Agentic Reasoning: A Streamlined Framework for Enhancing LLM Reasoning with Agentic Tools

Anonymous ACL submission

Abstract

We introduce Agentic Reasoning, a framework that enhances large language model (LLM) reasoning by integrating external tool-using agents. Agentic Reasoning dynamically leverages web search, code execution, and structured memory to address complex problems requiring deep research. A key innovation in our framework is the Mind-Map agent, which constructs a structured knowledge graph to store reasoning context and track logical relationships, ensuring coherence in long reasoning chains with extensive tool usage. Additionally, we conduct a comprehensive exploration of the Web-Search agent, leading to a highly effective search mechanism that surpasses all prior approaches. When deployed on DeepSeek-R1, our method achieves a new state-of-the-art (SOTA) among public models and delivers performance comparable to OpenAI Deep Research, the leading proprietary model in this domain. Extensive ablation studies validate the optimal selection of agentic tools and confirm the effectiveness of our Mind-Map and Web-Search agents in enhancing LLM reasoning. Our code and data are publicly available.

1 Introduction

Recently, large reasoning models, such as OpenAI's o1 (Jaech et al., 2024), Qwen-QwQ (Team), and DeepSeek-R1 (Team, 2024), have demonstrated impressive stepwise reasoning capabilities over long sequences through large-scale reinforcement learning. These advancements provide promising solutions to complex reasoning tasks (Wei et al., 2022; Lewkowycz et al., 2022; OpenAI) and have inspired foundational efforts to replicate o1-like reasoning patterns across a broader range of models (Qin et al., 2024; Huang et al., 2024; Zhang et al., 2024). It is recently revealed by DeepSeek-R1 that applying rule-based outcome rewards during training, such as evaluating whether a piece of code executes successfully, could yield remarkable reasoning capabilities equaling o1-level math and coding performance.

Model	**Accuracy (%)**
GPT-4o†	3.3
Grok-2†	3.8
Claude 3.5 Sonnet†	4.3
Gemini Thinking†	6.2
OpenAI o1†	9.1
DeepSeek-R1	**9.4**
OpenAI o3-mini (medium)†	10.5
OpenAI o3-mini (high)†	13.0
Agentic Reasoning w/ R1	**23.8** (+14.4)
Perplexity deep research †	*21.1*
OpenAI deep research †	*26.6*

† denotes proprietary models.

Table 1: On Humanity's Last Exam, we achieved a remarkable 23.8% with DeepSeek-R1, marking a 14.4% improvement over the base model. This narrows the gap to the proprietary OpenAI Deep Research to just 2.8%, which depends on a stronger internal reasoning model.

Although current reasoning methods excel in structured domains like math and code—where outcomes are easily verifiable—applying these techniques to less structured or knowledge-intensive tasks remains a significant challenge. As mentioned in DeepSeek-R1 (Team, 2024), not all problems benefit from formal reasoning approaches. Many fields, such as social sciences, ethics, or experiential disciplines, rely on abstract concepts, conventional wisdom, factual verification, understanding complex logical relationships, or moral reasoning. When models attempt to impose math- or coding-style reasoning onto such areas, they often produce flawed or overly rigid results. Developing approaches that account for these unique requirements is essential for advancing the applicability of reasoning model beyond their current domains.

Deep, thoughtful answers to open-ended questions often require extensive research, repeated verification, information retrieval, computational analysis, and the organization of complex logical relationships—steps fundamental to human reasoning. In this process, humans rely heavily on external tools, such as internet searches for gathering information, computational tools for quantitative analysis, or whiteboards and mind maps for organizing thoughts. This raises an intriguing question: can reasoning LLMs similarly leverage external tools to enhance their reasoning and tackle intensive knowledge work across diverse domains?

Previous efforts have attempted to integrate search or retrieval-augmented generation (RAG) into the reasoning process (Shao et al., 2024; Khaliq et al., 2024; Islam et al., 2024; Li et al., 2025), with notable examples including Gemini's and OpenAI's Deep Research. However, these models are proprietary, and their exact methodologies remain undisclosed. In contrast, open-source models primarily focus on retrieval and web-search integration during reasoning but still exhibit a notable performance gap compared to their closed-source counterparts.

We introduce Agentic Reasoning, a framework that enhances reasoning by integrating external LLM-based agents as tools. This approach allows LLMs to delegate specific tasks to auxiliary agents during the reasoning process, thereby improving their overall problem-solving capabilities. Through extensive experimentation with integrating various agents into the reasoning process, we identified three essential agents that prove highly effective for general reasoning across diverse problems. The Web-Search agent, which retrieves relevant information from the internet to supplement the model's knowledge. The Code agent, capable of performing computational analyses and coding tasks to support quantitative reasoning. Finally, the memory agent, which we call Mind-Map, constructs knowledge graphs based on the reasoning context, enabling the organization of complex logical relationships in a manner similar to human mind mapping. Together, these agents enhance the model's ability to tackle complex problems and do deep research with greater efficiency and precision.

We evaluated our model on general knowledge-intensive benchmarks requiring complex reasoning capabilities, categorized into two key areas: (1) solving expert-level questions and (2) conducting deep research on real-world expert-level tasks. For expert-level questions, we evaluate the model on Humanity's Last Exam (Phan et al., 2025), a recently released benchmark assessing AI performance across a broad range of subjects. As shown in Table 1, we achieves a new high of 23.8% accuracy, marking a 14.4% improvement over the raw model and narrowing the open-source vs. closed-source gap to just 2.8% compared to the proprietary OpenAI Deep Research, which benefits from a stronger internal reasoning model. For real-world expert-level tasks, Agentic Reasoning was assessed by domain experts, who found that it effectively automated complex manual investigation. This underscores its potential to streamline labor-intensive processes and boost productivity in knowledge-intensive domains.

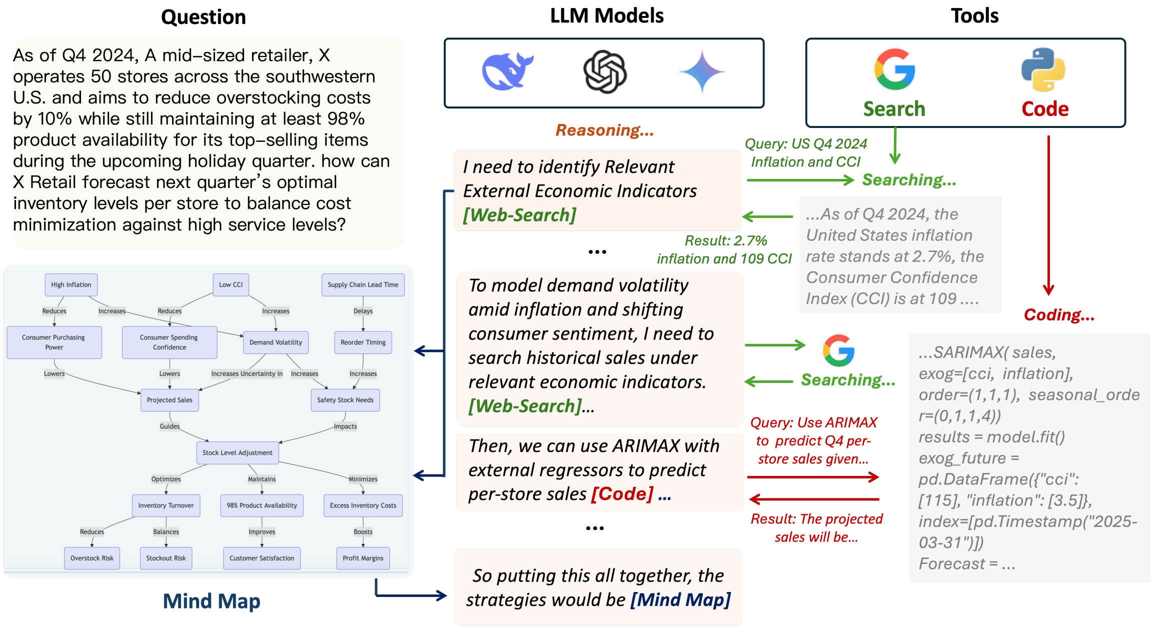
In brief, our contribution can be concluded as:

- We propose Agentic Reasoning, a streamlined framework that enhances reasoning by integrating external LLM-based agentic tools. We experimentally identify web-search, coding, and Mind-Map agents as three universally effective tools.
- We explore the design of the Web-Search agent and identify a strategy that outperforms previous search or RAG approaches.
- We develop a knowledge-graph-based Mind-Map to assist reasoning, improving the model's ability to handle complex logic and maintain coherence in long reasoning chains.
- We evaluate our approach on expert-level problem-solving and deep research tasks, achieving new SOTA results across several benchmarks and surpassing prior methods in human evaluations.

2 Method

2.1 Agentic Reasoning Pipeline

Our core idea is to enhance LLM reasoning by integrating external LLM-based agents into the process. During reasoning, the model can call these agents as tools to assist in problem-solving while maintaining a structured memory to store its reasoning context. In the overall process, we deploy a Web-Search agent and a Code agent as problem-solving tools, along with a knowledge-graph agent, called Mind-Map, to serve as memory.



As of Q4 2024, A mid–sized retailer, X operates 50 stores across the southwestern U.S. and aims to reduce overstocking costs by 10% while still maintaining at least 98% product availability for its top–selling items during the upcoming holiday quarter. how can X Retail forecast next quarter's optimal inventory levels per store to balance cost minimization against high service levels?

I need to identify Relevant External Economic Indicators ***[Web-Search]***

...As of Q4 2024, the United States inflation rate stands at 2.7%, the Consumer Confidence Index (CCI) is at 109 ....

To model demand volatility amid inflation and shifting consumer sentiment, I need to search historical sales under relevant economic indicators. ***[Web-Search]****...*

Then, we can use ARIMAX with external regressors to predict per-store sales ***[Code]*** ***...***

```
...SARIMAX( sales,
exog=[cci, inflation],
order=(1,1,1), seasonal_orde
r=(0,1,1,4))
results = model.fit()
exog_future =
pd.DataFrame({"cci":
[115], "inflation": [3.5]},
index=[pd.Timestamp("2025-
03-31")])
Forecast = ...
```

So putting this all together, the strategies would be ***[Mind Map]***

Figure 1: The overall workflow of Agentic Reasoning. Given a question, the reasoning LLM can invoke the Web-Search agent to retrieve external information, the Coding agent to perform quantitative computations, and the Mind-Map agent to structurally memorize the reasoning context, to provide a comprehensive solution.

Specifically, the reasoning LLM can dynamically determine when to call external agentic tools during its reasoning process. As shown in Figure 1, when needed, it embeds specialized tokens into its reasoning sequence, categorizing them as web-search tokens, coding tokens, or Mind-Map calling tokens. Alongside these tokens, the reasoning LLM generates a query as a message to the external agents. Upon detecting such a token, the reasoning process temporarily halts to extract the query and its reasoning context. These queries are then dispatched to the corresponding external agents. The agents would consider both the received query and the reasoning context to ensure the most pertinent results are returned to the main reasoning chain. These results are then reintegrated into the reasoning chain, allowing the model to continue its inference with an updated knowledge. This iterative retrieval-and-reasoning cycle continues as needed, enabling the model to dynamically refine its reasoning until it reaches a fully reasoned final answer.

2.2 Mind-Map Agent

We construct a Mind-Map to store and structure the real-time reasoning context of the reasoning model. This Mind-Map is built by transforming raw reasoning chains into a structured knowledge graph. Specifically, we use a graph-construction LLM to extract entities from the reasoning chain and identify semantic relationships between related entities, following a process similar to that used in GraphRAG (Edge et al., 2024).

The Mind-Map serves two primary functions. First, it clusters reasoning context into distinct groups and summarizes each of them. This is achieved by applying community clustering (Traag et al., 2019) on the knowledge graph and using an LLM to generate concise summaries for each group. Second, the knowledge graph can be queried with specific questions, such as "Who was Jason's maternal great-grandfather?" Using standard RAG on the knowledge graph (Edge et al., 2024), we retrieve and return the relevant information to response the query.

These functions integrate the Mind-Map into two key aspects of the Agentic Reasoning process. First, it provides reasoning context to external tools, enabling them to generate more context-aware responses. The context is generated by synthesizing the summaries of each clustered group, performed by an LLM. Additionally, when the reasoning model encounters uncertainty or loses track in an extended reasoning process, it can query the Mind-Map as an external memory to retrieve relevant information and continue reasoning seamlessly. This ensures the model maintains a long reasoning chain across multiple breakdown tasks and tool calls without missing critical information.

2.3 Web-Search agent

A search agent is invoked to retrieve the most relevant documents from the web. It consists of four key components: query breakdown, a search service, a re-ranking service, and RAG.

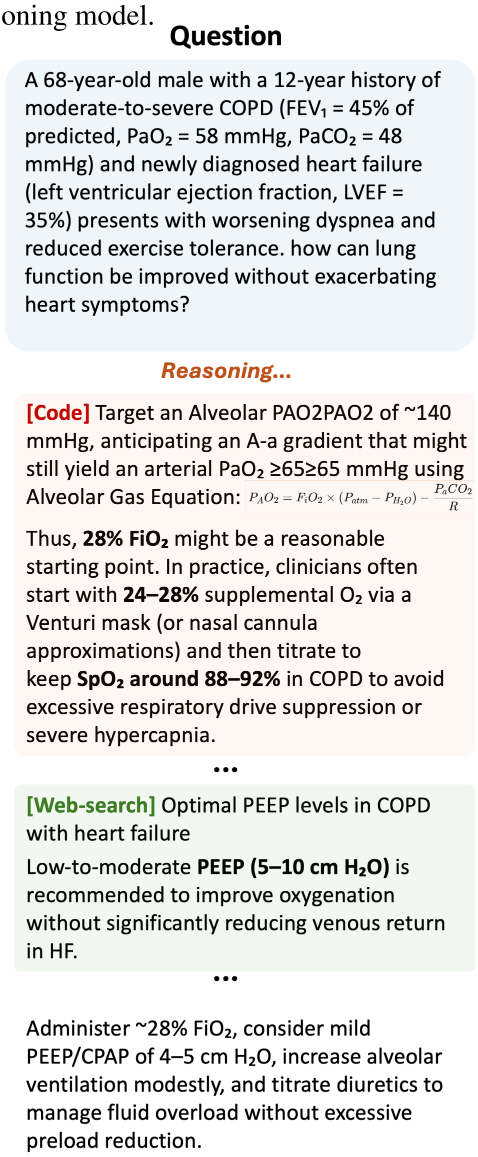
When the reasoning model generates a web-search query, it is sent to the Web-Search agent, which first reorganizes it into one or more search-optimized queries suitable for search engines like Google or Bing. The process involves sending the LLM the original query along with the reasoning context retrieved from the Mind-Map, prompting it to generate suitable refined search queries. For example, given the original query "Search the external economic indicators" and the context "We are looking for the optimal investing strategy for a retailer in the U.S. in Q4 2024", the Web-Search agent would break it down into more specific queries such as "U.S. Q4 2024 inflation rate" and "U.S. Q4 2024 CCI". These queries are then sequentially sent to the search engine, which returns related web pages.

After we retrieved the web pages for each refined query, we apply a re-ranking model to rank web pages based on their alignment with the original query and context. The average relevance score of the top 10 pages is computed, and if it falls below a predefined threshold, the Web-Search agent will iterate back to the last step and further refine the search query.

Once reranking is complete, web pages with relevance scores above the threshold are stored, and RAG is applied on them to extract meaningful insights. Each refined query undergoes RAG to generate a natural language response. Finally, an LLM synthesizes these responses into a cohesive final snippet, based on both the original query and reasoning context. This processed snippet is then integrated into the main reasoning process, ensuring that external insights enhance logical flow without causing disruption.

2.4 Coding Agent

Instead of prompting the reasoning model to generate code directly, we find it more efficient to delegate coding tasks to a specialized coding LLM. The reasoning model sends the relevant context and query message to the coding LLM, which then writes the required code, executes it via a compiler, and returns the results. This approach ensures that the reasoning model remains focused on its core reasoning process without being disrupted by coding tasks, allowing for longer and more coherent reasoning chains. Specifically, we format the coding request as follows: "Write code to perform <code message from reasoning model> given the context <reasoning context from Mind-Map> to answer the query <user query>." The coding LLM is instructed to always return its output in natural language, ensuring seamless integration with the reasoning model.



Question

A 68-year-old male with a 12-year history of moderate-to-severe COPD (FEV$_1$ = 45% of predicted, PaO$_2$ = 58 mmHg, PaCO$_2$ = 48 mmHg) and newly diagnosed heart failure (left ventricular ejection fraction, LVEF = 35%) presents with worsening dyspnea and reduced exercise tolerance. how can lung function be improved without exacerbating heart symptoms?

Reasoning...

[Code] Target an Alveolar PAO2PAO2 of ~140 mmHg, anticipating an A-a gradient that might still yield an arterial PaO$_2$ ≥65≥65 mmHg using Alveolar Gas Equation: $P_AO_2 = F_iO_2 \times (P_{atm} - P_{H_2O}) - \frac{P_aCO_2}{R}$

Thus, **28% FiO$_2$** might be a reasonable starting point. In practice, clinicians often start with **24–28%** supplemental O$_2$ via a Venturi mask (or nasal cannula approximations) and then titrate to keep **SpO$_2$ around 88–92%** in COPD to avoid excessive respiratory drive suppression or severe hypercapnia.

…

[Web-search] Optimal PEEP levels in COPD with heart failure

Low-to-moderate **PEEP (5–10 cm H$_2$O)** is recommended to improve oxygenation without significantly reducing venous return in HF.

…

Administer ~28% FiO$_2$, consider mild PEEP/CPAP of 4–5 cm H$_2$O, increase alveolar ventilation modestly, and titrate diuretics to manage fluid overload without excessive preload reduction.

Figure 2: Case study on a complex medical decision-making problem.

3 Experiments

3.1 Implementation Details

In our experiments, we use DeepSeek-R1 as the primary reasoning models by default. For the Web-Search agent, query breakdown and RAG are han-

Table 2: Performance comparison on GPQA dataset across Physics, Chemistry, and Biology.

Method	Phy.	Chem.	Bio.	All
Direct Reasoning				
QwQ-32B	75.6	39.8	68.4	58.1
Llama3.3-70B	54.7	31.2	52.6	43.4
DeepSeek-R1	86.8	56.1	63.8	71.5
GPT-4o†	59.5	40.2	61.6	50.0
o1†	92.8	64.7	69.2	78.0
o3-mini-low†	-	-	-	70.6
o3-mini-mid†	-	-	-	76.8
o3-mini-high†	-	-	-	79.7
Retrieve/Search in Reasoning				
RAgent w/QwQ-32B	76.7	46.2	68.4	61.6
RAgent w/DeepSeek-R1	87.7	58.2	65.7	72.9
SearchO1 w/QwQ-32B	77.9	47.3	78.9	63.6
SearchO1 w/DeepSeek-R1	90.2	61.3	71.4	74.6
Agentic Reasoning				
Ours w/QwQ-32B	**88.1**	**58.3**	**79.6**	**69.7**
Ours w/DeepSeekR1	**94.5**	**73.7**	**80.5**	**81.2**

dled by DeepSeek-V3 (Liu et al., 2024). We use a maximum of 32,768 tokens, temperature of 0.7, top_p of 0.8, top_k of 20, and a repetition penalty of 1.05 across all models for generation. We use Bing as the search engine, retrieving the top 20 most relevant pages. The re-ranking model is Cohere Rerank 3.5, with a top-10 average relevance score threshold of 0.7 to determine if iterative query refinement is needed, allowing a maximum of three iterations. Additionally, web pages with a relevance score above 0.7 are selected for RAG processing. For the Mind-Map agent, both knowledge graph construction and Graph-RAG retrieval are also performed using DeepSeek-V3. For the coding agent, we use claude-3.5-sonnet to generate code and Python 3.11 for execution. We report pass1 results by default.

3.2 Solving Expert-level Problems

Agentic Reasoning model is able to call external tools in its reasoning to solve expert-level problems, except Humanity's Last Exam we previously mentioned, we further evaluate it on two datasets: GPQA dataset (Rein et al., 2023), a PhD-level multiple-choice science QA benchmark, and GAIA (Mialon et al., 2023), a benchmark for AI agents that requires a set of abilities such as reasoning, web browsing, and tool-use proficiency.

As shown in Table 2, applying Agentic Reasoning to a strong reasoning model like DeepSeek-R1 achieves a new SOTA, surpassing even the best proprietary model, o3-mini-high. Compared to the base model DeepSeek-R1, our method boosts overall performance by nearly 10%. Compared to previous search-in-reasoning approaches (Li et al., 2025; Islam et al., 2024), Agentic Reasoning demonstrates superior reasoning enhancement, outperforming Search-O1 by approximately 5% overall. Furthermore, we show that this method is generally effective across different reasoning models, such as QwQ (Yang et al., 2024), where it improves base model accuracy by over 10%.

On GAIA (as shown in Table 3), Agentic Reasoning establishes a new SOTA among all publicly available methods. Compared to OpenAI's Deep Research, which leverages its proprietary high-performance reasoning models, our approach surpasses it on Level 1 and Level 2 tasks while narrowing the gap to 2.26% on Level 3. GAIA requires a combination of advanced reasoning, web browsing, and tool-use proficiency for successful completion. Our results demonstrate that Agentic Reasoning excels in handling complex tasks while maintaining strong generalization across diverse problem domains.

We also present a case study on a complex medical decision-making problem, as shown in Figure 2 The model autonomously executes code to compute the optimal FiO_2 (Fraction of Inspired Oxygen) for a patient, performs a web search to retrieve the most accurate PEEP (Positive End-Expiratory Pressure) value, and synthesizes both results to determine the best treatment plan.

GAIA	Level 1	Level 2	Level 3	Avg.
Langfun	58.06	51.57	24.49	49.17
InspectReAct	67.92	59.30	30.77	57.58
h2oGPTe	78.49	64.78	40.82	65.12
AgenticReasoning	**74.36**	**69.21**	**45.46**	**66.13**
Open AI Deep Research †	*74.29*	*69.06*	*47.60*	*67.36*

Table 3: Performance comparison on GAIA across different levels.

3.3 Deep Research

We evaluate the deep research capability of our approach using the FreshWiki dataset (Shao et al., 2024), which curates high-quality, recent Wikipedia articles. The model is prompted directly with the topic and asked to generate the article. Evaluation covers article quality, assessed via ROUGE and entity recall. This task needs a comprehensive analysis of long-form generation while highlighting key challenges like bias transfer and factual consistency.

Table 4: Comparison with Human-Written Articles

	ROUGE-1	ROUGE-L	Entity Recall
Direct Gen	27.32	13.13	6.11
RAG	29.14	14.23	8.84
RAgent	30.04	14.21	9.08
Search-O1	41.56	16.08	12.88
STORM	47.93	17.42	15.43
Ours	**54.10**	**19.62**	**18.77**

Table 5: Human Evaluation on Deep Research articles (Rate Range: 1-5).

	Interest Level	Organization	Relevance	Coverage
Direct Gen	1.2	1.6	1.2	1.7
RAG	1.4	2.1	1.9	2.3
RAgent	1.6	2.3	1.6	2.6
Search-O1	2.5	2.8	2.1	3.2
STORM	2.9	3.2	2.9	3.7
Gemini-DR†	2.7	2.5	2.3	3.0
Ours	**3.7**	**4.6**	**4.2**	**4.1**

We also conduct an evaluation of Agentic Reasoning for deep research in open deep research tasks. A group of PhD-level experts in finance, medicine, and law were asked to formulate 15 to 30 professional research questions closely related to their respective fields. These questions were designed to require at least 20 minutes of in-depth research to answer comprehensively. There are in total 56 questions were collected. The experts would review the generated articles on interest level, organization, relevance, and coverage. More details are in the appendix.

We evaluate our method using the same underlying reasoning model, DeepSeek-R1, and compare it against various search-enhanced reasoning approaches, including RAG, RAgent (Islam et al., 2024), and Search-O1 (Li et al., 2025), as well as STORM (Shao et al., 2024), which employs a more complex agent-based workflow. Additionally, we benchmark our approach against the proprietary Gemini Deep Research [1] on deep research tasks. As shown in Tables 4 and 5, our results demonstrate that Agentic Reasoning consistently outperforms all RAG and search-based methods, as well as Gemini Deep Research, across all benchmarks. These findings highlight the effectiveness of structured reasoning and tool-augmented frameworks in enabling more advanced and efficient deep research.

3.4 Analysis

3.4.1 Ablation on Toolbox

We conduct experiments to explore the impact of integrating different tools in Agentic Reasoning and find that tool quality is far more important than quantity. Specifically, the combination of web search, coding, and Mind-Map agents proves to be the most effective across various tasks, including those requiring expert-level proficiency. As shown in Figure 3, we evaluated performance on GPQA using Hugging Face's default agent toolbox with seven tools and LangChain with 109 tools. Surprisingly, adding more tools often degraded performance by increasing the risk of inappropriate tool selection. Many capabilities, such as translation or code interpretation, are already embedded within the reasoning model, making their external integration redundant. Moreover, inaccuracies in external tool outputs can negatively affect overall response quality.

Figure 3 also presents an ablation study on the three proposed tools in this paper. We tested different tool combinations to assess their individual contributions to agentic reasoning. Among single-tool deployments, web search performed the best, while coding and Mind-Map achieved comparable results. Notably, combining tools yielded a synergistic effect—web search + Mind-Map or web search + coding provided greater improvements than the sum of their individual gains. The best performance was achieved when integrating all three: web search, Mind-Map, and coding.

3.4.2 Ablation on Web-Search agent Design

Integrating web search into LLMs has been widely explored in recent research (Li et al., 2025; Lewis et al., 2020; Islam et al., 2024). In Agentic Reasoning, we investigate various web-search strategies to determine the most effective approach. Our ablation study primarily considers standard RAG and Knowledge Refinement, where retrieved sources are summarized for the response. Additionally, we incorporate Query Breakdown, Rerank, and Mind-Map Reasoning Context, key components in our Web-Search agent. Our findings reveal that Query Breakdown, Rerank, and Mind-Map Reasoning Context incrementally improve performance. Surprisingly, Knowledge Refinement, which is effective when used solely with RAG, becomes ineffective when combined with our three adopted components. This decline is primarily due to its redundancy with Rerank, which serves a similar

---

[1] OpenAI Deep Research experiments are currently restricted by a high paywall.

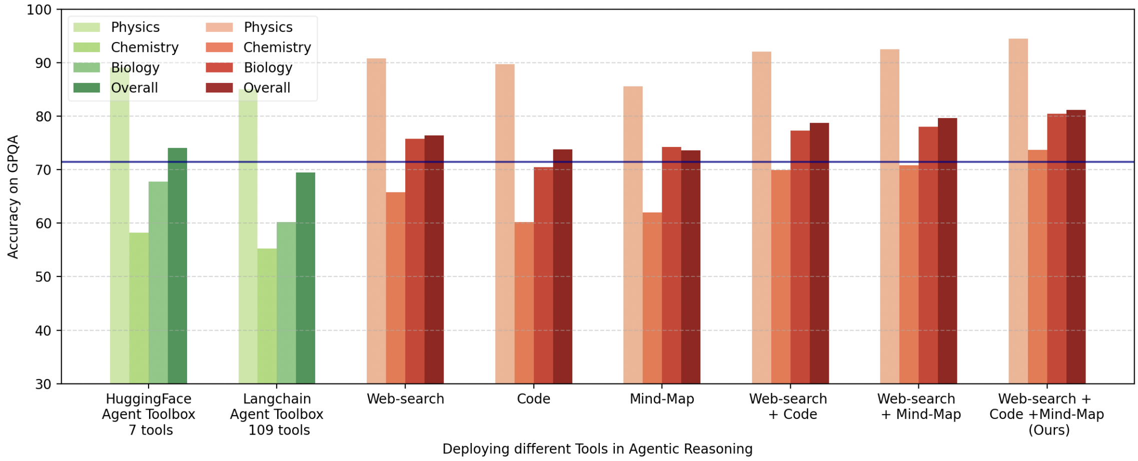
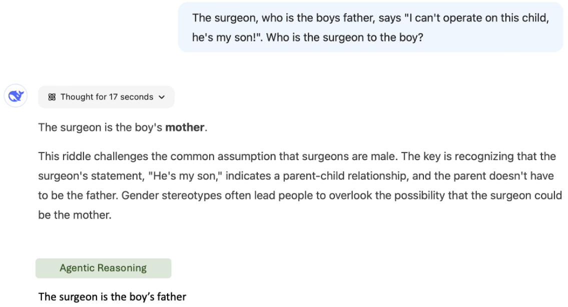


Figure 3: The ablation study examines the impact of different tools in reasoning. Green ones represent external toolboxes, red ones are combinations of our proposed tools. The blue line is the overall performance of the base reasoning model.

role but proves more effective in most cases. As a result, our final Web-Search agent includes RAG, Query Breakdown, Rerank, and Mind-Map Reasoning Context for optimal performance.

	RAG	Knowledge Refinement	**Query Breakdown**	**Rerank**	**Mind-Map Reasoning Context**	**GAPA**
Search-O1	✓	✓				74.6
Storm	✓					72.7
	✓		✓			73.3
	✓		✓	✓		75.2
	✓	✓	✓	✓	✓	76.2
	✓	✓	✓		✓	75.8
Agentic Reasoning	✓		✓	✓	✓	**76.4**

Table 6: Comparison of different web-search approaches.


The surgeon, who is the boys father, says "I can't operate on this child, he's my son!". Who is the surgeon to the boy?

Thought for 17 seconds

The surgeon is the boy's **mother**.

This riddle challenges the common assumption that surgeons are male. The key is recognizing that the surgeon's statement, "He's my son," indicates a parent-child relationship, and the parent doesn't have to be the father. Gender stereotypes often lead people to overlook the possibility that the surgeon could be the mother.

Agentic Reasoning

The surgeon is the boy's father

Figure 4: A tricky question that misleads most LLMs is correctly answered by us.

3.4.3 The Effect of Mind-Map

We have shown in Figure 3 that our quantitative results demonstrate that Mind-Map significantly enhances performance. In this section, we analyze its impact on reasoning in detail. Mind-Map proves particularly effective in maintaining long reasoning with tools and clarifying complex logical relationships.

We find that questions needs longer reasoning chains and more tool calls tend to be inherently more difficult, leading to lower accuracy, as shown in Figure 5. Examining the model's reasoning process, we find that LLMs often struggle to maintain coherence over long reasoning with massive tool calls. They may deviate from user queries, repeatedly call tools in the same way, or revisit previous errors, all of which negatively impact performance.

We introduced the Mind-Map agent to help the model manage its reasoning memory, ensuring coherent long reasoning and reducing errors. As shown in Figure 5, this mechanism significantly improves performance, particularly on questions requiring long reasoning chains and more tool calls. The structured memory provided by the Mind-Map agent preserves prior reasoning steps, mitigating common pitfalls in extended reasoning tasks.

Mind-Map is also especially helpful for the tasks heavily rely on logic relationships. We find it helps to correctly answer tricky logic-based questions that frequently fool LLMs. A well-known example is a modified riddle: "The surgeon, who is the boy's father, says 'I can't operate on this child, he's my son!' Who is the surgeon to the boy?" As shown in Figure 4, DeepSeek-R1 took 17 seconds to process this question but still produced the wrong answer, a failure also observed in models from the GPT and Gemini series models. These models often fall for a political-correct corpus contaminated response, failing to recognize the obvious logical structure. However, in our Agentic Reasoning framework, the use of a Mind-Map allows the model to explicitly analyze the logical relationships between the entities [surgeon], [boy], and [father], leading to the correct answer.

This property also enables Mind-Maps to enhance deductive reasoning in strategic games. We tested our approach in Werewolf, a classic social deduction game where players assume hidden roles as either villagers or werewolves. Villagers aim to identify the werewolves through discussion, while werewolves deceive the group and eliminate players without being caught. To evaluate performance, we invited seven experienced Werewolf players, each with over five years of experience, to compete against our Agentic Reasoning model. The results show that our model achieved an impressive 72% win rate, significantly surpassing both the expected statistical win rate and human performance in our experiment. In contrast, without Mind-Map, the model's win rate dropped to 36%. As the Mind-Map of the model's reasoning process shown in Figure 6, Mind-Map proved crucial in helping the model track relationships between players based on their spoken arguments. By maintaining a structured memory of interactions, it more effectively identified deception strategies, anticipated voting behaviors, and optimized its own disguise tactics. This result highlights that Mind-Map is not only a tool for structured logic but also a powerful enabler of strategic reasoning in dynamic, high-stakes environments.

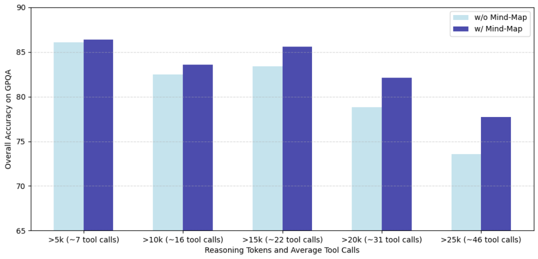


Figure 5: Mind-Map improves performance on questions need long reasoning.

4 Related Work

The concept of multi-agent collaboration in LLMs has gained attention with frameworks like AutoGPT (Yang et al., 2023) and LangChain Agents (Pandya and Holia, 2023), allowing models to interact with external APIs, search engines, and computational environments. While these frameworks introduce modular reasoning, they often lack optimized task delegation and structured integration, reducing their effectiveness in long-chain reasoning tasks. Recent research on Hierarchical Planning with LLMs (Luo et al., 2023) and Task-Specific AI Agents (Wu et al., 2024) explores structured agent cooperation for problem-solving. However, these approaches still do not deeply integrate agent tools within reasoning chains and fail to systematically explore optimal agent combinations that maximize reasoning performance.

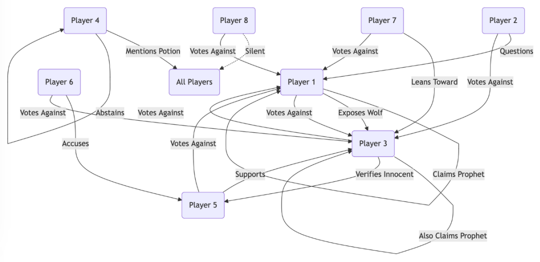


Figure 6: Mind-Map in playing werewolf game. The first round and the second round. Player8 is the model.

Previous studies focus a lot on integrating the search capability into LLMs. Recent agentic RAG systems(Khaliq et al., 2024; Islam et al., 2024; Li et al., 2025) have enabled models to autonomously determine when and what knowledge to retrieve, enhancing their planning and problem-solving capabilities. Additionally, research has explored (Li et al., 2025) integrating Web-Search agent into the reasoning model, like QwQ (Team) demonstrating the potential of search augmentation in structured reasoning. However, existing approaches have primarily focused on single-agent enhancements, neglecting the potential synergy of multiple agentic tools. Moreover, prior works have yet to integrate such tools with state-of-the-art reasoning models like DeepSeek-R1 or OpenAI-O1, limiting their effectiveness in solving highly complex tasks.

5 Conclusion

We introduced Agentic Reasoning, a framework that enhances LLM reasoning by integrating Mind-Map, web search, and coding. Our approach improves problem-solving and deep research capabilities, outperforming existing models in expert-level QA and real-world research tasks. Agentic Reasoning outperforms existing methods in both quantitative benchmarks and human evaluations. Future work will explore task-specific tools integration and test-time computing to further enhance AI's reasoning capabilities.

6 Limitations

Despite the strong performance of Agentic Reasoning, several limitations remain that warrant further research and refinement.

Computational Overhead and Efficiency. Integrating multiple external agents, including web search, Mind-Map, and code execution, significantly increases computational costs and inference latency. While these components enhance reasoning depth, their sequential invocation introduces bottlenecks, limiting real-time applicability. Future work would explore techniques such as agent prioritization, caching strategies, or adaptive invocation mechanisms to optimize efficiency without sacrificing accuracy.

Reliance on External Knowledge Sources. The effectiveness of Agentic Reasoning depends on the quality of retrieved knowledge, particularly in web search. The system lacks built-in verification mechanisms to assess the credibility of sources, making it susceptible to misinformation or biased content. Developing trust-aware retrieval mechanisms, such as fact-checking agents or weighted source reliability scores, could mitigate this risk and improve robustness in knowledge-intensive domains.

Interpretability and Trustworthiness. While the Mind-Map agent provides structured reasoning memory, the overall decision-making process remains highly dependent on LLMs. This reliance introduces the risk of hallucinations, which can derail the entire reasoning process, especially in complex, multi-step tasks. In high-stakes domains such as medical AI or legal reasoning, even minor inaccuracies can lead to significant consequences. Ensuring reliability requires additional safeguards, such as fact-verification mechanisms, confidence estimation, or human-in-the-loop oversight, to mitigate the risks associated with LLM-driven reasoning.

References

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Zhen Huang, Haoyang Zou, Xuefeng Li, Yixiu Liu, Yuxiang Zheng, Ethan Chern, Shijie Xia, Yiwei Qin, Weizhe Yuan, and Pengfei Liu. 2024. O1 replication journey–part 2: Surpassing o1-preview through simple distillation, big progress or bitter lesson? *arXiv preprint arXiv:2411.16489*.

Shayekh Bin Islam, Md Asib Rahman, KSM Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez. 2024. Open-rag: Enhanced retrieval-augmented reasoning with open-source large language models. *arXiv preprint arXiv:2410.01782*.

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

M Abdul Khaliq, P Chang, M Ma, Bernhard Pflugfelder, and F Miletić. 2024. Ragar, your falsehood radar: Rag-augmented reasoning for political fact-checking using multimodal large language models. *arXiv preprint arXiv:2404.12065*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857.

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

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Xusheng Luo, Shaojun Xu, and Changliu Liu. 2023. Obtaining hierarchy from human instructions: an llms-based approach. In *CoRL 2023 Workshop on Learning Effective Abstractions for Planning (LEAP)*.

Grégoire Mialon, Clémentine Fourrier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. Gaia: a benchmark for general ai assistants. *arXiv preprint arXiv:2311.12983*.

OpenAI. Learning to reason with LLMs.

Keivalya Pandya and Mehfuza Holia. 2023. Automating customer service using langchain: Building custom open-source gpt chatbot for organizations. *arXiv preprint arXiv:2310.05421*.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, Tung Nguyen, Daron Anderson, Imad Ali Shah, Mikhail Doroshenko, Alun Cennyth Stokes, Mobeen Mahmood, Jaeho Lee, Oleksandr Pokutnyi, Oleg Iskra, Jessica P. Wang, Robert Gerbicz, John-Clark Levin, Serguei Popov, Fiona Feng, Steven Y. Feng, Haoran Zhao, Michael Yu, Varun Gangal, Chelsea Zou, Zihan Wang, Mstyslav Kazakov, Geoff Galgon, Johannes Schmitt, Alvaro Sanchez, Yongki Lee, Will Yeadon, Scott Sauers, Marc Roth, Chidozie Agu, Søren Riis, Fabian Giska, Saiteja Utpala, Antrell Cheatom, Zachary Giboney, Gashaw M. Goshu, Sarah-Jane Crowson, Mohinder Maheshbhai Naiya, Noah Burns, Lennart Finke, Zerui Cheng, Hyunwoo Park, Francesco Fournier-Facio, Jennifer Zampese, John Wydallis, John B. Wydallis, Ryan G. Hoerr, Mark Nandor, Tim Gehrunger, Jiaqi Cai, Ben McCarty, Jungbae Nam, Edwin Taylor, Jun Jin, Gautier Abou Loume, Hangrui Cao, Alexis C Garretson, Damien Sileo, Qiuyu Ren, Doru Cojoc, Pavel Arkhipov, Usman Qazi, Aras Bacho, Lianghui Li, Sumeet Motwani, Christian Schroeder de Witt, Alexei Kopylov, Johannes Veith, Eric Singer, Paolo Rissone, Jaehyeok Jin, Jack Wei Lun Shi, Chris G. Willcocks, Ameya Prabhu, Longke Tang, Kevin Zhou, Emily de Oliveira Santos, Andrey Pupasov Maksimov, Edward Vendrow, Kengo Zenitani, Joshua Robinson, Aleksandar Mikov, Julien Guillod, Yuqi Li, Ben Pageler, Joshua Vendrow, Vladyslav Kuchkin, Pierre Marion, Denis Efremov, Jayson Lynch, Kaiqu Liang, Andrew Gritsevskiy, Dakotah Martinez, Nick Crispino, Dimitri Zvonkine, Natanael Wildner Fraga, Saeed Soori, Ori Press, Henry Tang, Julian Salazar, Sean R. Green, Lina Brüssel, Moon Twayana, Aymeric Dieuleveut, T. Ryan Rogers, Wenjin Zhang, Ross Finocchio, Bikun Li, Jinzhou Yang, Arun Rao, Gabriel Loiseau, Mikhail Kalinin, Marco Lukas, Ciprian Manolescu, Nate Stambaugh, Subrata Mishra, Ariel Ghislain Kemogne Kamdoum, Tad Hogg, Alvin Jin, Carlo Bosio, Gongbo Sun, Brian P Coppola, Haline Heidinger, Rafael Sayous, Stefan Ivanov, Joseph M Cavanagh, Jiawei Shen, Joseph Marvin Imperial, Philippe Schwaller, Shaipranesh Senthilkuma, Andres M Bran, Andres Algaba, Brecht Verbeken, Kelsey Van den Houte, Lynn Van Der Sypt, David Noever, Lisa Schut, Ilia Sucholutsky, Evgenii Zheltonozhskii, Qiaochu Yuan, Derek Lim, Richard Stanley, Shankar Sivarajan, Tong Yang, John Maar, Julian Wykowski, Martí Oller, Jennifer Sandlin, Anmol Sahu, Cesare Giulio Ardito, Yuzheng Hu, Felipe Meneguitti Dias, Tobias Kreiman, Kaivalya Rawal, Tobias Garcia Vilchis, Yuexuan Zu, Martin Lackner, James Koppel, Jeremy Nguyen, Daniil S. Antonenko, Steffi Chern, Bingchen Zhao, Pierrot Arsene, Sergey Ivanov, Rafał Poświata, Chenguang Wang, Daofeng Li, Donato Crisostomi, Ali Dehghan, Andrea Achilleos, John Arnold Ambay, Benjamin Myklebust, Archan Sen, David Perrella, Nurdin Kaparov, Mark H Inlow, Allen Zang, Kalyan Ramakrishnan, Daniil Orel, Vladislav Poritski, Shalev Ben-David, Zachary Berger, Parker Whitfill, Michael Foster, Daniel Munro, Linh Ho, Dan Bar Hava, Aleksey Kuchkin, Robert Lauff, David Holmes, Frank Sommerhage, Anji Zhang, Richard Moat, Keith Schneider, Daniel Pyda, Zakayo Kazibwe, Mukhwinder Singh, Don Clarke, Dae Hyun Kim, Sara Fish, Veit Elser, Victor Efren Guadarrama Vilchis, Immo Klose, Christoph Demian, Ujjwala Anantheswaran, Adam Zweiger, Guglielmo Albani, Jeffery Li, Nicolas Daans, Maksim Radionov, Václav Rozhoň, Vincent Ginis, Ziqiao Ma, Christian Stump, Jacob Platnick, Volodymyr Nevirkovets, Luke Basler, Marco Piccardo, Niv Cohen, Virendra Singh, Josef Tkadlec, Paul Rosu, Alan Goldfarb, Piotr Padlewski, Stanislaw Barzowski, Kyle Montgomery, Aline Menezes, Arkil Patel, Zixuan Wang, Jamie Tucker-Foltz, Jack Stade, Declan Grabb, Tom Goertzen, Fereshteh Kazemi, Jeremiah Milbauer, Abhishek Shukla, Hossam Elgnainy, Yan Carlos Leyva Labrador, Hao He, Ling Zhang, Alan Givré, Hew Wolff, Gözdenur Demir, Muhammad Fayez Aziz, Younesse Kaddar, Ivar Ängquist, Yanxu Chen, Elliott Thornley, Robin Zhang, Jiayi Pan, Antonio Terpin, Niklas Muennighoff, Hailey Schoelkopf, Eric Zheng, Avishy Carmi, Jainam Shah, Ethan D. L. Brown, Kelin Zhu, Max Bartolo, Richard Wheeler, Andrew Ho, Shaul Barkan, Jiaqi Wang, Martin Stehberger, Egor Kretov, Peter Bradshaw, JP Heimonen, Kaustubh Sridhar, Zaki Hossain, Ido Akov, Yury Makarychev, Joanna Tam, Hieu Hoang, David M. Cunningham, Vladimir Goryachev, Demosthenes Patramanis, Michael Krause, Andrew Redenti, David Aldous, Jesyin Lai, Shannon Coleman, Jiangnan Xu, Sangwon Lee, Ilias Magoulas, Sandy Zhao, Ning Tang, Michael K. Cohen, Micah Carroll, Orr Paradise, Jan Hendrik Kirchner, Stefan Steinerberger, Maksym Ovchynnikov, Jason O. Matos, Adithya Shenoy, Michael Wang, Yuzhou Nie, Paolo Giordano, Philipp Petersen, Anna Sztyber-Betley, Paolo Faraboschi, Robin Riblet, Jonathan Crozier, Shiv Halasyamani, Antonella Pinto, Shreyas Verma, Prashant Joshi, Eli Meril, Zheng-Xin Yong, Allison Tee, Jérémy Andréoletti, Orion Weller, Raghav Singhal, Gang Zhang, Alexander Ivanov, Seri Khoury, Nils Gustafsson, Hamid Mostaghimi, Kunvar Thaman, Qijia Chen, Tran Quoc Khánh, Jacob Loader, Stefano Cavalleri, Hannah Szlyk, Zachary Brown, Himanshu Narayan, Jonathan Roberts, William Alley, Kunyang Sun, Ryan Stendall, Max Lamparth, Anka Reuel, Ting Wang, Hanmeng Xu, Pablo Hernández-Cámara, Freddie Martin, Thomas Preu, Tomek Korbak, Marcus Abramovitch, Dominic Williamson, Ida Bosio, Ziye Chen, Biró Bálint, Eve J. Y. Lo, Maria Inês S. Nunes, Yibo Jiang, M Saiful Bari, Peyman Kassani, Zihao Wang, Behzad Ansarinejad, Yewen Sun, Stephane Durand, Guillaume Douville, Daniel Tordera, George Balabanian, Earth Anderson, Lynna Kvistad, Alejandro José Moyano, Hsiaoyun Milliron, Ahmad Sakor, Murat Eron, Isaac C. McAlister, Andrew Favre D. O., Shailesh Shah, Xiaoxiang Zhou, Firuz Kamalov, Ronald Clark, Sherwin Abdoli, Tim Santens, Harrison K Wang, Evan

Chen, Alessandro Tomasiello, G. Bruno De Luca, Shi-Zhuo Looi, Vinh-Kha Le, Noam Kolt, Niels Mündler, Avi Semler, Emma Rodman, Jacob Drori, Carl J Fossum, Luk Gloor, Milind Jagota, Ronak Pradeep, Honglu Fan, Tej Shah, Jonathan Eicher, Michael Chen, Kushal Thaman, William Merrill, Moritz Firsching, Carter Harris, Stefan Ciobâcă, Jason Gross, Rohan Pandey, Ilya Gusev, Adam Jones, Shashank Agnihotri, Pavel Zhelnov, Siranut Usawasutsakorn, Mohammadreza Mofayezi, Alexander Piperski, Marc Carauleanu, David K. Zhang, Kostiantyn Dobarskyi, Dylan Ler, Roman Leventov, Ignat Soroko, Thorben Jansen, Scott Creighton, Pascal Lauer, Joshua Duersch, Vage Taamazyan, Dario Bezzi, Wiktor Morak, Wenjie Ma, William Held, Tran Đuc Huy, Ruicheng Xian, Armel Randy Zebaze, Mohanad Mohamed, Julian Noah Leser, Michelle X Yuan, Laila Yacar, Johannes Lengler, Katarzyna Olszewska, Hossein Shahrtash, Edson Oliveira, Joseph W. Jackson, Daniel Espinosa Gonzalez, Andy Zou, Muthu Chidambaram, Timothy Manik, Hector Haffenden, Dashiell Stander, Ali Dasouqi, Alexander Shen, Emilien Duc, Bita Golshani, David Stap, Mikalai Uzhou, Alina Borisovna Zhidkovskaya, Lukas Lewark, Miguel Orbegozo Rodriguez, Mátyás Vincze, Dustin Wehr, Colin Tang, Shaun Phillips, Fortuna Samuele, Jiang Muzhen, Fredrik Ekström, Angela Hammon, Oam Patel, Faraz Farhidi, George Medley, Forough Mohammadzadeh, Madellene Peñaflor, Haile Kassahun, Alena Friedrich, Claire Sparrow, Rayner Hernandez Perez, Taom Sakal, Omkar Dhamane, Ali Khajegili Mirabadi, Eric Hallman, Kenchi Okutsu, Mike Battaglia, Mohammad Maghsoudimehrabani, Alon Amit, Dave Hulbert, Roberto Pereira, Simon Weber, Handoko, Anton Peristyy, Stephen Malina, Samuel Albanie, Will Cai, Mustafa Mehkary, Rami Aly, Frank Reidegeld, Anna-Katharina Dick, Cary Friday, Jasdeep Sidhu, Hassan Shapourian, Wanyoung Kim, Mariana Costa, Hubeyb Gurdogan, Brian Weber, Harsh Kumar, Tong Jiang, Arunim Agarwal, Chiara Ceconello, Warren S. Vaz, Chao Zhuang, Haon Park, Andrew R. Tawfeek, Daattavya Aggarwal, Michael Kirchhof, Linjie Dai, Evan Kim, Johan Ferret, Yuzhou Wang, Minghao Yan, Krzysztof Burdzy, Lixin Zhang, Antonio Franca, Diana T. Pham, Kang Yong Loh, Joshua Robinson, Abram Jackson, Shreen Gul, Gunjan Chhablani, Zhehang Du, Adrian Cosma, Jesus Colino, Colin White, Jacob Votava, Vladimir Vinnikov, Ethan Delaney, Petr Spelda, Vit Stritecky, Syed M. Shahid, Jean-Christophe Mourrat, Lavr Vetoshkin, Koen Sponselee, Renas Bacho, Florencia de la Rosa, Xiuyu Li, Guillaume Malod, Leon Lang, Julien Laurendeau, Dmitry Kazakov, Fatimah Adesanya, Julien Portier, Lawrence Hollom, Victor Souza, Yuchen Anna Zhou, Julien Degorre, Yiğit Yalın, Gbenga Daniel Obikoya, Luca Arnaboldi, Rai, Filippo Bigi, M. C. Boscá, Oleg Shumar, Kaniuar Bacho, Pierre Clavier, Gabriel Recchia, Mara Popescu, Nikita Shulga, Ngefor Mildred Tanwie, Denis Peskoff, Thomas C. H. Lux, Ben Rank, Colin Ni, Matthew Brooks, Alesia Yakimchyk, Huanxu, Liu, Olle Häggström, Emil Verkama, Hans Gundlach, Leonor Brito-Santana, Brian Amaro, Vivek Vajipey, Rynaa Grover, Yiyang Fan, Gabriel Poesia Reis e Silva, Linwei Xin, Yosi Kratish, Jakub Łucki, Wen-Ding Li, Sivakanth Gopi, Andrea Caciolai, Justin Xu, Kevin Joseph Scaria, Freddie Vargus, Farzad Habibi, Long, Lian, Emanuele Rodolà, Jules Robins, Vincent Cheng, Tony Fruhauff, Brad Raynor, Hao Qi, Xi Jiang, Ben Segev, Jingxuan Fan, Sarah Martinson, Erik Y. Wang, Kaylie Hausknecht, Michael P. Brenner, Mao Mao, Xinyu Zhang, David Avagian, Eshawn Jessica Scipio, Alon Ragoler, Justin Tan, Blake Sims, Rebeka Plecnik, Aaron Kirtland, Omer Faruk Bodur, D. P. Shinde, Zahra Adoul, Mohamed Zekry, Ali Karakoc, Tania C. B. Santos, Samir Shamseldeen, Loukmane Karim, Anna Liakhovitskaia, Nate Resman, Nicholas Farina, Juan Carlos Gonzalez, Gabe Maayan, Sarah Hoback, Rodrigo De Oliveira Pena, Glen Sherman, Elizabeth Kelley, Hodjat Mariji, Rasoul Pouriamanesh, Wentao Wu, Sandra Mendoza, Ismail Alarab, Joshua Cole, Danyelle Ferreira, Bryan Johnson, Mohammad Safdari, Liangti Dai, Siriphan Arthornthurasuk, Alexey Pronin, Jing Fan, Angel Ramirez-Trinidad, Ashley Cartwright, Daphiny Pottmaier, Omid Taheri, David Outevsky, Stanley Stepanic, Samuel Perry, Luke Askew, Raúl Adrián Huerta Rodríguez, Ali M. R. Minissi, Sam Ali, Ricardo Lorena, Krishnamurthy Iyer, Arshad Anil Fasiludeen, Sk Md Salauddin, Murat Islam, Juan Gonzalez, Josh Ducey, Maja Somrak, Vasilios Mavroudis, Eric Vergo, Juehang Qin, Benjámin Borbás, Eric Chu, Jack Lindsey, Anil Radhakrishnan, Antoine Jallon, I. M. J. McInnis, Pawan Kumar, Laxman Prasad Goswami, Daniel Bugas, Nasser Heydari, Ferenc Jeanplong, Archimedes Apronti, Abdallah Galal, Ng Ze-An, Ankit Singh, Joan of Arc Xavier, Kanu Priya Agarwal, Mohammed Berkani, Benedito Alves de Oliveira Junior, Dmitry Malishev, Nicolas Remy, Taylor D. Hartman, Tim Tarver, Stephen Mensah, Javier Gimenez, Roselynn Grace Montecillo, Russell Campbell, Asankhaya Sharma, Khalida Meer, Xavier Alapont, Deepakkumar Patil, Rajat Maheshwari, Abdelkader Dendane, Priti Shukla, Sergei Bogdanov, Sören Möller, Muhammad Rehan Siddiqi, Prajvi Saxena, Himanshu Gupta, Innocent Enyekwe, Ragavendran P V, Zienab EL-Wasif, Aleksandr Maksapetyan, Vivien Rossbach, Chris Harjadi, Mohsen Bahaloohoreh, Song Bian, John Lai, Justine Leon Uro, Greg Bateman, Mohamed Sayed, Ahmed Menshawy, Darling Duclosel, Yashaswini Jain, Ashley Aaron, Murat Tiryakioglu, Sheeshram Siddh, Keith Krenek, Alex Hoover, Joseph McGowan, Tejal Patwardhan, Summer Yue, Alexandr Wang, and Dan Hendrycks. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. 2024. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

DeepSeek Team. 2024. Deepseek-r1-lite-preview is now live: unleashing supercharged reasoning power.

Qwen Team. Qwq: Reflect deeply on the boundaries of the unknown, november 2024. *URL https://qwenlm. github. io/blog/qwq-32b-preview*.

Vincent A Traag, Ludo Waltman, and Nees Jan Van Eck. 2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12.

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.

Junde Wu, Jiayuan Zhu, Yunli Qi, Jingkun Chen, Min Xu, Filippo Menolascina, and Vicente Grau. 2024. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Hui Yang, Sifu Yue, and Yunzhong He. 2023. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, et al. 2024. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*.

7 Appendix

7.1 Participant Demographics

We recruited a total of 23 PhD students from medical, law, and business schools, ensuring a diverse academic and geographical representation. The distribution of participants across disciplines and countries is as follows:

Academic Background

- **Medical School**: 10 students (43%)
- **Law School**: 7 students (30%)
- **Business School**: 6 students (26%)

Geographical Distribution

- **China**: 9 students (39%)
- **United Kingdom (UK)**: 8 students (35%)
- **United States (USA)**: 6 students (26%)

All participants were PhD candidates with at least three years of research experience in their respective fields. Their expertise spanned across clinical research, AI ethics, legal reasoning, and financial modeling. Participation was entirely voluntary, driven by academic interest in AI-driven reasoning systems.

7.2 Human Evaluation Survey

Please assess each response generated by the model based on the following criteria. Provide your rating on a scale from 1 to 5, where 1 is the lowest and 5 is the highest. You may also leave optional comments to clarify your reasoning.

1. **Interest Level (Int.)**
   - How engaging and intellectually stimulating is the generated response?
   - Rating Scale: **1:** Not engaging - fails to capture interest. **2:** Somewhat uninteresting - lacks depth or novelty. **3:** Neutral - informative but not particularly engaging. **4:** Engaging - provides depth and insight. **5:** Highly engaging - deep and thought-provoking.
   - **Optional Comment**: What aspects of the response contributed to or detracted from its interest level?

2. **Organization (Org.)**
   - How well-structured and logically organized is the response?
   - Rating Scale: **1:** Very disorganized - hard to follow. **2:** Somewhat disorganized - requires effort to understand. **3:** Neutral - moderately structured but could be clearer. **4:** Well-organized - logical and easy to follow. **5:** Exceptionally structured - very clear and logically ordered.
   - **Optional Comment**: Are there any areas where the response could be better structured?

3. **Relevance (Rel.)**
   - How relevant is the response to the research question posed?
   - Rating Scale: **1:** Not relevant - off-topic or misleading. **2:** Somewhat relevant - partially addresses the question. **3:** Neutral - addresses the question but with some tangents. **4:** Mostly relevant - minor deviations but generally on point. **5:** Highly relevant - fully addresses the question.
   - **Optional Comment**: Did the response stay on topic? If not, how did it deviate?

4. **Coverage (Cov.)**
   - How comprehensively does the response cover the question?
   - Rating Scale: **1:** Superficial - lacks depth and critical information. **2:** Somewhat incomplete - covers only basic aspects. **3:** Neutral - adequate coverage but missing key details. **4:** Mostly complete - only minor gaps. **5:** Fully comprehensive - deeply covers all necessary aspects.
   - **Optional Comment**: Are there any areas where additional information would improve the response?

Thank you for your participation!