extending current benchmarks to encompass multi-dimensional trust metrics beyond mere accuracy.

8 Conclusion

This survey charts the rapid convergence of retrieval and LLM reasoning. We reviewed three evolutionary stages: (1) Reasoning-Enhanced RAG, which uses multi-step reasoning to refine each stage of RAG; (2) RAG-Enhanced Reasoning, which leverages retrieved knowledge to bridge factual gaps during long CoT; and (3) Synergized RAG-Reasoning systems, where single- or multi-agents iteratively refine both search and reasoning, exemplified by "Deep Research". Collectively, these lines demonstrate that tight retrieval–reasoning coupling improves factual grounding, logical coherence, and adaptability beyond one-way enhancement. Looking forward, we identify research avenues toward synergized RAG-Reasoning systems that are more effective, multimodally-adaptive, trustworthy, and human-centric.

Limitations

While this survey synthesizes over 200 research papers across RAG and reasoning with large language models, its scope favors breadth over depth. In striving to provide a unified and comprehensive taxonomy, we may not delve deeply into the technical nuances or implementation details of individual methods-especially within specialized subfields of either RAG (e.g., sparse vs. dense retrieval, memory-augmented retrievers) or reasoning (e.g., formal logic solvers, symbolic methods, or long-context reasoning). Moreover, our categorization framework (reasoning-enhanced RAG, RAG-enhanced reasoning, and synergized RAG and reasoning) abstracts across diverse methodologies. While this facilitates a high-level understanding of design patterns, it may obscure the finer-grained trade-offs, assumptions, and limitations unique to each class of approach.

Acknowledgment

Hai-Tao Zheng and Yangning Li is supported by National Natural Science Foundation of China(Grant No.62276154), Research Center for ComputerNetwork (Shenzhen) Ministry of Education, the Natural Science Foundation of Guangdong Province(Grant No.2023A1515012914 and 440300241033100801770), Basic Research Fund of Shenzhen City (Grant No.JCYJ20210324120012033, JCYJ20240813112009013 and GJHZ20240218113603006), the Major Key Projectof PCL for Experiments and Applications (PCL2023A09). This work is also supported in part by NSF under grants III-2106758, and POSE-2346158.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. Topiocqa: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. Crisismmd: Multimodal twitter datasets from natural disasters. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

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

Anonymous. 2025. DynQR: Dynamic uncertainty-guided query rewriting for effective retrieval-augmented generation. In *Submitted to ACL Rolling Review - December 2024*. Under review.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-RAG: Self-reflective retrieval augmented generation. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

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

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

Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. 2012. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

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

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

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

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

Yanfei Chen, Jinsung Yoon, Devendra Sachan, Qingze Wang, Vincent Cohen-Addad, Mohammadhossein Bateni, Chen-Yu Lee, and Tomas Pfister. 2024a. Re-invoke: Tool invocation rewriting for zero-shot tool retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4705–4726.

Zehui Chen, Kuikun Liu, Qiuchen Wang, Jiangning Liu, Wenwei Zhang, Kai Chen, and Feng Zhao. 2024b. Mindsearch: Mimicking human minds elicits deep ai searcher. *arXiv preprint arXiv:2407.20183*.

Zhongwu Chen, Chengjin Xu, Dingmin Wang, Zhen Huang, Yong Dou, Xuhui Jiang, and Jian Guo. 2024c. Rulerag: Rule-guided retrieval-augmented generation with language models for question answering. *arXiv preprint arXiv:2410.22353*.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu

Wei, Weiwei Deng, and Qi Zhang. 2023. Uprise: Universal prompt retrieval for improving zero-shot evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12318–12337.

Rong Cheng, Jinyi Liu, Yan Zheng, Fei Ni, Jiazhen Du, Hangyu Mao, Fuzheng Zhang, Bo Wang, and Jianye Hao. 2025. Dualrag: A dual-process approach to integrate reasoning and retrieval for multi-hop question answering. *arXiv preprint arXiv:2504.18243*.

Zheng Chu, Jingchang Chen, Qianglong Chen, Haotian Wang, Kun Zhu, Xiyuan Du, Weijiang Yu, Ming Liu, and Bing Qin. 2024. Beamaggr: Beam aggregation reasoning over multi-source knowledge for multi-hop question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1229–1248.

Debrup Das, Debopriyo Banerjee, Somak Aditya, and Ashish Kulkarni. 2024. Mathsensei: A tool-augmented large language model for mathematical reasoning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 942–966.

Chao Deng, Jiale Yuan, Pi Bu, Peijie Wang, Zhong-Zhi Li, Jian Xu, Xiao-Hui Li, Yuan Gao, Jun Song, Bo Zheng, et al. 2024. Longdocurl: a comprehensive multimodal long document benchmark integrating understanding, reasoning, and locating. *arXiv preprint arXiv:2412.18424*.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 3563–3578.

Avik Dutta, Mukul Singh, Gust Verbruggen, Sumit Gulwani, and Vu Le. 2024. Rar: Retrieval-augmented retrieval for code generation in low resource languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21506–21515.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024a. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6491–6501.

Yue Fan, Hu Zhang, Ru Li, Yujie Wang, Hongye Tan, and Jiye Liang. 2024b. Frva: Fact-retrieval and verification augmented entailment tree generation for explainable question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9111–9128.

Jinyuan Fang, Zaiqiao Meng, and Craig Macdonald. 2024. Trace the evidence: Constructing knowledge-grounded reasoning chains for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8472–8494.

Weizhi Fei, Xueyan Niu, Guoqing Xie, Yanhua Zhang, Bo Bai, Lei Deng, and Wei Han. 2024. Retrieval meets reasoning: Dynamic in-context editing for long-text understanding. *arXiv preprint arXiv:2406.12331*.

Wenfeng Feng, Chuzhan Hao, Yuewei Zhang, Jingyi Song, and Hao Wang. 2025. Airrag: Activating intrinsic reasoning for retrieval augmented generation via tree-based search. *arXiv preprint arXiv:2501.10053*.

James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi. 2020. Iirc: A dataset of incomplete information reading comprehension questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1137–1147.

Zafeirios Fountas, Martin A Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. 2024. Human-like episodic memory for infinite context llms. *arXiv preprint arXiv:2407.09450*.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023a. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

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

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Xinyan Guan, Jiali Zeng, Fandong Meng, Chunlei Xin, Yaojie Lu, Hongyu Lin, Xianpei Han, Le Sun, and Jie Zhou. 2025. Deeprag: Thinking to retrieval step by step for large language models. *arXiv preprint arXiv:2502.01142*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Siwei Han, Peng Xia, Ruiyi Zhang, Tong Sun, Yun Li, Hongtu Zhu, and Huaxiu Yao. 2025. Mdocagent: A multi-modal multi-agent framework for document understanding. *arXiv preprint arXiv:2503.13964*.

Shibo Hao, Tianyang Liu, Zhen Wang, and Zhiting Hu. 2023. Toolkengpt: Augmenting frozen language models with massive tools via tool embeddings. In *Advances in Neural Information Processing Systems*, volume 36, pages 45870–45894.

Bolei He, Nuo Chen, Xinran He, Lingyong Yan, Zhenkai Wei, Jinchang Luo, and Zhen-Hua Ling. 2024a. Retrieving, rethinking and revising: The chain-of-verification can improve retrieval augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10371–10393.

Jiashu He, Mingyu Derek Ma, Jinxuan Fan, Dan Roth, Wei Wang, and Alejandro Ribeiro. Give: Structured reasoning of large language models with knowledge graph inspired veracity extrapolation. In *Forty-second International Conference on Machine Learning*.

Jie He, Nan Hu, Wanqiu Long, Jiaoyan Chen, and Jeff Z Pan. 2024b. Mintqa: A multi-hop question answering benchmark for evaluating llms on new and tail knowledge. *arXiv preprint arXiv:2412.17032*.

Xiaoxin He, Yijun Tian, Yifei Sun, Nitesh Chawla, Thomas Laurent, Yann LeCun, Xavier Bresson, and Bryan Hooi. 2024c. G-retriever: Retrieval-augmented generation for textual graph understanding and question answering. *Advances in Neural Information Processing Systems*, 37:132876–132907.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Chan-Wei Hu, Yueqi Wang, Shuo Xing, Chia-Ju Chen, and Zhengzhong Tu. 2025a. mrag: Elucidating the design space of multi-modal retrieval-augmented generation. *arXiv preprint arXiv:2505.24073*.

Minda Hu, Licheng Zong, Hongru Wang, Jingyan Zhou, Jingjing Li, Yichen Gao, Kam-Fai Wong, Yu Li, and Irwin King. 2024. Serts: Self-rewarding tree search for biomedical retrieval-augmented generation. *arXiv preprint arXiv:2406.11258*.

Yunhai Hu, Yilun Zhao, Chen Zhao, and Arman Cohan. 2025b. Mcts-rag: Enhancing retrieval-augmented generation with monte carlo tree search. *arXiv preprint arXiv:2503.20757*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. 2024. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *Artificial Intelligence Review*, 57(7):175.

Yulong Hui, Yao Lu, and Huanchen Zhang. 2024. Uda: A benchmark suite for retrieval augmented generation in real-world document analysis. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Michael Iannelli, Sneha Kuchipudi, and Vera Dvorak. 2024. Sla management in reconfigurable multi-agent rag: A systems approach to question answering. *arXiv preprint arXiv:2412.06832*.

Shayekh Islam, Md Asib Rahman, KSM Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-rag: Enhanced retrieval augmented reasoning with open-source large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14231–14244.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7029–7043.

Yixin Ji, Kaixin Wu, Juntao Li, Wei Chen, Mingjie Zhong, Xu Jia, and Min Zhang. 2024. Retrieval and reasoning on kgs: Integrate knowledge graphs into large language models for complex question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7598–7610.

Mingyi Jia, Junwen Duan, Yan Song, and Jianxin Wang. 2025. Find: Fine-grained information density guided adaptive retrieval-augmented generation for disease diagnosis. *arXiv preprint arXiv:2502.14614*.

Jinhao Jiang, Jiayi Chen, Junyi Li, Ruiyang Ren, Shijie Wang, Wayne Xin Zhao, Yang Song, and Tao Zhang. 2024. Rag-star: Enhancing deliberative reasoning with retrieval augmented verification and refinement. *arXiv preprint arXiv:2412.12881*.

Pengcheng Jiang, Jiacheng Lin, Lang Cao, Runchu Tian, SeongKu Kang, Zifeng Wang, Jimeng Sun, and Jiawei Han. 2025. Deepretrieval: Hacking real search engines and retrievers with large language models via reinforcement learning. *arXiv preprint arXiv:2503.00223*.

Bowen Jin, Chulin Xie, Jiawei Zhang, Kashob Kumar Roy, Yu Zhang, Zheng Li, Ruirui Li, Xianfeng Tang, Suhang Wang, Yu Meng, et al. 2024. Graph chain-of-thought: Augmenting large language models by reasoning on graphs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 163–184.

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

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Tomoyuki Kagaya, Thong Jing Yuan, Yuxuan Lou, Jayashree Karlekar, Sugiri Pranata, Akira Kinose, Koki Oguri, Felix Wick, and Yang You. 2024. Rap: Retrieval-augmented planning with contextual memory for multimodal llm agents. *arXiv preprint arXiv:2402.03610*.

Mohammed Khaliq, Paul Chang, Mingyang Ma, Bernhard Pflugfelder, and Filip Miletić. 2024. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 280–296.

Gangwoo Kim, Sungdong Kim, Byeongguk Jeon, Joonsuk Park, and Jaewoo Kang. 2023. Tree of clarifications: Answering ambiguous questions with retrieval-augmented large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 996–1009.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7740–7754.

Heiko Koziolek, Sten Grüner, Rhaban Hark, Virendra Ashiwal, Sofia Linsbauer, and Nafise Eskandani. 2024. Llm-based and retrieval-augmented control code generation. In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 22–29.

Satyapriya Krishna, Kalpesh Krishna, Anhad Mohananey, Steven Schwarcz, Adam Stambler, Shyam Upadhyay, and Manaal Faruqui. 2024. Fact, fetch, and reason: A unified evaluation of retrieval-augmented generation. *arXiv preprint arXiv:2409.12941*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Sung-Min Lee, Eunhwan Park, Donghyeon Jeon, Inho Kang, and Seung-Hoon Na. 2024. Radcot: Retrieval-augmented distillation to specialization models for generating chain-of-thoughts in query expansion. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13514–13523.

Zhicheng Lee, Shulin Cao, Jinxin Liu, Jiajie Zhang, Weichuan Liu, Xiaoyin Che, Lei Hou, and Juanzi Li. 2025. Rearag: Knowledge-guided reasoning enhances factuality of large reasoning models with iterative retrieval augmented generation. *arXiv preprint arXiv:2503.21729*.

Dawei Li, Shu Yang, Zhen Tan, Jae Baik, Sukwon Yun, Joseph Lee, Aaron Chacko, Bojian Hou, Duy Duong-Tran, Ying Ding, et al. 2024a. Dalk: Dynamic co-augmentation of llms and kg to answer alzheimer's disease questions with scientific literature. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2187–2205.

Guanghua Li, Wensheng Lu, Wei Zhang, Defu Lian, Kezhong Lu, Rui Mao, Kai Shu, and Hao Liao. 2024b. Re-search for the truth: Multi-round retrieval-augmented large language models are strong fake news detectors. *arXiv preprint arXiv:2403.09747*.

Guozheng Li, Peng Wang, Wenjun Ke, Yikai Guo, Ke Ji, Ziyu Shang, Jiajun Liu, and Zijie Xu. 2024c. Recall, retrieve and reason: towards better in-context relation extraction. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 6368–6376.

Huayang Li, Pat Verga, Priyanka Sen, Bowen Yang, Vijay Viswanathan, Patrick Lewis, Taro Watanabe, and Yixuan Su. 2024d. Alr2: A retrieve-then-reason framework for long-context question answering. *arXiv preprint arXiv:2410.03227*.

Jia Li, Xianjie Shi, Kechi Zhang, Lei Li, Ge Li, Zhengwei Tao, Fang Liu, Chongyang Tao, and Zhi Jin. 2025a. Coderag: Supportive code retrieval on bigraph for real-world code generation. *arXiv preprint arXiv:2504.10046*.

Minghan Li, Honglei Zhuang, Kai Hui, Zhen Qin, Jimmy Lin, Rolf Jagerman, Xuanhui Wang, and Michael Bendersky. 2024e. Can query expansion improve generalization of strong cross-encoder rankers? In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2321–2326.

Shilong Li, Yancheng He, Hangyu Guo, Xingyuan Bu, Ge Bai, Jie Liu, Jiaheng Liu, Xingwei Qu, Yangguang Li, Wanli Ouyang, et al. 2024f. Graphreader: Building graph-based agent to enhance long-context abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12758–12786.

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

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

Yangning Li, Yinghui Li, Xinyu Wang, Yong Jiang, Zhen Zhang, Xinran Zheng, Hui Wang, Hai-Tao Zheng, Fei Huang, Jingren Zhou, and Philip S. Yu. 2025d. Benchmarking multimodal retrieval augmented generation with dynamic VQA dataset and self-adaptive planning agent. In *The Thirteenth International Conference on Learning Representations*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zhi Li, Yicheng Li, Hequan Ye, and Yin Zhang. 2024g. Towards autonomous tool utilization in language models: A unified, efficient and scalable framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16422–16432.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025e. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Zhuoqun Li, Xuanang Chen, Haiyang Yu, Hongyu Lin, Yaojie Lu, Qiaoyu Tang, Fei Huang, Xianpei Han, Le Sun, and Yongbin Li. 2024h. Structrag: Boosting knowledge intensive reasoning of llms via inference-time hybrid information structurization. *arXiv preprint arXiv:2410.08815*.

Zijing Liang, Yanjie Xu, Yifan Hong, Penghui Shang, Qi Wang, Qiang Fu, and Ke Liu. 2024. A survey of multimodel large language models. In *Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering*, pages 405–409.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. 2023. Ra-dit: Retrieval-augmented dual instruction tuning. In *The Twelfth International Conference on Learning Representations*.

Wang Ling, Dani Yogatama, Chris Dyer, and Phil Blunsom. 2017. Program induction by rationale generation: Learning to solve and explain algebraic word problems. *arXiv preprint arXiv:1705.04146*.

Aiwei Liu, Leyi Pan, Yijian Lu, Jingjing Li, Xuming Hu, Xi Zhang, Lijie Wen, Irwin King, Hui Xiong, and Philip Yu. 2024. A survey of text watermarking in the era of large language models. *ACM Computing Surveys*, 57(2):1–36.

Pei Liu, Xin Liu, Ruoyu Yao, Junming Liu, Siyuan Meng, Ding Wang, and Jun Ma. 2025. Hm-rag: Hierarchical multi-agent multimodal retrieval augmented generation. *arXiv preprint arXiv:2504.12330*.

Chang Han Low, Ziyue Wang, Tianyi Zhang, Zhitao Zeng, Zhu Zhuo, Evangelos B Mazomenos, and Yueming Jin. 2025. Surgraw: Multi-agent workflow with chain-of-thought reasoning for surgical intelligence. *arXiv preprint arXiv:2503.10265*.

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

Pan Lu, Swaroop Mishra, Tanglin 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. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, et al. 2025a. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

Man Luo, Xin Xu, Zhuyun Dai, Panupong Pasupat, Mehran Kazemi, Chitta Baral, Vaiva Imbrasaite, and Vincent Y Zhao. 2023. Dr. icl: Demonstration-retrieved in-context learning. *arXiv preprint arXiv:2305.14128*.

Ne Luo, Aryo Pradipta Gema, Xuanli He, Emile van Krieken, Pietro Lesci, and Pasquale Minervini. 2025b. Self-training large language models for tool-use without demonstrations. *arXiv preprint arXiv:2502.05867*.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. 2024a. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. *arXiv preprint arXiv:2407.10805*.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, and Aixin Sun. 2024b. Sciagent: Tool-augmented language models for scientific reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15701–15736.

Yubo Ma, Yuhang Zang, Liangyu Chen, Meiqi Chen, Yizhu Jiao, Xinze Li, Xinyuan Lu, Ziyu Liu, Yan Ma, Xiaoyi Dong, et al. 2025. Mmlongbench-doc: Benchmarking long-context document understanding with visualizations. *Advances in Neural Information Processing Systems*, 37:95963–96010.

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

Maria Marina, Nikolay Ivanov, Sergey Pletenev, Mikhail Salnikov, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Alexander Panchenko, and Viktor Moskvoretskii. 2025. Llm-independent adaptive rag: Let the question speak for itself. *arXiv preprint arXiv:2505.04253*.

Costas Mavromatis and George Karypis. 2024. Gnn-rag: Graph neural retrieval for large language model reasoning. *arXiv preprint arXiv:2405.20139*.

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. In *The Twelfth International Conference on Learning Representations*.

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

Shashi Narayan, Shay B Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807.

Xuan-Phi Nguyen, Shrey Pandit, Senthil Purushwalkam, Austin Xu, Hailin Chen, Yifei Ming, Zixuan Ke, Silvio Savarese, Caiming Xong, and Shafiq Joty. 2024. Sfr-rag: Towards contextually faithful llms. *arXiv preprint arXiv:2409.09916*.

Cheng Niu, Yang Guan, Yuanhao Wu, Juno Zhu, Juntong Song, Randy Zhong, Kaihua Zhu, Siliang Xu, Shizhe Diao, and Tong Zhang. 2024. Veract scan: Retrieval-augmented fake news detection with justifiable reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 266–277.

Yasumasa Onoe, Michael J.Q. Zhang, Eunsol Choi, and Greg Durrett. 2021. Creak: A dataset for commonsense reasoning over entity knowledge. *OpenReview*.

Richard Yuanzhe Pang, Alicia Parrish, Nitish Joshi, Nikita Nangia, Jason Phang, Angelica Chen, Vishakh Padmakumar, Johnny Ma, Jana Thompson, He He, et al. 2022. Quality: Question answering with long input texts, yes! In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5336–5358.

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

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

Shuofei Qiao, Honghao Gui, Chengfei Lv, Qianghuai Jia, Huajun Chen, and Ningyu Zhang. 2024. Making language models better tool learners with execution feedback. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3550–3568.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, et al. 2023. Toolllm: Facilitating large language models to master 16000+ real-world apis. *arXiv preprint arXiv:2307.16789*.

Leonardo Ranaldi, Marco Valentino, and Andrè Freitas. 2024. Eliciting critical reasoning in retrieval-augmented language models via contrastive explanations. *arXiv preprint arXiv:2410.22874*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Aniruddha Salve, Saba Attar, Mahesh Deshmukh, Sayali Shivpuje, and Arnab Mitra Utsab. 2024. A collaborative multi-agent approach to retrieval-augmented generation across diverse data. *arXiv preprint arXiv:2412.05838*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023.

Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36:68539–68551.

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

Thomas Schmied, Fabian Paischer, Vihang Patil, Markus Hofmarcher, Razvan Pascanu, and Sepp Hochreiter. 2024. Retrieval-augmented decision transformer: External memory for in-context rl. *arXiv preprint arXiv:2410.07071*.

Johannes Schneider. 2025. Generative to agentic ai: Survey, conceptualization, and challenges. *arXiv preprint arXiv:2504.18875*.

Eva Sharma, Chen Li, and Lu Wang. 2019. Bigpatent: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*.

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

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

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

Chuanneng Sun, Songjun Huang, and Dario Pompili. 2024a. Retrieval-augmented hierarchical in-context reinforcement learning and hindsight modular reflections for task planning with llms. *arXiv preprint arXiv:2408.06520*.

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2380–2390.

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

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. 2024b. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *The Twelfth International Conference on Learning Representations*.

Qiang Sun, Tingting Bi, Sirui Li, Eun-Jung Holden, Paul Duuring, Kai Niu, and Wei Liu. 2025b. Symbioticrag: Enhancing document intelligence through human-llm symbiotic collaboration. *arXiv preprint arXiv:2505.02418*.

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

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 641–651.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Yicheng Tao, Haotian Liu, Shanwen Wang, and Hongteng Xu. 2025. Assisting mathematical formalization with a learning-based premise retriever. *arXiv preprint arXiv:2501.13959*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

SM Tonmoy, SM Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *arXiv preprint arXiv:2401.01313*.

Hieu Tran, Zonghai Yao, Junda Wang, Yifan Zhang, Zhichao Yang, and Hong Yu. 2024. Rare: Retrieval-augmented reasoning enhancement for large language models. *arXiv preprint arXiv:2412.02830*.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. ♫ musique: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554.

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

Nikolaos Tsantalis, Ameya Ketkar, and Danny Dig. 2020. Refactoringminer 2.0. *IEEE Transactions on Software Engineering*, 48(3):930–950.

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

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

Junjie Wang, Mingyang Chen, Binbin Hu, Dan Yang, Ziqi Liu, Yue Shen, Peng Wei, Zhiqiang Zhang, Jinjie Gu, Jun Zhou, et al. 2024b. Learning to plan for retrieval-augmented large language models from knowledge graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7813–7835.

Song Wang, Zihan Chen, Chengshuai Shi, Cong Shen, and Jundong Li. 2024c. Mixture of demonstrations for in-context learning. *Advances in Neural Information Processing Systems*, 37:88091–88116.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. 2025a. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Yu Wang, Nedim Lipka, Ryan A Rossi, Alexa Siu, Ruiyi Zhang, and Tyler Derr. 2024d. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19206–19214.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2025b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Yujing Wang, Hainan Zhang, Liang Pang, Binghui Guo, Hongwei Zheng, and Zhiming Zheng. 2025c. Maferw: Query rewriting with multi-aspect feedbacks for retrieval-augmented large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25434–25442.

Zheng Wang, Shu Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024e. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1966–1978.

Zhengren Wang, Jiayang Yu, Dongsheng Ma, Zhe Chen, Yu Wang, Zhiyu Li, Feiyu Xiong, Yanfeng Wang, Linpeng Tang, Wentao Zhang, et al. 2025d. Rare: Retrieval-augmented reasoning modeling. *arXiv preprint arXiv:2503.23513*.

Zihao Wang, Shaofei Cai, Anji Liu, Yonggang Jin, Jinbing Hou, Bowei Zhang, Haowei Lin, Zhaofeng He, Zilong Zheng, Yaodong Yang, et al. 2024f. Jarvis-1: Open-world multi-task agents with memory-augmented multimodal language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zihao Wang, Anji Liu, Haowei Lin, Jiaqi Li, Xiaojian Ma, and Yitao Liang. 2024g. Rat: Retrieval augmented thoughts elicit context-aware reasoning in long-horizon generation. *arXiv preprint arXiv:2403.05313*.

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. *arXiv preprint arXiv:2411.04368*.

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

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

Jiaqi Wei, Hao Zhou, Xiang Zhang, Di Zhang, Zijie Qiu, Wei Wei, Jinzhe Li, Wanli Ouyang, and Siqi Sun. 2025b. Alignrag: An adaptable framework for resolving misalignments in retrieval-aware reasoning of rag. *arXiv preprint arXiv:2504.14858*.

Zhihua Wen, Zhiliang Tian, Wei Wu, Yuxin Yang, Yanqi Shi, Zhen Huang, and Dongsheng Li. 2023. Grove: A retrieval-augmented complex story generation framework with a forest of evidence. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3980–3998.

Nirmalie Wiratunga, Ramitha Abeyratne, Lasal Jayawardena, Kyle Martin, Stewart Massie, Ikechukwu Nkisi-Orji, Ruvan Weerasinghe, Anne Liret, and Bruno Fleisch. 2024. Cbr-rag: case-based reasoning for retrieval augmented generation in llms for legal question answering. In *International Conference on Case-Based Reasoning*, pages 445–460. Springer.

Feijie Wu, Zitao Li, Fei Wei, Yaliang Li, Bolin Ding, and Jing Gao. 2025a. Talk to right specialists: Routing and planning in multi-agent system for question answering. *arXiv preprint arXiv:2501.07813*.

Jialong Wu, Wenbiao Yin, Yong Jiang, Zhenglin Wang, Zekun Xi, Runnan Fang, Linhai Zhang, Yulan He, Deyu Zhou, Pengjun Xie, et al. 2025b. Webwalker: Benchmarking llms in web traversal. *arXiv preprint arXiv:2501.07572*.

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

Shirley Wu, Shiyu Zhao, Qian Huang, Kexin Huang, Michihiro Yasunaga, Kaidi Cao, Vassilis Ioannidis, Karthik Subbian, Jure Leskovec, and James Y Zou. 2024. Avatar: Optimizing llm agents for tool usage via contrastive reasoning. *Advances in Neural Information Processing Systems*, 37:25981–26010.

Heming Xia, Yongqi Li, Chak Tou Leong, Wenjie Wang, and Wenjie Li. 2025a. Tokenskip: Controllable chain-of-thought compression in llms. *arXiv preprint arXiv:2502.12067*.

Yuan Xia, Jingbo Zhou, Zhenhui Shi, Jun Chen, and Haifeng Huang. 2025b. Improving retrieval augmented language model with self-reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 25534–25542.

Guangzhi Xiong, Qiao Jin, Xiao Wang, Yin Fang, Haolin Liu, Yifan Yang, Fangyuan Chen, Zhixing Song, Dengyu Wang, Minjia Zhang, et al. 2025. Rag-gym: Optimizing reasoning and search agents with process supervision. *arXiv preprint arXiv:2502.13957*.

Kehan Xu, Kun Zhang, Jingyuan Li, Wei Huang, and Yuanzhuo Wang. 2024. Crp-rag: A retrieval-augmented generation framework for supporting complex logical reasoning and knowledge planning. *Electronics*, 14(1):47.

Kehan Xu, Kun Zhang, Jingyuan Li, Wei Huang, and Yuanzhuo Wang. 2025a. Crp-rag: A retrieval-augmented generation framework for supporting complex logical reasoning and knowledge planning. *Electronics (2079-9292)*, 14(1).

Ran Xu, Wenqi Shi, Yuchen Zhuang, Yue Yu, Joyce C Ho, Haoyu Wang, and Carl Yang. 2025b. Collab-rag: Boosting retrieval-augmented generation for complex question answering via white-box and black-box llm collaboration. *arXiv preprint arXiv:2504.04915*.

Chen Yang, Chenyang Zhao, Quanquan Gu, and Dongruo Zhou. 2024a. Cops: Empowering llm agents with provable cross-task experience sharing. *arXiv preprint arXiv:2410.16670*.

Rui Yang. 2024. Casegpt: a case reasoning framework based on language models and retrieval-augmented generation. *arXiv preprint arXiv:2407.07913*.

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

Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Gui, Ziran Jiang, Ziyu Jiang, et al. 2024b. Crag-comprehensive rag benchmark. *Advances in Neural Information Processing Systems*, 37:10470–10490.

Yahe Yang and Chengyue Huang. 2025. Tree-based rag-agent recommendation system: A case study in medical test data. *arXiv preprint arXiv:2501.02727*.

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

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

Shunyu Yao, Howard Chen, John Yang, and Karthik Narasimhan. 2022. Webshop: Towards scalable real-world web interaction with grounded language agents. *Advances in Neural Information Processing Systems*, 35:20744–20757.

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

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

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Jaeseok Yoo, Hojae Han, Youngwon Lee, Jaejin Kim, and Seung-won Hwang. 2025. Perc: Plan-as-query example retrieval for underrepresented code generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7982–7997.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *ICLR 2024 Workshop on Large Language Model (LLM) Agents*.

Hong Qing Yu and Frank McQuade. 2025. Rag-kg-il: A multi-agent hybrid framework for reducing hallucinations and enhancing llm reasoning through rag and incremental knowledge graph learning integration. *arXiv preprint arXiv:2503.13514*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14672–14685.

Jing Zhang, Xiaokang Zhang, Jifan Yu, Jian Tang, Jie Tang, Cuiping Li, and Hong Chen. 2022a. Subgraph retrieval enhanced model for multi-hop knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5773–5784.

Jinghan Zhang, Xiting Wang, Weijieying Ren, Lu Jiang, Dongjie Wang, and Kunpeng Liu. 2025a. Ratt: A thought structure for coherent and correct llm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26733–26741.

Jintian Zhang, Yuqi Zhu, Mengshu Sun, Yujie Luo, Shuofei Qiao, Lun Du, Da Zheng, Huajun Chen, and Ningyu Zhang. 2025b. Lightthinker: Thinking step-by-step compression. *arXiv preprint arXiv:2502.15589*.

Miao Zhang, Zhenlong Fang, Tianyi Wang, Qian Zhang, Shuai Lu, Junfeng Jiao, and Tianyu Shi. 2025c. A cascading cooperative multi-agent framework for on-ramp merging control integrating large language models. *arXiv preprint arXiv:2503.08199*.

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

Tianjun Zhang, Shishir G Patil, Naman Jain, Sheng Shen, Matei Zaharia, Ion Stoica, and Joseph E Gonzalez. 2024a. Raft: Adapting language model to domain specific rag. In *First Conference on Language Modeling*.

Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, et al. 2025e. Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap. *arXiv preprint arXiv:2501.01945*.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, et al. 2025f. From web search towards agentic deep research: Incentivizing search with reasoning agents. *arXiv preprint arXiv:2506.18959*.

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

Xikun Zhang, Antoine Bosselut, Michihiro Yasunaga, Hongyu Ren, Percy Liang, Christopher D Manning, and Jure Leskovec. 2022b. Greaselm: Graph reasoning enhanced language models. In *International Conference on Learning Representations*.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, et al. 2024b. ∞ bench: Extending long context evaluation beyond 100k tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277.

Yize Zhang, Tianshu Wang, Sirui Chen, Kun Wang, Xingyu Zeng, Hongyu Lin, Xianpei Han, Le Sun, and Chaochao Lu. 2025h. ARise: Towards knowledge-augmented reasoning via risk-adaptive search. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10978–10995, Vienna, Austria. Association for Computational Linguistics.

Yusen Zhang, Ruoxi Sun, Yanfei Chen, Tomas Pfister, Rui Zhang, and Sercan Arik. 2024c. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237.

Zhebin Zhang, Xinyu Zhang, Yuanhang Ren, Saijiang Shi, Meng Han, Yongkang Wu, Ruofei Lai, and Zhao Cao. 2023. Iag: Induction-augmented generation framework for answering reasoning questions. *arXiv preprint arXiv:2311.18397*.

Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. 2024a. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924*.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024b.

Pacar: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12564–12573.

Xinping Zhao, Dongfang Li, Yan Zhong, Boren Hu, Yibin Chen, Baotian Hu, and Min Zhang. 2024c. Seer: Self-aligned evidence extraction for retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3027–3041.

Kunhao Zheng, Jesse Michael Han, and Stanislas Polu. 2021. Minif2f: a cross-system benchmark for formal olympiad-level mathematics. *arXiv preprint arXiv:2109.00110*.

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 preprint arXiv:2504.03160*.

Jiawei Zhou and Lei Chen. 2025. Openrag: Optimizing rag end-to-end via in-context retrieval learning. *arXiv preprint arXiv:2503.08398*.

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

Yifei Zhou, Song Jiang, Yuandong Tian, Jason Weston, Sergey Levine, Sainbayar Sukhbaatar, and Xian Li. 2025b. Sweet-rl: Training multi-turn llm agents on collaborative reasoning tasks. *arXiv preprint arXiv:2503.15478*.

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

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

A Full Benchmark

Section 6 introduces representative benchmarks for different RAG-reasoning tasks. This appendix complements that discussion with a comprehensive list of benchmarks organized by task and domain. Table 2 details each benchmark's attributes, including the publication venue, code repository, task category, domain, primary knowledge sources, knowledge type, and reasoning capabilities. By consolidating these attributes into a single table, we facilitate the selection and comparison of benchmarks, enabling researchers to identify the most suitable datasets for future studies on RAG-enhanced reasoning.

Our benchmark compilation is primarily derived from the methods surveyed in Sections 3 to 5 of this paper, with a particular focus on synergized approaches discussed in Section 5. We deliberately targeted benchmarks that require both external knowledge retrieval and internal deep reasoning, as this dual requirement reflects real-world scenarios where models must not only access relevant information but also integrate and reason over it effectively. For example, in the QA domain, we include datasets that necessitate synthesizing evidence across multiple documents to answer questions that cannot be resolved through single-sentence retrieval. HotpotQA (Yang et al., 2018) exemplifies this challenge, requiring reasoning across different Wikipedia articles. In coding tasks, benchmarks such as LiveCodeBench (Jain et al., 2024) and Refactoring Oracle (Tsantalis et al., 2020) extend beyond pure algorithmic problem-solving by demanding retrieval of external code snippets and documentation. Similarly, in mathematics, benchmarks like MATH (Hendrycks et al., 2021) and AQUA-RAT (Das et al., 2024) assess not only computational proficiency but also the retrieval of relevant theorems and formulas, testing the model's ability to integrate external mathematical knowledge with internal reasoning processes.

In addition to established benchmarks, we have incorporated newer and more challenging datasets that better mirror real-world applications. These datasets often demand extensive retrieval processes combined with expert-level or domain-specific reasoning, as seen in Humanity's Last Exam (HLE) (Phan et al., 2025) and web search evaluation tasks like BrowseComp (Wei et al., 2025a). Overall, our collection encompasses **46** benchmarks covering **13** distinct tasks across **12** domains, each explicitly annotated with features such as knowledge source, knowledge type, and reasoning capacity. This breadth ensures coverage of diverse domains and task types, forming a solid foundation for evaluating the interplay between retrieval and reasoning in RAG systems.

Within this benchmark set, single-hop QA datasets like TriviaQA (Joshi et al., 2017) focus on precise retrieval and fact recall, requiring models to locate and synthesize a single piece of evidence. In contrast, multi-hop QA benchmarks such as HotpotQA (Yang et al., 2018) and MuSiQue (Trivedi et al., 2022) challenge models to chain information from multiple documents and employ deductive reasoning to bridge disparate facts into coherent answers. Structured knowledge benchmarks, such as GraphQA (He et al., 2024c), require reasoning over relational graph representations, integrating nodes and edges to resolve complex queries beyond plain text retrieval. Complementing these open-ended tasks, multiple-choice evaluations like MMLU-Pro (Wang et al., 2025b) test domain-specific knowledge in areas such as science, history, or law, assessing the model's ability to perform various reasoning styles, including inductive and abductive inference. Multimodal QA benchmarks, like WebShop (Yao et al., 2022), test a model's capacity to align textual and visual information to determine the correct answer. Long-form QA datasets such as ∞BENCH (Zhang et al., 2024b) evaluate models' ability to maintain logical consistency and perform inductive reasoning over lengthy contexts. Collectively, these benchmarks establish a comprehensive evaluation chain for systematically assessing RAG-reasoning capabilities.

Beyond text-based QA, RAG-augmented benchmarks span diverse tasks involving long-form generation, interactive reasoning, and domain-specific challenges in mathematics and programming. Mathematics benchmarks such as MATH (Hendrycks et al., 2021) draw from competition-level problems to assess arithmetic and symbolic reasoning. Summarization tasks like XSum (Narayan et al., 2018) evaluate a model's ability to condense entire news articles into concise summaries while preserving factual correctness. Fact-checking benchmarks, such as FEVER (Thorne et al., 2018), test the capacity for evidence retrieval and claim verification. Code-focused evaluations, including LiveCodeBench (Jain et al., 2024), examine deductive and abductive reasoning in the context of algo-

Dataset	*Venue*	Resource	Task	Domain	Knowledge Source	Knowledge Type	Reasoning Capability	Size	Input	Output
Code										
LiveCodeBench (Jain et al., 2024)	*Arxiv'24*	Link	Code	General	Internet	Logical	Deductive, Abductive	1,055	Question/Text, Code, Instruction	Code Instance, Test Output
Refactoring Oracle (Tsantalis et al., 2020)	*IEEE'22*	Link	Code	Software	Internet, Human	Logical	Deductive	7,226	Code, Instruction	Code Instance
ColBench (Zhou et al., 2025b)	*Arxiv'25*	Link	Code	Software	LLM, Human	Logical	Abductive, Inductive	10,000+	Question/Text, Links/Sources, Code	Code Instance
Math										
MATH (Hendrycks et al., 2021)	*NeurIPS'21*	Link	Domain-specific QA	Math	Exam/Competition	Logical, Arithmetic	Deductive	12,500	Question/Text, Equations	Number, Natural Language
MiniF2F (Zheng et al., 2021)	*ICLR'22*	Link	Domain-specific QA	Math	Exam/Competition, Books	Logical, Arithmetic	Deductive	488	Question/Text, Equations	Number, Natural Language
AQuA (Ling et al., 2017)	*Arxiv'17*	Link	Domain-specific QA	Math	Previous Source, Exam/Competition, Internet	Arithmetic, Logical	Deductive	100,000	Question/Text, Options, Equations	Natural Language, Options/Labels
Fact Checking										
CRAG (Yang et al., 2024b)	*NeurIPS'24*	Link	Fact Checking	General	Internet	Commonsense	Deductive, Abductive	4,409	Question/Text	Natural Language
CREAK (Onoe et al., 2021)	*NeurIPS'21*	Link	Fact Checking	General	Human	Commonsense	Deductive, Abductive, Analogical	13,000	Question/Text	Options/Labels, Natural Language
Fever (Thorne et al., 2018)	*ACL'18*	Link	Fact Checking	General	Internet	Logical	Deductive, Abductive	185,445	Question/Text, Links/Sources	Natural Language, Options/Labels
PubHealth (Kotonya and Toni, 2020)	*EMNLP'20*	Link	Fact Checking	Health	Internet	Commonsense, Logical	Abductive, Deductive	11,800	Question/Text	Natural Language, Options
Graph QA										
GraphQA (He et al., 2024c)	*NeurIPS'24*	Link	Graph QA	General	Previous Source	Commonsense, Multimodal	Deductive, Abductive	107,503	Question/Text	Natural Language
GRBENCH (Jin et al., 2024)	*ACL'24*	Link	Graph QA	General	LLM, Human	Logical	Deductive, Inductive	1,740	Question/Text	Natural Language
Long-form QA										
∞ BENCH (Zhang et al., 2024b)	*Arxiv'24*	Link	Long-form QA	General	Internet, Human	Multimodal, Logical	Inductive, Abductive	3,946	Question/Text, Code, Equations	Natural Language, Number, Code Instance
Multimodal QA										
CrisisMMD (Alam et al., 2018)	*Arxiv'18*	Link	Multimodal QA	Crisis Response	Media, Internet	Commonsense, Multimodal	Abductive	16,097	Question/Text, Figure/Image	Options, Natural Language
ALFWORLD (Shridhar et al.)	*ICLR'21*	Link	Multimodal QA	Game	Previous Source	Multimodal	Deductive, Abductive	3,827	Question/Text, Figure/Image	Natural Language
MMLongBench-DOC (Ma et al., 2025)	*NeurIPS'24*	Link	Multimodal QA	Narrative	Previous Source, Internet	Multimodal	Deductive, Abductive	1,082	Figure/Image, Question/Text, Documents	Natural Language, Number
LongDocURL (Deng et al., 2024)	*Arxiv'24*	Link	Multimodal QA	Narrative	Internet, Previous Source, LLM	Multimodal	Deductive, Abductive	2,325	Figure/Image, Question/Text, Documents	Natural Language, Number
UDA (Hui et al., 2024)	*NIPS'24*	Link	Multimodal QA	Narrative	Internet, Paper/Report	Multimodal	Deductive	29,590	Documents, Question/Text	Natural Language, Number
SCIENCEQA (Lu et al., 2022)	*NeurIPS'22*	Link	Multimodal QA	Science	Human	Logical, Multimodal	Deductive	21,000	Question/Text, Options, Figure/Image	Options, Natural Language, Number
WebShop (Yao et al., 2022)	*NeurIPS'22*	Link	Multimodal QA	E-commerce	Internet	Multimodal	Inductive, Abductive	12,087	Instruction, Question/Text	Natural Language, Image/Figure
SurgCoTBench (Low et al., 2025)	*Arxiv'25*	—	Multimodal QA	Health	Human	Multimodal, Logical	Abductive, Deductive	14,176	Question/Text, Figure/Image, Options	Options, Natural Language, Number

Table 2: Full representative knowledge and reasoning intensive benchmarks across diverse task categories (Part 1).

Dataset	*Venue*	Resource	Task	Domain	Knowledge Source	Knowledge Type	Reasoning Capability	Size	Input	Output
Multi-choice QA										
Bamboogle (Press et al., 2023)	*EMNLP'23*	Link	Multi-choice QA	General	Internet	Logical	Deductive, Abductive	125	Question/Text	Natural Language
BIG-Bench (Srivastava et al., 2022)	*Arxiv'22*	Link	Multi-choice QA	General	Internet	Commonsense, Logical	Deductive, Abductive, Inductive, Analogical	204	Question/Text, Options	Natural Language, Number, Options/Labels
ADQA (Li et al., 2024a)	*EMNLP'24*	Link	Multi-choice QA	Health	Previous Source	Commonsense, Logical	Deductive, Abductive	446	Question/Text, Options	Options
QuALITY (Pang et al., 2022)	*NAACL'22*	Link	Multi-choice QA	Narrative	Books	Commonsense, Logical	Deductive, Abductive	6,737	Question/Text, Options	Options
MMLU-Pro (Wang et al., 2025b)	*NeurIPS'24*	Link	Multi-choice QA	Science	Previous Source, Internet	Arithmetic, Commonsense, Logical	Deductive, Inductive	12,032	Question/Text, Options	Natural Language, Number, Options
Multi-hop QA										
FRAMES (Krishna et al., 2024)	*Arxiv'24*	Link	Multi-hop QA	General	Internet	Commonsense, Logical, Arithmetic	Deductive	824	Question/Text	Natural Language
HotpotQA (Yang et al., 2018)	*EMNLP'18*	Link	Multi-hop QA	General	Internet	Commonsense	Deductive	113,000	Question/Text	Natural Language
GPQA (Rein et al., 2024)	*Arxiv'24*	Link	Multi-hop QA	Science	Human	Logical	Deductive, Abductive	448	Question/Text, Options	Natural Language, Number, Options
HLE (Phan et al., 2025)	*Arxiv'25*	Link	Multi-hop QA	Science	Human	Logical, Arithmetic, Multimodal	Deductive, Abductive	2,500	Question/Text, Options, Figure/Image	Natural Language, Number, Options
CWQ (Talmor and Berant, 2018)	*NAACL'18*	Link	Multi-hop QA	General	Internet	Commonsense	Deductive	34,689	Question/Text	Natural Language
IIRC (Ferguson et al., 2020)	*EMNLP'20*	Link	Multi-hop QA	General	Internet	Commonsense, Logical	Deductive	13,000+	Question/Text, Links/Sources	Number, Natural Language
MINTQA (He et al., 2024b)	*Arxiv'24*	Link	Multi-hop QA	General	Internet	Commonsense, Logical	Deductive	10,479	Question/Text	Natural Language
MuSiQue (Trivedi et al., 2022)	*ACL'22*	Link	Multi-hop QA	General	Previous Source, Internet	Commonsense, Logical	Deductive	25,000	Question/Text	Natural Language
TopiOCQA (Adlakha et al., 2022)	*TACL'22*	Link	Multi-hop QA	General	Internet	Commonsense, Logical	Deductive	54,494	Question/Text	Natural Language
2WikiMultiHopQA (Ho et al., 2020)	*COLING'20*	Link	Multi-hop QA	General	Internet	Commonsense, Logical	Deductive	192,606	Question/Text	Natural Language
Multi-step QA										
StrategyQA (Geva et al., 2021)	*TACL'21*	Link	Multi-step QA	General	Internet	Commonsense, Logical	Deductive	2,780	Question/Text	Natural Language
Single-hop QA										
SimpleQA (Wei et al., 2024)	*Arxiv'24*	Link	Single-hop QA	General	LLM, Human	Commonsense	Deductive	4,326	Question/Text	Natural Language
TriviaQA (Joshi et al., 2017)	*ACL'17*	Link	Single-hop QA	General	Internet	Commonsense, Logical	Deductive	650,000+	Question/Text	Natural Language
NQ (Kwiatkowski et al., 2019)	*ACL'19*	Link	Single-hop QA	General	Internet	Commonsense, Logical	Deductive	307,373	Question/Text	Natural Language
Text Summarization										
XSum (Narayan et al., 2018)	*EMNLP'18*	Link	Text Summarization	Narrative	Internet, Media	Logical, Commonsense	Abductive	226,711	Question/Text	Natural Language
BIGPATENT (Sharma et al., 2019)	*ACL'19*	Link	Text Summarization	Patent	Internet	Commonsense, Logical	Abductive	1.3 M	Question/Text	Natural Language
Web Browsing										
BrowseComp (Wei et al., 2025a)	*Arxiv'25*	Link	Web Browsing	General	Human, Internet	Commonsense, Logical	Deductive	1,266	Question/Text	Natural Language
BrowseComp-ZH (Zhou et al., 2025a)	*Arxiv'25*	Link	Web Browsing	General	Human, Internet	Commonsense, Logical	Deductive	289	Question/Text	Natural Language
GAIA (Mialon et al., 2023)	*ICLR'23*	Link	Web Browsing	General	Internet, TooL	Commonsense, Logical	Deductive	466	Question/Text, Image/File/Code	Natural Language
WebWalkerQA (Wu et al., 2025b)	*Arxiv'25*	Link	Web Browsing	General	Human, LLM	Commonsense, Logical	Deductive	680	Question/Text	Natural Language
Dialog										
DailyDialog (Li et al., 2017)	*Arxiv'17*	Link	Dialog	General	Internet	Commonsense, Logical	–	13,118	Question/Text	Natural Language

Table 3: Full epresentative knowledge and reasoning intensive benchmarks across diverse task categories (Part 2, continued).

Benchmark	Domain	Primary Retrieval Challenge	Primary Reasoning Challenge
TriviaQA, NQ	General	**Scale & Noise:** Retrieval from massive, noisy corpora.	**Ambiguity:** Handling real-world queries that are often underspecified or ambiguous.
HotpotQA, 2WikiMultiHopQA, MuSiQue, HLE	General	**Multi-document / High-dependency Synthesis:** Requires finding and connecting evidence scattered across multiple Wikipedia articles.	**Multi-hop Deduction:** Explicitly designed to test the ability to link two or more discrete facts into a coherent reasoning path.
MMLU-Pro, QUALITY	Science, Narrative	**Expert-level Retrieval:** Requires accessing deep specialized knowledge from academic or densely written narrative sources.	**Complex & Long-form Reasoning:** MMLU-Pro demands expert-level problem-solving over rote memorization. QUALITY uniquely requires comprehension of very long texts (often >5,000 tokens).
MATH, AQUA-RAT	Math	**Formal Knowledge Retrieval:** Locating precise mathematical theorems, lemmas, or formulas in formal corpora.	**Symbolic & Deductive Reasoning:** Involves performing precise, multi-step logical and algebraic operations where each step must be correct. AQUA-RAT is unique in providing natural language rationales, thus testing the model's ability to explain its formal reasoning.
LiveCodeBench	Code	**Structural & Modal Heterogeneity:** Must retrieve from diverse, heterogeneous sources such as code repositories, documentation, and community forums like Stack Overflow.	**Tool Use & Self-correction Reasoning:** Requires applying retrieved code snippets/APIs, executing code, and reasoning based on test outputs to debug and iteratively improve solutions.
BrowseComp, WebWalkerQA	General (Web)	**Dynamism, Interactivity, and Long-tail Retrieval:** Tests agentic planning and tool use in live, unstructured web environments. BrowseComp requires creative, persistent navigation to locate hard-to-find, intertwined information, while WebWalkerQA focuses on systematic traversal of a website's subpages.	**Agentic & Strategic Reasoning:** Requires planning and executing multi-step strategies (e.g., searching, clicking, extracting) in dynamic and unpredictable contexts to achieve a defined goal.

Table 4: The primary retrieval and reasoning challenges for different RAG-Reasoning benchmarks.

rithmic problem-solving. Web-based tasks, exemplified by BrowseComp (Wei et al., 2025a), emulate real-world search behavior, requiring iterative query formulation and navigation across multiple webpages.

In addition to cataloging datasets, Table 4 provides a synthesized overview of the primary retrieval and reasoning challenges associated with each benchmark discussed in this survey. This comparative analysis reveals critical gaps in current benchmark coverage that future research must address. From a **domain perspective**, most benchmarks still focus on a limited set of general or academic scenarios, with few tackling real-world, realistic industrial or vertical-domain tasks where retrieval sources might be personalized, proprietary or highly specialized. Regarding **retrieval capabilities**, existing benchmarks rarely test systems' ability to handle heterogeneous or multimodal content, nor do they systematically evaluate robustness against noisy, evolving, or conflicting information within a unified framework for trustworthiness. In terms of **reasoning capabilities**, current benchmarks primarily assess deductive reasoning, leaving underexplored more complex forms such as deep causal reasoning, counterfactual thinking, decision-oriented reasoning, or analogical reasoning in specialized domains. Moreover, there is a lack of standardized benchmarks and metrics for evaluating the entire reasoning-retrieval trajectory, including the efficiency of retrieval steps, the quality of intermediate queries, and the logical consistency of multi-step reasoning chains.

B Deep Research Implementations

In this section, we extend the discussion of the agentic paradigm introduced in Section 5.2, in which RAG systems adopt the role of active researchers who plan multistep queries, interleave retrieval with reasoning, and coordinate specialized tools or agents. These characteristics collectively define what we refer to as deep research, representing the ability of a system to autonomously break down complex questions, iteratively gather diverse evidence, and synthesize information through multiple reasoning steps. This paradigm seeks to enhance autonomy, reduce hallucinations, and improve factual accuracy in open-domain tasks.

Such deep research systems can be realized through either single-agent or multi-agent architectures. Single-agent systems rely on a single model to manage the entire process of question decomposition, retrieval, and synthesis, offering simplicity and shared context but facing limitations in handling highly specialized or multi-modal tasks. In contrast, multi-agent systems distribute these responsibilities among specialized agents, enabling modularity and potentially greater robustness. However, this collaborative design introduces

Name	Base Model	Optimizatio	Reward	Retriever	Agent Architecture	Train Data	Evaluation Data	Link
Agentic Reasoning (Wu et al., 2025c)	N/A	Prompting	N/A	Web Search	Centralized	N/A	GPQA	Link
gpt-researcher		Prompting	N/A	Web Search, Local Retrieval	Centralized	N/A	N/A	Link
deep-searcher	Deepseek, , Claude, Gemini, Qwen	Prompting	N/A	Web Search	Hierarchical	N/A	N/A	Link
Search-R1 (Jin et al., 2025)	Qwen2.5-7B-Instruct, Qwen2.5-7B-Base, Qwen-2.5-3B-Instruct, Qwen-2.5-3B-Base	GRPO, PPO	Exact Match	Web Search	Single	NQ, HotpotQA	NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, Bamboogle	Link
ZeroSearch (Sun et al., 2025a)	Qwen2.5-3B-Base, Qwen2.5-7B-Base, Qwen2.5-7B-Instruct, Qwen2.5-3B-Instruct, LLaMA3.2-3B-Instruct, LLaMA3.2-3B-Base	GRPO, PPO, Reinforce	Exact Match	Web Search	Single	NQ, HotpotQA	NQ, TriviaQA, PopQA, HotpotQA, 2WikiMultiHopQA, MuSiQue, Bamboogle	Link
Webthinker (Li et al., 2025c)	GPT-o1, GPT-o3, Deepseek-R1, QwQ-32B, Qwen2.5-32B-Instruct	DPO	Preference Pairs	Web Search	Single	SuperGPQA, WebWalkerQA, OpenThoughts, NaturalReasoning, NuminaMath	GPQA, GAIA, WebWalkerQA, Humanity's Last Exam	Link
nanoDeepResearch	OpenAI series, Claude	Prompting	N/A	Web Search	Centralized	N/A	N/A	Link
DeerFlow	Qwen,	Prompting	N/A	Web Search	Decentralized	N/A	N/A	Link
deep-research	Deepseek,	Prompting	N/A	Web Search	Single	N/A	N/A	Link
open-deep-research	OpenAI series, Deepseek, Claude, Gemini	Prompting	N/A	Web Search	Single	N/A	N/A	Link
DeepResearcher (Zheng et al., 2025)	Qwen2.5-7B-Instruct	GRPO	Format	Web Search	Decentralized	NQ, TQ, HotpotQA, 2WikiMultiHopQA	MuSiQue, Bamboogle, PopQA, NQ, TQ, HotpotQA, 2WikiMultiHopQA	Link
R1-Searcher (Song et al., 2025)	Qwen2.5-7B-Base, Llama3.1-8B-Instruct	GRPO, Reinforce++, SFT	Retrieval, Format	Web Search, Local Retrieval	Single	HotpotQA, 2WikiMultiHopQA	HotpotQA, 2WikiMultiHopQA, MuSiQue, Bamboogle	Link
ReSearch (Chen et al., 2025a)	Qwen2.5-7B-Instruct, Qwen2.5-32B-Instruct	GRPO	Format, Answer	Web Search	Single	MuSiQue	HotpotQA, 2WikiMultiHopQA, MuSiQue, Bamboogle	Link
Search-o1 (Li et al., 2025b)	QwQ-32B-Preview	Prompting	N/A	Web Search	Single	N/A	GPQA, MATH500, AMC2023, AIME2024, LiveCodeBench, Natural Questions, TriviaQA, HotpotQA, 2Wiki, MuSiQue, Bamboogle	Link
r1-reasoning-rag	Deepseek	Prompting	N/A	Local Retrieval, Web Search	Single	N/A	N/A	Link
Open Deep Search (Alzubi et al., 2025)	Llama3.1-70B, Deepseek-R1	Prompting	N/A	Web Search	Single	N/A	SimpleQA, FRAME	Link
node-DeepResearch	Gemini,	Prompting	N/A	Web Search	Single	N/A	N/A	Link
deep-research	Gemini, OpenAI series, Deepseek, Claude, Grok	Prompt	N/A	Local Retrieval, Web Search	Single	N/A	N/A	Link

Table 5: Overview of deep research implementations.

additional complexity in coordination and communication, as well as higher computational costs.

Alongside these developments in agent orchestration, the nature of retrievers used in deep research has also evolved significantly. Early RAG systems relied on sparse keyword-based retrieval, later surpassed by dense retrievers employing bi-encoder architectures for semantic matching. More recent deep research systems increasingly integrate web search-based retrievers, allowing real-time access to open-domain information. Some retrievers have also been transformed into LLM-callable tools for flexible invocation. This evolution of retrievers has played a crucial role in enabling the sophisticated information-gathering processes required for deep research.

C Comparison of Reasoning Workflows and Agent Orchestration Strategies

Table 6 summarizes the diverse reasoning workflows and agent orchestration strategies employed in Synergized RAG-Reasoning systems, highlighting their respective strengths, limitations, and suitable application scenarios. Reasoning workflows vary from linear chain-based approaches, which are efficient but vulnerable to error propagation, to more complex tree-based and graph-based methods that offer higher recall and transparency at the cost of increased computational overhead. Similarly, agent orchestration strategies range from single-agent setups to multi-agent systems that distribute specialized roles among agents, enhancing robustness and scalability. However, these advanced designs often introduce additional communication overhead and complexity in conflict resolution. This comparison illustrates the trade-offs inherent in choosing particular workflows or orchestration architectures and underscores the need for adaptive systems that can dynamically balance efficiency, accuracy, and resource constraints in real-world applications.

Category	Sub-category	Strengths	Limitations	Suitable Scenarios
Reasoning Workflow	Chain-based	One retrieval per reasoning step; low latency and token cost. Easy to cache and monitor.	An early wrong sub-query propagates; context grows fast on long chains.	Single-hop or short multi-hop QA where each intermediate fact is easy to access.
	Tree-based (ToT)	High recall: explores multiple branches in parallel, hedges against early errors. Transparent what-if traces.	Quadratic cost; tree branches require many retrieval calls.	Ambiguous or "multiple plausible paths" tasks (e.g., HotpotQA, legal reasoning) where missing one clue kills accuracy.
	Tree-based (MCTS)	Budget-aware exploration: focuses calls on promising branches; graceful anytime stopping.	Tuning-heavy and may converge to a suboptimal subtree.	Deep-search problems under tight API-call or token budgets (e.g., biomedical QA).
	Graph-based (Walk-on-Graph)	Efficient in explicit KG/document graphs; short reasoning paths on KGs.	Requires high-quality KGs; fails if graphs lack explicit edges; less flexible for open-web contexts.	Enterprise or domain-specific QA where a curated KG exists (e.g., product catalogs).
	Graph-based (Think-on-Graph)	Adaptive and verifiable; LLM updates a live evidence graph, allowing node-level citation checks and high factual accuracy.	Higher latency; many micro-tool calls; search space can explode without pruning.	Open-domain "deep research" or fact-dense synthesis tasks (e.g., BrowseComp, systematic reviews).
Agent Orchestration	Single-agent (Prompt-only)	Simple implementation via a ReAct loop; low resource overhead.	Constrained by prompt engineering and system design flexibility.	Prototyping demos and small-scale applications where simplicity outweighs performance.
	Single-agent (SFT)	Clear, well-defined RAG and reasoning patterns; higher precision than prompt-only approaches.	Requires large synthetic data; may overfit tool schemas, reducing out-of-domain generalization.	Production chatbots with stable APIs and predictable query formats (e.g., internal customer support).
	Single-agent (RL)	Adaptive RAG and reasoning yields high recall and accuracy; learns when to retrieve and reason.	Challenging to define suitable reward signals; computationally expensive to train.	Open-domain research or long-form QA where call costs are high and optimal stop conditions matter.
	Multi-agent (Decentralized)	High recall via parallel domain experts; robustness to noisy or diverse corpora.	High communication and consensus overhead; conflicting answers require resolution.	Large-scale evidence aggregation across heterogeneous sources (e.g., meta-analysis, news tracking).
	Multi-agent (Centralized/Hierarchical)	Budget-efficient: manager avoids duplicate searches and ensures a clear provenance chain. Scales horizontally without exponential cost growth.	Manager prompts or policies can become a single-point bottleneck, limiting performance.	Complex tasks requiring coordinated subtasks under strict API-call budgets.

Table 6: Comparison of reasoning workflows and agent orchestration in Synergized RAG-Reasoning systems.