

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 TAMING DATA CHAOS: AGENTIC KNOWLEDGE WAREHOUSING FOR CONTEXTUAL INTELLIGENCE

Anonymous authors

Paper under double-blind review

ABSTRACT

Information seeking can be viewed as bridging the knowledge gap between a query and its answer. While large language models (LLMs) perform strongly across diverse tasks, their capacity to fill this gap is bounded by pretraining data and deteriorates on queries requiring specialized or up-to-date knowledge. A common solution is to augment LLMs with external knowledge, either by injecting retrieved evidence into the context or by interleaving retrieval with reasoning. The former restricts exploration of layered dependencies, whereas the latter is constrained by context length, limiting both efficiency and scalability. Yet complex tasks often involve intricate dependencies and may require processing large volumes of raw text, under which both strategies become inadequate.

To tackle this bottleneck, we present Agentic Knowledge Warehouse (AWARE), a retrieval paradigm that transforms vast unstructured data into minimal, task-specific knowledge consumable by LLMs. Rather than simply returning raw information, AWARE curates knowledge through an agentic process that plans, explores, and synthesizes evidence into coherent context. Specifically, it organizes raw corpora with document-level gist memory for global coverage, applies diffusion-based exploration with vertical exploitation to recover layered dependencies, and employs map-reduce inspired synthesis to integrate large-scale evidence into a compact, LLM-ready context. This design enables both in-depth exploration and scalable integration, reconstructing the knowledge space needed to address task-specific knowledge gaps. Experiments on GAIA, WebWalker, and BrowseComp show that AWARE outperforms baselines, validating its effectiveness and generality. Our codes are available in this *anonymous repository*.

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. As a result, 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 data sources involve long-form, heterogeneous, unstructured, noisy, and redundant content, such as web pages and PDF files (Zhu et al., 2024). In such settings, useful information is sparsely embedded within large volumes of retrieved text, leading to a very low signal-to-noise ratio. As a result, LLM context windows are quickly saturated, leaving it unable to isolate and assemble the crucial evidence needed to bridge the knowledge gaps. Effective information seeking

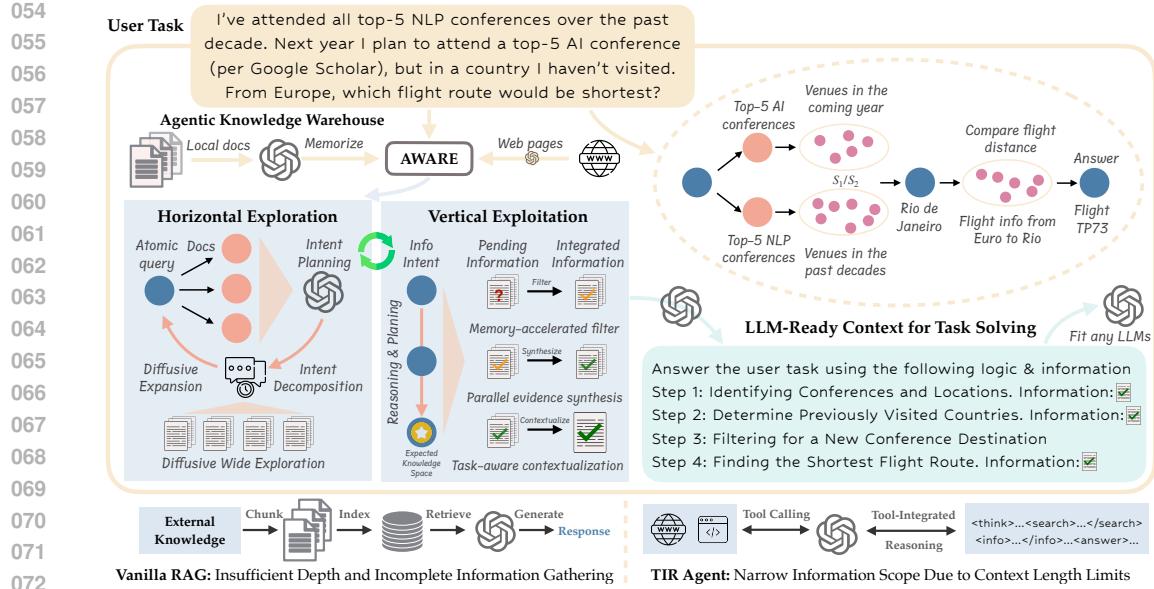


Figure 1: Illustration of a complex information-seeking task, where the answer depends on satisfying multiple conditions through horizontal exploration and vertical exploitation. AWARE addresses this agentically by formulating intents, decomposing them into atomic queries, and expanding coverage via diffusion to gather raw documents. Resolved intents advance iteratively to the next, with documents processed in parallel and synthesized through a map-reduce procedure into subspace knowledge, which is then transformed into an LLM-ready context.

therefore requires more than extracting isolated facts; it demands reorganizing fragmented raw text into coherent 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 are restricted in breadth by context length (Zhao et al., 2024a; Li et al., 2025c). In 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 vast 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 employ existing retrieval tools, AWARE instead reconceptualizes retrieval itself through an agentic process that organizes unstructured corpora into structured, task-specific knowledge and delivers it in an LLM-ready form. At the corpus level, AWARE ingests unstructured sources and constructs a structured warehouse. Each document is paired with a *gist memory* that abstracts its overall theme and structure, and indexed through a hybrid scheme combining dense memory representations with sparse raw text for both global awareness and fine-grained access. At the task level, given a complex query, AWARE plans information intents, decomposes them into atomic sub-queries, and retrieves the corresponding atomic knowledge units. It expands coverage through diffusion-based horizontal exploration and resolves layered dependencies through vertical exploitation. To maintain efficiency at scale, this process leverages *memory-accelerated exploration* to filter documents via *gist memories* and executes in a map-reduce-style parallel evidence synthesis outside the main reasoning loop. Finally, the collected evidence is synthesized into a coherent, task-specific knowledge chain and reformulated as a compact, LLM-ready context. Through this design, AWARE reconstructs the minimal sufficient knowledge space required to bridge complex information gaps, delivering it in a form that LLMs can readily comprehend and leverage at scale.

To evaluate the effectiveness of AWARE, we conduct extensive experiments on three challenging information-seeking benchmarks: General AI Assistant (GAIA), WebWalker, and BrowseComp. The results show that AWARE consistently outperforms baselines. Our contributions are threefold: (1) We introduce the Agentic Knowledge Warehouse, a new retrieval paradigm that transforms unstructured data into structured, task-specific knowledge in an LLM-ready form, seamlessly applicable to any standalone model. (2) We instantiate AWARE through a flexible framework that integrates document-level gist memory with an agentic decision process, combining diffusion-based horizontal exploration and vertical exploitation to realize contextual intelligence. (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). 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., 2025), 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; Wei et al., 2025).

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; Chen et al., 2025; 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; Geng et al., 2025; Wu et al., 2025a), 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 typically small and can often be resolved by the model’s pretrained knowledge or a single retrieval step. In contrast, complex tasks create a much larger and more intricate gap. Recovering the expected 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*:

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

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

162 formed by combining multiple *atomic knowledge spaces*:

$$164 \quad \mathcal{K}_i = \mathcal{S}_{i,1} \cap \mathcal{S}_{i,2} \cap \cdots \cap \mathcal{S}_{i,n_i}, \quad (2)$$

165 where each $\mathcal{S}_{i,j}$ represents a minimal unit of knowledge that can be directly retrieved through a
166 single query q , typically consisting of a set of relevant documents $\{D\}$ ¹.

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

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

$$176 \quad H(\mathcal{Y} | \mathcal{X}) > H(\mathcal{Y} | \mathcal{X}, \mathcal{K}_1) > \cdots > H(\mathcal{Y} | \mathcal{X}, \mathcal{K}_1, \dots, \mathcal{K}_t) = 0. \quad (3)$$

178 Here, *breadth* aggregates multiple atomic sources that jointly constrain uncertainty, while *depth* re-
179 flects the sequential dependencies through which uncertainty is progressively eliminated. In this
180 view, the *knowledge chain* functions as an information channel that transmits the missing bits re-
181 quired to close the gap between input and answer.

182 **Data Chaos: The Bottleneck for LLM Context.** We refer to *Data Chaos* as the state in which
183 essential knowledge is entangled within vast amounts of unstructured, redundant, and noisy data
184 that an LLM cannot directly exploit. Suppose the universal knowledge space is denoted by \mathcal{S} . In
185 principle, one could hope to isolate a *minimal but sufficient* subset $\mathcal{K}^* \subset \mathcal{S}$ that contains exactly the
186 information required for producing the answer \mathcal{Y} and nothing more. Such a representation would be
187 ideal, as it would minimize entropy and maximize the signal-to-noise ratio of the input context.

188 In practice, however, retrieval for complex tasks produces large volumes of raw text (e.g., hundreds
189 of web pages or thousands of PDF documents) that remain far from this ideal. The data are dom-
190 inated by formatting artifacts, boilerplate language, and irrelevant content, with useful knowledge
191 sparsely embedded within. Feeding such raw collections into an LLM is both inefficient and in-
192 effective: the context window is quickly saturated, and the high entropy of the input obscures the
193 information truly relevant to the task. In this sense, retrieved data in their raw form are not *LLM-*
194 *ready*, but rather exemplify the disorder of *Data Chaos*.

196 3.2 THE PROPOSED METHOD: AGENTIC KNOWLEDGE WAREHOUSE

198 To tackle the challenge of Data Chaos, we propose *Agentic Knowledge Warehouse* (AWARE), a
199 retrieval framework that iteratively explores and exploits external knowledge, distilling high-entropy
200 and noisy retrieval results into a minimal yet sufficient knowledge space, which is then transformed
201 into task-specific, LLM-ready context for enabling contextual intelligence. AWARE operates in two
202 stages: *data indexing* and *agentic knowledge discovery*, which we detail in the following sections.

203 3.2.1 DATA INDEXING WITH GIST MEMORY

205 Indexing real-world text data is challenging: document lengths vary from short news snippets to full
206 academic papers, and structures are highly diverse, as in web pages with templates, metadata, and
207 mixed formats. Direct applying dense retrieval to such raw text struggles under these conditions:
208 for very long documents, it lacks global awareness because the encoder can only process limited
209 windows; for structurally complex documents, it fails to capture implicit layout or organizational
210 cues that are not explicitly encoded in raw text.

211 To overcome these limitations, AWARE introduces an intermediate representation that makes im-
212 plicit global and structural information explicit. For each document D , a lightweight long-context
213 model produces a textual abstraction \mathcal{D} that verbalizes the document’s high-level topics and struc-
214 tural cues while omitting details. For example, a journal issue page may be represented by its

215 ¹Here, a “document” is used in a broad sense and may refer to a web page, a PDF file, or a full text piece.

216 title, scope, and the categories of included articles, without enumerating the individual titles. This
 217 abstraction parallels the way humans process long texts: after reading, we tend to retain only a *gist*-
 218 level memory that preserves the overall theme and structure while discarding fine-grained details.
 219 We therefore refer to \mathcal{D} as a form of *gist memory*. Unlike standard dense embeddings, which are
 220 typically derived from limited text windows and thus capture primarily local semantics, *gist memory*
 221 encodes a document’s global theme and structure, such as topical hierarchy and organizational flow.
 222 This richer abstraction allows retrieval methods not only to locate broadly relevant documents but
 223 also to filter and prioritize them more effectively, accessing cues that would otherwise be overlooked
 224 by dense representations alone.

225 Indexing then proceeds in a hybrid manner. Each *gist* representation \mathcal{D} is encoded into a dense
 226 vector $\mathbf{z}_D \in \mathbb{R}^d$ to capture global semantics, while the raw document D is indexed using a sparse
 227 scheme to retain fine-grained evidence. Given a query q , relevance is computed as:

$$228 \quad \text{Rel}(q, D) = \alpha \cdot \text{sim}_{\text{dense}}(q, \mathbf{z}_D) + (1 - \alpha) \cdot \text{sim}_{\text{sparse}}(q, D), \quad (4)$$

229 where $\alpha \in [0, 1]$ balances semantic coherence from dense matching with detail sensitivity from
 230 sparse matching. This *gist-memory based hybrid index* ensures that retrieval remains both globally
 231 AWARE and locally precise, enabling effective navigation of heterogeneous knowledge sources.
 232

233 3.2.2 AGENTIC KNOWLEDGE DISCOVERY

235 By the definition of Eq. (1), the expected knowledge space \mathcal{K} for a complex task cannot be ob-
 236 tained in a single step but must be assembled hierarchically. Specifically, \mathcal{K} is composed of a se-
 237 quence of subspaces $\{\mathcal{K}_1, \mathcal{K}_2, \dots, \mathcal{K}_t\}$, where each \mathcal{K}_i captures the evidence needed to resolve one
 238 stage of reasoning. In turn, every \mathcal{K}_i is constructed from a collection of atomic knowledge spaces
 239 $\{\mathcal{S}_{i,1}, \mathcal{S}_{i,2}, \dots, \mathcal{S}_{i,n_i}\}$, each of which can be retrieved by issuing a single query q .

240 AWARE constructs this knowledge in an agentic manner. Given a task \mathcal{X} , the system reasons over it
 241 to issue an initial information intent I_1 , identifies the underlying knowledge gaps, and decomposes
 242 I_1 into atomic sub-queries:

$$243 \quad I_1 \mapsto \{q_{1,1}, q_{1,2}, \dots, q_{1,n_1}\}. \quad (5)$$

245 Each sub-query $q_{1,j}$ corresponds to a concrete retrieval action defined in Eq. (4), yielding an atomic
 246 knowledge space $\mathcal{S}_{1,j}$. Collectively, these atomic spaces form the subspace \mathcal{K}_1 . Once \mathcal{K}_1 provides
 247 sufficient evidence to resolve I_1 , the process advances to the next intent I_2 , producing \mathcal{K}_2 in the
 248 same manner. This sequential procedure continues until all subspaces are constructed, yielding an
 249 approximation of the expected knowledge space \mathcal{K} required to solve the task. This process can be
 250 expressed recursively as:

$$251 \quad \mathcal{Y} = \Theta_t \left(\dots \Theta_2 \left(\Theta_1(\mathcal{X} \mid I_1, \mathcal{K}_1) \mid I_2, \mathcal{K}_2 \right) \dots \mid I_t, \mathcal{K}_t \right). \quad (6)$$

252 where I_t denotes the information intent at step t , \mathcal{K}_t is the corresponding subspace constructed from
 253 atomic spaces, and Θ_t is the reasoning operation that advances once I_t is resolved.

254 Building on this general definition, we instantiate AWARE with three core mechanisms: *Diffusive*
 255 *Wide Exploration* for knowledge coverage, *Memory-Guided Parallel Synthesis* for processing effi-
 256 ciency, and *Task-Aware Contextualization* for synthesizing LLM-ready context. These components
 257 will be introduced in detail below.

259 **Diffusive Wide Exploration.** A central challenge in constructing a subspace \mathcal{K}_i for a given intent
 260 I_i lies in *intent alignment*: the description of I_i may be biased or incomplete, so its initial sub-queries
 261 may fail to cover the expected subspace. To mitigate this, AWARE employs a *Diffusion Search*
 262 strategy designed to maximize the coverage of intent-relevant knowledge. After executing the initial
 263 queries and obtaining atomic spaces, the agent evaluates whether the accumulated evidence suffices
 264 for I_i . If not, it expands the search frontier by generating additional queries conditioned on past
 265 results, thereby progressively enlarging the retrieved knowledge space:

$$266 \quad \{q_{i,1}, \dots, q_{i,n_i}\} \mapsto \{\mathcal{S}_{i,1}, \dots, \mathcal{S}_{i,n_i}\} \mapsto \begin{cases} \mathcal{K}_i, & \text{if sufficient,} \\ \{q_{i,n_i+1}, \dots\}, & \text{otherwise.} \end{cases} \quad (7)$$

268 This recursive expansion allows AWARE to iteratively refine and broaden the evidence pool, ensur-
 269 ing the resulting subspace \mathcal{K}_i captures the full scope of knowledge necessary to resolve the intent.

270 **Memory-Guided Parallel Synthesis.** A second challenge is *scalability*: as diffusion expands, the
 271 number of queries and retrieved documents grows rapidly, making exhaustive analysis prohibitively
 272 costly. To address this, AWARE employs *memory-guided parallel evidence synthesis*, inspired by
 273 the map–reduce paradigm. Each retrieved document D , equipped with its gist memory \mathcal{D} , is first
 274 processed by a filtering operator \mathcal{F} that performs lightweight relevance checks based solely on \mathcal{D} ,
 275 discarding irrelevant candidates without accessing the full text. The surviving documents are then
 276 mapped in parallel by an extraction operator \mathcal{E} into fine-grained evidence units, which are subse-
 277 quently reduced by a synthesis operator \mathcal{R} into the constructed subspace \mathcal{K}_i for intent I_i :

$$\mathcal{K}_i = \mathcal{R} \left(\{\mathcal{E}(D) \mid D \in \mathcal{F}(\{D\}, I_i, \{\mathcal{D}\})\} \right). \quad (8)$$

280 All operators $\mathcal{F}, \mathcal{E}, \mathcal{R}$ are powered by an auxiliary lightweight LLM, enabling AWARE to filter
 281 aggressively, extract in parallel, and synthesize compactly. This design ensures scalability and effi-
 282 ciency while preserving broad evidence coverage with manageable reasoning cost.

283 **Task-Aware Contextualization.** After evidence collection, AWARE organizes the gathered infor-
 284 mation into a structured, task-specific context. Formally, given a task \mathcal{X} , the generated intents $\{I_i\}$
 285 and the constructed subspaces $\{\mathcal{K}_i\}$, the system assembles an organized knowledge chain:

$$\mathcal{C} = \mathcal{X} \cup (I_1 \rightarrow \mathcal{K}_1) \rightarrow (I_2 \rightarrow \mathcal{K}_2) \rightarrow \dots \rightarrow (I_t \rightarrow \mathcal{K}_t), \quad (9)$$

288 which explicitly encodes the reasoning trajectory and its supporting evidence. In essence, \mathcal{C} repre-
 289 sents the *LLM-ready form* of the expected knowledge space \mathcal{K} : compact, structured, and directly
 290 consumable by an LLM. This task-specific context can then be fed into any downstream LLM
 291 to generate the final answer \mathcal{Y} . We refer to this transformation from raw retrieval to structured,
 292 reasoning-ready context as AWARE’s *contextual intelligence*.

293 In summary, AWARE is a retrieval framework that leverages an agentic paradigm to construct min-
 294 imal yet sufficient LLM-ready knowledge representations for diverse tasks. Further implementation
 295 details are provided in Appendix A.1, and Table 2 illustrates the major processes via a case study.

297 4 EXPERIMENTS

300 4.1 DATASETS AND BASELINES.

301 **Datasets.** We evaluate AWARE on three challenging benchmarks for complex information-
 302 seeking. **GAIA** (General AI Assistant) comprises over 450 real-world queries spanning multi-step
 303 reasoning, multimodal understanding, and tool use (Mialon et al., 2023). Following prior work (Li
 304 et al., 2025c; Wu et al., 2025a), we use 103 text-only validation questions. **WebWalkerQA** includes
 305 680 queries across domains such as conferences and organizations, requiring agents to traverse sub-
 306 pages and integrate dispersed evidence, which makes it a long-horizon reasoning challenge (Wu
 307 et al., 2025b). **BrowseComp** consists of 1,266 questions whose answers, although short and verifi-
 308 able, are deliberately hidden beyond top search results (Wei et al., 2025). Because this benchmark
 309 is extremely difficult and often involves hundreds of page visits per query, it imposes substantial
 310 evaluation overhead. We therefore evaluate on two topics, *Art* and *History*, totaling 252 questions.

311 **Baselines.** We compare AWARE against three groups of baselines. (1) *Direct Reasoning*: strong
 312 standalone LLMs used without external tools, including Qwen2.5-32B, Qwen2.5-32B, QwQ-32B,
 313 GPT-4o, Gemini-2.5-Flash and DeepSeek-R1-671B (DeepSeek-AI, 2025; Gemini Team, 2025; Ope-
 314 nAI, 2024). (2) *Retrieval-Augmented Generation*: methods that inject retrieved evidence, such as
 315 vanilla RAG and enhanced variants with query planning or iterative refinement (Shao et al., 2023;
 316 Chan et al., 2024). (3) *Tool-Integrated Reasoning*: approaches that interleave retrieval with rea-
 317 soning, including ReAct, Search-o1, and WebThinker (Yao et al., 2023; Li et al., 2025b;c). Ap-
 318 pendix A.1 provides implementation details for AWARE and baselines.

319 4.2 MAIN RESULTS

321 Table 1 reports the performance of AWARE and baseline. Our key findings are as follows:

322 (1) Under direct reasoning without retrieval, all models handle GAIA tasks more readily, yet their
 323 accuracy remains modest. By contrast, accuracy drops sharply on WebWalkerQA and BrowseComp,

324
 325 Table 1: Main experimental results. Best scores are shown in bold, and second-best are underlined.
 326 Following the official settings, we report Exact Match (EM) for GAIA and BrowseComp, and LLM
 327 Equivalence Accuracy for WebWalkerQA.

Method	General AI Assistant					WebWalkerQA			BrowseComp		
	Level 1	Level 2	Level 3	Avg.	Easy	Medium	Hard	Avg.	Art	History	Avg.
<i>Direct Reasoning (w/o Retrieval)</i>											
Qwen2.5-32B	20.5	9.6	8.3	13.6	3.8	2.5	3.3	3.1	0.0	0.0	0.0
Qwen3-32B	15.4	7.7	0.0	9.7	3.1	1.4	2.5	2.2	0.0	0.0	0.0
QwQ-32B	25.6	9.6	16.7	16.5	7.5	2.1	3.8	4.0	0.0	0.8	0.4
GPT4o	23.1	15.4	8.3	17.5	6.7	6.0	4.2	5.5	0.8	0.8	0.8
Gemini-2.5-Flash	33.3	11.5	0.0	18.5	16.3	7.9	5.8	9.1	0.0	0.0	0.0
DeepSeek-R1-671B	43.6	26.9	8.3	31.1	5.0	11.8	11.3	10.0	0.0	0.0	0.0
<i>Retrieval-Augmented Generation (RAG)</i>											
Vanilla RAG (Qwen2.5-32B)	12.8	11.8	8.3	11.8	23.1	14.3	11.3	15.3	0.0	0.0	0.0
Vanilla RAG (QwQ-32B)	33.3	36.5	8.3	32.0	36.9	26.1	33.5	31.2	2.4	1.6	2.0
Query Planning (Qwen2.5-32B)	30.8	17.3	0.0	20.4	29.4	36.4	25.0	30.7	0.0	0.0	0.0
Query Planning (QwQ-32B)	48.7	25.0	8.3	32.0	28.8	35.7	30.8	32.5	1.6	0.8	1.2
Iterative RAG (Qwen2.5-32B)	35.9	19.2	8.3	24.3	30.6	35.7	25.4	30.9	0.0	0.0	0.0
Iterative RAG (QwQ-32B)	51.3	28.8	8.3	35.0	29.4	32.9	31.3	31.5	0.8	0.0	0.4
<i>Tool-Integrated Reasoning (TIR)</i>											
ReAct (Qwen2.5-32B)	46.1	44.2	8.3	40.7	44.3	46.7	29.2	38.4	0.0	0.0	0.0
ReAct (QwQ-32B)	48.7	34.6	16.7	37.8	35.6	29.1	13.2	24.1	0.8	0.8	0.8
ReAct (GPT4o)	51.2	34.6	8.3	34.6	34.6	42.0	23.9	33.8	2.4	1.6	1.9
Search-o1-32B	53.8	44.2	16.7	39.8	43.1	35.0	27.1	34.1	1.6	2.4	1.9
WebThinker-32B-Base	53.8	44.2	16.7	44.7	47.5	41.1	39.2	41.9	2.4	2.4	2.3
WebThinker-32B-RL	56.4	50.0	16.7	48.5	58.8	44.6	40.4	46.5	2.4	3.1	2.7
AWARE (QwQ-32B)	61.5	<u>46.2</u>	33.3	50.5	<u>53.1</u>	55.0	50.8	53.1	8.7	8.0	8.3

347 confirming that these benchmarks demand recent and long-tail knowledge rarely captured in model
 348 parameters. Interestingly, Qwen3-32B, although more recent, underperforms both Qwen2.5-32B
 349 and QwQ-32B, suggesting that Qwen3’s hybrid reasoning design might compromise efficacy. Based
 350 on these observations, we use Qwen2.5-32B and QwQ-32B as the backbone models for RAG and
 351 tool-integrated reasoning baselines.

352 (2) AWARE consistently outperforms not only vanilla RAG but also advanced variants that incor-
 353 porate query rewriting or iterative refinement, validating the robustness of its retrieval paradigm.
 354 Unlike these pre-inference schemes that often leave evidence fragmented or incomplete, AWARE
 355 employs diffusion-based exploration and memory-guided synthesis to recover layered dependencies
 356 while filtering noise at scale. This design yields substantial gains on tasks that demand multi-hop
 357 reasoning and long-horizon synthesis, where RAG methods struggle to provide coherent context.

358 (3) AWARE surpasses TIR baselines, including workflow-based methods (e.g., Search-o1) and end-
 359 to-end optimized systems (e.g., WebThinker). Although WebThinker benefits from large-scale in-
 360 domain training, AWARE without task-specific optimization still outperforms WebThinker-Base
 361 across all dimensions and achieves dataset-level gains over WebThinker-32B-RL, falling only on
 362 two levels within the dataset hierarchy. This highlights AWARE’s diffusive and parallel exploration,
 363 which enables broader coverage and reduces the risk of missing critical evidence.

364 4.3 ABLATION STUDY

366 AWARE is designed as an integrated retrieval framework, operating as a unified system with inter-
 367 dependent components, making it more meaningful to analyze as a whole rather than in isolation.
 368 Accordingly, our ablation study examines three dimensions: (1) the role of different LLMs
 369 as AWARE’s central reasoning agent, (2) the generalizability of AWARE-generated context across
 370 diverse models, and (3) the dynamics of agentic retrieval, with a focus on diffusion search depth and
 371 evidence synthesis efficiency. Figure 2 summarizes the results, which we discuss below.

373 **Impact of Reasoning LLM Selection.** AWARE relies critically on the capabilities of its central
 374 reasoning model. As shown in Figure 2 (a), it consistently outperforms the tool-integrated reasoning
 375 baseline (Search-o1) across different LLMs, demonstrating the robustness of its design. Nonethe-
 376 less, the strength of the reasoning model plays a decisive role. Reasoning-oriented models such as
 377 QwQ-32B and Qwen3-30B-A3B achieve the best results, clearly surpassing Qwen3-32B, a hybrid
 378 model with diluted reasoning capacity. When paired with Gemini2.5-Flash, AWARE also delivers

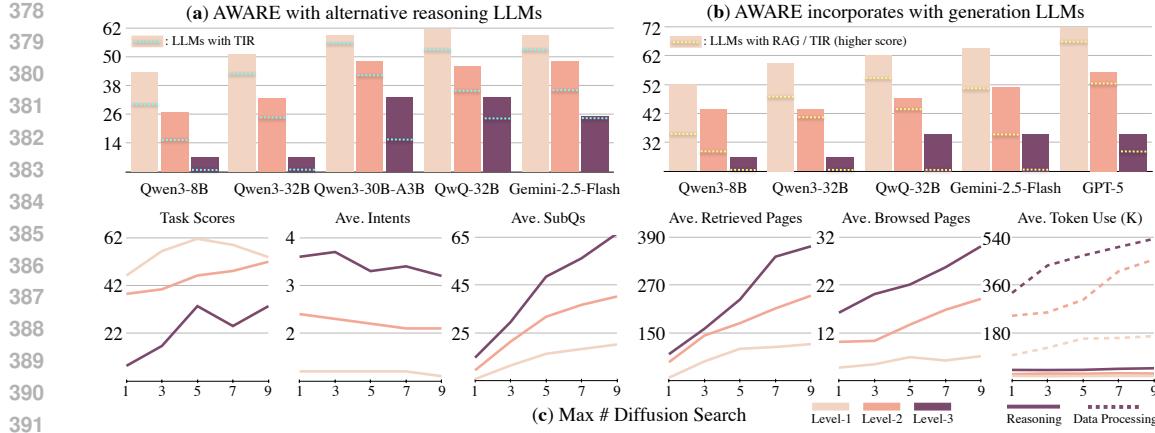


Figure 2: Analysis of AWARE on three perspectives: (a) effect of the central reasoning LLM, comparing AWARE with a TIR baseline (Search-01); (b) transferability of AWARE’s LLM-ready context, compared with RAG and TIR across downstream LLMs for answer generation; and (c) impact of the diffusion-search budget on performance and resulting retrieval dynamics.

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 Diffusion Search Depths. Our analysis in Figure 2 (c) highlights how AWARE’s tailored techniques jointly contribute to effective and scalable retrieval. First, diffusion search proves critical: increasing its depth expands the evidence pool, improving task performance from depth 1 to 5 before fluctuating as information saturates. We also observe that deeper diffusion reduces the number of required search intents, easing the reasoning workload and accelerating convergence to answers. Second, deeper diffusion inevitably increases sub-queries and retrieved pages, especially for complex information-seeking tasks. Here, AWARE’s memory-guided parallel evidence synthesis is validated: it filters out nearly 90% of irrelevant pages using gist memory and processes the remainder in a map-reduce manner, demonstrating strong scalability. Finally, token usage analysis shows that reasoning accounts for a small fraction of total cost compared to large-scale data processing. This validates AWARE’s design of assigning heavy reasoning to strong central models while outsourcing bulk data handling to lightweight auxiliary models, achieving an effective balance between efficiency and performance.

4.4 CASE STUDY

Table 2 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 foundation model’s intrinsic knowledge to establish the target genus, which anchors subsequent information intents. For the first intent, retrieving the precise titles

432
433
434
435
436

Table 2: 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.

437
438

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

439

Initial Reasoning: Identify the scientific genus named for Copenhagen. → **Hafnia**

440

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).

443

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]

449

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.

455

466

of the three papers, AWARE executes diffusive parallel exploration: issuing multiple atomic queries, gathering 36 candidate pages, filtering them with gist-level 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 neatly within the LLM’s context window.

467

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. As a result, AWARE achieves efficient exploration of massive raw text, effective evidence consolidation, and robust downstream reasoning without overloading the model with noise.

471

5 CONCLUSION

473

In this work, we introduced the Agentic Knowledge Warehouse (AWARE), a retrieval paradigm that enriches LLMs with external knowledge in a structured, task-specific form. At the corpus level, AWARE abstracts vast unstructured sources into gist memories, offering global semantic coverage and encoding implicit structural cues often missed by conventional dense retrieval. At the task level, it engages in an agentic reasoning process that decomposes complex queries into layered intents, conducts diffusion-based horizontal exploration with vertical exploitation, and synthesizes the results into a coherent, compact, and LLM-ready context. This design enables AWARE to reconstruct the minimal yet sufficient knowledge space needed to close task-specific knowledge gaps.

482

Extensive experiments on challenging information-seeking benchmarks, complemented by ablation studies across multiple dimensions and a detailed case study, validate the soundness, scalability, and adaptability of AWARE. Taken together, these results show that AWARE constitutes a practical, robust, and broadly applicable solution for enabling contextual intelligence in standalone LLMs. A discussion of limitations and future directions is provided in Appendix A.2.

486 REPRODUCIBILITY STATEMENT
487

488 Our method is implemented using open-source models (Qwen series), open frameworks (Lang-
489 Graph), and publicly available datasets. All codes and prompts are released in an anonymous repos-
490 itory. The repository also contains experiment logs with intermediate outputs, including search
491 intents, atomic queries, browsed web pages, refined evidence, and organized information. These
492 records facilitate reproduction of our results and provide additional case studies that demonstrate
493 the behavior of our approach. We also provide implementation details in Appendix A.1.

494
495 REFERENCES
496

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

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

503 Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-embedding:
504 Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge
505 distillation. In *Findings of the Association for Computational Linguistics ACL 2024*, pp. 2318–
506 2335, 2024.

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

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

512 Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt,
513 and Jonathan Larson. From local to global: A graph rag approach to query-focused summariza-
514 tion, 2024.

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

518 Google Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality,
519 long context, and next generation agentic capabilities, 2025.

520 Xinyu Geng, Peng Xia, Zhen Zhang, Xinyu Wang, Qiuchen Wang, Ruixue Ding, Chenxi Wang,
521 Jialong Wu, Yida Zhao, Kuan Li, et al. Webwatcher: Breaking new frontiers of vision-language
522 deep research agent. *arXiv preprint arXiv:2508.05748*, 2025.

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

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

531 Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-
532 r1: Training llms to reason and leverage search engines with reinforcement learning. *CoRR*,
533 abs/2503.09516, 2025. doi: 10.48550/ARXIV.2503.09516.

540 Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal,
 541 Heinrich Kütller, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented genera-
 542 tion for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:
 543 9459–9474, 2020.

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

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

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

555 Yuchen Li, Hengyi Cai, Rui Kong, Xinran Chen, Jiamin Chen, Jun Yang, Haojie Zhang, Jiayi Li,
 556 Jiayi Wu, Yiqun Chen, et al. Towards ai search paradigm. *arXiv preprint arXiv:2506.17188*,
 557 2025d.

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

561 OpenAI. Gpt-4 technical report, 2024.

563 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
 564 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-
 565 low instructions with human feedback. *Advances in neural information processing systems*, 35:
 566 27730–27744, 2022.

567 Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. Lewis, Majid Yazdani, Nicola De
 568 Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim
 569 Rocktäschel, and Sebastian Riedel. KILT: a benchmark for knowledge intensive language tasks. In
 570 Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven
 571 Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou (eds.), *NAACL-HLT 2021, On-
 572 line, June 6-11, 2021*, pp. 2523–2544. Association for Computational Linguistics, 2021.

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

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

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

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

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

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

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

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

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

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

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

608 Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. Retrieval aug-
 609 mented generation (rag) and beyond: A comprehensive survey on how to make your llms use
 610 external data more wisely, 2024a.

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

615 Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongan Liu, Wenhan Liu, Chenlong Deng, Haonan
 616 Chen, Zhicheng Dou, and Ji-Rong Wen. Large language models for information retrieval: A
 617 survey, 2024.

618 A APPENDIX

619 A.1 IMPLEMENTATION DETAILS

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

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

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

633 During the *data indexing process*, we employ BGE-M3 as the dense embedding model (Chen et al.,
 634 2024), complemented by a BM25 index constructed over the full web content. All retrieval oper-
 635 ations are instantiated using ElasticSearch, which provides a stable and scalable infrastructure for
 636 large-scale search. For online retrieval, we rely on Google’s Custom Search JSON API to identify
 637 relevant pages, and utilize Jina AI’s Web Reader to extract full web content. For all baselines, we
 638 either report results directly from their original papers or reproduce them using official imple-
 639 ments. All experiments are conducted on a node of eight NVIDIA A100-40G GPUs.

640 To ensure transparency and reproducibility, we release all prompts used in AWARE along with full
 641 experiment logs, including intermediate artifacts such as search intents, atomic queries, retrieved

648 and browsed pages, refined evidence, and organized knowledge. These materials are available in
 649 *this anonymous repository*.
 650

651 **A.2 LIMITATIONS AND FUTURE DIRECTIONS**
 652

653 Although AWARE demonstrates strong performance across diverse benchmarks, several limitations
 654 of this work should be acknowledged.
 655

656 **Method Scope.** AWARE is proposed as a general retrieval paradigm that operates independently
 657 of specific models and can integrate seamlessly with both open-source and closed-source LLMs.
 658 However, unlike data-driven approaches that train task-specific models, AWARE does not incorpo-
 659 rate optimization strategies tailored to particular domains. This limitation arises from objective con-
 660 straints. The agentic framework of AWARE involves multiple capabilities such as planning, intent
 661 generation, evidence refinement, and synthesis, all of which would require carefully annotated or
 662 synthetically generated data for supervised optimization. While reinforcement learning could serve
 663 as an end-to-end alternative, producing large-scale, high-quality training data and running optimiza-
 664 tion for large models would demand significant resources. For example, reinforcement learning on
 665 a 32B model reasonably requires at least 32 H100 80G GPUs, which remain beyond reach.
 666

667 Despite this, we argue that AWARE can continually benefit from improvements in general-purpose
 668 LLMs. The very skills required within AWARE, including reasoning, planning, and synthesis, fall
 669 within the optimization scope of mainstream models. Thus, even without explicit task-specific fine-
 670 tuning, AWARE achieves strong results. Moreover, excessive specialization on narrow domains may
 671 harm the generality of large models, introducing overfitting risks and reducing adaptability.
 672

673 **Experimental Scope.** Given that AWARE is currently designed for text-only scenarios, we follow
 674 prior work and exclude the non-textual portion of GAIA, which prevents us from evaluating on the
 675 full benchmark. For BrowseComp, the evaluation is further constrained by the high cost of API
 676 usage. Even when restricted to two topics, *Art* and *History*, constructing the corresponding web
 677 page collection required crawling nearly 50,000 pages and issuing close to 10,000 Google queries.
 678 Across all three benchmarks in this paper, the combined use of the Google Search API and the
 679 Jina web page crawling API has already incurred costs exceeding 1,200 USD, placing considerable
 680 pressure on the experimental budget and limiting our ability to scale the evaluation further. Future
 681 work may explore more cost-efficient pipelines to enable broader coverage of these benchmarks.
 682

683 **Baseline Coverage.** We strive to ensure that baselines in our main experiments are comparable in
 684 model size, open-sourcing status, and implementation feasibility. Nonetheless, it is not possible to
 685 evaluate against all related baselines. Some rely on substantially different model sizes, others are
 686 not fully released, and some require resources that are unavailable in our setting.
 687

688 **Future Directions.** These limitations primarily reflect constraints in data availability, computa-
 689 tional resources, and experimental scope, rather than methodological shortcomings. Addressing
 690 them naturally opens several avenues for future work. One direction is to explore lightweight op-
 691 timization strategies, such as reinforcement learning with synthetic data, to adapt AWARE more
 692 closely to domain-specific tasks. Another is to extend AWARE beyond the text-only setting toward
 693 multimodal benchmarks, where information is distributed across heterogeneous modalities. Finally,
 694 exploring more precise strategies for constructing web page collections would reduce the crawling
 695 of irrelevant pages, thereby lowering overall API costs and enabling more efficient evaluation.
 696

697 **B AI USAGE DISCLOSURE**
 698

699 In this work, AI assistants were used exclusively for polishing the manuscript, including grammar
 700 checking and language refinement. The initial draft was prepared manually by the authors, and only
 701 selected sections were refined with AI assistance.
 702

703 AI assistants did not contribute to any other part of the research, including ideation, literature review,
 704 or figure preparation.
 705