From Underspecified Queries to Clear Research Scope: Context-Aware Planning for Deep Research

Anonymous ACL submission

Abstract

Current deep research typically generate research brief by expanding user queries using large language model (LLM). This expansion-only approach often produces underspecified formulations that lack clear scope and contextual grounding, creating a boundary ambiguity challenge that leads to redundant exploration and inefficient tool usage. While existing methods rely on iterative user clarification, they often fail when users cannot provide precise domain constraints. In this paper, we propose that robust research planning requires proactively exposing implicit constraints, rather than relying solely on query expansion or user clarification. We introduce the Context-Aware Planning Framework (CAPF), which utilizes a pre-search phase to surface contextual signals and analyze boundary ambiguities. By integrating these insights into the planning, CAPF generates grounded research briefs that better constrain downstream exploration. Experiments across four complex fact-seeking tasks demonstrate that CAPF consistently outperforms recent deep research agents and agentic retrieval systems. Our analysis further reveals that exposing the agent to the inherent difficulties of a question during planning stage is a critical factor for achieving higher accuracy while simultaneously reducing resource consumption.

1 Introduction

The paradigm of deep research (Zheng et al., 2025; Xu and Peng, 2025; Java et al., 2025) has emerged as a significant advancement over traditional Retrieval Augmented Generation (RAG) systems (Lee et al., 2019; Karpukhin et al., 2020; Mao et al., 2021) for its ability to plan, search, react and synthesize over multiple rounds. Unlike passive RAG systems (Singh et al., 2025) that simply consume retrieved information, deep research systems (Shi et al., 2025) continuously revise the initial plan based on newly acquired evidence and decide how

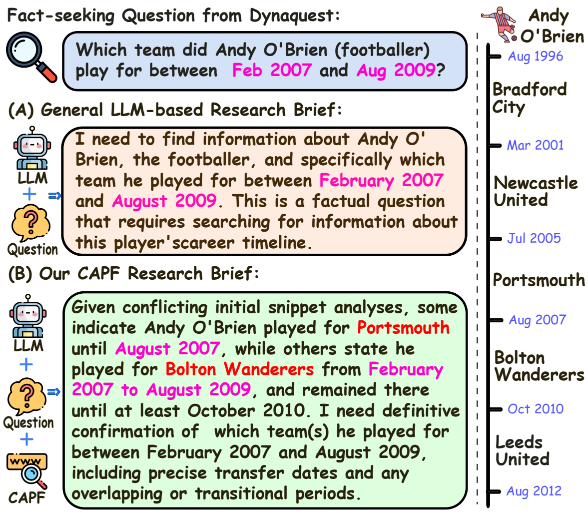

Figure 1: Comparison between an LLM-based research brief and a CAPF-generated research brief. By detecting the transfer conflict, CAPF formulates a precise scope to guide downstream exploration.

to adjust next steps. However, this operational flexibility introduces a critical vulnerability: boundary ambiguity. Without explicitly defined search constraints, agents are prone to over-exploring unconstrained information spaces where irrelevant or boundary-external content becomes entangled with core evidence (Lei et al., 2025; Pham et al., 2025). In such settings, contradictory statements, outdated data, or vague messages do not merely degrade individual responses but propagate through the iterative planning loop. This interference shifts the agent's trajectory, leading to divergent reasoning (Lin et al., 2025b) and excessive resource consumption (Zhang et al., 2025b). For instance, when answering the query "*What is the most recent film to join the top 10 highest-grossing films of all time?*", the correct answer is Ne Zha 2 (2025), but the agent is frequently misled by abundant references to Inside Out 2 (2024).

Existing deep research frameworks (Sun et al., 2025; Wan et al., 2025) primarily rely on iterative user clarification to resolve boundary ambiguity. This strategy is inherently constrained by the user's prior knowledge and their ability to discern latent noise within the problem space. When users cannot provide precise domain constraints, agents remain vulnerable to underconstrained exploration. In this paper, we propose that effective boundary definition should be grounded in intrinsic signals surfaced prior to formal exploration rather than solely on external user feedback. We introduce the Context-Aware Planning Framework (CAPF), which employs a pre-search to expose and analyze implicit constraints and potential ambiguities during the planning stage. As illustrated in Figure 1, this proactive exposure enables the agent to identify critical contextual pivot points, specifically the transfer event in August 2007 involving "Portsmouth" and "Bolton Wanderers". By integrating these identified entity and temporal boundaries into the research plan, the agent effectively constrains the subsequent search space, ensuring more targeted and robust downstream exploration.

We evaluate our framework on four recent question answering (QA) benchmarks, including RealtimeQA (Kasai et al., 2023), HoHQA (Ouyang et al., 2025), DynaQuest (Lin et al., 2025a), and SealQA (Pham et al., 2025). These datasets mirror realistic information environments where ground-truth answers are often obscured by temporal drift, conflicting claims, or latent ambiguities. As a result, simple information retrieval frequently returns documents that are misleading or only superficially relevant but unhelpful, which makes it difficult for conventional systems to identify correct answers. Existing deep research agents also struggle with these underconstrained explorations, as user-provided clarification typically lacks the necessary contextual or temporal granularity. This process may result in excessive tool usage, unnecessary computational cost, and degraded answer quality. Our experimental results demonstrate that by integrating proactive context-aware signals into the planning stage, our framework consistently outperforms state-of-the-art (SOTA) deep research agents and agentic retrieval systems in accuracy. Furthermore, our analysis indicates that CAPF significantly optimizes resource efficiency by constraining the exploration process, it reduces redundant tool usage and computational overhead while maintaining superior performance across all benchmarks.

Our contributions are summarized as follows:

- We identify the boundary ambiguity challenge in current deep research, where underspecified queries and inadequate user clarification lead to inefficient exploration.
- We introduce CAPF, a context-aware planning framework that utilizes pre-search phrase to surface and analyze latent contextual signals and implicit constraints, ensuring the generation of grounded research briefs.
- Evaluations on four realistic QA benchmarks show that CAPF consistently outperforms SOTA deep research agents and agentic retrieval systems.
- Further analysis demonstrates that proactive boundary exposure during planning is critical for enhancing answer accuracy while simultaneously reducing computational overhead.

2 Related Work

2.1 Agentic Reasoning Systems

The emergence of agentic deep research systems (Xu and Peng, 2025; Shi et al., 2025) represents a substantive shift from passive retrieval to active exploration in addressing complex information-seeking tasks. Rather than merely answering a single query, these systems (Java et al., 2025; Wan et al., 2025) are required to operate through structured and iterative cycles of planning, retrieval, reasoning, and revision. A defining characteristic (Sun et al., 2025) of these systems is their reliance on the internal reasoning of LLMs to interpret retrieved information, evaluate its relevance or correctness, filter out noise, and revise their search plans accordingly. This agentic workflow (Li et al., 2025b) allows systems to handle multi-step reasoning and cross-source document synthesis, surpassing the capabilities of static RAG pipelines (Dong et al., 2025; Li et al., 2025c).

However, recent analyses (Lan et al., 2025; Huang et al., 2025) reveal a persistent weakness that current deep research agents often exhibit intermediate reasoning traces marked by frequent backtracking, shallow exploration, and an inability to anticipate challenging subtasks when faced with difficult tasks. One of the bottlenecks lies in the construction of the research plan. Existing agents (Lei et al., 2025; Zheng et al., 2025) typically respond to errors or noisy evidence only after such issues emerge within the exploration cycle, relying on the LLM's internal reasoning to detect and correct mistakes. In other words, they seldom define a clear research scope and anticipate potential pitfalls during the planning stage, leaving the subsequent search process vulnerable to avoidable failures. In contrast, our approach takes a proactive and anticipatory perspective by surfacing contextual signals at the outset of the research process. Making these signals explicit early allows the agent to recognize upcoming difficulties and incorporate them into the plan, which produces more robust and efficient exploration.

2.2 Noisy Evidence in Context

The role of noisy evidence (Zhang et al., 2025a; Li et al., 2025a) is typically framed as a negative factor in reasoning systems, as low-quality information is assumed to disrupt exploration by degrading answer accuracy and increasing resource consumption. In traditional information retrieval research, mechanisms for conflict detection or factual verification (Zheng et al., 2024) are usually applied only after the retrieval stage. In current deep research agents (Java et al., 2025; Xu and Peng, 2025), noise handling is largely shifted to the research exploration phase, where LLMs rely on their internal reasoning ability to assess credibility and resolve contradictions. As a result, noisy or contradictory evidence is handled in a reactive manner, being filtered or corrected only after it has already influenced the agent's workflow. This reactive treatment can easily misguide the agent, leading to two major consequences. First, it increases the likelihood of producing incorrect conclusions (Wu et al., 2025). Second, it amplifies the overall cost of the search process by triggering unnecessary exploration and tool usage (Zhang et al., 2025b).

Recent research has begun to reconsider the role of noisy evidence, highlighting several ways in which weak signals can benefit reasoning systems. For instance, studies on conflict-aware evaluation demonstrate that partially incorrect evidence can reveal model vulnerabilities and promote more reliable decision-making (Wu et al., 2024). In parallel, verification-feedback frameworks show that low-confidence or weak evidence can strengthen retrieval by guiding subsequent evidence selection and encouraging iterative refinement (Deng et al., 2025). Despite this growing recognition, prior work has not explored how research agents might incorporate potential noise during planning. We address this gap by treating weak signals as critical cues to define a clear research boundary. By surfacing noisy snippets before execution, our framework allows the agent to anticipate difficulty, adjust its plan proactively, and improve robustness and efficiency on complex tasks.

3 Methodology

In this section, we present our CAPF, a context-aware planning framework that exposes the agent to the contextual boundaries of a question during the planning stage, enabling more effective deep research exploration on complex tasks.

3.1 Preliminaries

The goal of a deep research agent is to generate a comprehensive, evidence-grounded report that addresses the user's query. Given a query $\mathcal{Q}$, the agent typically begins by using an LLM to construct a research plan $\mathcal{O}$ that guides the subsequent research process:

$$\mathcal{O}_{\text{base}} \sim \mathcal{P}_{\text{LLM}}(\mathcal{O} \mid \mathcal{Q}), \tag{1}$$

where $\mathcal{P}_{\text{LLM}}$ denotes the planning distribution induced by the LLM when conditioned solely on the original query $\mathcal{Q}$. Then the research supervisor examines the plan $\mathcal{O}_{\text{base}}$, decomposes it into manageable tasks $\{\mathcal{O}^1, \mathcal{O}^2, ..., \mathcal{O}^N\}$, and delegates these tasks to the corresponding research sub-agents:

$$\mathcal{D}(\mathcal{O}_{\text{base}}) = \{\mathcal{O}^1, \mathcal{O}^2, ..., \mathcal{O}^N\}, \tag{2}$$

where $\mathcal{D}(\mathcal{O}_{\text{base}})$ denotes the decomposition of the plan into N manageable sub-tasks, and $\mathcal{O}^n$ refers to the n-th sub-task.

For each sub-task $\mathcal{O}^n$, the research supervisor initiates an iterative research process carried out by a pool of sub-agents. At t-th iteration, the research topic τ_t^n is defined as:

$$\tau_t^n = \mathcal{T}_{\text{LLM}}(\mathcal{O}^n, \mathcal{F}_{t-1}^n), \tag{3}$$

where $\mathcal{T}_{\text{LLM}}$ denotes an LLM-based topic generator that produces the research topic conditioned on the sub-task and past findings. $\mathcal{F}_{t-1}^n$ represents the set of findings produced by the sub-agents in the previous iteration based on the retrieved information and their corresponding analyses.

Finally, once all sub-tasks have been completed, the research agent integrates findings $\mathcal{F}$ with the original research plan $\mathcal{O}_{\text{base}}$ and query $\mathcal{Q}$. The final report is generated as:

$$\mathcal{R} = \mathcal{G}_{\text{LLM}}\left(\mathcal{Q}, \mathcal{O}_{\text{base}}, \bigcup_{n=1}^{N}\bigcup_{t=1}^{T}\mathcal{F}_t^n\right), \tag{4}$$

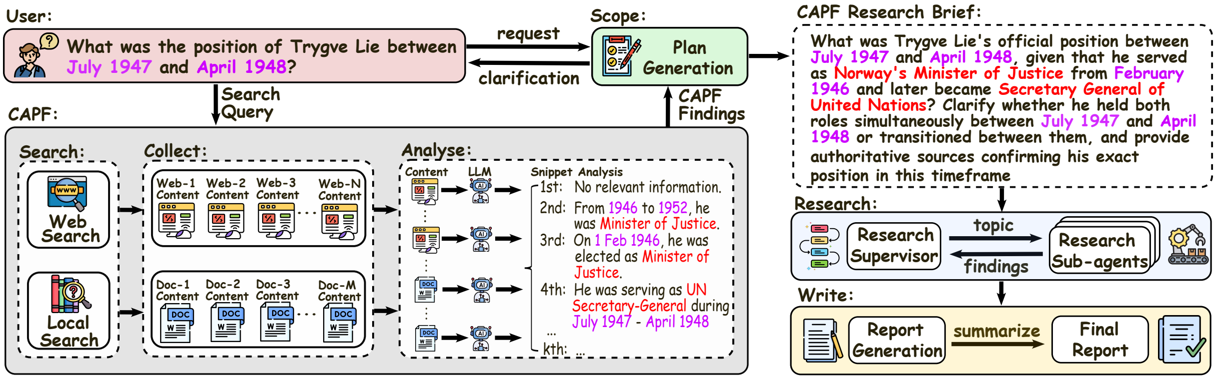

Figure 2: Overview of our CAPF framework, which performs fast pre-search over web and local sources to surface weak signals before research exploration begins. By analyzing noisy snippets and clarifying the research scope, CAPF enables the agent to generate a more accurate and robust research plan.

where $\mathcal{G}_{\text{LLM}}$ denotes the LLM-based report generator, and $\bigcup_{t=1}^{T} \mathcal{F}_t^n$ represents the complete collection of findings accumulated for sub-task n over all iterations until its termination.

3.2 CAPF

CAPF operates primarily at the research planning stage, producing a more informed and resilient plan that subsequently strengthens the entire deep research workflow. As illustrated in Figure 2, it conducts a fast pre-search over web and local sources using the user's query to surface weak signals before the main exploration process begins. Specifically, given a user query $\mathcal{Q}$, CAPF retrieves an initial set of noisy or ambiguous snippets from external and internal sources:

$$\mathcal{S}_{\text{pre}} = \mathcal{R}_{\text{web}}(\mathcal{Q}) \cup \mathcal{R}_{\text{local}}(\mathcal{Q}), \tag{5}$$

where $\mathcal{S}_{\text{pre}}$ denotes the weak-signal evidence collected during the pre-search stage, and $\mathcal{R}(\cdot)$ represents the search function.

For each collected snippet s_i, CAPF processes it individually and sends it to the LLM to extract potential cues from its content:

$$c_i = \mathcal{A}_{\text{LLM}}(s_i), \quad s_i \in \mathcal{S}_{\text{pre}}, \tag{6}$$

and aggregates all extracted analyses into:

$$\mathcal{C}_{\text{pre}} = \{\, c_i \mid s_i \in \mathcal{S}_{\text{pre}} \,\}, \tag{7}$$

where $\mathcal{A}_{\text{LLM}}$ denotes the LLM-based evidence analyzer, and c_i is the analysis produced for snippet s_i.

Using the extracted cue set $\mathcal{C}_{\text{pre}}$, CAPF augments the planning stage by conditioning the LLM on both the original query and the surfaced weak signals. The enhanced research plan is generated as:

$$\mathcal{O}_{\text{capf}} \sim \mathcal{P}_{\text{LLM}}(\mathcal{O} \mid \mathcal{Q}, \mathcal{C}_{\text{pre}}), \tag{8}$$

where $\mathcal{P}_{\text{LLM}}$ denotes the planning distribution induced by the LLM when provided with additional contextual cues extracted during pre-search. By incorporating $\mathcal{C}_{\text{pre}}$, the planner can explicitly account for early indications of ambiguity, missing information, or potential conflicts, thereby producing a more informed and resilient plan.

With CAPF, the extracted cue set $\mathcal{C}_{\text{pre}}$ influences not only the construction of the research plan but also the downstream task decomposition and iterative topic generation through $\mathcal{O}_{\text{capf}}$. The resulting sub-tasks are obtained by decomposing the enhanced plan as:

$$\mathcal{D}(\mathcal{O}_{\text{capf}}) = \{\mathcal{O}^1_{\text{capf}}, \ldots, \mathcal{O}^{N_{\text{capf}}}_{\text{capf}}\}, \tag{9}$$

where the decomposition is shaped by the weak signals surfaced during pre-search. During iterative exploration, the topic for the t-th iteration of sub-task n is generated as:

$$\tau^n_{t,\text{capf}} = \mathcal{T}_{\text{LLM}}\left(\mathcal{O}^n_{\text{capf}}, \mathcal{F}^n_{t-1}\right), \tag{10}$$

which enables the agent to incorporate weak signal cues into sub-task refinement, priority decisions, and the resolution of early ambiguities.

By integrating these contextual signal cues throughout both planning and execution, CAPF reshapes the entire deep research workflow, enabling the agent to anticipate difficulties, refine its strategy, and conduct exploration more effectively. The full workflow is detailed in Algorithm 1.

Algorithm 1 CAPF for Deep Research

Require: user query $\mathcal{Q}$, web search function $\mathcal{R}_{\text{web}}$, local search function $\mathcal{R}_{\text{local}}$, snippet analyzer $\mathcal{A}_{\text{LLM}}$, LLM planner $\mathcal{P}_{\text{LLM}}$, plan decomposer $\mathcal{D}(\cdot)$, topic generator $\mathcal{T}_{\text{LLM}}$, report generator $\mathcal{G}_{\text{LLM}}$
Ensure: final research report $\mathcal{R}_{\text{final}}$

$\mathcal{S}_{\text{pre}} \leftarrow \mathcal{R}_{\text{web}}(\mathcal{Q}) \cup \mathcal{R}_{\text{local}}(\mathcal{Q})$ ▷ Fast pre-search to collect weak-signal snippets
$\mathcal{C}_{\text{pre}} \leftarrow \emptyset$
for each snippet $s_i \in \mathcal{S}_{\text{pre}}$ **do**
  $c_i \leftarrow \mathcal{A}_{\text{LLM}}(s_i)$ ▷ Analyze snippet and extract cues
  $\mathcal{C}_{\text{pre}} \leftarrow \mathcal{C}_{\text{pre}} \cup \{c_i\}$
end for
CAPF-enhanced plan $\mathcal{O}_{\text{capf}} \sim \mathcal{P}_{\text{LLM}}(\mathcal{O} \mid \mathcal{Q}, \mathcal{C}_{\text{pre}})$
$\{\mathcal{O}^1_{\text{capf}}, \ldots, \mathcal{O}^{N_{\text{capf}}}_{\text{capf}}\} \leftarrow \mathcal{D}(\mathcal{O}_{\text{capf}})$ ▷ Decompose plan into sub-tasks
for $n = 1$ to N_{capf} **do**
  $\mathcal{F}^n_0 \leftarrow \emptyset$
  $t \leftarrow 1$
  while $|\mathcal{F}^n_{t-1}| < \theta$ **do**
    $\tau^n_{t,\text{capf}} \leftarrow \mathcal{T}_{\text{LLM}}(\mathcal{O}^n_{\text{capf}}, \mathcal{F}^n_{t-1})$
    $f^n_t \leftarrow \text{EXECUTE}(\tau^n_{t,\text{capf}})$
    $\mathcal{F}^n_t \leftarrow \mathcal{F}^n_{t-1} \cup \{f^n_t\}$
    $t \leftarrow t + 1$
  end while
end for
$\mathcal{F}_{\text{all}} \leftarrow \bigcup_{n=1}^{N_{\text{capf}}} \mathcal{F}^n$ ▷ Aggregate findings across all sub-tasks
$\mathcal{R}_{\text{final}} \leftarrow \mathcal{G}_{\text{LLM}}(\mathcal{Q}, \mathcal{O}_{\text{capf}}, \mathcal{F}_{\text{all}})$ ▷ Synthesize final evidence-grounded report
return $\mathcal{R}_{\text{final}}$

4 Experiments

4.1 Experimental Setup

Evaluation Datasets. To evaluate the effectiveness of our framework, we select datasets whose answers are highly susceptible to noisy or misleading evidence, including RealtimeQA (Kasai et al., 2023), HoHQA (Ouyang et al., 2025), DynaQuest (Lin et al., 2025a), and SealQA (Pham et al., 2025). **Note:** Many questions in these benchmarks are extremely time-sensitive, and must be updated to ensure the correctness of the test. We provide details of the dataset updates and verification process in Appendix A.

Comparison and Metrics. We compare CAPF with both proprietary and open-source deep research agents, including the official deep research systems in GPT-5.1 (OpenAI, 2025), Qwen3-Max (Qwen AI, 2025), and Kimi-K2 (Moonshot AI, 2025), as well as a locally deployed LangChain-AI open deep research agent (LangChain AI, 2025). We also include a range of retrieval baselines, from naive-RAG (Karpukhin et al., 2020) and web-search pipelines (Tavily AI, 2024) to agentic RAG systems such as Madam-RAG (Wang et al., 2025) and FastGPT (FastGPT AI, 2024) (web search enabled). Evaluation uses both rule-based and model-based metrics: EM and F1 scores (Rajpurkar et al., 2016), and an LLM-as-a-Judge setup (Zheng et al., 2023) with DeepSeek-R1 (Guo et al., 2025). Since deep research systems generate long-form reports, we summarize each report using the base model to obtain the final answer for evaluation. The efficiency of different systems is analyzed separately in Section 5. Additional experimental details are provided in Appendix B.

4.2 Experimental Results

Table 1 presents the overall performance of representative deep research agents and information retrieval systems against our CAPF, leading to the following observations:

(i) **Access to up-to-date information is critical for accuracy**. Systems equipped with web search capabilities consistently outperform those relying solely on local knowledge bases, as the latter are prone to producing incorrect answers due to information staleness. This trend is clearly reflected in the superior performance of Web-search compared to Naive-RAG. We further observe that more recently trained models tend to achieve higher accuracy than earlier versions. A plausible explanation is that newer models have been exposed to more recent news and web corpora during training, whereas older models have not.

(ii) **Deep research agents outperform traditional information retrieval systems**. Across all evaluated settings, deep research agents consistently achieve better performance than traditional information retrieval systems, including web-enabled systems such as FastGPT. We attribute this performance gap to the iterative exploration nature of deep research agents. This performance gap can be attributed to the iterative exploration paradigm adopted by deep research agents. By repeatedly incorporating intermediate feedback and dynamically adjusting their research strategies during execution, these agents exhibit greater robustness to noisy evidence and a stronger capacity for factual verification.

(iii) **Exposing the agent to a clear contextual scope during planning is crucial for improving accuracy**. Our framework enables a controlled comparison with the LangChain-AI open deep research system, as both approaches are locally deployed agents built on LangGraph. The primary distinction is the integration of the CAPF mod-

Method	RealtimeQA			HoHQA			Dynaquest			SealQA		
	EM	F1	MBE	EM	F1	MBE	EM	F1	MBE	EM	F1	MBE
LLM Direct												
Qwen3-235B (Yang et al., 2025)	13.11	25.30	24.39	5.11	17.15	9.85	9.14	24.16	15.14	15.35	19.82	18.89
DeepSeek-V3 (Liu et al., 2024)	29.87	40.65	43.29	5.83	15.10	9.12	18.09	30.79	28.07	14.57	17.40	15.35
DeepSeek-V3.2 (Liu et al., 2025)	31.70	44.04	45.73	7.66	20.82	15.32	20.82	38.08	34.38	18.50	23.66	25.19
GPT-5.1 (OpenAI, 2025)	34.15	49.57	50.91	10.22	24.97	25.55	22.08	40.80	37.53	20.87	28.13	28.74
Information Retrieval System												
Naive-RAG (Karpukhin et al., 2020)	9.45	17.33	26.52	4.01	18.04	22.37	3.15	13.14	21.14	9.06	11.26	12.99
Web-Search (Tavily AI, 2024)	38.41	54.12	59.15	28.47	41.16	38.32	28.39	41.76	38.48	13.39	18.00	14.17
Madam-RAG (Wang et al., 2025)	20.12	28.35	30.48	13.87	30.08	28.10	16.08	27.55	23.66	21.65	27.80	31.89
FastGPT (FastGPT AI, 2024)	42.07	59.46	60.06	21.17	43.64	40.51	33.75	48.49	50.78	23.23	26.58	31.49
Deep Research Agent												
Open DR (LangChain AI, 2025)	39.33	56.87	58.84	32.48	52.53	51.46	37.54	51.62	52.68	18.90	25.68	28.34
Kimi-K2 DR (Moonshot AI, 2025)	49.09	67.60	70.73	42.33	62.41	64.60	41.50	54.13	54.57	31.50	36.47	32.68
Qwen3-Max DR (Qwen AI, 2025)	45.12	59.70	64.63	38.69	59.40	63.87	43.40	56.80	53.31	30.31	35.90	35.83
GPT-5.1 DR (OpenAI, 2025)	57.62	72.03	76.22	46.72	66.76	**71.17**	45.30	59.91	57.72	34.25	40.55	42.52
Our framework												
CAPF DR$_{Qwen3-235B}$	46.95	61.24	67.07	40.14	56.47	58.02	40.38	53.67	52.36	31.89	35.24	33.46
CAPF DR$_{DeepSeek-V3}$	59.75	73.95	78.35	48.54	64.29	68.61	**46.69**	**60.45**	**58.99**	33.85	41.39	39.76
CAPF DR$_{DeepSeek-V3.2}$	55.49	67.58	74.46	41.97	61.25	66.79	44.48	55.38	53.31	29.92	36.23	32.68
CAPF DR$_{GPT-5.1}$	**61.89**	**75.82**	**79.87**	**48.91**	**67.52**	69.34	44.79	58.19	58.35	**35.04**	**43.34**	**44.49**

Table 1: Performance comparison across four fact-seeking benchmarks (RealtimeQA, HoHQA, DynaQuest, and SealQA). MBE represents the model-based evaluation score obtained via an LLM-as-a-Judge protocol. DR denotes Deep Research agent.

ule into the planning stage. By strengthening the research brief with a clearer and more informed contextual scope, CAPF guides downstream exploration more effectively and leads to a substantial improvement in answer accuracy. Notably, this improvement allows our approach to match or exceed the performance of several widely used proprietary deep research agents, including the official deep research functionalities in GPT-5.1, Qwen3-Max, and Kimi-K2. Additional case studies that reveal further insights are provided in Appendix C.

4.3 Ablation Study

We present an ablation study that quantifies the contribution of individual CAPF components to answer accuracy across multiple datasets. As shown in Table 2, the agent without CAPF exhibits the weakest performance, although it still marginally outperforms directly prompting an LLM to answer the question, highlighting the benefit of structured research workflows even without context-aware planning. Enabling either local or web pre-search leads to consistent performance gains, with web pre-search providing a larger improvement than local search. This difference can be attributed to the ability of web search to surface more recent and diverse evidence, which helps the agent identify temporal constraints and conflicting signals during planning. As a result, the research brief becomes more precise and better scoped, leading to improved downstream exploration and answer generation. Finally, the full CAPF framework achieves the best performance among all variants, showing absolute improvements of +15.36% in F1 and +17.52% in MBE over the agent without CAPF, demonstrating the complementary benefits of integrating context-aware pre-search signals into the planning process.

5 Analysis

In this section, we present a comprehensive analysis of how CAPF improves the efficiency of deep research by comparing agents with and without CAPF. Our analysis focuses on four aspects: resource utilization, message quality, evolution of the search tree, and synthesis coherence. A qualitative comparison between CAPF and proprietary deep research agents is provided in Appendix C.

5.1 Resource Utilization

We analyze the impact of CAPF on resource utilization from three perspectives: the number of search API calls (S-API), total execution time (E-Time), and the number of tool invocations (T-Call).

Method	RealtimeQA			HoHQA			Dynaquest			SealQA		
	EM	F1	MBE	EM	F1	MBE	EM	F1	MBE	EM	F1	MBE
LLM Direct$_{DeepSeek-V3}$	29.87	40.65	43.29	5.83	15.10	9.12	18.09	30.79	28.07	14.57	17.40	15.35
No CAPF DR$_{DeepSeek-V3}$	38.41	56.70	54.87	30.29	47.81	51.09	32.18	53.87	50.79	16.93	20.25	18.90
CAPF$_{Local}$ DR$_{DeepSeek-V3}$	41.16	59.76	60.67	32.11	53.65	54.74	35.96	54.73	51.10	20.87	24.41	27.16
CAPF$_{Web}$ DR$_{DeepSeek-V3}$	50.91	64.13	66.76	37.23	58.59	60.21	38.48	56.78	57.10	30.71	34.56	34.25
CAPF DR$_{DeepSeek-V3}$	59.75	73.95	78.35	48.54	64.29	68.61	46.69	60.45	58.99	33.85	41.39	39.76

Table 2: Ablation results of CAPF on four fact-seeking benchmarks. CAPF$_{Local}$ and CAPF$_{Web}$ enable only local and web search, respectively, while CAPF enables both. All experiments use DeepSeek-V3 as the base model.

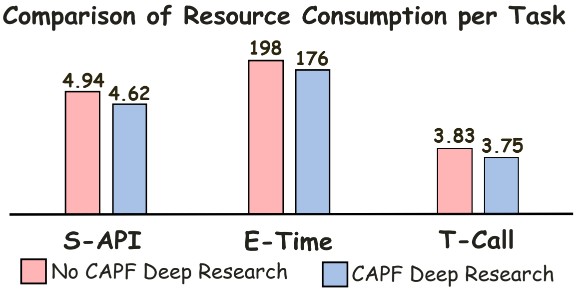

Figure 3: Comparison of resource consumption per task with and without CAPF. Best viewed by zooming in.

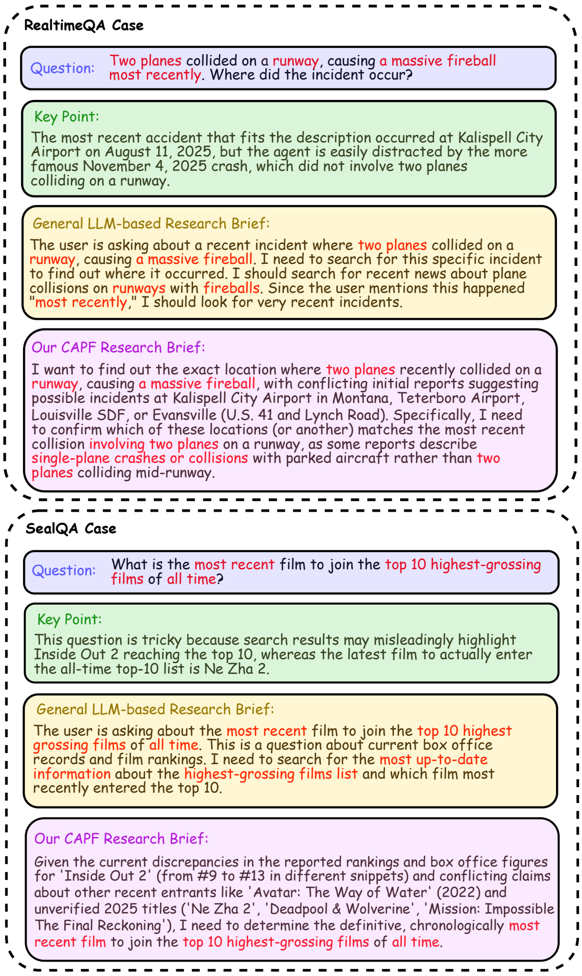

Figure 4: Case studies comparing general LLM-based research briefs and our CAPF-enhanced research briefs. Best viewed by zooming in.

Figure 3 compares a deep research agent with and without CAPF across these metrics. It shows that incorporating CAPF consistently reduces resource consumption along all three dimensions. In particular, the S-API per task decreases by 0.32, and the E-Time is reduced by 22 seconds. These improvements indicate that exposing the agent to a clear research scope during the planning stage enables more efficient exploration by avoiding unnecessary searches and redundant execution steps. A comparison of resource consumption across different backbone models is provided in Appendix D.

5.2 Message Quality

We investigate the quality of messages propagated within deep research agent systems, as these messages play a critical role in guiding the entire research process. Figure 4 presents two examples from RealtimeQA and SealQA, illustrating how CAPF enhances user queries by both constraining the problem scope to relevant candidates and explicitly surfacing critical pitfalls inherent in the questions. To enable systematic comparison, we manually select a set of challenging cases from the four benchmarks and annotate key points corresponding to the core aspects. We then conduct a human evaluation of the research briefs and subtopics generated with and without CAPF, measuring how well they align with these annotated key points. As shown in Figure 5, CAPF improves message quality by surfacing inherent difficulties through pre-search, leading to higher key-point coverage in research briefs and sub-topics. Additional qualitative examples are provided in Appendix E.

5.3 Evolution of the Search Tree

We explore how CAPF influences the evolution of the search tree during research exploration from both global and local perspectives. At the global level, we focus on search tree depth (STD), which

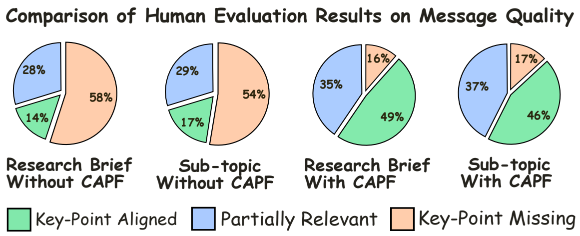

Figure 5: Comparison of human evaluation results on message quality with and without CAPF. Best viewed by zooming in.

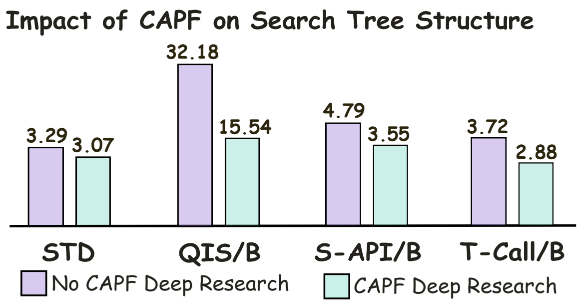

Figure 6: Comparison of search tree structure with and without CAPF. Best viewed by zooming in.

reflects how efficiently the search process converges. At the local level, we examine branch-level efficiency using three complementary metrics: query intent shift rate per branch (QIS/B), search API calls per branch (S-API/B), and tool invocations per branch (T-Call/B). These metrics characterize how efficiently each branch conducts focused exploration and reaches resolution. As shown in Figure 6, CAPF consistently optimizes the structure of the search tree, with particular improvements in branch-level efficiency metrics. These results suggest that by surfacing inherent difficulties through pre-search, CAPF provides more detailed guidance for each branch, enabling more focused and efficient exploration and verification. Additional qualitative examples are provided in Appendix F.

5.4 Synthesis Coherence

We observe that CAPF also improves the coherence of the final report by shaping how retrieved evidence is interpreted and integrated. Importantly, we find that retrieving critical evidence during research exploration does not guarantee that the agent will correctly utilize it in the final report. Figure 3 illustrates a representative case in which both systems correctly identify a player transfer that occurred in August 2007 based on the retrieved evidence. However, without a CAPF-enhanced research brief and sub-topics that highlight this key point, the report incorrectly infers affiliation with Bolton Wanderers, overlooking the previous transfer. As a result, the report misinterprets the temporal implications of the transfer event. In contrast, CAPF explicitly surfaces this temporal inconsistency during the planning stage. This guidance is reflected in the final report, which correctly accounts for the player's club affiliation both before and after the transfer, generating a more coherent synthesis.

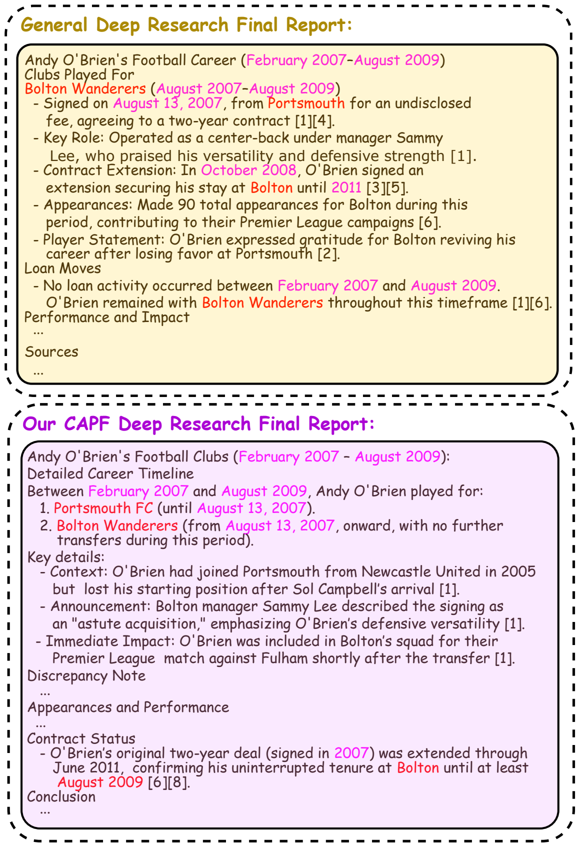

Figure 7: Case studies comparing the general deep research final report and our CAPF deep research final report. Best viewed by zooming in.

6 Conclusion

In this work, we identify a key limitation of current deep research systems, which relies on shallow research briefs from underspecified query expansion that fail to anticipate research scope. To address this issue, we introduce CAPF, a context-aware planning framework that incorporates contextual signals into the planning stage to establish clearer research boundaries and guide more focused exploration. Experiments on four recent realistic QA benchmarks show that CAPF consistently outperforms SOTA deep research and agentic retrieval systems. Further analysis indicates that exposing the agent to the inherent difficulty of a question during the planning stage is critical for improving both accuracy and efficiency in deep research agents.

Limitations

Although our approach achieves SOTA performance across a wide range of realistic fact-seeking benchmarks while also demonstrating improved resource efficiency, it has two notable limitations: (i) **Multi-entity Queries** and (ii) **Search Tree Optimization**.

First, our framework is primarily designed for questions whose answers are obscured by noisy or misleading search evidence. By introducing a context-aware planning stage, CAPF allows the agent to surface potential ambiguities and anticipate difficulties during planning. However, for multi-entity queries that mainly rely on user clarification and semantic decomposition, such as *"What are the investment philosophies of Duan Yongping, Warren Buffett, and Charlie Munger?"*, the core challenge lies less in handling noisy evidence and more in decomposing the query into coherent subproblems and conducting structured analysis for each component. In such scenarios, the pre-search signals provided by CAPF offer limited additional benefits. Table 3 presents a comparison of different deep research systems on DeepResearch Bench (Du et al., 2025). Compared with deep research agents without CAPF, our framework shows a clear improvement in overall performance. Although it still lags behind Cellcog, it nevertheless surpasses agents such as OpenAI Deep Research and Claude Research. We note that DeepResearch Bench does not provide a local text corpus. Our evaluation adopts a web-only variant of CAPF, denoted as CAPF_{Web}, which enables web pre-search during the CAPF stage.

Method	overall	comp.	insight	inst.	read.
Cellcog	51.94	52.17	51.9	51.37	51.94
CAPF_{Web} DR	49.59	49.24	47.06	51.96	52.14
OpenAI DR	46.45	46.46	43.73	49.39	47.22
Claude Research	45.43	45.34	42.79	47.58	44.66
No CAPF DR	44.94	44.75	42.83	46.75	47.86

Table 3: Comparison of different deep research systems on DeepResearch Bench.

Second, while CAPF effectively shapes the early-stage research plan and constrains the exploration space, it does not perfectly optimize the each step of the exploration trajectory. Through our case studies, we identify several directions that could further improve downstream research efficiency. The first issue arises in the transition from the research brief to sub-topic generation. This process is currently driven by a fixed system prompt in conjunction with an LLM. In some cases, the resulting decomposition is suboptimal. As illustrated in the second case study in Appendix F, the downstream sub-topic decomposition is constrained by the inherent capabilities of the LLM. From a human perspective, the research brief could be more cleanly decomposed into two independent sub-tasks, one focusing on the Prime Minister role and the other on the Vice President role. In contrast, the model-generated sub-topics remain partially entangled, which complicates subsequent exploration. A second issue concerns query generation conditioned on sub-topics. We observe that the LLM tends to produce queries that are highly similar in semantics. Figure 14 shows two representative examples: *"Mercedes Aráoz roles Second Vice President Prime Minister Peru 2017–2018"* and *"Mercedes Aráoz acting president Peru 2017–2018"*. Due to their close semantic overlap, these queries yield largely redundant web search results, leading to unnecessary resource consumption.

References

Xingyu Deng, Xi Wang, and Mark Stevenson. 2025. + verirel: Verification feedback to enhance document retrieval for scientific fact checking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 4706–4711.

Guanting Dong, Jiajie Jin, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2025. Rag-critic: Leveraging automated critic-guided agentic workflow for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3551–3578.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *arXiv preprint arXiv:2506.11763*.

FastGPT AI. 2024. Fastgpt: Open-source retrieval-augmented generation framework. https://doc.tryfastgpt.ai. Accessed: 2025-02-01.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Yuxuan Huang, Yihang Chen, Haozheng Zhang, Kang Li, Huichi Zhou, Meng Fang, Linyi Yang, Xiaoguang Li, Lifeng Shang, Songcen Xu, and 1 others. 2025.

Deep research agents: A systematic examination and roadmap. *arXiv preprint arXiv:2506.18096*.

Abhinav Java, Ashmit Khandelwal, Sukruta Midigeshi, Aaron Halfaker, Amit Deshpande, Navin Goyal, Ankur Gupta, Nagarajan Natarajan, and Amit Sharma. 2025. Characterizing deep research: A benchmark and formal definition. *arXiv preprint arXiv:2508.04183*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Jungo Kasai, Keisuke Sakaguchi, Ronan Le Bras, Akari Asai, Xinyan Yu, Dragomir Radev, Noah A Smith, Yejin Choi, Kentaro Inui, and 1 others. 2023. Real-time qa: What's the answer right now? *Advances in neural information processing systems*, 36:49025–49043.

Tian Lan, Bin Zhu, Qianghuai Jia, Junyang Ren, Haijun Li, Longyue Wang, Zhao Xu, Weihua Luo, and Kaifu Zhang. 2025. Deepwidesearch: Benchmarking depth and width in agentic information seeking. *arXiv preprint arXiv:2510.20168*.

LangChain AI. 2025. Langchain open_deep_research agent documentation. https://python.langchain.com/docs/agents/agent_types/deep_research. Accessed: 2025-02-01.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Yu Lei, Shuzheng Si, Wei Wang, Yifei Wu, Gang Chen, Fanchao Qi, and Maosong Sun. 2025. Rhinoinsight: Improving deep research through control mechanisms for model behavior and context. *arXiv preprint arXiv:2511.18743*.

Chunyang Li, Weiqi Wang, Tianshi Zheng, and Yangqiu Song. 2025a. Patterns over principles: The fragility of inductive reasoning in llms under noisy observations. *arXiv preprint arXiv:2502.16169*.

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

Zhiyuan Li, Haisheng Yu, Guangchuan Guo, Nan Zhou, and Jiajun Zhang. 2025c. Muisqa: Multi-intent retrieval-augmented generation for scientific question answering. *arXiv preprint arXiv:2511.16283*.

Qian Lin, Junyi Li, and Hwee Tou Ng. 2025a. Dynaquest: A dynamic question answering dataset reflecting real-world knowledge updates. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 26918–26936, Vienna, Austria. Association for Computational Linguistics.

Xixun Lin, Yucheng Ning, Jingwen Zhang, Yan Dong, Yilong Liu, Yongxuan Wu, Xiaohua Qi, Nan Sun, Yanmin Shang, Pengfei Cao, and 1 others. 2025b. Llm-based agents suffer from hallucinations: A survey of taxonomy, methods, and directions. *arXiv preprint arXiv:2509.18970*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

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

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Generation-augmented retrieval for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4089–4100, Online. Association for Computational Linguistics.

Moonshot AI. 2025. Kimi k2 model card. https://www.moonshot.cn/kimi. Accessed: 2025-02-01.

OpenAI. 2025. Gpt-5.1. https://platform.openai.com/docs/models. Accessed: 2025-02-01.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. 2025. Hoh: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6036–6063, Vienna, Austria. Association for Computational Linguistics.

Thinh Pham, Nguyen Nguyen, Pratibha Zunjare, Weiyuan Chen, Yu-Min Tseng, and Tu Vu. 2025. Sealqa: Raising the bar for reasoning in search-augmented language models. *arXiv preprint arXiv:2506.01062*.

Qwen AI. 2025. Qwen3-max technical report. https://qwenlm.github.io/Qwen3. Accessed: 2025-02-01.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Zhengliang Shi, Yiqun Chen, Haitao Li, Weiwei Sun, Shiyu Ni, Yougang Lyu, Run-Ze Fan, Bowen Jin, Yixuan Weng, Minjun Zhu, Qiujie Xie, Xinyu Guo, Qu Yang, Jiayi Wu, Jujia Zhao, Xiaqiang Tang, Xinbei Ma, Cunxiang Wang, Jiaxin Mao, and 7 others. 2025. Deep research: A systematic survey. *arXiv preprint arXiv:2512.02038*.

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

Shuang Sun, Huatong Song, Yuhao Wang, Ruiyang Ren, Jinhao Jiang, Junjie Zhang, Fei Bai, Jia Deng, Wayne Xin Zhao, Zheng Liu, and 1 others. 2025. Simpledeepsearcher: Deep information seeking via web-powered reasoning trajectory synthesis. *arXiv preprint arXiv:2505.16834*.

Tavily AI. 2024. Tavily web search api documentation. https://docs.tavily.com. Accessed: 2025-02-01.

Haiyuan Wan, Chen Yang, Junchi Yu, Meiqi Tu, Jiaxuan Lu, Di Yu, Jianbao Cao, Ben Gao, Jiaqing Xie, Aoran Wang, and 1 others. 2025. Deepresearch arena: The first exam of llms' research abilities via seminar-grounded tasks. *arXiv preprint arXiv:2509.01396*.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025. Retrieval-augmented generation with conflicting evidence. *arXiv preprint arXiv:2504.13079*.

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. 2025. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of rag noise in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5019–5039.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an llm's internal prior and external evidence. *Advances in neural information processing systems*, 37:33402–33422.

Renjun Xu and Jingwen Peng. 2025. A comprehensive survey of deep research: Systems, methodologies, and applications. *arXiv preprint arXiv:2506.12594*.

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

Qianchi Zhang, Hainan Zhang, Liang Pang, Ziwei Wang, Hongwei Zheng, Yongxin Tong, and Zhiming Zheng. 2025a. Finefilter: A fine-grained noise filtering mechanism for retrieval-augmented large language models. *arXiv preprint arXiv:2502.11811*.

Yuxin Zhang, Yan Wang, Yongrui Chen, Shenyu Zhang, Xinbang Dai, Sheng Bi, and Guilin Qi. 2025b. Magic mushroom: A customizable benchmark for fine-grained analysis of retrieval noise erosion in rag systems. *arXiv preprint arXiv:2506.03901*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Liwen Zheng, Chaozhuo Li, Xi Zhang, Yu-Ming Shang, Feiran Huang, and Haoran Jia. 2024. Evidence retrieval is almost all you need for fact verification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9274–9281.

Yuxiang Zheng, Dayuan Fu, Xiangkun Hu, Xiaojie Cai, Lyumanshan Ye, Pengrui Lu, and Pengfei Liu. 2025. Deepresearcher: Scaling deep research via reinforcement learning in real-world environments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 414–431, Suzhou, China. Association for Computational Linguistics.

Appendix

A Evaluation Datatsets

To assess the effectiveness of our framework, we evaluate it on a set of benchmarks designed to challenge systems with noisy and potentially misleading evidence that cannot be effectively resolved through user clarification. Specifically, we adopt RealtimeQA (Kasai et al., 2023), HoHQA (Ouyang et al., 2025), DynaQuest (Lin et al., 2025a), and SealQA (Pham et al., 2025), all of which include questions whose correct answers are easily confounded by outdated, conflicting, or ambiguous information.

However, we observe that these benchmarks are highly time-sensitive. Many questions were correct at the time of dataset construction, become partially or fully incorrect after only a few months due to real-world changes. For example, SealQA originally annotated the answer to *"Which company most recently surpassed a $1 trillion market capitalization for the first time in its history?"* as Broadcom, based on its valuation in December 2024. Subsequent market developments render this annotation outdated, as Eli Lilly surpassed the $1 trillion threshold on November 21, 2025.

In addition to temporal drift, some questions exhibit intrinsic ambiguity that can lead to systematic misinterpretation by agents. For instance, RealtimeQA includes the question *"What is the name of the storm that pummelled Britain this week?"*. The phrase *"this week"* is underspecified and depends on the evaluation time, which introduces uncertainty during question analysis. To ensure a reliable evaluation, we replace such vague temporal expressions with explicit references, such as *"January 2025"*, and update the corresponding answer to *"Storm Éowyn"*.

To quantify the extent of these issues, we construct a statistical analysis that compares the revision rates across the four benchmarks. Figure 8 reports the proportion of questions and answers that require updates to ensure temporal correctness and clarity. We observe that the majority of revisions are concentrated in RealtimeQA, as the dataset was created in 2023 and contains a substantial number of entries whose underlying facts or annotations have since become outdated. The remaining three benchmarks also exhibit non-negligible revision needs, with approximately 15% of their questions or answers requiring modification to maintain evaluation validity.

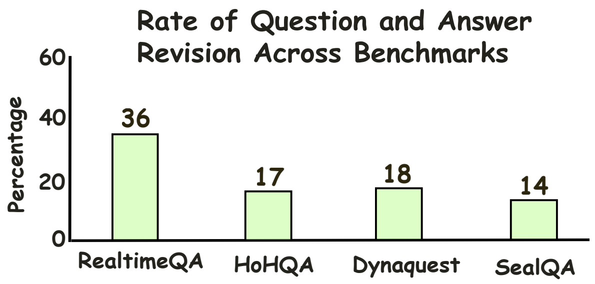

Figure 8: Proportion of question and answer requiring updates across four benchmarks.

B Experimental Setup

Baseline and Hyperparameters. We implement our framework with LangGraph and incorporate CAPF into the planning node of the system. CAPF incorporates two complementary retrieval modules: a web-search component powered by the Tavily search API (Tavily AI, 2024), and a local retrieval component implemented as a customized RAG system (Karpukhin et al., 2020) built from the noisy context passages supplied by each dataset. We evaluate our approach on four widely deployed LLMs, including Qwen3-235B (Yang et al., 2025), DeepSeek-V3 (Liu et al., 2024), DeepSeek-V3.2 (Liu et al., 2025), and GPT-5.1 (OpenAI, 2025), which serve as the research models for our agent. For hyperparameters, we set "max_concurrent_research_units" to 5, "max_researcher_iterations" to 6, "research_model_max_tokens" to 10000, and "max_react_tool_calls" to 10.

System Prompts. This section presents the system prompts used in our framework. The prompts are organized according to the three main stages of the deep research pipeline: research planning, research exploration, and report synthesis. While most prompts are specifically designed for CAPF, a subset is adapted from the open deep research agent proposed by LangChain-AI (LangChain AI, 2025). Figure 9 illustrates the key prompts associated with CAPF. In particular, "CAPF_snippet_analysis_prompt" focuses on the processing of retrieved text snippets. It is designed to extract only the core factual information from each snippet that is directly relevant to the user's question, while explicitly prohibiting interpretation, inference, or correctness judgment. By enforcing strict constraints on the extraction behavior, this prompt ensures that noisy, incomplete, or potentially misleading evidence is preserved in its original form, rather than prematurely filtered or corrected. In contrast, "Research_brief_prompt" focuses on transforming the extracted snippet-level cues into a concrete and actionable research objective. It takes as input the user's original query, the conversation context, and the weak signals surfaced during the CAPF snippet analysis stage, and synthesizes them into a single, well-scoped research question. Rather than treating snippet-derived information as confirmed facts, this prompt explicitly frames such information as leads to be verified, and requires potential conflicts, ambiguities, or missing attributes to be articulated in the research brief itself. As a result, the generated research question not only specifies what information needs to be found, but also highlights which aspects require careful validation or disambiguation, thereby providing clearer guidance for subsequent exploration.

C Main Experiment

In this section, we provide a qualitative analysis of representative proprietary deep research systems, including the official deep research functionalities in GPT-5.1 (OpenAI, 2025), Qwen3-Max (Qwen AI, 2025), and Kimi-K2 (Moonshot AI, 2025). Since these systems are closed-source, our analysis goes beyond the performance differences reported in Table 1 and focuses on practical usage experience, illustrative case studies, and qualitative comparisons with our CAPF framework.

For GPT-5.1 (OpenAI, 2025) and Kimi-K2 (Moonshot AI, 2025), the deep research agents operate in a largely closed manner. These systems do not expose explicit research briefs or intermediate sub-topics to the user. Instead, they provide partial intermediate reflections with limited transparency into how subsequent actions are selected during execution. Figure 10 presents an example from GPT-5.1 for the query *"What was the position of Trygve Lie between Jul 1947 and Apr 1948?"*. The generated output offers an extensive narrative that aggregates a wide range of contextual details related to Trygve Lie during the target period. While such thoroughness can be informative, it is often unnecessarily verbose for fact-seeking queries, where a precise, well-scoped research objective and targeted verification would suffice. We also observe that both systems exhibit relatively long response times, with a single request frequently exceeding ten minutes. In addition, Kimi-K2 tends to produce excessively long intermediate outputs that include substantial information not directly relevant to the original question. For relatively simple fact-seeking tasks, this lack of adaptive control can lead to redundant exploration and reduced overall efficiency.

For Qwen3-Max (Qwen AI, 2025), the research team has not yet released a dedicated deep research mode. However, by enabling both the thinking and search options, the system can approximate a deep research style workflow. Under this configuration, Qwen3-Max is able to decompose the original question into multiple sub-questions and conduct independent exploration for each sub-task. We consider this setup to be a reasonable proxy for comparison. Unlike GPT-5.1 and Kimi-K2, Qwen3-Max exposes part of its research process, particularly the generated research brief. Figure 11 presents a comparison between the research briefs produced by Qwen3-Max and those generated by CAPF. While Qwen3-Max relies primarily on direct question decomposition by the language model, it lacks an explicit mechanism for anticipating potential difficulties or ambiguities in advance. In contrast, CAPF incorporates weak signals surfaced during pre-search to proactively identify challenging aspects of the question, resulting in a more informed and focused research plan.

Backbone Model	E-Time	S-API	T-Call
DeepSeek-V3	176	4.62	3.75
DeepSeek-V3.2	397	7.16	9.28
Qwen3-235B	274	6.54	7.73
GPT-5.1	246	11.25	8.48

Table 4: Comparison of resource utilization of different backbone models with CAPF.

D Resource Utilization

The main body of the paper analyzes the impact of CAPF on resource utilization by comparing agents with and without CAPF. In this section, we focus on a complementary analysis that examines how different backbone models affect resource consumption under the same CAPF framework. We adopt the same evaluation metrics as in the main experiments, including the number of search API calls (S-API), total execution time (E-Time), and the number of tool invocations (T-Call).

Table 4 reports the resource utilization of four widely deployed large language models, namely Qwen3-235B (Yang et al., 2025), DeepSeek-V3 (Liu et al., 2024), DeepSeek-V3.2 (Liu et al., 2025), and GPT-5.1 (OpenAI, 2025), when all are equipped with CAPF under identical experimental settings. The results indicate that resource consumption does not exhibit a strict positive correlation with answer accuracy. For example, DeepSeek-V3 consumes the least resources, using nearly half the resources required by DeepSeek-V3.2, while achieving higher answer accuracy. In contrast, GPT-5.1 incurs the highest S-API cost but also attains the best overall performance. Qwen3-235B and DeepSeek-V3.2 fall between these extremes, reflecting different trade-offs between efficiency and effectiveness under the same planning framework.

E Message Quality

In this section, we provide a detailed analysis of how CAPF improves message quality in deep research systems by strengthening the formulation of research briefs and sub-topics. Specifically, CAPF enhances user queries along two complementary dimensions. First, it constrains the problem scope by explicitly enumerating plausible candidates, temporal ranges, or organizational entities implicated by the query, thereby reducing ambiguity and preventing unfocused exploration. Second, it surfaces critical pitfalls inherent in the question, such as competing interpretations, entity confusion, or potentially misleading assumptions suggested by early evidence, and incorporates these issues directly into the planning stage.

Figure 12 presents two representative cases that illustrate these effects in practice. In the first case, CAPF transforms an underspecified historical query into a research brief that explicitly distinguishes between alternative roles and geographic contexts, ensuring that each plausible interpretation is independently verified. In the second case, CAPF identifies multiple potential ownership structures implied by the query and organizes them into structured sub-topics, allowing the agent to systematically investigate competing hypotheses rather than prematurely committing to a single narrative. Across both examples, the resulting research briefs are more precise, better aligned with the true difficulty of the questions, and provide clearer guidance for downstream exploration.

F Evolution of Search Tree

In this section, we analyze how CAPF influences the evolution of the search tree during deep research. Our analysis focuses on how planning signals introduced by CAPF shape the branching structure, depth progression, and query refinement behavior throughout the research process. Figures 13, 14, 15, 16 present four representative case studies that illustrate the resulting search tree structures under CAPF.

In the first case, CAPF enables the agent to separate alternative career paths and timeline intersections at an early stage. This results in shallow yet well-targeted branches that converge quickly toward resolution. By explicitly structuring competing explanations into parallel branches, the agent is able to systematically compare alternatives instead of prematurely committing to a single narrative, thereby reducing unnecessary exploration.

The second case further illustrates both the strengths and current limitations of CAPF. During the pre-search stage, CAPF successfully identifies a key difficulty of the question that Mercedes Aráoz may have simultaneously held the positions of Prime Minister and Vice President during the period from 2017 to 2018. Based on this observation, CAPF generates a research brief with a clear scope and explicit verification goals. However, we also observe that the downstream decomposition into sub-topics is constrained by the capabilities of the LLM. From a human perspective, the research brief could be more effectively decomposed into two independent sub-tasks, one focusing on the Prime Minister role and the other on the Vice President role. In contrast, the model-generated sub-topics remain partially entangled, which leads to an additional round of search that may not be strictly necessary.

Overall, these case studies demonstrate that exposing weak signals and potential pitfalls during planning not only improves message quality but also fundamentally alters the dynamics of exploration by shaping a more structured and interpretable search tree. At the same time, they highlight a promising direction for future research, namely how to better align LLMs with explicit investigative structures in order to further reduce redundant exploration and improve resource efficiency.

CAPF_snippet_analysis_prompt

You will be given the user's question and a retrieved snippet of text.
Your goal is only to extract the core information from the snippet that directly relates to the question.
CRITICAL: Always respond in {language}.
The user's question is:
<Question>
{user_question}
</Question>
The retrieved snippet is:
<SNIPPET>
{snippet}
</SNIPPET>
Do NOT interpret, infer, expand, reason, or judge correctness.
Do NOT mention contradictions, uncertainties, or irrelevant content.
If the snippet contains NO information that meaningfully helps answer the question, output exactly: "No relevant information"
Otherwise, output only the directly relevant information, quoted or summarized very briefly.

Research_brief_prompt

You will be given a set of messages that have been exchanged so far between yourself and the user, along with additional clues from initial research snippets related to the user's query.
Your job is to translate this information into a more detailed and concrete research question that will guide the next stage of research.
CRITICAL: Always respond in {language}. All your responses and the final research question must be in {language}.
The messages exchanged so far are:
<Messages>
{messages}
</Messages>
The following are clues from initial snippet analysis related to the question:
<SnippetsAnalysis>
{snippets_analysis}
</SnippetsAnalysis>
Today's date is {date}.
You will return a single, well-formulated research question that will be used to guide further research.
Guidelines:
1. **Maximize Specificity and Detail**: Incorporate all relevant details from the user's messages *and* the snippet clues. Include known user preferences and key attributes mentioned by the user. Ensure the research question is specific about what needs to be found, investigated, or confirmed.
2. **Include All Relevant Dimensions**: If the user has provided certain criteria or context, make sure these are included. If there are essential attributes missing from the user's query, explicitly state them as open-ended or to be determined (e.g., "for all regions" if no region is specified).
3. **Avoid Unwarranted Assumptions**: Do not invent or assume details that the user hasn't provided. If something is not specified, indicate that it's not specified and may require exploration. Do not treat the snippet clues as confirmed facts—treat them as leads or evidence to follow up on.
4. **Use the User's Perspective (First Person)**: Phrase the research question from the user's perspective, as if the user is stating what they want. For example, "I want to find out ..." or "What is the ...".
5. **Source Prioritization**: If the user has indicated preferences for certain sources or if the query type suggests an ideal source, mention it. (For instance, for a product query, suggesting looking at official sites or reputable reviews; for academic queries, focusing on original research papers; for person queries, official profiles like LinkedIn, etc. Ensure sources align with the user's language if specified.)
6. **Leverage Snippet Clues**: Use the information from the snippet analysis to enrich the research question.
 - If the clues suggest possible answers or important details (e.g., specific names, dates, or facts), include these details in the question as context or points to verify.
 - If the clues reveal conflicting information or uncertainties, explicitly mention these issues in the research question so that the research can address them (for example, "some sources indicate X while others suggest Y").
 - Treat snippet-derived information as leads that need confirmation; the research question should reflect the need to investigate those leads rather than taking them at face value.
7. **Clarity and Focus**: The final research question should be clear and focused, leaving no ambiguity about what needs to be researched. It should be detailed enough to guide a researcher or sub-agents in finding the necessary information, including resolving any conflicts or gaps identified.
Your output should be a single, standalone research question (in one sentence or a short paragraph) that integrates the user's request with the snippet-derived clues, ready to be passed to the research phase.

Figure 9: The CAPF system prompts.

What was the position of Trygve Lie between Jul 1947 and Apr 1948?

Step1: Research Planning
Research Brief and Sutopics

Step2: Detailed Exploration

2-1: I'm verifying Trygve Lie's official UN role from July 1947 to April 1948. He should be the Secretary-General during this period, based on the timeline.

Searched for Trygve Lie Secretary-General details

2-2: Considering Trygve Lie's tenure as UN Secretary-General from July 1947 to April 1948, noting his exit from Norwegian position and involvement in key events like the Palestine partition plan and Cold War initiatives.

W Read en.wikipedia.org

...

2-N: I'm thinking about how the Secretary-General of the UN handles political initiatives and secures roles, typically detailed in the UN Charter and official pages.

Searched for chief administrative officer United Nations

Step3: Final Report

Trygve Lie's Official Position (July 1947 – April 1948)

Position and Title (1947–1948)
Between July 1947 and April 1948, Trygve Lie was serving as the Secretary-General of the United Nations, the highest administrative official of the UN. In fact, he was the first person to hold this post,
...

Role and Responsibilities
As UN Secretary-General, Trygve Lie's role was essentially to act as the organization's top administrator and an international diplomat. The United Nations Charter defines the Secretary-General
...

Notable Actions and Events
During the period in question, Secretary-General Lie was involved in several significant international issues, given that the late 1940s were a very turbulent time. Some of the notable actions and events he dealt with include:
...

Figure 10: The detailed response of GPT5.1 deep research agent on a fact-seeking question.

Qwen3-Max Research Brief:

This question is asking about the position held by Trygve Lie between July 1947 and April 1948. I need to search for information about Trygve Lie and his positions during this specific time period.

Let me break this down:

1. I need to find out who Trygve Lie was

2. I need to determine what position he held specifically between July 1947 and April 1948

This seems like a straightforward historical fact question. I should search for information about Trygve Lie and his roles during this time period.

Let me start with a search about Trygve Lie and his position in 1947-194

CAPF Research Brief:

What was Trygve Lie's official position between July 1947 and April 1948, given that he served as Norway's Minister of Justice from February 1946 and later became Secretary General of United Nations? Clarify whether he held both roles simultaneously between July 1947 and April 1948 or transitioned between them, and provide authoritative sources confirming his exact position in this timeframe

Figure 11: The comparison of Qwen3-Max research brief and our CAPF research brief.

Question: What position did Joseph Martin (Canadian politician) take in Jun 1910?

Research Brief:
I want to find out what political position Joseph Martin (Canadian politician) held in June 1910, as initial snippet analysis suggests he may have been serving as a British MP for St Pancras East from 1910-1918 or possibly as Attorney General and Acting Provincial Secretary in Canada at that time. The research should clarify whether he was in the UK or Canada during this period and confirm his specific role(s).

Sub-topics:
(1) Investigate whether Joseph Martin (Canadian politician) was serving as a British MP for St Pancras East in June 1910, including confirming the date range of his tenure and whether he was actively in the UK during that period.

(2) Investigate whether Joseph Martin (Canadian politician) was serving as Attorney General and Acting Provincial Secretary in Canada in June 1910, including confirming the timeline of his service in these roles and his location during that period.

Question: Which organization owned Ampad between Dec 2006 and Aug 2009?

Research Brief:
Find out which organization owned Ampad between December 2006 and August 2009, given that Crescent Capital Investments (later renamed Esselte) acquired the company in 2003, and verify whether ownership changed during this period or remained under Esselte/Crescent. Also investigate any potential connection to American Tissue or Bain Capital (given their historical ties to Ampad) during this timeframe.

Sub-topics:
(1) Investigate the ownership of Ampad between December 2006 and August 2009, specifically whether Esselte (formerly Crescent Capital Investments) retained ownership during this period. Include any known changes in ownership or corporate restructuring during this timeframe.

(2) Explore any historical or business connections between Ampad and American Tissue during the period from December 2006 to August 2009, including any shared ownership, partnerships, or financial ties.

(3) Investigate any potential ties between Ampad and Bain Capital during the period from December 2006 to August 2009, including ownership stakes, financial involvement, or corporate affiliations.

Figure 12: Two case studies illustrating how CAPF enhances user queries into structured research briefs and sub-topics.

```
"Query": "What was Francis H. Case 's occupation before Oct 1920?",
"CAPF_info": {
    "rag_info": [
        "Francis H . Case :  Francis Higbee Case ( December 9 , 1896June 22 , 1962 ) was  an American
        journalist and politician who served for 25 years as a member of United States Congress...",
        "Biography :  Case was born in Everly , Iowa , the son of Mary Ellen ( née Grannis ) and the
         Reverend Herbert Llywellen Case . He moved with his parents to Sturgis...",
        ],
     "web_info": [
         "Francis H. Case - South Dakota Historical Society Press: Dakota Images Francis Highbee
          Case was born in Everly, Iowa, on 9 December 1896. When he was thirteen, his family...",
         "Case Collection - Archives - McGovern Library at Dakota Wesleyan ...: Francis H. Case was
          born in Everly, Iowa, on December 9, 1896, and moved with his parents to Sturgis, South...",
         "CASE, Francis Higbee (1896-1962) - Voteview: # UCLA Presents voteview.com beta\n\n##
          CASE, Francis Higbee (1896-1962)\n\n#### Republican of South Dakota\n\n#### 0th...",
         "..."
        ]
       "snippet_analysis": [
          "[Snippet 1 Analysis]  He received his B.A. from Dakota Wesleyan University in 1918 and his
                                 M.A.  from Northwestern University in 1920.",
         "[Snippet 2 Analysis]  An assistant editor at the Epworth Herald in Chicago and later the
                                telegraph writer and editorial writer for the Daily Journal.",
         "[Snippet 3 Analysis]  ...",
         "..."
         ]},
"Research_brief":  "I want to find out Francis H. Case's occupation before October 1920, considering
that he graduated with an M.A. from Northwestern University in 1920, served as a private in the U.S.
Marine Corps during World War I, and may have worked as a journalist (specifically as an assistant
editor at the Epworth Herald in Chicago or in other editorial roles). Did he hold any specific jobs or
titles prior to October 1920, and how did his military service and education timeline intersect with
his  early career?",
"Research_process": {
"Identify Francis H. Case's roles in journalism before October 1920, specifically whether he worked
as an assistant editor at the Epworth Herald in Chicago or held other editorial positions. Gather
details about these roles, including job titles, responsibilities, and dates of employment.": {
    "iterations": [
       {
        "iteration": 1,
        "queries": [
           "Francis H. Case journalism roles before October 1920",
           "Francis H. Case assistant editor Epworth Herald Chicago",
           "Francis H. Case editorial positions pre-1920" ],
        "results": [
           "Search results: SOURCE 1: Francis Case - Wikipedia URL: https://en.wikipedia.org/wiki/
            Francis_Case\n\nSUMMARY:\n<summary>\nFrancis Higbee Case (December 9, 1896
            – June 22, 1962) was an America..."] } ] },
"Examine the timeline of Francis H. Case's military service during World War I and his  education
 (M.A. from Northwestern University in 1920) to determine how these intersected with his early
career.": {
    "iterations": [
       {
        "iteration": 1,
        "queries": [
             "Francis H. Case military service World War I timeline",
             "Francis H. Case education Northwestern University 1920",
             "Francis H. Case early career before October 1920" ],
        "results": [
           "Search results: SOURCE 1: Francis Higbee Case (1896 - 1962) - Genealogy URL: https://
            www.geni.com/people/Francis-H-Case-U-S-Senator/6000000016303748055\n\nSUMMARY:
            Meanwhile, during World Wa..." ] } ] } },
"compressed_research": [
"List of Queries and Tool Calls Made  1. Search for: Francis H. Case journalism roles before...",
"List of Queries and Tool Calls Made 1. Initial search: Francis H. Case roles before October..."
"..."]
"findings": "Reflection recorded: To research Francis H. Case's occupation before October 1920...",
"final_report": "Francis H. Case’s Occupation Before October 1920:  A Comprehensive Analysis..."
```


Figure 13: Case 1 illustrates the structure of the search tree with CAPF.

```
"Query": "What was the position of Mercedes Aráoz between Dec 2017 and Mar 2018?",
"CAPF_info": {
    "rag_info": [
        "Vice presidency :  In August 2016 , Aráoz was appointed as chair of the APEC Peru 2016 High
         Level Commission . The commission was made to determine necessary guidelines...",
        "Mercedes Aráoz :  At the 2016 general election , Aráoz ran for the Second Vice Presidency
         along Pedro Pablo Kuczynski and Martin Vizcarra in the Peruvians for Change ticket...",
        "Academic career :  She is Principal Professor of International Economy at the University
         of the Pacific and member of the Research Center . She is also professor at the...",
        ],
    "web_info": [
        "Mercedes Aráoz: Mercedes Rosalba Aráoz Fernández (born 5 August 1961) is a Peruvian
         economist, professor and politician.\n\nShe was the second Vice President of Peru of...",
        "Mercedes Aráoz: Council of Women World Leaders\nCouncil of Women World Leaders
         Prime Minister of Peru (2017-2018)\n\nMercedes Aráoz became Peru's Prime...",
        "Mercedes Aráoz - Crop Trust:  Vice-Chair of the Executive Board Mercedes Araoz headshot
         H.E. Mercedes Aráoz is a Peruvian economist and politician. Currently, she is a...",
        "...",
        ]
    "snippet_analysis": [
        "[Snippet 1 Analysis]  In the Kuczynski administration, she was appointed Prime Minister,
                               serving from 17 September 2017 to 2 April 2018.",
        "[Snippet 2 Analysis]  Mercedes Aráóz was the Second Vice President of Peru from 28 July
                               2016 to 7 May 2020 and Prime Minister from 17 September 2017 to
                               2 April.",
        "[Snippet 3 Analysis]  Professor Aráoz served as the Acting President and Prime Minister of
                               Peru from 2017 to 2018.",
        "..."
        ]},
"Research_brief":  "I want to confirm the exact political position(s) held by Mercedes Aráoz between
December  2017 and March 2018, particularly whether she served as Prime Minister of Peru during
this entire period (17 September 2017 to 2 April 2018, per multiple snippets) and whether she
simultaneously held other roles like Second Vice President (mentioned July 2016-May 2020) or
'acting president' (as one snippet suggests). Sources should prioritize official government records,
accredited news reports from that timeframe, and her parliamentary/LinkedIn profiles to resolve",
potential inconsistencies about additional titles during this tenure.",
"Research_process": {
"Please research the exact political positions held by Mercedes Aráoz between December 2017
and March 2018, focusing on whether she served as Prime Minister of Peru during the entire period
from 17 September 2017 to 2 April 2018. Additionally, verify if she simultaneously held other roles
such as Second Vice President (as mentioned from July 2016 to May 2020) or 'acting president'
during this tenure. Sources should prioritize official government records, accredited news reports
from that timeframe, and her parliamentary/LinkedIn profiles to clarify any inconsistencies about
 additional titles during this period.": {
    "iterations": [
      {
        "iteration": 1,
        "queries": [
            "Mercedes Aráoz political positions September 2017 to April 2018",
            "Mercedes Aráoz roles Second Vice President Prime Minister Peru 2017-2018",
            "official government records Mercedes Aráoz Prime Minister Peru 2017-2018",
            "Mercedes Aráoz acting president Peru 2017-2018" ],
        "results": [
            "Search results: SOURCE 1: Mercedes Aráoz - Simple English Wikipedia, the free
            encyclopedia URL: https://simple.wikipedia.org/wiki/MercedesSUMMARY:
             Mercedes Rosalba A..."] } ,
      {
        "iteration": 2,
        "queries": [
            "Mercedes Aráoz political positions December 2017 to March 2018",
             "Peru government records Mercedes Aráoz Prime Minister 2017-2018",,
            "Mercedes Aráoz parliamentary speeches December 2017" ],
        "results": [
            "Search results:  SOURCE 1: Mercedes Aráoz - Wikipedia URL: https://en.wikipedia.org/wiki
             /Mercedes_Ar%C3%A1oz\n\nSUMMARY: <summary>Mercedes Rosalba Aráoz Fernández is
             a Peruvian economist, pr..." ] } ] } },
"compressed_research": [
"List of Queries and Tool Calls Made:**\n1. Initial search for Mercedes Aráoz's political positions ...",
"List of Queries and Tool Calls Made:**\n1. Initial search for Mercedes Aráoz's political positions ...",
"..."]
"findings": "Reflection recorded: The user is asking about Mercedes Aráoz's political positions...",
"final_report": "Political Positions of Mercedes Aráoz (December 2017 – March 2018)  Key Roles..."
```


Figure 14: Case 2 illustrates the structure of the search tree with CAPF.

```
"Query": "Who was the head coach of the team Gold Coast Suns between Sep 2016 and Aug 2017?",
"CAPF_info": {
    "rag_info": [
        "Players acquired through concessions :  Initial 2011 playing squad recruitment 2011–2014 :
         McKenna era . Guy McKenna would continue to coach the team throughout the 2011...",
        "Inaugural TAC Cup team :  In November 2009 the team signed twelve 17-year-olds around
         the country to compete in the Victorian Football League ( VFL ) year through...",
        "Brisbane :  The Gold Coast Suns have a rivalry with fellow Queensland AFL team the
         Brisbane Lions . The two teams contest the QClash twice each season . The first QClash ...",
        "...",
        ],
    "web_info": [
        "Where it all went wrong for Rodney Eade at the Gold Coast Suns: Expectations were raised
         when Rodney Eade, who had a track record of turning the Sydney...",
        "Rodney Eade sacked by Gold Coast Suns as AFL coach - ABC News: Eade had endured a
         testing reign as Suns coach, with the club's inconsistent on-field form mixed in with a ...",
        "Damien Hardwick joins the Gold Coast SUNS: “We have been unequivocal in our desire to
         deliver success for our football club, for our loyal and growing group of members and ...",
        "...",
        ]
      "snippet_analysis": [
        "[Snippet 1 Analysis]  Guy McKenna was the inaugural coach of the Gold Coast Suns until
                               he was sacked at the end of the 2014 season. Rodney Eade was the
                               head coach from 2015 and was sacked during the 2017 season.",
        "[Snippet 2 Analysis]  No relevant information.",
        "[Snippet 3 Analysis]  No relevant information",
        "[Snippet 4 Analysis]  Rodney Eade was the head coach of the Gold Coast Suns during the
                               specified period.",
        "..."
        ]},
"Research_brief":  "Confirm who served as the head coach of the Gold Coast Suns between September
2016 and August 2017, specifically verifying if Rodney Eade held this position until his reported sacking
in August 2017, and identifying any interim or permanent replacements during this timeframe.
Prioritize sourcing this information from official club announcements, reputable sports news outlets, or
the team's historical coaching records to resolve any potential conflicts or gaps in the available
information.",
"Research_process": {
"Investigate Rodney Eade's tenure as head coach of the Gold Coast Suns between September 2016
and August 2017. Verify if he held this position until his reported sacking in August 2017 and identify
any interim or permanent replacements during this timeframe. Prioritize sourcing from official club
announcements, reputable sports news outlets, or the team's historical coaching records to ensure
accuracy.": {
    "iterations": [
      {
        "iteration": 1,
        "queries": [
           "Rodney Eade Gold Coast Suns head coach tenure September 2016 August 2017",
           "Gold Coast Suns coach replacement August 2017",
           "official Gold Coast Suns coaching records 2016-2017" ],
        "results": [
           "Search results: SOURCE 1: Rodney Eade - Wikipedia URL: https://en.wikipedia.org/wiki/
            Rodney_Eade\n\nSUMMARY:\n<summary>\nRodney Eade, born April 4, 1958 is a former
            Australian rules footballe..."] } ] } },
"compressed_research": [
"List of Queries and Tool Calls Made  1. Search for: Rodney Eade Gold Coast Suns head coach tenure...",
"List of Queries and Tool Calls Made  1. Search for: Rodney Eade Gold Coast Suns head coach tenure...",
"..."]
"findings": "Reflection recorded: To accurately answer this question, we need to focus on verifying...",
"final_report": "Head Coach of the Gold Coast Suns (September 2016 - August 2017): Confirmation..."
```


Figure 15: Case 3 illustrates the structure of the search tree with CAPF.

```
"Query": "Who was the spouse of Robert Hossein before Aug 1957?",
"CAPF_info": {
    "rag_info": [
        "Foreign honours :  - : Commander of the Order of Cultural Merit ( 2006 ) Member of
         Eurasian Academy ( 2016 ) .",
        "Robert Hossein :  Robert Hossein ( 30 December 1927 – 31 December 2020 ) was a French
        film actor , director , and writer . He directed the 1982 adaptation of Les Misérables...",
        "Death :  Hossein died of COVID-19 during the COVID-19 pandemic in France on 31 December
        2020, one day after his 93rd birthday ."
        "...",
        ],
    "web_info": [
        "Robert Hossein - European Film Star Postcards: The role of the jaded criminal stuck with him
         in the coming decades. Hossein also started directing with the thriller Les salauds vont...",
        "Robert Hossein: Robert Hossein's grandfather was born in Iran.[better source needed] His
         father was André Hossein, a composer of Iranian Azerbaijani origin and his mother...",
        "Marina Vlady: | Marina Vlady | Vlady in 2009 |  Born | Marina Catherine de Poliakoff-Baydaroff
        (1938-05-10) 10 May 1938 (age 87) Clichy, Hauts-de-Seine, France...",
        "...",
        ]
      "snippet_analysis": [
        "[Snippet 1 Analysis]  No relevant information.",
        "[Snippet 2 Analysis]  No relevant information.",
        "[Snippet 3 Analysis]  No relevant information",
        "[Snippet 4 Analysis]  He was married first to Candice Patou (then Marina Poliakoff; on 23
                               December 1955).",
        "[Snippet 5 Analysis]  The spouse of Robert Hossein before Aug 1957 was Marina Vlady.",
        "..."
        ]},
"Research_brief":  "What is the name of Robert Hossein's spouse before August 1957, given conflicting
information from initial snippets suggesting it could be either Candice Patou (also referred to as Marina
Poliakoff, married on 23 December 1955) or Marina Vlady (married in 1955, with varying divorce years
 mentioned as 1959 or 1960), and additional mention of Marina de Poliakoff Baydaroff as another name?
The research should aim to clarify these discrepancies by verifying the correct identity and marital
timeline.",
"Research_process": {
"Identify Robert Hossein's spouse before August 1957, including the correct name (Candice Patou,
Marina Poliakoff, or Marina Vlady) and the exact marriage and divorce dates. Verify the information
through official records, biographies, or reputable historical sources to resolve discrepancies in reported
names and timelines.": {
    "iterations": [
      {
        "iteration": 1,
        "queries": [
           "Robert Hossein spouse before August 1957",
           "Robert Hossein and Marina Vlady marriage date",
           "Robert Hossein and Candice Patou relationship",
           "Robert Hossein and Marina Poliakoff marriage"],
        "results": [
           "Search results: SOURCE 1: Robert Hossein - Timenote -  URL: https://timenote.info/en/
            Robert-Hossein\n\nSUMMARY:\n<summary>\nRobert Hossein (30 December 1927 – 31
            December 2020) was a prominent Fr..."] } ] } },
"compressed_research": [
"Fully Comprehensive Findings on Robert Hossein's Spouse (Before August 1957): 1. Marriage to...",
"Fully Comprehensive Findings on Robert Hossein's Spouse (Before August 1957): 1. Marriage to...",
"..."]
"findings": "Reflection recorded: The research question involves verifying Robert Hossein's spouse...",
"final_report": "Clarification of Robert Hossein's Spouse Before August 1957 \n Summary of Findings..."
```


Figure 16: Case 4 illustrates the structure of the search tree with CAPF.