AWARE: Agentic Knowledge Warehousing for Contextual Intelligence

Hongjin Qian[1], **Siqi Bao**[2*], **Zhao Cao**[3], **Zheng Liu**[1,4*]
[1] Beijing Academy of Artificial Intelligence
[2] Hong Kong University of Science and Technology
[3] Renmin University of China
[4] Hong Kong Polytechnic University
{chienqhj,zhengliu1026}@gmail.com, sbao@connect.ust.hk

Abstract

Information seeking bridges the knowledge gap between a query and its answer. Although LLMs perform well broadly, their ability to close this gap is limited by pretraining and degrades on specialized or up-to-date queries. A common remedy augments LLMs with external knowledge, either by injecting retrieved evidence into context or interleaving retrieval with reasoning. The former limits exploration of layered dependencies, while the latter is bounded by context length, constraining efficiency and scalability. For complex tasks with intricate dependencies and large text volumes, both approaches become inadequate.

To tackle this bottleneck, we present AWARE (Agentic Knowledge Warehouse), an agentic knowledge warehousing framework that transforms heterogeneous, unstructured data into minimal, task-conditioned knowledge representations consumable by LLMs. Rather than exposing raw text, AWARE constructs knowledge through intent planning, *online multi-threaded exploration*, and *map–reduce evidence integration*, producing compact, LLM-ready context under finite budgets. Specifically, it applies *offline document structuring* to generate document headers that support controlled access, performs exploration with targeted refinement to recover layered information dependencies, and integrates distributed evidence into task-aware representations for downstream answer generation. Experiments on GAIA, WebWalker, and BrowseComp-Plus show improvements over all baselines[1].

1 Introduction

Recently, large language models (LLMs) have excelled in information-seeking tasks, generating coherent responses to queries ranging from simple to complex reasoning (Ouyang et al., 2022; Gemini Team, 2025; DeepSeek-AI, 2025). Yet their knowledge, fixed to the training corpus, is limited in coverage and timeliness. Hence, performance might deteriorate on knowledge-intensive tasks requiring specialized or up-to-date information (Zhao et al., 2024b; Huang et al., 2025).

To mitigate this limitation, LLMs are often augmented with external sources such as the web or local knowledge bases, most commonly through retrieval-augmented generation (RAG), where retrieved information is injected into the model's context to ground responses beyond pretrained knowledge (Lewis et al., 2020; Gao et al., 2024). While effective in many cases, this pre-inference retrieval scheme often falls short on complex tasks that require multi-step reasoning to uncover interdependent evidence (Zhao et al., 2024a). Tool-integrated reasoning methods improve on this by interleaving retrieval with reasoning, allowing agents to iteratively refine queries, interact with tools, and synthesize evidence (Jin et al., 2025; Li et al., 2025c), but its reliance on in-context evidence makes it inefficient and limited in scalability as context length grows.

Beyond these limitations, a more fundamental challenge remains: when LLMs rely on real-world data, they inevitably face *Data Chaos*, where the underlying sources involve long-form, heterogeneous, unstructured, noisy, and redundant content, such as web pages and PDF files (Zhu et al., 2024). A defining property of this regime is *low knowledge density*: answer-relevant information is sparsely embedded in retrieved text, so the marginal information gained per token is small. As a result, LLM context windows are quickly saturated by low-density content, making it difficult to isolate and assemble the critical evidence needed to bridge the knowledge gap. Effective information seeking therefore requires more than extracting isolated facts; it demands reorganizing fragmented raw text into coherent, task-conditioned evidence that an LLM can reliably exploit (Zhang et al., 2024).

This challenge becomes particularly evident in complex tasks, which often require processing large volumes of raw text to uncover the information needed. As shown in Figure 1, identifying the "shortest flight route" requires reasoning under multiple constraints, with relevant evidence scattered across many web pages. Solving such tasks demands both vertical exploitation for depth and horizontal exploration for breadth. Classical RAG lacks the depth for multi-layered reasoning, whereas tool-integrated reasoning methods provide depth but is restricted in breadth by context length (Zhao et al., 2024a; Li et al., 2025c). In

*Corresponding author.
[1]Our codes are in this *repository*

Findings of the Association for Computational Linguistics: ACL 2026, pages 2530–2540
July 2-7, 2026

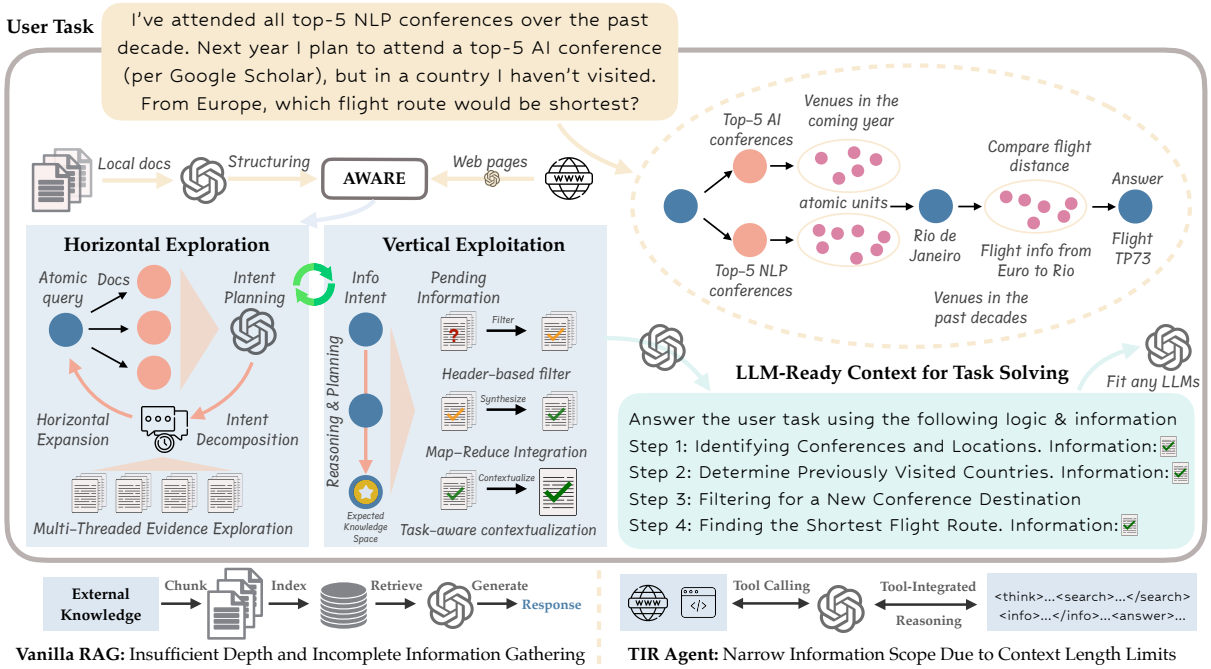

Figure 1: Overview of AWARE for complex information seeking. Given a user task, AWARE first applies *offline document structuring* to generate document headers that enable access control over heterogeneous sources (e.g., local files and web pages). At the task level, the agent decomposes the query into a sequence of intents, and alternates between *online multi-threaded exploration* for broad evidence coverage and vertical refinement for intent resolution. To remain scalable, retrieved documents are screened at the header level and processed via *map–reduce evidence integration*, extracting evidence in parallel and aggregating it into intent-specific subspace knowledge. Finally, these subspaces are organized and contextualized into a compact, LLM-ready representation that preserves the intent progression and supporting evidence for downstream answer generation.

practice, addressing these tasks requires a new retrieval paradigm that can operate effectively under data chaos, distill task-relevant signals from large-scale raw text, and assemble a minimal yet sufficient evidence space to bridge the knowledge gap.

In this paper, we introduce *Agentic Knowledge Warehouse* (AWARE), a retrieval framework that transforms large-scale unstructured sources into structured, task-specific knowledge consumable by LLMs. As shown in Figure 1, whereas single-pass retrieval and tool-integrated reasoning focus on how LLMs invoke existing retrieval tools, AWARE reconceptualizes retrieval as an agentic knowledge construction process that produces compact, LLM-ready representations.

At the corpus level, AWARE applies *offline document structuring* to build a structured corpus interface. For each document or web page, it generates a compact *document header* that encodes high-level semantic scope and structural organization, enabling access control under low knowledge density. These headers support coarse-grained navigation and prioritization, while the original text remains available for selective, fine-grained evidence extraction when warranted.

At the task level, given a query, AWARE plans a sequence of information intents, decomposes each intent into atomic sub-queries, and constructs intent-specific knowledge subspaces from the retrieved atomic units. It broadens coverage through *multi-threaded exploration* and resolves layered dependencies through targeted refinement. To maintain scalability, AWARE performs *map–reduce evidence integration*, first screening candidates at the header level and then extracting and aggregating evidence in parallel outside the reasoning loop. Finally, the constructed subspaces are organized into a task-aware knowledge chain and contextualized as a compact, LLM-ready representation for downstream answer generation. Through this design, AWARE reconstructs the task-sufficient knowledge required to bridge complex information gaps while respecting context budgets.

To evaluate the effectiveness of AWARE, we conduct extensive experiments on three challenging information-seeking benchmarks: General AI Assistant (GAIA), WebWalker, and BrowseComp-Plus. The results show that AWARE consistently outperforms baselines. Our contributions are threefold:

(1) We introduce *AWARE*, an agentic knowledge construction framework that converts heterogeneous, unstructured sources into compact, task-conditioned knowledge representations that are directly consumable by standalone LLMs.

(2) We instantiate AWARE as a modular pipeline with *offline document structuring* using document headers for access control, *online multi-threaded explo-*

ration with targeted refinement, *map–reduce evidence integration* for scalability, and *task-aware contextualization* for LLM-ready context.

(3) We provide extensive empirical validation showing that AWARE effectively connects large-scale external knowledge with LLM reasoning, offering a scalable and general solution to enhance standalone LLMs.

2 Related Work

Incorporating large language models (LLMs) with external tools for knowledge augmentation has emerged as a crucial paradigm for extending their capabilities beyond static pretraining (Zhao et al., 2024b; Li et al., 2025d). Early studies explored retrieval-augmented generation (RAG) (Lewis et al., 2020), where an input query retrieves relevant evidence before inference, and the retrieved content is injected into the model's context (Zhao et al., 2024a; Qian et al., 2025b,c). Subsequent enhancements have since been proposed, including query rewriting (Chan et al., 2024), self-critique mechanisms (Asai et al., 2024), memory augmentation (Qian et al., 2025c), and graph-based retrieval strategies (Edge et al., 2024). While effective in many settings, these pre-inference schemes face limitations when information needs are multi-layered or sparsely embedded across sources (Zhang et al., 2024; Qian and Liu, 2025b).

To address this, agentic search methods have recently gained traction. Most build on the ReAct framework (Yao et al., 2023), further optimized through expert-designed workflows (Li et al., 2025b; Qian and Liu, 2025a; Qiu et al., 2025) or via end-to-end reinforcement learning (Jin et al., 2025; Sun et al., 2025; Shi et al., 2025). Beyond classical knowledge-intensive tasks such as those in Wikipedia-based datasets (Petroni et al., 2021), recent work has also shifted attention to challenging information-seeking benchmarks like GAIA and BrowseComp (Mialon et al., 2023; Wei et al., 2025), which demand deep reasoning and long-horizon planning. Representative approaches include TTD-DR (Han et al., 2025), WebThinker (Li et al., 2025c) and the WebAgent series (Li et al., 2025a; Qian et al., 2025a; Wu et al., 2025a; Xia et al., 2025), which emphasize reasoning-intensive exploration across real-world web environments.

Overall, prior efforts have largely focused on how to leverage existing retrievers in different ways (Zhang et al., 2025). In contrast, our proposed AWARE establishes a new retrieval framework that directly constructs minimal yet sufficient knowledge for complex tasks, delivering curated LLM-ready context as a foundation for downstream reasoning.

3 Method

3.1 Preliminary

Complex Information-Seeking Task. Solving a task with an LLM can be formalized as $\mathcal{Y} = \Theta(\mathcal{X} \mid \mathcal{K})$, where Θ denotes the model's generative function and $\mathcal{K}$ captures the knowledge required to bridge the gap between input and output. In this view, producing the correct answer amounts to filling a *knowledge gap* that separates $\mathcal{X}$ from $\mathcal{Y}$. When the task is simple fact-based or commonsense in nature, the gap is small and can often be resolved by the model's pretrained knowledge or a single retrieval step. In contrast, complex tasks create a larger and more intricate gap. Recovering the knowledge space $\mathcal{K}$ in such cases is challenging, requiring multi-layered exploration and iterative decision-making in which evidence is progressively gathered, refined, and integrated until the gap is sufficiently closed to yield a reliable answer.

For *complex information-seeking tasks*, solving the problem typically unfolds as a multi-step reasoning process, where the required knowledge emerges in stages rather than all at once. In this setting, the knowledge gap $\mathcal{K}$ can be formally represented as a sequential *knowledge chain*:

$$\mathcal{K} = (\mathcal{K}_1 \rightarrow \mathcal{K}_2 \rightarrow \cdots \rightarrow \mathcal{K}_t), \tag{1}$$

where $\mathcal{K}_i$ denotes the crucial knowledge required at the i-th reasoning step, and the arrow $\rightarrow$ indicates the sequential dependency among steps. Each intermediate knowledge space $\mathcal{K}_i$ is itself formed by combining multiple *atomic knowledge spaces*:

$$\mathcal{K}_i = \mathcal{S}_{i,1} \cap \mathcal{S}_{i,2} \cap \cdots \cap \mathcal{S}_{i,n_i}, \tag{2}$$

where each $\mathcal{S}_{i,j}$ represents a minimal unit of knowledge that can be directly retrieved through a single query q, typically consisting of a set of relevant documents $\{D\}$[2].

This formulation emphasizes two complementary dimensions of reasoning. *Depth* arises from the sequential composition of the knowledge chain, while *breadth* comes from the conjunction of multiple atomic knowledge spaces within each step. Special cases follow naturally: when $t = 1, n_i = 1$, the task reduces to a single-hop factual query; when $t > 1$ but each $n_i = 1$, it corresponds to simple multi-hop reasoning over independent facts.

From an information-theoretic perspective, the *knowledge gap* $\mathcal{K}$ quantifies the additional information required to determine the correct answer $\mathcal{Y}$ given input $\mathcal{X}$. Solving a complex task can thus be seen as an iterative reduction of conditional entropy,

$$\begin{aligned} H(\mathcal{Y} \mid \mathcal{X}) > \; & H(\mathcal{Y} \mid \mathcal{X}, \mathcal{K}_1) \; > \; \cdots \\ & > \; H(\mathcal{Y} \mid \mathcal{X}, \mathcal{K}_1, \ldots, \mathcal{K}_t) \; = \; 0. \end{aligned}$$

Here, *breadth* aggregates multiple atomic sources that jointly constrain uncertainty, while *depth* reflects the sequential dependencies through which uncertainty is progressively eliminated. In this view, the *knowledge*

[2]Here, a "document" is used in a broad sense and may refer to a web page, a PDF file, or a full text piece.

chain functions as an information channel that transmits the missing bits required to close the gap between input and answer.

Data Chaos: Low Knowledge Density under Finite Context Budgets. We formalize *Data Chaos* as a context bottleneck arising from *low knowledge density* in raw evidence for complex information-seeking tasks, where task-relevant information is sparsely embedded in large volumes of heterogeneous text and cannot be packed into a budget-feasible, task-sufficient context.

Let $\mathcal{S}$ denote the universal knowledge space, and let $\mathcal{K}^* \subset \mathcal{S}$ be a *minimal sufficient knowledge subset* for a task $(\mathcal{X}, \mathcal{Y})$ such that

$$H(\mathcal{Y} \mid \mathcal{X}, \mathcal{K}^*) = 0,$$

while any strict subset of $\mathcal{K}^*$ is insufficient to determine the correct answer. Ideally, an LLM would be conditioned on a low-entropy context that is information-equivalent to $\mathcal{K}^*$.

In practice, complex tasks yield a large raw evidence set $\mathcal{R} = \{D_1, \ldots, D_m\} \subset \mathcal{S}$, where task-relevant evidence is sparse, distributed, and entangled with boilerplate, formatting artifacts, and weakly related content. To quantify this sparsity, we define the *knowledge density* of any context representation C as the amount of answer-relevant information per unit length:

$$\delta(C) \triangleq \frac{I(\mathcal{Y}; C \mid \mathcal{X})}{|C|},$$

where $|\cdot|$ denotes the token length (or any consistent length measure) and $I(\cdot;\cdot \mid \cdot)$ is conditional mutual information. For raw evidence, $\delta(\mathcal{R})$ is typically low, meaning that each additional token contributes little marginal information about $\mathcal{Y}$ beyond $\mathcal{X}$.

Under a finite context budget B, consider any context construction (or compression) mapping ϕ that produces $\tilde{\mathcal{K}} = \phi(\mathcal{R})$ with $|\tilde{\mathcal{K}}| \le B$. Let

$$\delta^*(\mathcal{R}) \triangleq \sup_{\phi} \delta(\phi(\mathcal{R}))$$

denote the maximal achievable knowledge density attainable from $\mathcal{R}$ by admissible transformations. Then any budget-feasible representation satisfies

$$I(\mathcal{Y}; \phi(\mathcal{R}) \mid \mathcal{X}) \le |\phi(\mathcal{R})| \cdot \delta^*(\mathcal{R}) \le B \cdot \delta^*(\mathcal{R}).$$

We say *Data Chaos* occurs when the budgeted information capacity falls short of what is needed to resolve the task:

$$B \cdot \delta^*(\mathcal{R}) < H(\mathcal{Y} \mid \mathcal{X}).$$

Equivalently, no budget-compliant transformation can be task-sufficient:

$$\forall\, \phi \text{ with } |\phi(\mathcal{R})| \le B, \qquad H(\mathcal{Y} \mid \mathcal{X}, \phi(\mathcal{R})) > 0.$$

That is, although $\mathcal{R}$ may be information-rich in aggregate, its low knowledge density prevents the required answer-relevant information from being concentrated into an LLM-ready context under finite budgets, making raw retrieval outputs an unreliable information channel for long-horizon reasoning.

3.2 The proposed method: AWARE

To tackle Data Chaos, we propose *Agentic Knowledge Warehouse* (AWARE), a retrieval framework that constructs task-sufficient, LLM-ready context from **low-density** noisy external sources. AWARE proceeds in two stages: *offline document structuring* and *hierarchical knowledge construction*.

3.2.1 Offline Document Structuring

For each document or web page D, AWARE constructs a compact *document header* $\mathcal{H}_D$ that encodes its high-level semantic scope and structural organization while omitting fine-grained content. By inspecting $\mathcal{H}_D$, the system can efficiently assess whether a document warrants deeper inspection, without committing context budget to the full text.

Formally, AWARE decomposes corpus interaction into two stages with distinct roles. At the *structured level*, decision-making operates over the set of document headers $\{\mathcal{H}_D\}$ to estimate relevance, identify redundancy, and prioritize candidates for inspection, producing a document-level control state $\mathcal{C}$. At the *content level*, this control state governs selective access to the original document D, where fine-grained evidence is extracted only when necessary. This separation is summarized as:

$$\begin{aligned} \text{Structured control}: &\quad \{\mathcal{H}_D\} \mapsto \mathcal{C}, \\ \text{Content access}: &\quad \mathcal{C} \mapsto D. \end{aligned} \tag{3}$$

By decoupling structured control from content access, AWARE avoids indiscriminate exposure of low-density raw text and supports cost-aware reasoning over large and heterogeneous corpora. Document headers serve as access control primitives that determine when to inspect, when to defer, and when to terminate exploration, improving the effective use of limited context budgets for scalable knowledge construction.

3.2.2 Hierarchical Knowledge Construction

By Eq. (1), the expected knowledge space $\mathcal{K}$ for a complex task cannot be obtained in a single step but must be assembled hierarchically. We model $\mathcal{K}$ as a sequence of subspaces $\{\mathcal{K}_1, \mathcal{K}_2, \ldots, \mathcal{K}_t\}$, where each $\mathcal{K}_i$ contains the evidence required to resolve one stage of reasoning. Each $\mathcal{K}_i$ is further constructed from a collection of atomic knowledge spaces $\{\mathcal{S}_{i,1}, \mathcal{S}_{i,2}, \ldots, \mathcal{S}_{i,n_i}\}$, where each $\mathcal{S}_{i,j}$ corresponds to the evidence obtained from a single information access action.

AWARE constructs this hierarchy in an agentic manner. Given a task $\mathcal{X}$, the system forms an initial information intent I_1 by identifying the primary knowledge gap, and decomposes I_1 into a set of sub-queries:

$$I_1 \mapsto \{q_{1,1}, q_{1,2}, \ldots, q_{1,n_1}\}. \tag{4}$$

Each sub-query $q_{1,j}$ triggers a concrete information access operation over the structured corpus interface, yielding an atomic knowledge space $\mathcal{S}_{1,j}$. Collectively, these atomic spaces are integrated into the subspace

$\mathcal{K}_1$. Once $\mathcal{K}_1$ provides sufficient evidence to resolve I_1, the system advances to the next intent I_2 and repeats the same procedure to construct $\mathcal{K}_2$. This process continues until the intent sequence is resolved, producing an approximation to the expected knowledge space $\mathcal{K}$ required for the task. We express the overall prediction as conditioning a reasoning operator on the constructed intent-evidence chain:

$$\mathcal{Y} = \Theta\Big(\mathcal{X} \mid \{(I_i, \mathcal{K}_i)\}_{i=1}^{t}\Big), \tag{5}$$

where $(I_i, \mathcal{K}_i)$ denotes the intent at stage i and its corresponding constructed subspace, and $\Theta(\cdot)$ represents the central reasoning operation that advances as intents are resolved.

Building on this formulation, we instantiate AWARE with three core mechanisms: *Online Multi-Threaded Exploration* for broad evidence coverage, *Map–Reduce Evidence Integration* for scalable processing, and *Task-Aware Contextualization* for constructing compact, LLM-ready context. We describe these components in detail below.

Online Multi-Threaded Exploration. A core challenge in constructing a subspace $\mathcal{K}_i$ for a given intent I_i lies in *intent alignment*. The description of I_i may be biased or incomplete, such that an initial set of sub-queries fails to surface all relevant evidence. Rather than treating this as a fixed retrieval problem, AWARE addresses intent alignment through *online exploration* that is guided by structured, document-level signals.

After executing an initial batch of sub-queries and obtaining atomic knowledge spaces, the agent forms a provisional evidence state for I_i. Crucially, sufficiency is assessed without immediately committing to full-text processing: the agent first inspects document headers $\{\mathcal{H}_D\}$ of the newly retrieved candidates to estimate topical coverage, detect redundancy, and identify missing aspects of the intent. When salient gaps remain, AWARE expands exploration in multiple threads by generating additional sub-queries conditioned on the current evidence state and the semantic and structural cues exposed by headers, targeting uncovered facets of I_i rather than reiterating prior exploration.

This procedure can be summarized as:

$$\mathcal{K}_i = \begin{cases} \bigcup_j \mathcal{S}_{i,j}, & \text{if the evidence is sufficient for } I_i, \\ \emptyset, & \text{otherwise,} \end{cases} \tag{6}$$

where $\{\mathcal{S}_{i,j}\}$ are obtained from the current set of sub-queries, and an empty outcome triggers additional query generation.

By conditioning expansion decisions on structured assessment over accumulated evidence and document headers, AWARE progressively broadens the evidence pool in a controlled manner. The resulting subspace $\mathcal{K}_i$ therefore reflects deliberate intent-driven exploration, rather than unstructured iterative querying.

Map–Reduce Evidence Integration. A second challenge concerns scalability. As online exploration broadens the evidence pool, exhaustively processing every retrieved document at full resolution quickly becomes impractical. AWARE addresses this challenge via *map–reduce evidence integration*, which separates coarse screening from fine-grained extraction and aggregates evidence in a parallel, budget-aware manner.

For each retrieved document D, AWARE first performs a lightweight screening step using its document header $\mathcal{H}_D$ to assess coarse relevance to the current intent I_i. This structured-level filtering eliminates clearly irrelevant candidates without accessing full text, substantially reducing downstream cost. Documents that pass screening are then processed in parallel to extract fine-grained evidence units, which are finally aggregated into the constructed subspace $\mathcal{K}_i$.

Formally, the procedure can be expressed as

$$\mathcal{K}_i = \mathcal{R}\Big(\{\mathcal{E}(D) \mid D \in \mathcal{F}(\mathcal{D}_i,\ I_i,\ \{\mathcal{H}_D\})\}\Big), \tag{7}$$

where $\mathcal{D}_i$ denotes the set of candidates retrieved for intent I_i, $\mathcal{F}$ is a header-based filter, $\mathcal{E}$ is a parallel evidence extractor over full text, and $\mathcal{R}$ is a reducer that integrates extracted units into $\mathcal{K}_i$.

All operators $\mathcal{F}$, $\mathcal{E}$, and $\mathcal{R}$ are instantiated using a lightweight auxiliary language model. By separating header-level screening from text-level extraction and enabling parallel integration, AWARE achieves scalable evidence synthesis while preserving broad coverage with controlled reasoning cost.

Task-Aware Contextualization. After evidence integration, AWARE organizes the constructed subspaces into a structured, task-conditioned context that can be consumed efficiently by a downstream LLM. Given a task $\mathcal{X}$, the intent sequence $\{I_i\}_{i=1}^{t}$, and the corresponding subspaces $\{\mathcal{K}_i\}_{i=1}^{t}$, the system assembles an ordered context representation

$$\mathcal{C} = \big(\mathcal{X},\ (I_1, \mathcal{K}_1),\ (I_2, \mathcal{K}_2),\ \ldots,\ (I_t, \mathcal{K}_t)\big), \tag{8}$$

which records the intent progression together with the supporting evidence at each stage.

Rather than exposing raw documents or unstructured evidence, $\mathcal{C}$ serves as a compact representation of the constructed knowledge for the task, preserving relevance, inter-intent dependencies, and evidential grounding while remaining budget-feasible. Importantly, under the knowledge-density view (Sec. 3.1), contextualization increases the effective density of answer-relevant information by concentrating evidence into a shorter representation, i.e., $\delta(\mathcal{C}) \gg \delta(\mathcal{R})$ for the same task, making the limited context budget more informative per token. A downstream LLM then conditions on $\mathcal{C}$ to generate the final answer $\mathcal{Y}$.

Table 1: Main experimental results. Best scores are shown in bold, and second-best are underlined. Following the official settings, we report Exact Match (EM) for GAIA, and LLM Equivalence Accuracy for WebWalkerQA.

Method	General AI Assistant				WebWalkerQA			
	Level 1	Level 2	Level 3	Avg.	Easy	Medium	Hard	Avg.
Direct Reasoning (w/o Retrieval)								
Qwen2.5-32B	20.5	9.6	8.3	13.6	3.8	2.5	3.3	3.1
Qwen3-32B	15.4	7.7	0.0	9.7	3.1	1.4	2.5	2.2
QwQ-32B	25.6	9.6	16.7	16.5	7.5	2.1	3.8	4.0
GPT-4o	23.1	15.4	8.3	17.5	6.7	6.0	4.2	5.5
Gemini-2.5-Flash	33.3	11.5	0.0	18.5	16.3	7.9	5.8	9.1
DeepSeek-R1-671B	43.6	26.9	8.3	31.1	5.0	11.8	11.3	10.0
Retrieval-Augmented Generation								
Vanilla RAG (Qwen2.5-32B)	12.8	11.8	8.3	11.8	23.1	14.3	11.3	15.3
Vanilla RAG (QwQ-32B)	33.3	36.5	8.3	32.0	36.9	26.1	33.5	31.2
Query Planning (Qwen2.5-32B)	30.8	17.3	0.0	20.4	29.4	36.4	25.0	30.7
Query Planning (QwQ-32B)	48.7	25.0	8.3	32.0	28.8	35.7	30.8	32.5
Iterative RAG (Qwen2.5-32B)	35.9	19.2	8.3	24.3	30.6	35.7	25.4	30.9
Iterative RAG (QwQ-32B)	51.3	28.8	8.3	35.0	29.4	32.9	31.3	31.5
Tool-Integrated Reasoning								
ReAct (Qwen2.5-32B)	46.1	44.2	8.3	40.7	44.3	46.7	29.2	38.4
ReAct (QwQ-32B)	48.7	34.6	16.7	37.8	35.6	29.1	13.2	24.1
ReAct (GPT-4o)	51.2	34.6	8.3	34.6	34.6	42.0	23.9	33.8
Search-o1-32B	53.8	34.6	16.7	39.8	43.1	35.0	27.1	34.1
WebThinker-32B	53.8	44.2	16.7	44.7	47.5	41.1	39.2	41.9
AWARE (QwQ-32B)	**61.5**	**46.2**	**33.3**	**50.5**	**53.1**	**55.0**	**50.8**	**53.1**

4 Experiments

4.1 Datasets and Baselines.

Datasets. We evaluate AWARE on three challenging benchmarks for complex information-seeking, where relevant information is typically scattered across heterogeneous sources such as web pages and unstructured documents (e.g., PDF and TXT files), reflecting the data chaos encountered in real-world settings.

GAIA (General AI Assistant) comprises over 450 real-world queries spanning multi-step reasoning, multimodal understanding, and tool use (Mialon et al., 2023). Following prior work (Li et al., 2025c; Wu et al., 2025a), we use 103 text-only validation questions. **WebWalkerQA** includes 680 queries across domains such as conferences and organizations, requiring agents to traverse subpages and integrate dispersed evidence, which makes it a long-horizon reasoning challenge (Wu et al., 2025b). **BrowseComp-Plus** consists of 830 complex questions whose answers are short and verifiable (Chen et al., 2025), but each question typically requires large-scale web search and multi-document reading to identify, gather, and verify the necessary evidence. Appendix A introduce baselines and implementation details.

4.2 Main Results

Table 1 reports the performance of AWARE and baseline. Our key findings are as follows: (1) Under direct reasoning without retrieval, all models handle GAIA tasks more readily, yet their accuracy remains modest. By contrast, accuracy drops sharply on WebWalkerQA, confirming that these benchmarks demand recent and long-tail knowledge rarely captured in model parameters. Interestingly, Qwen3-32B, although more recent, underperforms both Qwen2.5-32B and QwQ-32B, suggesting that Qwen3's hybrid reasoning design might compromise efficacy. Based on these observations, we use Qwen2.5-32B and QwQ-32B as the backbones for RAG and agentic-search baselines. (2) AWARE consistently outperforms not only vanilla RAG but also advanced variants that incorporate query rewriting or iterative refinement, validating the robustness of its retrieval paradigm. Unlike these pre-inference schemes that often leave evidence fragmented or incomplete, AWARE employs multi-threaded exploration and map-reduce integration to recover layered dependencies while filtering noise at scale. This design yields gains on tasks that demand multi-hop reasoning and long-horizon synthesis, where RAG methods struggle to provide coherent context. (3) AWARE surpasses agentic-search baselines, including workflow-based methods (e.g., Search-o1) and end-to-end optimized systems (e.g., WebThinker). Although WebThinker benefits from large-scale in-domain train-

Table 2: Performance and search calls on the BrowseComp-Plus benchmark using BM25 retrieval.

Model	Accuracy (%)	Search Calls
Gemini 2.5 Flash	15.54	10.56
Gemini 2.5 Pro	19.04	7.44
Sonnet 4	14.34	9.95
GPT-4.1	14.58	10.35
GPT-5	55.90	22.96
Qwen3-32B	3.49	0.92
Search-R1-32B	3.86	1.78
AWARE	29.9	16.41

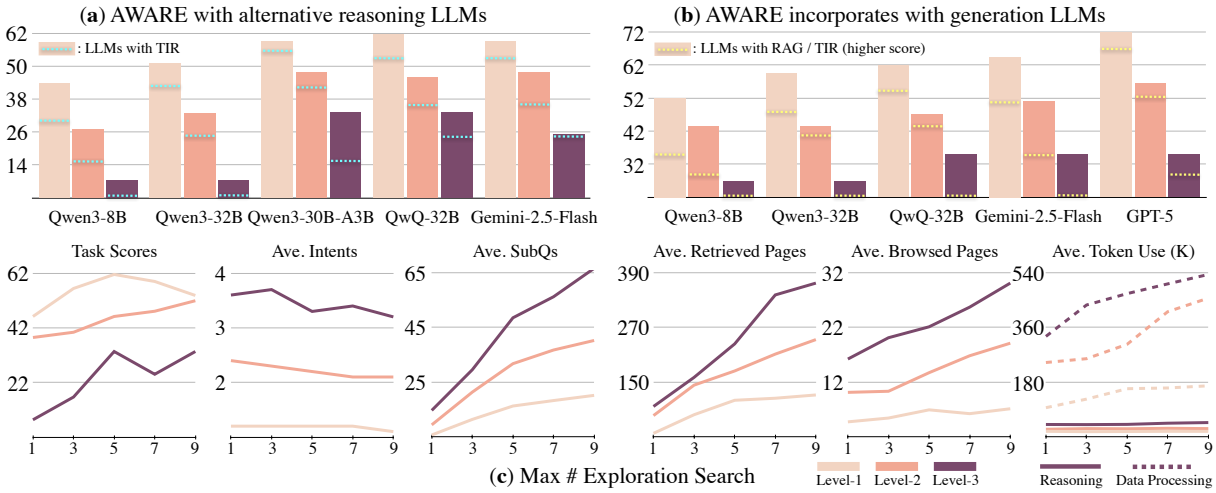

Figure 2: Analysis of AWARE on three perspectives: (a) effect of the central reasoning LLM, comparing AWARE with a TIR baseline (Search-o1); (b) transferability of AWARE's LLM-ready context, compared with RAG and TIR across downstream LLMs for answer generation; and (c) impact of the multi-threaded exploration budget on performance and resulting retrieval dynamics.

ing, AWARE without task-specific optimization still outperforms across all dimensions. This highlights AWARE's multi-threaded exploration, which enables broader coverage and reduces the risk of missing critical evidence.

4.3 Results on BrowseComp-Plus

Table 2 reports answer accuracy and average search calls on BrowseComp-Plus. Overall, models exhibit a performance versus search-intensity trade-off: strong proprietary systems (e.g., GPT-5) achieve higher accuracy but rely on substantially more search calls, while smaller open-source models make few calls and often fail to gather sufficient evidence, resulting in low accuracy. AWARE lies between these extremes. It improves accuracy over open-source baselines with increased exploration, yet uses fewer searches than the strongest proprietary model, suggesting its gains are not driven solely by higher search frequency. This pattern aligns with AWARE's design goal of allocating search and reading effort selectively via document access control rather than indiscriminate querying.

4.4 Ablation Study

AWARE is designed as an integrated framework, operating as a system, making it meaningful to analyze as a whole rather than in isolation. Accordingly, our ablation study examines three dimensions: (1) the role of different LLMs as AWARE's central reasoning agent, (2) the generalizability of AWARE-generated context across diverse models, and (3) the dynamics of agentic retrieval, with a focus on multi-threaded exploration depth and evidence synthesis efficiency. Figure 2 summarizes the results, which we discuss below.

Impact of Reasoning LLM Selection. AWARE relies on the capabilities of its central reasoning model. As shown in Figure 2 (a), it consistently outperforms the tool-integrated reasoning baseline (Search-o1) across different LLMs, demonstrating the robustness of its design. Nonetheless, the strength of the reasoning model plays a decisive role. Reasoning-oriented models such as QwQ-32B and Qwen3-30B-A3B achieve the best results, surpassing Qwen3-32B, a hybrid model with diluted reasoning capacity. When paired with Gemini2.5-Flash, AWARE also delivers competitive results through a dynamic strategy: the model enables "thinking mode" for complex planning steps while producing direct outputs for simpler ones, striking a balance between efficiency and accuracy. Overall, these findings show that while AWARE adapts well to diverse LLMs, its performance scales with the depth and quality of reasoning in the central agent.

Generalizability of AWARE-Generated Context. AWARE serves as a retrieval framework that produces reliable, task-specific context in an LLM-ready form, which can be seamlessly applied to any model. As illustrated in Figure 2 (b), supplying this curated context to different generation models consistently outperforms both TIR and RAG baselines, underscoring the robustness and generalizability of AWARE. For smaller models such as Qwen3-8B, the performance gains are especially pronounced, showing that AWARE can effectively compensate for the limited reasoning and knowledge capacity of lightweight LLMs. Conversely, when applied to stronger models such as GPT-5, the curated context is leveraged even more effectively, yielding further improvements and demonstrating AWARE's scalability across model strengths.

Agentic Behavior across Multi-Threaded Exploration Depths. Our analysis in Figure 2 (c) shows how AWARE's components jointly enable effective and scalable retrieval. First, multi-threaded explo-

Table 3: Case study on a Level-3 sample from GAIA. The reasoning agent within AWARE first addresses the initial knowledge gap using intrinsic knowledge, then conducts agentic external knowledge exploration. The resulting task-specific context reconstructs a minimal yet sufficient knowledge space and delivers it in an LLM-ready form for solving the input task.

Task: What animals that were mentioned in both Ilias Lagkouvardos's and Olga Tapia's papers on the alvei species of the genus named for Copenhagen outside the bibliographies were also present in the 2021 article cited on the alvei species' Wikipedia page about a multicenter, randomized, double-blind study?
Ground-truth answer: Mice

Initial Reasoning: Identify the scientific genus named for Copenhagen. → Hafnia
Information Intent $\mathcal{I}_1$:
- Find scientific papers by Ilias Lagkouvardos concerning Hafnia alvei.
- Find scientific papers by Olga Tapia concerning Hafnia alvei.
- Find the Wikipedia page for Hafnia alvei to locate the 2021 article cited about a multicenter, randomized, double-blind study.
4 atomic queries issued → 36 pages retrieved → 13 pages browsed; Tokens: 8.9K (reasoning), 227.2K (processing).
Knowledge Subspace $\mathcal{K}_1$: Ilias Lagkouvardos authored a paper "An Insight into Goat Cheese: The Tales of Artisanal and Industrial Gidotyri Microbiota" on H. alvei were identified, while Olga Tapia authored a 2022 paper in Molecules titled "The Molecular Weaponry Produced by the Bacterium Hafnia alvei in Foods", discussing the bacterium's role in food systems. The Wikipedia page for Hafnia alvei cites a 2021 multicenter, randomized, double-blind study published in Nutrients—"The probiotic strain H. alvei HA4597® improves weight loss in overweight subjects under moderate hypocaloric diet".
Information Intent $\mathcal{I}_2$:
- Search for the full text of Olga Tapia's 2022 paper "The Molecular Weaponry Produced by the Bacterium Hafnia alvei in Foods" published in Molecules and identify any animals mentioned within the text (excluding bibliographies). ..... [omitted]

Task-Specific Context $\mathcal{C}$: The task requires the following steps and information:
Step 1: Identify the scientific genus named for Copenhagen and confirm the relevant alvei species. Info: The genus is Hafnia, derived from the Latin name for Copenhagen; relevant species: Hafnia alvei.
Step 2: Identify the 2021 article cited on the Hafnia alvei Wikipedia page. Info: Nutrients (2021): "The probiotic strain H. alvei HA4597® improves weight loss in overweight subjects under moderate hypocaloric diet" (multicenter, randomized, double-blind).
..... [omitted]
Step 5: Identify animals mentioned in the 2021 Nutrients study. Info: Human participants (n=236); also referenced preclinical studies using obese mice (ob/ob, HFD-fed).
Step 6: Determine overlap of animals across sources. Info: The only animal shared in the papers is mice.

ration is crucial: increasing depth enlarges the evidence pool and improves performance from depth 1 to 5, then fluctuates as evidence saturates. Deeper exploration also reduces the number of required intents, easing the reasoning workload and accelerating convergence. Second, deeper exploration increases subqueries and retrieved pages, especially on complex tasks. Here, map–reduce evidence integration proves effective: it removes nearly 90% of irrelevant pages using document headers and processes the remainder in a map–reduce fashion, supporting scalability. Finally, token analysis indicates that reasoning contributes only a small fraction of total cost relative to large-scale data processing, validating AWARE's division of labor between a strong central model for reasoning and lightweight auxiliaries for bulk processing.

4.5 Case Study

Table 3 presents a case study that illustrates how AWARE constructs an LLM-ready context for a complex task. The query requires first identifying a *scientific genus* and then consulting academic papers, with the final answer derived by intersecting the animals discussed across these sources.

AWARE begins by leveraging the model's intrinsic knowledge to establish the target genus, which anchors subsequent information intents. For the first intent, retrieving the titles of the three papers, AWARE executes multi-threaded exploration: issuing multiple atomic queries, gathering 36 candidate pages, filtering them with relevance checks, and browsing only 13 to distill the relevant evidence. This step demonstrates AWARE's ability to maximize coverage while keeping processing efficient. The process then advances to the next intent, shifting from paper discovery to full-text analysis, with new sub-queries generated adaptively to uncover the animals mentioned. Once all necessary evidence is accumulated, AWARE synthesizes the results into a structured, task-specific context that integrates both retrieved knowledge and intermediate reasoning steps. The resulting representation is compact yet sufficient, capturing logical dependencies across intents, minimizing redundancy, and fitting within the LLM's context window.

Notably, the case highlights AWARE's efficiency: reasoning consumes only 8.9K tokens, while large-scale evidence processing consumes 227K tokens. This demonstrates AWARE's balanced design, where strong central models focus on reasoning while lightweight auxiliaries handle bulk text processing.

5 Conclusion

We presented AWARE, an agentic knowledge warehousing framework that equips LLMs with external knowledge in a structured, task-conditioned form. AWARE applies offline document structuring to build a controlled corpus interface, and performs hierarchical knowledge construction that decomposes complex queries into intents, explores and refines evidence, and integrates it via map–reduce synthesis into compact, LLM-ready context. This design enables the reconstruction of task-sufficient knowledge under constrained context budgets. Experiments on challenging benchmarks, together with ablations and case studies, show that AWARE is scalable and adaptable across tasks and model settings. Overall, these results indicate that AWARE provides a practical route to contextual intelligence for standalone LLMs.

Acknowledgement

This work was supported by National Natural Science Foundation of China No. 62502049.

Limitations

Although AWARE demonstrates strong performance across diverse benchmarks, several limitations of this work should be acknowledged.

Method Scope. AWARE is proposed as a general retrieval paradigm that operates independently of specific models and can integrate seamlessly with both open-source and closed-source LLMs. However, unlike data-driven approaches that train task-specific models, AWARE does not incorporate optimization strategies tailored to particular domains. This limitation arises from objective constraints. The agentic framework of AWARE involves multiple capabilities such as planning, intent generation, evidence refinement, and synthesis, all of which would require carefully annotated or synthetically generated data for supervised optimization. While reinforcement learning could serve as an end-to-end alternative, producing large-scale, high-quality training data and running optimization for large models would demand significant resources. For example, reinforcement learning on a 32B model reasonably requires at least 32 H100 80G GPUs, which remain beyond reach. Despite this, we argue that AWARE can continually benefit from improvements in general-purpose LLMs. The very skills required within AWARE, including reasoning, planning, and synthesis, fall within the optimization scope of mainstream models. Thus, even without explicit task-specific fine-tuning, AWARE achieves strong results. Moreover, excessive specialization on narrow domains may harm the generality of large models, introducing overfitting risks and reducing adaptability.

Experimental Scope. Given that AWARE is currently designed for text-only tasks, we follow prior work and exclude the non-textual portion of GAIA, which prevents us from evaluating on the full benchmark. Across all three benchmarks in this paper, the use of the Google Search API and the Jina web page crawling API has incurred costs exceeding 1,200 USD, placing considerable pressure on the experimental budget and limiting our ability to scale the evaluation further. Future work may explore more cost-efficient ways to enable wider coverage of these benchmarks.

Baseline Coverage. We strive to ensure that baselines in our main experiments are comparable in model size, open-sourcing status, and implementation feasibility. Nonetheless, it is not possible to evaluate against all related baselines. Some rely on substantially different model sizes, others are not fully released, and some require resources that are unavailable in our setting.

References

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

Chi-Min Chan, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, and Jie Fu. 2024. Rq-rag: Learning to refine queries for retrieval augmented generation. In *First Conference on Language Modeling*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2318–2335.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Sahel Sharifymoghaddam, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. 2025. Browsecomp-plus: A more fair and transparent evaluation benchmark of deep-research agent. *CoRR*, abs/2508.06600.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

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. *Preprint*, arXiv:2404.16130.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Google Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Rujun Han, Yanfei Chen, Zoey CuiZhu, Lesly Miculicich, Guan Sun, Yuanjun Bi, Weiming Wen, Hui Wan, Chunfeng Wen, Solène Maître, George Lee, Vishy Tirumalashetty, Emily Xue, Zizhao Zhang, Salem Haykal, Burak Gokturk, Tomas Pfister, and Chen-Yu Lee. 2025. Deep researcher with test-time diffusion. *Preprint*, arXiv:2507.16075.

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

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

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.

Kuan Li, Zhongwang Zhang, Huifeng Yin, Liwen Zhang, Litu Ou, Jialong Wu, Wenbiao Yin, Baixuan Li, Zhengwei Tao, Xinyu Wang, et al. 2025a. Websailor: Navigating super-human reasoning for web agent. *arXiv preprint arXiv:2507.02592*.

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

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

Yuchen Li, Hengyi Cai, Rui Kong, Xinran Chen, Jiamin Chen, Jun Yang, Haojie Zhang, Jiayi Li, Jiayi Wu, Yiqun Chen, et al. 2025d. Towards ai search paradigm. *arXiv preprint arXiv:2506.17188*.

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

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *NAACL-HLT 2021, Online, June 6-11, 2021*, pages 2523–2544. Association for Computational Linguistics.

Hongjin Qian and Zheng Liu. 2025a. Metaagent: Toward self-evolving agent via tool meta-learning. *Preprint*, arXiv:2508.00271.

Hongjin Qian and Zheng Liu. 2025b. Scent of knowledge: Optimizing search-enhanced reasoning with information foraging. *arXiv preprint arXiv:2505.09316*.

Hongjin Qian, Zheng Liu, Chao Gao, Yankai Wang, Defu Lian, and Zhicheng Dou. 2025a. Hawkbench: Investigating resilience of rag methods on stratified information-seeking tasks. *arXiv preprint arXiv:2502.13465*.

Hongjin Qian, Zheng Liu, Peitian Zhang, Zhicheng Dou, and Defu Lian. 2025b. Boosting long-context information seeking via query-guided activation refilling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9453–9464.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025c. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2366–2377, New York, NY, USA. Association for Computing Machinery.

Jiahao Qiu, Xuan Qi, Tongcheng Zhang, Xinzhe Juan, Jiacheng Guo, Yifu Lu, Yimin Wang, Zixin Yao, Qihan Ren, Xun Jiang, Xing Zhou, Dongrui Liu, Ling Yang, Yue Wu, Kaixuan Huang, Shilong Liu, Hongru Wang, and Mengdi Wang. 2025. Alita: Generalist agent enabling scalable agentic reasoning with minimal predefinition and maximal self-evolution. *Preprint*, arXiv:2505.20286.

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

Yaorui Shi, Sihang Li, Chang Wu, Zhiyuan Liu, Junfeng Fang, Hengxing Cai, An Zhang, and Xiang Wang. 2025. Search and refine during think: Autonomous retrieval-augmented reasoning of llms. *Preprint*, arXiv:2505.11277.

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

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

Jialong Wu, Baixuan Li, Runnan Fang, Wenbiao Yin, Liwen Zhang, Zhengwei Tao, Dingchu Zhang, Zekun Xi, Gang Fu, Yong Jiang, et al. 2025a. Webdancer: Towards autonomous information seeking agency. *arXiv preprint arXiv:2505.22648*.

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

Ziyi Xia, Kun Luo, Hongjin Qian, and Zheng Liu. 2025. Open data synthesis for deep research. *arXiv preprint arXiv:2509.00375*.

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

Weinan Zhang, Junwei Liao, Ning Li, Kounianhua Du, and Jianghao Lin. 2024. Agentic information retrieval. *arXiv preprint arXiv:2410.09713*.

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

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. 2024b. A survey of large language models. *Preprint*, arXiv:2303.18223.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Haonan Chen, Zhicheng Dou, and Ji-Rong Wen. 2024. Large language models for information retrieval: A survey. *Preprint*, arXiv:2308.07107.

A Appendix

A.1 Baselines

We compare AWARE against three groups of baselines. (1) *Direct Reasoning*: strong standalone LLMs used without external tools, including Qwen2.5-32B, Qwen2.5-32B, QwQ-32B, GPT-4o, Gemini-2.5-Flash and DeepSeek-R1-671B (DeepSeek-AI, 2025; Gemini Team, 2025; OpenAI, 2024). (2) *Retrieval-Augmented Generation*: methods that inject retrieved evidence, such as vanilla RAG and enhanced variants with query planning or iterative refinement (Shao et al., 2023; Chan et al., 2024). (3) *Tool-Integrated Reasoning*: approaches that interleave retrieval with reasoning, including ReAct, Search-o1, and WebThinker (Yao et al., 2023; Li et al., 2025b,c).

A.2 Implementation Details

In the main experiments, AWARE adopts QwQ-32B as the central reasoning model, supported by Qwen2.5-7B as an auxiliary processor for parallel data synthesis. For each query, AWARE curates a task-specific context, which is then directly fed into standalone LLMs for answer generation. Unless otherwise specified, the maximum depth for the *multi-threaded exploration* is set to 5.

To construct the web page collection used in the benchmarks, we first run AWARE directly with a search engine. In this initialization step, the local index is replaced by the search engine, and the use of document header is approximated by the first 1,024 tokens of each retrieved page. For every query, the top-20 web pages are collected. After five full runs on each benchmark, this procedure yields approximately 100K web pages in total. These pages are then processed to generate document headers, which serve as the basis for the subsequent indexing process within AWARE.

Notably, evaluation with the online search engine proves both *slow* and *unstable*, since each sample requires executing multiple sub-queries and crawling tens to hundreds of web pages. The performance in this setting is also substantially lower than that achieved with the local index.

During the *data indexing process*, we employ BGE-M3 as the dense embedding model (Chen et al., 2024), complemented by a BM25 index constructed over the full web content. All retrieval operations are instantiated using ElasticSearch, which provides a stable and scalable infrastructure for large-scale search. For online retrieval, we rely on Google's Custom Search JSON API to identify relevant pages, and utilize Jina AI's Web Reader to extract full web content. For all baselines, we either report results directly from their original papers or reproduce them using official implementations. All experiments are conducted on a node of eight NVIDIA A100-40G GPUs.

To ensure transparency and reproducibility, we release all prompts used in AWARE along with full experiment logs, including intermediate artifacts such as search intents, atomic queries, retrieved and browsed pages, refined evidence, and organized knowledge. These materials are available in *this anonymous repository*.

B AI Usage Disclosure

In this work, AI assistants were used exclusively for polishing the manuscript, including grammar checking and language refinement. The initial draft was prepared manually by the authors, and only selected sections were refined with AI assistance.

AI assistants did not contribute to any other part of the research, including ideation, literature review, or figure preparation.