

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 DEMYSTIFYING AND ENHANCING THE EFFICIENCY OF LARGE LANGUAGE MODEL BASED SEARCH AGENTS

Anonymous authors

Paper under double-blind review

ABSTRACT

Large Language Model (LLM)-based search agents have shown remarkable capabilities in solving complex tasks by dynamically decomposing problems and addressing them through interleaved reasoning and retrieval. However, this interleaved paradigm introduces substantial efficiency bottlenecks. First, we observe that both highly accurate and overly approximate retrieval methods degrade system efficiency: exact search incurs significant retrieval overhead, while coarse retrieval requires additional reasoning steps during generation. Second, we identify inefficiencies in system design, including improper scheduling and frequent retrieval-induced stalls, which lead to cascading latency—where even minor delays in retrieval amplify end-to-end inference time. To address these challenges, we introduce *SearchAgent-X*, a high-efficiency inference framework for LLM-based search agents. *SearchAgent-X* leverages high-recall approximate retrieval and incorporates two key techniques: priority-aware scheduling and non-stall retrieval. Extensive experiments demonstrate that *SearchAgent-X* consistently outperforms state-of-the-art systems such as vLLM and HNSW-based retrieval across diverse tasks, achieving up to $3.4\times$ higher throughput and $5\times$ lower latency, without compromising generation quality. Code is available at <https://anonymous.4open.science/r/SearchAgent-X>.

1 INTRODUCTION

Traditional Retrieval-Augmented Generation (RAG) typically uses a sequential retrieve-then-generate approach (Fan et al., 2024; Gao et al., 2024; Gupta et al., 2024; Huang & Huang, 2024; Wu et al., 2024; Yu et al., 2024; Zhao et al., 2024a;b), which limits dynamic interaction with knowledge bases. Recent advancements have ushered in RAG 2.0, known as *Search Agents* (Trivedi et al., 2022; Singh et al., 2025; Li et al., 2025; Jin et al., 2025a; Chen et al., 2025; Song et al., 2025; xAI, 2025; OpenAI, 2025). This paradigm leverages the strong reasoning capabilities of Large Language Models (LLMs), allowing for the dynamic and adaptive interleaving of reasoning steps with retrieval calls throughout the generation process. Instead of a fixed pipeline, search agents can decide *when* and *what* to retrieve based on LLM’s ongoing reasoning, leading to significant improvements in the quality and depth of the generated responses. Leveraging post-training techniques similar to DeepSeek-R1, some pioneering models can even autonomously initiate retrieval actions during reasoning without intermediate supervision (Jin et al., 2025a; Chen et al., 2025; Song et al., 2025).

However, the improved generation quality achieved by search agents often comes at the cost of efficiency—an overhead that is nontrivial in practical deployments. In reasoning-with-search scenarios, achieving low-latency responses is critical for ensuring a seamless user experience (Ray et al., 2024; Jin et al., 2024). Moreover, during post-training of LLM-based search agents, efficient model rollouts over large-scale training corpora are essential to support scalable learning. While recent systems incorporate advanced inference optimizations—such as sequence concatenation (Jin et al., 2025a; Chen et al., 2025; Song et al., 2025) and prefix caching (Jin et al., 2024; Kwon et al., 2023; Zheng et al., 2024)—these techniques are not specifically designed to address the unique computational challenges posed by the tight interleaving of multi-step reasoning and dynamic retrieval.

To this end, we first conduct a systematic analysis of the efficiency factors governing LLM-based search agents, uncovering insights that diverge from the understanding of naive RAG. Our in-depth analysis reveals two key observations: First, we demonstrate a non-monotonic relationship

054 between retrieval accuracy and end-to-end efficiency. Both excessively high (e.g., exact search) and
 055 excessively low retrieval recall degrade overall efficiency. While aiming for perfect recall incurs
 056 unnecessary computational overhead in the retrieval phase, low recall necessitates more retrieval
 057 iterations and longer reasoning paths by the LLM to compensate (as shown in Figure 1). This
 058 highlights that search agent systems benefit from high-recall approximate search that effectively
 059 supports reasoning without unnecessary retrieval costs. Second, we find that search agent systems
 060 are highly sensitive to retrieval latency. Unlike naive RAG where retrieval is largely amortized, even
 061 minor increases in retrieval time in the search agent system can cause a disproportionately large
 062 increase in end-to-end latency (Figure 2a). We attribute this magnification effect to two primary
 063 root causes: *improper scheduling*, where standard policies like FCFS fail to prioritize requests that
 064 would benefit most from KV-cache reuse (Figure 2b), and *retrieval-induced stalls*, where timing
 065 misalignments between asynchronous retrieval and token generation force requests to wait, leading to
 066 unnecessary recomputation (Figure 2c).

067 Motivated by these findings, we propose **SearchAgent-X**, an inference system dedicated for efficient
 068 search agents. **SearchAgent-X** is designed to optimize end-to-end system throughput and latency
 069 by smoothly coordinating the interleaving of self-reasoning and retrieval. Since both overly low and
 070 high retrieval efforts lead to degraded efficiency, **SearchAgent-X** chooses to build upon a high-recall
 071 approximate retrieval method. To tackle the problem of improper scheduling, **SearchAgent-X**
 072 schedules requests with priority awareness through their real-time status to enhance KV-cache
 073 utilization. Moreover, in order to overcome frequent retrieval-induced stalls, **SearchAgent-X**
 074 proposes a non-stall retrieval mechanism through an adaptive strategy that allows generation to
 075 proceed without unnecessary waiting while ensuring sufficient retrieval quality.

076 Our extensive experiments demonstrate that **SearchAgent-X** consistently and significantly outper-
 077 forms state-of-the-art baseline systems across various operational settings. In both offline and online
 078 inference scenarios, **SearchAgent-X** achieves substantial improvements in system performance (e.g.,
 079 $1.3\text{-}3.4\times$ higher throughput) by improving LLM KV-cache utilization (from 0.07 to 0.65), all while
 080 maintaining the high generation quality characteristic of search agents with exact retrieval.

081 2 BACKGROUND AND MOTIVATION

082 2.1 PRELIMINARY: LLM-BASED SEARCH AGENT SYSTEMS

083 LLM-based search agent systems are designed to tackle complex requests by decomposing problems
 084 into a series of interleaved, multi-turn reasoning and information retrieval steps. This allows the LLM
 085 to adaptively seek and integrate external knowledge throughout its reasoning process. Appendix A
 086 shows an example of the process of a LLM-based search agent.

087 **Supporting Multi-Turn Reasoning.** Search agent systems often build on LLM inference frameworks
 088 like vLLM (Jin et al., 2025a). They use *Sequence Concatenation* for dynamic retrieval: during
 089 inference, the system monitors model output for retrieval signals. Upon such a signal, LLM decoding
 090 pauses, a query is issued, and retrieved results are concatenated with previously generated tokens to
 091 form a new, extended *Sequence*. This is then re-injected into the LLM to resume reasoning.

092 To enhance efficiency, *Prefix Cache* is commonly leveraged (Jin et al., 2024; Zheng et al., 2024).
 093 This technique stores key-value (KV) pairs from the LLM’s attention mechanism for prior tokens,
 094 allowing efficient reuse in subsequent generations. This is particularly advantageous in search agents,
 095 as the concatenated sequence’s prefix, excluding newly retrieved tokens, overlaps with the previous
 096 generation. Furthermore, shared system prompts across search agent requests can be cached and
 097 reused. In our evaluation, enabling prefix caching saved over 24% of token recomputation costs.

098 **Sequence Scheduling.** Efficient scheduling is vital for high throughput. Modern LLM inference
 099 frameworks utilize *Iteration-Level Scheduling*, where GPU scheduling decisions occur at the gran-
 100 ularity of the single token generation step (Kwon et al., 2023; Zheng et al., 2024). Compared to
 101 sequence-level scheduling (NVIDIA, 2019; 2021), iteration-level scheduling avoids waiting for all
 102 sequences in a batch to complete, thus preventing bubble problems and becoming a leading solution.
 103 Frameworks like vLLM typically employ a First-Come-First-Serve (FCFS) scheduling policy.

104 **Retrieval Mechanism.** On the retrieval side, semantic search techniques efficiently locate relevant
 105 external knowledge. Queries are usually encoded into dense vector representations for searching

108 in vector space. The two primary approaches are exact nearest neighbor (ENN) search (Dasgupta
 109 & Sinha, 2013) and approximate nearest neighbor (ANN) search (Malkov & Yashunin, 2018; Guo
 110 et al., 2020). Graph-based ANN methods like HNSW (Malkov & Yashunin, 2018) offer a favorable
 111 speed/accuracy trade-off, making them suitable for large knowledge bases.
 112

113 2.2 KEY INSIGHTS: FACTORS GOVERNING EFFICIENCY

115 Despite significant progress in high-performance LLM inference and retrieval, the LLM-based search
 116 agent’s efficiency remains poorly understood. In this section, we analyze the influence of two key
 117 factors: 1) retrieval accuracy and 2) retrieval latency, and examine how they contribute to severe
 118 inefficiencies in current solutions. For retrieval, we assume a local search with a *fixed dense encoder*.
 119 **Experiments in this section use Search-R1 7B on MuSiQue and the default vLLM configuration.**

120 2.2.1 IMPACT OF RETRIEVAL ACCURACY

122 **Insight 1:** *Both overly high and overly low retrieval recall degrade end-to-end efficiency. High
 123 recall increases retrieval overhead, while low recall leads to longer reasoning steps.*

125 We first investigate the impact of different retrieval accuracies on the system efficiency of search agents. In-
 126 tuitively, lower retrieval accuracy means lower retrieval
 127 overhead, thus higher system efficiency. However, we
 128 observe a “less is more” phenomenon for LLM-based
 129 search agents. Low-recall retrievals may result in sub-
 130 optimal context, forcing the model to compensate by is-
 131 suing additional retrievals and extending the reasoning
 132 length. Figure 1 shows how varying the ANN search
 133 range affects throughput and average retrieval counts.
 134 When the search range is too small (e.g., 10), the model
 135 fails to retrieve useful documents, resulting in longer
 136 reasoning steps and an average of 6.5 retrievals per re-
 137 quest. This reduces throughput to just above 2.1. As
 138 the search range increases to 500, retrieval quality im-
 139 proves, and the model completes reasoning with fewer
 140 retrievals (around 5.7), boosting throughput to over 3.2.

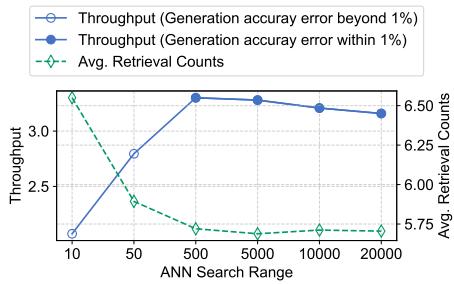
141 However, further increasing the search range (e.g., be-
 142 yond 10,000) yields diminishing returns. While average
 143 retrieval counts decrease slightly, throughput declines
 144 due to the higher cost of very high-recall ANN searches.
 145 This suggests that simply maximizing retrieval recall
 146 is not the optimal strategy for search agent efficiency. Once retrieval quality sufficiently supports
 147 reasoning, additional search efforts offer marginal benefits and can even harm overall efficiency.

148 2.2.2 IMPACT OF RETRIEVAL LATENCY

150 **Insight 2:** *Compared to naive RAG, search agents are much more sensitive to retrieval latency due
 151 to ignoring inter-request priorities and retrieval-induced stalls.*

152 For naive RAG, all requests are retrieved before generation. Retrieval latency (millisecond level) is
 153 negligible compared to the total request latency (second level), so it is insensitive to retrieval latency.
 154 However, for search agents, retrieval occurs during self-reasoning, where the time scale of a single
 155 token generation and retrieval latency are comparable. Minor retrieval latency can cause requests to
 156 be inserted into different token generation iterations, leading to different system behaviors.

157 Figure 2 shows the average end-to-end latency of search agents and RAG under different retrieval
 158 latency (controlled by different search ranges), with a request rate of 5 requests/second and a test
 159 duration of 10 minutes. For fair comparison, we assume RAG generates the same length of tokens
 160 with search agent, and its end-to-end latency t^{e2e} is calculated as $t_0^{e2e} + \bar{t}_{ret} \cdot \bar{n}_{ret}$, with t_0^{e2e} as the
 161 token generation time without retrieval, \bar{t}_{ret} as the average retrieval time, and \bar{n}_{ret} as the average
 retrieval counts per request. The results indicate that search agents suffer from drastic efficiency



152 **Figure 1: Impact of Retrieval Accuracy on Search Agent Efficiency.** Higher ANN
 153 search range means higher-recall retrieval. Throughput marks the number of requests
 154 completed per second (higher is better). Retrieval count indicates the number of re-
 155 trievals called per request. End-to-end generation accuracy error is calculated by com-
 156 parison with an exact retrieval method.

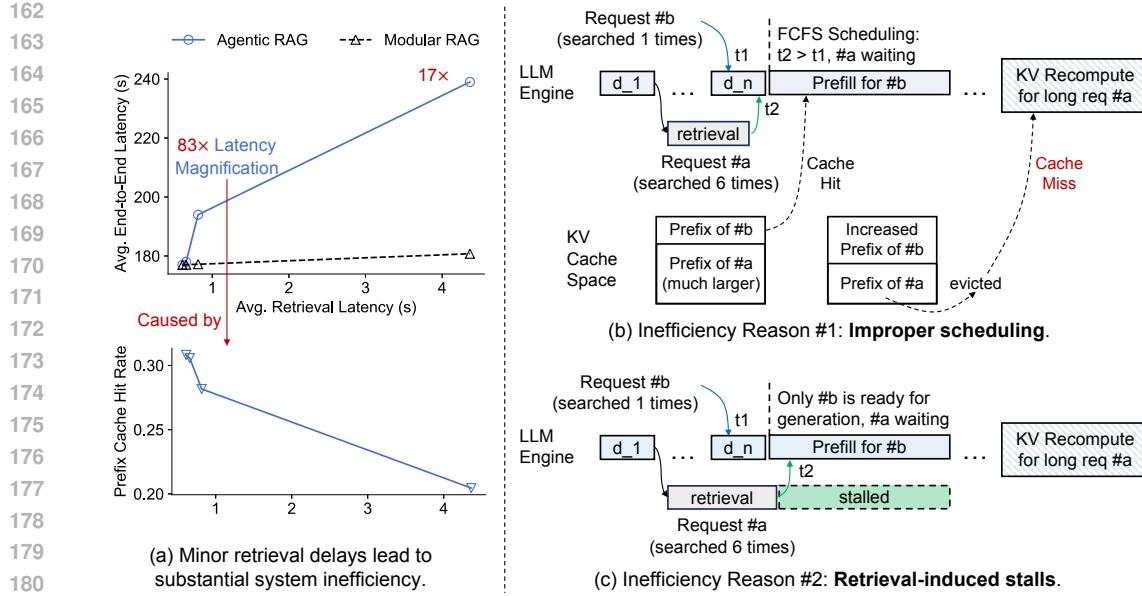


Figure 2: **Impact of Retrieval Latency on Search Agent Efficiency.** (a) Search agents exhibit significantly higher retrieval latency sensitivity than naive RAG (up to $83\times$ magnification), linked to lower prefix KV-cache hit rates. Here, the latency magnification (e.g., $83\times$ at search range 10k) is computed as the ratio between the increase in agentic RAG end-to-end latency and the corresponding increase in retrieval latency, both measured relative to the 500-range modular RAG baseline. (b, c) Root causes include: (b) improper scheduling, where serving shorter requests first evicts cache for longer ones, causing recomputation; and (c) retrieval-induced stalls, where requests missing a scheduling point must wait, risking cache eviction.

degradation under even minor retrieval delays. As average retrieval latency increases from 0.6s to 4.4s, the end-to-end latency of the search agent is magnified by over $83\times$, while RAG remains largely stable. This severe magnification in search agent is strongly correlated with a sharp decrease in the prefix KV-cache hit rate, dropping from over 30% to under 21%, which forces frequent and costly KV recomputations (Figure 2a). Note that in Figure 2a, each point corresponds to a different setting of ANN parameters (i.e., different `efSearch` values).

We identify two root causes for this observed behavior, both contributing to unnecessary KV recomputation, particularly for longer, multi-turn requests: **improper scheduling** and **retrieval-induced stalls**. Figure 2b illustrates the issue of improper scheduling. Consider request #a, which involves a longer reasoning path with 6 retrievals, and request #b, which just completes a single retrieval. Even if request #a arrives first, if its retrieval completes slightly later than that of #b ($t_2 > t_1$), a standard FCFS scheduler may choose to serve #b first in the next iteration. As #b proceeds with its generation, it occupies valuable KV-cache space, potentially leading to the eviction of the prefix KV-cache belonging to #a. When request #a eventually resumes, it encounters a *cache miss* and must recompute its entire prefix from scratch, significantly increasing its latency. Our measurements highlight the high cost of such improper scheduling: 55.9% of tokens were unnecessarily recomputed in affected cases, leading to more than a 108% increase in computation time per request.

Even with improved scheduling, another significant inefficiency risk comes from reasoning stalls, depicted in Figure 2c. The asynchronous execution of retrieval and generation can lead to subtle timing misalignments. If a long request like #a completes its retrieval only slightly after the deadline for inclusion in the next generation step, it misses the current scheduling batch and is forced to wait until the subsequent one. We term this unproductive waiting period "retrieval-induced stalls." During this stall, shorter requests (e.g., #b) that are ready can continue executing. Their execution may further displace #a's prefix from the KV-cache, once again resulting in costly recomputation upon #a's eventual resumption. Our data shows that, on average, more than 25% of sequences experience such stalls after completing their retrieval across various scenarios.

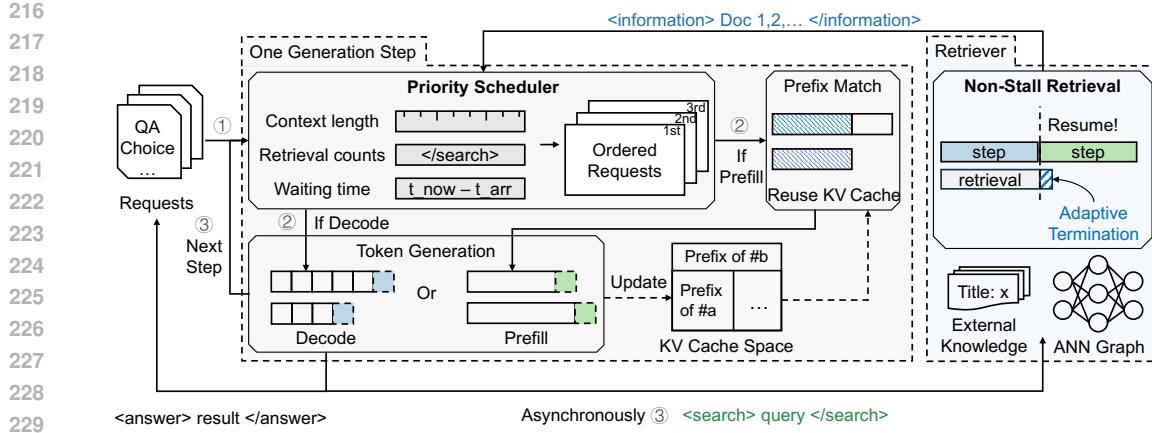


Figure 3: **SearchAgent-X’s Architecture.** Requests are scheduled with priorities. Reasoning and retrieval are interleaved, with a non-stall retrieval mechanism to avoid unnecessary waiting.

Limitations of Existing Solutions. Our analysis highlights key limitations in current search agent systems. ENN retrieval, despite full recall, incurs prohibitive retrieval overhead. While high-recall ANN search is more suitable, it suffers from retrieval-induced stalls due to asynchronous execution. Furthermore, prevalent FCFS scheduling in LLM inference frameworks (Kwon et al., 2023; Abhyankar et al., 2024) disregards the search agent’s unique request priorities, leading to suboptimal cache utilization and costly recomputation.

3 DESIGN OF SEARCHAGENT-X

3.1 OVERALL ARCHITECTURE

Drawing upon the above insights, we propose SearchAgent-X, a simple yet efficient inference system that is easy to deploy, explicitly designed to optimize end-to-end efficiency for search agent workloads by smoothly interleaving self-reasoning and retrieval. Figure 3 shows SearchAgent-X’s architecture, a tightly integrated system processing search agent requests at the token generation level. At each LLM output step, the system checks for special tags that trigger the Retriever for an ANN-based search (e.g., <search>) or request completion (e.g., <answer>), respectively.

To optimize GPU resource usage, SearchAgent-X incorporates a **priority scheduler**. It dynamically prioritizes concurrent requests using real-time collected metrics like retrieval count and waiting time, aiming to enhance KV-cache reuse by processing higher-priority requests first. During prefill, prefix matching reuses existing KV pairs from cache, significantly reducing computational overhead; new KV states are computed if caching is inapplicable or a miss occurs. Retrieval and generation operate asynchronously to enhance throughput. When retrieval is triggered, the system queries a pre-built ANN graph index. To proactively avoid retrieval-induced stalls, SearchAgent-X employs **non-stall retrieval** with adaptive search termination, allowing generation to proceed without unnecessary waiting while ensuring sufficient retrieval quality.

3.2 PRIORITY SCHEDULING

SearchAgent-X employs a priority-based scheduling mechanism to efficiently and fairly manage concurrent generation requests. As introduced earlier, each search agent request i involves a list of generation sequences $[s_{i,0}, s_{i,1}, \dots, s_{i,r_i}]$, where $s_{i,0}$ is the initial sequence and $s_{i,j}$ ($j > 0$) represents a sequence resumed after the j -th retrieval. Let r_i denote the current number of retrievals performed for request i , and s_{i,r_i} be the sequence currently being processed.

As discussed earlier, requests that have undergone more retrieval steps (i.e., higher r_i) benefit more significantly from prefix cache reuse due to longer shared prefixes. Prioritizing such requests can therefore enhance overall cache efficiency and reduce redundant computation. However, scheduling

270 solely based on retrieval count risks starving requests with fewer or no retrievals, leading to increased
 271 end-to-end latency and reduced fairness.
 272

273 To mitigate these issues, **SearchAgent-X** utilizes a hierarchical scheduler that dynamically prioritizes
 274 requests based on a combination of three key metrics associated with request i : (1) the number of
 275 retrievals completed $R_i = r_i$; (2) the context length of the current sequence $C_i = L_{\text{seq},i}$; and (3)
 276 the waiting time of the initial request $W_i = t_{\text{now}} - t_{\text{arr},i}$. The first two metrics implicitly prioritize
 277 sequences with longer reusable prefixes, while the last ensures fairness by giving preference to
 278 requests that have been waiting longer overall.
 279

280 Instead of combining these diverse metrics into a single weighted score, which would require tedious
 281 and potentially task-specific tuning of weights, **SearchAgent-X** discretizes each metric into G
 282 distinct priority levels. For a given metric $M \in \{R, W, C\}$, the threshold defining the lower bound
 283 for level k is calculated as:
 284

$$T_{M,k} = \min(M) + \frac{k}{G} \cdot (\max(M) - \min(M)), \quad 0 \leq k < G \quad (1)$$

285 A request i is then assigned to the highest priority level k for which at least one of its metric values
 286 (R_i, W_i, C_i) exceeds the corresponding threshold $T_{M,k}$:
 287

$$k = \max \{j \in [0, G-1] \mid R_i > T_{R,j} \vee W_i > T_{W,j} \vee C_i > T_{C,j}\} \quad (2)$$

288 Requests that do not meet any threshold are assigned to the base level 0.
 289

290 Within each assigned priority level, active sequences are further sorted according to their current
 291 queueing time, defined as $W_i^{\text{cur}} = t_{\text{now}} - t_{\text{arr},i}^{(r_i)}$, where $t_{\text{arr},i}^{(r_i)}$ is the time when the sequence s_{i,r_i}
 292 becomes ready for processing (e.g., after retrieval completes). Sorting by W_i^{cur} in descending order
 293 ensures that among requests of similar priority level, those that have been waiting longest for their
 294 current step are processed first, mitigating the risk of KV-cache eviction during extended waits.
 295

296 Finally, **SearchAgent-X** determines the execution order by traversing the priority levels from highest
 297 to lowest and processing the sequences within each level based on their sorted W_i^{cur} .
 298

3.3 NON-STALL RETRIEVAL

300 To mitigate inefficiencies from retrieval latency and prevent retrieval-induced stalls (Section 2.2.2),
 301 **SearchAgent-X** incorporates a flexible, non-stall early termination strategy for Approximate Nearest
 302 Neighbor (ANN) search. Unlike traditional ANN search that iteratively refines candidates until
 303 meeting pre-set criteria (e.g., explored nodes, list stability) and can thus cause pipeline stalls if
 304 retrieval is slow, **SearchAgent-X** adaptively concludes the search. This adaptive termination is
 305 based on two key conditions: the maturity of retrieval results and the readiness of the LLM engine,
 306 ensuring generation proceeds without unnecessary blocking.
 307

308 At the core of this strategy is the concept of a *soft limit* for the retrieval process. This soft limit represents
 309 a checkpoint where search results are likely to have achieved sufficient quality for the generation
 310 task. **SearchAgent-X** estimates retrieval maturity by monitoring returns in quality improvement
 311 during the ANN search. While retrieval quality generally improves with more explored neighbors,
 312 we find that the rate of improvement diminishes significantly after a certain point, exhibiting a
 313 "knee" where newly found points contribute less to the overall quality. **SearchAgent-X** exploits this
 314 observation. A normalized metric RQ_t is used to evaluate the quality of newly discovered candidates
 315 at step t , defined as: $(d_t - d_{\text{best}})/(d_{\text{worst}} - d_{\text{best}})$, where d_t is the new candidate's distance to the
 316 query, while d_{best} and d_{worst} are the distances of the best and worst candidates currently in ANN
 317 algorithm's list. A high RQ_t value suggests the new candidate offers little improvement over existing
 318 ones, indicating diminishing returns from further search (see details in Appendix B.2).
 319

320 The maturity exit criterion is met when this smoothed quality signal (derived from RQ_t) indicates
 321 a plateau (i.e., exceeds a threshold τ) *and* the LLM engine is ready for its next token generation
 322 operation. Upon meeting both conditions, **SearchAgent-X** halts the retrieval and provides the current,
 323 sufficiently mature results to the LLM; otherwise, retrievals stop naturally. This adaptive alignment of
 324 asynchronous retrieval and generation significantly reduces end-to-end latency without compromising
 325 the quality of the retrieved context, contrasting with traditional fixed stopping criteria that may not
 326 optimally synchronize with the dynamic state of the generation pipeline. **SearchAgent-X**'s complete
 327 execution process and implementation details can be found in Appendix B.
 328

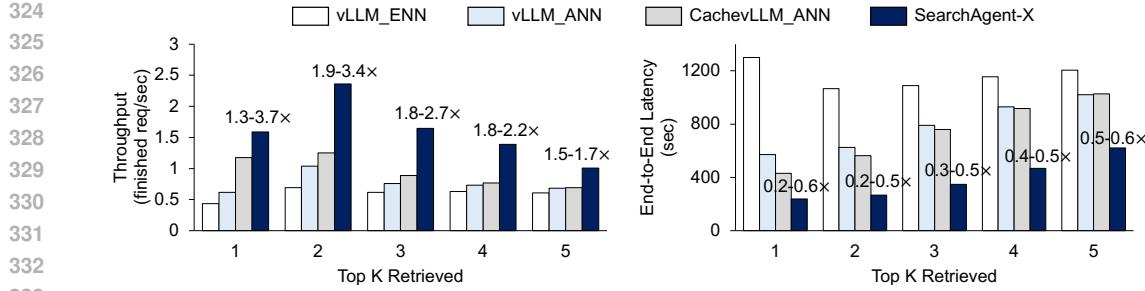


Figure 4: **End-to-End Efficiency of Offline Inference.** Left: Requests completed per second (higher is better). Right: Average end-to-end latency (lower is better).

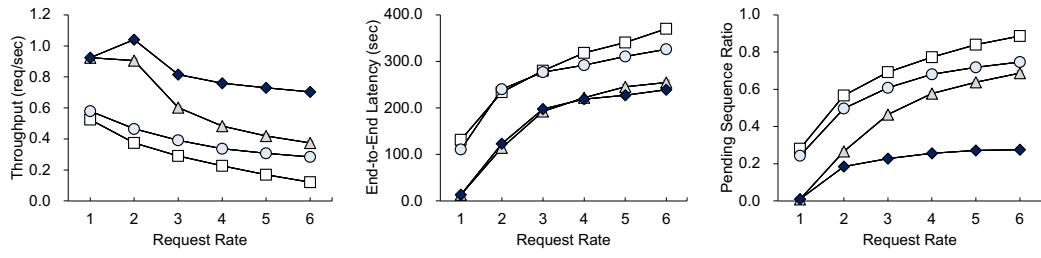


Figure 5: **End-to-End Efficiency of Online Inference.** Left: Throughput. Middle: Latency. Right: Pending Sequence Ratio, the percentage of sequences initiated but not completed within the test period. Lower is better, indicating reasonable workload scheduling.

4 EVALUATION

4.1 EXPERIMENTAL SETUP

Models and Datasets. We evaluate our method on two different search agent models from Search-R1 (Jin et al., 2025a) and ReCall (Chen et al., 2025). For retrieval, we adopt a chunked Wikipedia dataset as the knowledge base, using an ANN index constructed with HNSWlib (Malkov & Yashunin). Note that our approach is model-agnostic and readily generalizes to other reasoning models/ANN methods.

Testbed. For the 7B model, we use a single NVIDIA L20 GPU with 48GB memory. For the 14B model, we use two A100 GPUs with 40GB memory each, connected via PCIe 3.0. The retrieval system runs on 22 CPU cores with 120GB of RAM.

Baselines. We compare the performance of four methods: 1) vLLM_ENN: the vanilla vLLM with exact retrieval. 2) vLLM_ANN: vanilla vLLM system (Kwon et al., 2023) with approximate retrieval. 3) CachevLLM_ANN: vanilla vLLM with approximate retrieval and prefix cache. 4) SearchAgent-X: our proposed system. Refer to Appendix B.3 for detailed setup.

4.2 END-TO-END PERFORMANCE

We first evaluate the end-to-end performance of different methods. For efficiency measurement, we use Musique (Jin et al., 2025b) and HotpotQA (Yang et al., 2018), two datasets of complex multi-hop queries. Two scenarios are tested: (1) offline inference, where all requests arrive at the start; and (2) online inference, where requests arrive at a fixed rate. In the offline setting, we process 1000 requests and measure efficiency upon completion. In the online setting, requests arrive at rates from 1 to 6 over a 10-minute window. **Results for the 7B Search-R1 model are shown in Figures 4 and 5; full results across all metrics and models are in Appendix C.1 and C.2.**

In offline scenarios, SearchAgent-X consistently outperforms all baselines in terms of system throughput and per-request latency. As shown in Figure 4, SearchAgent-X achieves 1.3-3.4 \times higher throughput and only 0.2-0.6 \times the latency compared to the baselines across different top- k values. Even in the most challenging case of top- k =5, SearchAgent-X still beats the best baseline

378

379

Table 1: Generation Quality of SearchAgent-X and Exact Retrieval.

380

381

Dataset	Musique	NQ	2Wiki	HotpotQA	Bamboogle	StrategyQA	Avg.
Generation Accuracy							
Exact Retrieval	0.203	0.316	0.371	0.429	0.472	0.788	0.430
SearchAgent-X	0.203	0.320	0.370	0.428	0.472	0.789	0.430

385

CachevLLM_ANN, with a significant margin ($1.5\times$ in throughput and $0.6\times$ in latency). We attribute this improvement to the SearchAgent-X’s high-recall ANN, and the design of efficient scheduling and non-stall retrieval mechanisms. vLLM_ENN performs poorly in this scenario, as it incurs excessive retrieval overhead and hinders end-to-end reasoning efficiency. vLLM_ANN employs a high-recall ANN and performs obviously better than vLLM_ENN, but it still suffers from the inefficiencies of large amounts of recomputation due to the lack of prefix cache. CachevLLM_ANN uses prefix cache to reduce recomputation, but it still fails to wisely manage the scheduling of requests and avoid retrieval-induced stalls, leading to a significant performance gap compared to SearchAgent-X.

393

We also find that the performance of all methods first increases then decreases with the increase of top- k values. This aligns well with our previous observation that both overly high and overly low retrieval quality degrade end-to-end efficiency. When the top- k value is too small, the model may fail to retrieve useful documents, leading to longer reasoning sequences and lower throughput. Conversely, when the top- k value is too large, the concatenated sequence becomes too long, resulting in longer prefill time. However, we note that SearchAgent-X consistently outperforms all baselines across all top- k values, indicating its robustness to different retrieval settings.

400

401

We also observe that SearchAgent-X often yields larger gains on harder tasks. As shown in Table 3 in Appendix C.1, the HotpotQA dataset (a easier dataset according to the Search-R1 analysis) yields lower benefits compared to the Musique dataset in our main experiments (e.g., 1.52x of throughput improvement on easy dataset vs. 1.84x of throughput improvement on hard dataset, for top- k =3).

404

In online scenarios, SearchAgent-X utilizes computing resources more efficiently than baselines, completing more requests in the same amount of time. As shown in Figure 5, SearchAgent-X completes at least $1.5\times$, and up to $3.5\times$ more requests on average than the baseline, within the request rate range of 1 to 6. Further, we record the pending sequence ratio that measures the resource utility of the system, defined as the percentage of sequences that are initiated but not completed within the test period. As shown in Figure 5 (right), SearchAgent-X achieves stable, small pending sequence ratios (about 0.2), while the baselines experiences dramatic increases with higher request rates (more than 0.6), indicating the effectiveness of SearchAgent-X’s scheduling.

412

SearchAgent-X achieves comparable generation quality to exact retrieval. We evaluate the generation quality of SearchAgent-X and exact retrieval (vLLM_ENN) on six representative datasets. We use the *Exact Match* metric as generation accuracy to measure the correctness of the generated answers (Jin et al., 2025b). As shown in Table 1, SearchAgent-X achieves similar generation accuracy, retrieval counts and output length (see Appendix C.2 for details) as exact retrieval across all datasets, indicating that it does not compromise generation accuracy for efficiency. Another interesting finding is that SearchAgent-X may even achieve higher generation accuracy on some datasets, such as NQ (0.320 vs. 0.316). The results could be attributed to two aspects. First, full ANN recall does not necessarily mean optimal generation accuracy; the correct answer document may not always be captured by semantic similarity. Second, the search agent model has the adaptability to adjust the reasoning length. ANN might lead search agents to perform an extra reasoning step (e.g., 2.292 vs. 2.288 for NQ), adjusting the retrieval query, which in turn improves generation accuracy.

424

425

4.3 ABLATION STUDY

426

427

The priority scheduling and non-stall retrieval of SearchAgent-X help improve the prefix cache utility, thus enhancing end-to-end efficiency. In this section, we use Search-R1 7B model on Musique dataset. Figure 6 (left) shows the end-to-end performance of different techniques for offline inference with top- k = 5. We have several observations. First, the advantages of prefix cache are diminished in this challenging scenario. With top- k =5, it’s only $1.01\times$ that of vLLM_ANN, compared to $1.91\times$ with top- k =1 (see all prefix caching improvement data in Figure 4). This validates that

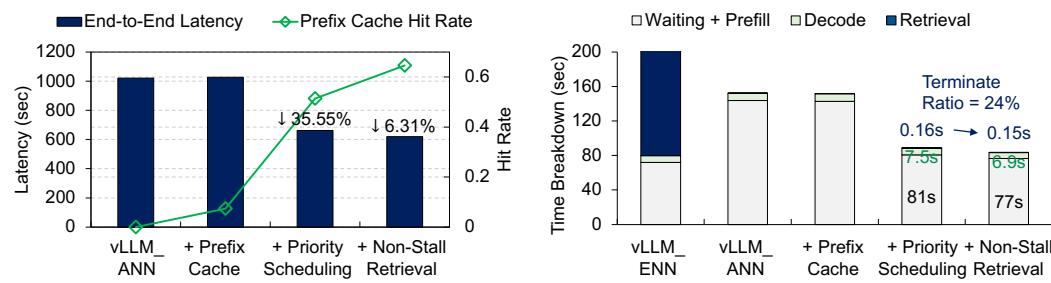


Figure 6: **SearchAgent-X Technique Breakdown for End-to-End Performance (Left) and Per-Sequence Generation Time (Right).** Each bar/scatter adds one technique over its left bar/scatter, with the leftmost being vanilla vLLM and the rightmost being the full SearchAgent-X.

the benefits of prefix cache still require appropriate scheduling and retrieval methods to unleash its potential. Second, SearchAgent-X’s priority scheduling reduces the end-to-end latency by 35.55% based on prefix cache. This is because the priority of requests is properly managed, maximizing the utilization of GPU resources. In addition, the prefix cache hit rate increases from 0.07 to 0.51, verifying the effectiveness of the technique. Third, SearchAgent-X’s non-stall retrieval further improves the hit rate to 0.65, leading to a further 6.3% reduction in latency. This shows that the adaptive termination strategy fully utilizes the “free lunch” of asynchronous execution, timely recalling mature retrieval results, thereby improving system processing efficiency.

Figure 6 (right) further demonstrates the per-sequence generation time of different parts. We have more observations. First, for vLLM_ENN, the retrieval time is the largest component, accounting for over 60% of the total time. Instead, its prefill time is the lowest across different techniques, since its reasoning requires waiting for long-time retrieval, thus reducing the pressure on token generation. Second, for priority scheduling, we note that it reduces not only the prefill time (due to more prefix cache utilized), but also the decode time, showing a better system processing capability. This is because by improving KV-cache utilization, it avoids recomputation of long requests, freeing up GPU space earlier for better decode parallelism. Third, non-stall retrieval actually only reduces 0.01s of retrieval time (from 0.16s to 0.15s), with about 24% of the retrievals being early terminated, but significantly reduces the end-to-end latency (41s, **the end-to-end latency shown in Figure 6 (left)**). This aligns well with our previous observation that minor retrieval latency can cause drastic efficiency degradation (as shown in Figure 2). Non-stall retrieval adaptively terminates only a small set of retrievals when necessary, yet yields the significant benefit of better cache utilization. More experiments can be found in Appendix C.3 to C.7.

5 RELATED WORK

Several recent systems optimize one- or multi-round RAG pipelines, such as TELERAG (Lin et al., 2025), RAGO (Jiang et al., 2025), PATCHWORK (Hu et al., 2025), RAGServe (Ray et al., 2024), RAGCache (Jin et al., 2024), AquaPipe (Yu et al., 2025), AutoRAG (Fu et al., 2024), and PipeRAG (Jiang et al., 2024). These methods either prefetch or pipeline retrieval results, or tune retrieval/generation hyperparameters, but they still treat retrieval and generation as largely separated stages and do not analyze the root causes of inefficiency in the dynamically interleaved search–reasoning pattern targeted by SearchAgent-X.

Techniques that could change sequence prefix, such as context trimming, summarization, or token sparsification (Jiang et al., 2023) are also compatible with our design: as long as they are applied in a prefix-preserving fashion during decoding, prefix caching remains effective and our scheduler can simply use the effective cached prefix length (after sparsification) as a scheduling feature.

Meanwhile, broader agent workflow optimizations, such as auto-tuning (He et al., 2025), KV-cache management (Abhyankar et al., 2024), and partial tool execution (Xu et al., 2024), improve overall efficiency but overlook the specific challenges of retrieval accuracy and latency in search agents. In contrast, SearchAgent-X directly addresses these challenges by tightly coupling priority-scheduled reasoning with non-stall retrieval, yielding improved efficiency. Notably, our approach is orthogonal

486 to, and can potentially be combined with, other RAG optimization techniques such as context
 487 compression (Shi et al., 2024) and retrieval reranking (Glass et al., 2022).
 488

489 **6 CONCLUSION**
 490

491 LLM reasoning-driven search agents offer great potential for complex problems, but face severe,
 492 distinct efficiency challenges. This paper highlights the non-trivial impact of retrieval accuracy and
 493 the latency sensitivity caused by scheduling deficiencies and retrieval-induced stalls. Our proposed
 494 system, SearchAgent-X, designed based on these insights, demonstrates substantial improvement
 495 in system efficiency, all while maintaining high generation quality. This study provides important
 496 insights for practical deployments of high-efficiency LLM-based search agents, and the proposed
 497 solutions are easily adaptable to other ANN retrieval methods and LLM reasoning models.
 498

499 **ETHICS STATEMENT**
 500

501 This work adheres to the ICLR Code of Ethics. Our study does not involve human subjects, personally
 502 identifiable information, sensitive data, or applications with potential for misuse. Therefore, we
 503 believe there are no ethical concerns related to this paper.
 504

505 **REPRODUCIBILITY STATEMENT**
 506

507 We provide the complete source code of SearchAgent-X in an anonymous repository at <https://anonymous.4open.science/r/SearchAgent-X>. The repository is self-contained with instruc-
 508 tions and scripts to replicate all experiments, along with dataset specifications and pre-trained model
 509 references.
 510

512 **REFERENCES**
 513

514 Reyna Abhyankar, Zijian He, Vikranth Srivatsa, Hao Zhang, and Yiying Zhang. InferCept: Efficient
 515 Intercept Support for Augmented Large Language Model Inference, May 2024. URL <http://arxiv.org/abs/2402.01869> [cs].
 516

517 Mingyang Chen, Tianpeng Li, Haoze Sun, Yijie Zhou, Chenzheng Zhu, Haofen Wang, Jeff Z. Pan,
 518 Wen Zhang, Huajun Chen, Fan Yang, Zenan Zhou, and Weipeng Chen. ReSearch: Learning to
 519 Reason with Search for LLMs via Reinforcement Learning, March 2025. URL <http://arxiv.org/abs/2503.19470>. arXiv:2503.19470 [cs].
 520

521 Sanjoy Dasgupta and Kaushik Sinha. Randomized partition trees for exact nearest neighbor search.
 522 In Shai Shalev-Shwartz and Ingo Steinwart (eds.), *Proceedings of the 26th Annual Conference
 523 on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 317–337,
 524 Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL <https://proceedings.mlr.press/v30/Dasgupta13.html>.
 525

526 Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and
 527 Qing Li. A survey on rag meeting llms: Towards retrieval-augmented large language models. In
 528 *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp.
 529 6491–6501, 2024.
 530

531 Jia Fu, Xiaoting Qin, Fangkai Yang, Lu Wang, Jue Zhang, Qingwei Lin, Yubo Chen, Dongmei Zhang,
 532 Saravan Rajmohan, and Qi Zhang. AutoRAG-HP: Automatic Online Hyper-Parameter Tuning
 533 for Retrieval-Augmented Generation, June 2024. URL <http://arxiv.org/abs/2406.19251>.
 534 arXiv:2406.19251 [cs].

535 Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng
 536 Wang, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey,
 537 March 2024. URL <http://arxiv.org/abs/2312.10997>. arXiv:2312.10997 [cs].
 538

539 Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan
 Cai, and Alfio Gliozzo. Re2g: Retrieve, rerank, generate. *arXiv preprint arXiv:2207.06300*, 2022.

540 Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar.
 541 Accelerating large-scale inference with anisotropic vector quantization. In *International Conference*
 542 *on Machine Learning*, pp. 3887–3896. PMLR, 2020.

543

544 Shailja Gupta, Rajesh Ranjan, and Surya Narayan Singh. A Comprehensive Survey of Retrieval-
 545 Augmented Generation (RAG): Evolution, Current Landscape and Future Directions, October
 546 2024. URL <http://arxiv.org/abs/2410.12837>. arXiv:2410.12837 [cs].

547

548 Zijian He, Reyna Abhyankar, Vikranth Srivatsa, and Yiying Zhang. Cognify: Supercharging gen-ai
 549 workflows with hierarchical autotuning. *arXiv preprint arXiv:2502.08056*, 2025.

550

551 Charles C. Holt. Forecasting seasonals and trends by exponentially weighted moving averages.
552 International Journal of Forecasting, 20(1):5–10, 2004. ISSN 0169-2070. doi: <https://doi.org/10.1016/j.ijforecast.2003.09.015>. URL <https://www.sciencedirect.com/science/article/pii/S0169207003001134>.

553

554 Bodun Hu, Luis Pabon, Saurabh Agarwal, and Aditya Akella. Patchwork: A unified framework
 555 for RAG serving. *CoRR*, abs/2505.07833, 2025. doi: 10.48550/arXiv.2505.07833. URL <https://arxiv.org/abs/2505.07833>.

556

557 Yizheng Huang and Jimmy Huang. A Survey on Retrieval-Augmented Text Generation for Large Lan-
 558 guage Models, August 2024. URL <http://arxiv.org/abs/2404.10981>. arXiv:2404.10981
 559 [cs].

560

561 Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing
 562 prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference*
 563 *on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics,
 564 2023. doi: 10.48550/arXiv.2310.05736. URL <https://arxiv.org/abs/2310.05736>.

565

566 Wenqi Jiang, Shuai Zhang, Boran Han, Jie Wang, Bernie Wang, and Tim Kraska. Piperag: Fast
 567 retrieval-augmented generation via algorithm-system co-design. *arXiv preprint arXiv:2403.05676*,
 568 2024.

569

570 Wenqi Jiang, Suvinay Subramanian, Cat Graves, Gustavo Alonso, Amir Yazdanbakhsh, and Vidushi
 571 Dadu. Rago: Systematic performance optimization for retrieval-augmented generation serving. In
 572 *Proceedings of the 52nd Annual International Symposium on Computer Architecture (ISCA '25)*,
 573 pp. 974–989. ACM, 2025. doi: 10.1145/3695053.3731093. URL <https://arxiv.org/abs/2503.14649>.

574

575 Bowen Jin, Hansi Zeng, Zhenrui Yue, Dong Wang, Hamed Zamani, and Jiawei Han. Search-R1:
 576 Training LLMs to Reason and Leverage Search Engines with Reinforcement Learning, March
 577 2025a. URL <http://arxiv.org/abs/2503.09516>. arXiv:2503.09516 [cs].

578

579 Chao Jin, Zili Zhang, Xuanlin Jiang, Fangyue Liu, Xin Liu, Xuanzhe Liu, and Xin Jin. RAGCache:
 580 Efficient Knowledge Caching for Retrieval-Augmented Generation, April 2024. URL <http://arxiv.org/abs/2404.12457>. arXiv:2404.12457 [cs].

581

582 Jiajie Jin, Yutao Zhu, Guanting Dong, Yuyao Zhang, Xinyu Yang, Chenghao Zhang, Tong Zhao, Zhao
 583 Yang, Zhicheng Dou, and Ji-Rong Wen. FlashRAG: A Modular Toolkit for Efficient Retrieval-
 584 Augmented Generation Research, February 2025b. URL <http://arxiv.org/abs/2405.13576>.
 585 arXiv:2405.13576 [cs].

586

587 Jeff Johnson, Matthijs Douze, and Herve Jegou. Billion-scale similarity search with gpus. *IEEE*
 588 *Transactions on Big Data*, pp. 535–547, Jul 2021. doi: 10.1109/tbdata.2019.2921572. URL
 589 <http://dx.doi.org/10.1109/tbdata.2019.2921572>.

590

591 Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph
 592 Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model
 593 serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems*
Principles, pp. 611–626, 2023.

594 Conglong Li, Minjia Zhang, David G. Andersen, and Yuxiong He. Improving approximate nearest
 595 neighbor search through learned adaptive early termination. In *Proceedings of the 2020 ACM*
 596 *SIGMOD International Conference on Management of Data*, SIGMOD '20, pp. 2539–2554,
 597 New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450367356. doi:
 598 10.1145/3318464.3380600. URL <https://doi.org/10.1145/3318464.3380600>.

599
 600 Xiaoxi Li, Guanting Dong, Jiajie Jin, Yuyao Zhang, Yujia Zhou, Yutao Zhu, Peitian Zhang, and
 601 Zhicheng Dou. Search-o1: Agentic search-enhanced large reasoning models. *arXiv preprint*
 602 *arXiv:2501.05366*, 2025.

603 Chien-Yu Lin, Keisuke Kamahori, Yiyu Liu, Xiaoxiang Shi, Madhav Mahesh Kashyap, Yile Gu,
 604 Rulin Shao, Zihao Ye, Kan Zhu, Stephanie Wang, Arvind Krishnamurthy, Rohan Kadekodi,
 605 Luis Ceze, and Baris Kasikci. Telerag: Efficient retrieval-augmented generation inference with
 606 lookahead retrieval. *CoRR*, abs/2502.20969, 2025. doi: 10.48550/arXiv.2502.20969. URL
 607 <https://arxiv.org/abs/2502.20969>.

608
 609 Yu. A. Malkov and D. A. Yashunin. Hnswlib - fast approximate nearest neighbor search. <https://github.com/nmslib/hnswlib>. Accessed: 2025-05-15.

610
 611 Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using
 612 hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine*
 613 *intelligence*, 42(4):824–836, 2018.

614
 615 Yury Malkov, Alexander Ponomarenko, Andrey Logvinov, and Vladimir Krylov. Approximate
 616 nearest neighbor algorithm based on navigable small world graphs. *Information Systems*, 45:
 617 61–68, 2014. ISSN 0306-4379. doi: <https://doi.org/10.1016/j.is.2013.10.006>. URL <https://www.sciencedirect.com/science/article/pii/S0306437913001300>.

618
 619 NVIDIA. FasterTransformer. <https://github.com/NVIDIA/FasterTransformer>, 2019. Ac-
 620 cessed: 2025-05-14.

621
 622 NVIDIA. Triton Inference Server: Dynamic Batching. https://docs.nvidia.com/deeplearning/triton-inference-server/user-guide/docs/user_guide/model_configuration.html#dynamic-batcher, 2021. Accessed: 2025-05-14.

623
 624 OpenAI. Introducing deep research. <https://openai.com/index/introducing-deep-research/>, 2025. Accessed: 2025-05-15.

625
 626 Siddhant Ray, Rui Pan, Zhuohan Gu, Kuntai Du, Ganesh Ananthanarayanan, Ravi Netravali, and
 627 Junchen Jiang. RAGServe: Fast Quality-Aware RAG Systems with Configuration Adaptation,
 628 December 2024. URL <http://arxiv.org/abs/2412.10543>. arXiv:2412.10543 [cs].

629
 630 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks.
 631 In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
 632 Association for Computational Linguistics, 11 2019. URL <https://arxiv.org/abs/1908.10084>.

633
 634 Z. Ren, K. Doeckemeijer, P. Apparao, and A. Trivedi. Storage-based approximate nearest neighbor
 635 search: What are the performance, cost, and i/o characteristics? In *2025 IEEE International*
 636 *Symposium on Workload Characterization (IISWC)*. IEEE, oct 2025.

637
 638 Kaize Shi, Xueyao Sun, Qing Li, and Guandong Xu. Compressing long context for enhancing rag
 639 with amr-based concept distillation. *arXiv preprint arXiv:2405.03085*, 2024.

640
 641 Aditi Singh, Abul Ehtesham, Saket Kumar, and Tala Talaei Khoei. Agentic Retrieval-Augmented
 642 Generation: A Survey on Agentic RAG, February 2025. URL <http://arxiv.org/abs/2501.09136>. arXiv:2501.09136 [cs].

643
 644 Huatong Song, Jinhao Jiang, Yingqian Min, Jie Chen, Zhipeng Chen, Wayne Xin Zhao, Lei Fang,
 645 and Ji-Rong Wen. R1-Searcher: Incentivizing the Search Capability in LLMs via Reinforcement
 646 Learning, March 2025. URL <http://arxiv.org/abs/2503.05592>. arXiv:2503.05592 [cs].

648 Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval
 649 with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint*
 650 *arXiv:2212.10509*, 2022.

651

652 M. Wang, X. Xu, Q. Yue, and Y. Wang. A comprehensive survey and experimental comparison
 653 of graph-based approximate nearest neighbor search. *Proceedings of the VLDB Endowment*, 14
 654 (11):1964–1978, 2021. doi: 10.14778/3476249.3476255. URL <https://doi.org/10.14778/3476249.3476255>.

655

656 Shangyu Wu, Ying Xiong, Yufei Cui, Haolun Wu, Can Chen, Ye Yuan, Lianming Huang, Xue
 657 Liu, Tei-Wei Kuo, Nan Guan, and Chun Jason Xue. Retrieval-Augmented Generation for Nat-
 658 ural Language Processing: A Survey, July 2024. URL <http://arxiv.org/abs/2407.13193>.
 659 *arXiv:2407.13193* [cs].

660

661 xAI. Grok agents: Combining reasoning and tool use. <https://x.ai/news/grok-3#grok-agents-combining-reasoning-and-tool-use>, 2025. Accessed: 2025-05-15.

662

663 Yechen Xu, Xinhao Kong, Tingjun Chen, and Danyang Zhuo. Conveyor: Efficient tool-aware llm
 664 serving with tool partial execution. *arXiv preprint arXiv:2406.00059*, 2024.

665

666 Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov,
 667 and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question
 668 answering. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

669

670 Hao Yu, Aoran Gan, Kai Zhang, Shiwei Tong, Qi Liu, and Zhaofeng Liu. Evaluation of retrieval-
 671 augmented generation: A survey. In *CCF Conference on Big Data*, pp. 102–120. Springer, 2024.

672

673 Runjie Yu, <https://orcid.org/0009-0002-8006-5316>, View Profile, Weizhou Huang,
 674 <https://orcid.org/0000-0002-8024-0789>, View Profile, Shuhan Bai, <https://orcid.org/0000-0002-0842-7602>, View Profile, Jian Zhou, <https://orcid.org/0000-0001-5295-4680>, View
 675 Profile, Fei Wu, <https://orcid.org/0000-0001-9746-4714>, and View Profile. AquaPipe: A
 676 Quality-Aware Pipeline for Knowledge Retrieval and Large Language Models. *Proceedings*
 677 *of the ACM on Management of Data*, 3(1):1–26, February 2025. doi: 10.1145/3709661. URL
 678 <https://dlnext.acm.org/doi/10.1145/3709661>. Publisher: Association for Computing
 679 Machinery.

680

681 Penghao Zhao, Hailin Zhang, Qinhuan Yu, Zhengren Wang, Yunteng Geng, Fangcheng Fu, Ling Yang,
 682 Wentao Zhang, Jie Jiang, and Bin Cui. Retrieval-Augmented Generation for AI-Generated Content:
 683 A Survey, June 2024a. URL <http://arxiv.org/abs/2402.19473>. *arXiv:2402.19473* [cs].

684

685 Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K. Qiu, and Lili Qiu. Retrieval Augmented
 686 Generation (RAG) and Beyond: A Comprehensive Survey on How to Make your LLMs use
 687 External Data More Wisely, September 2024b. URL <http://arxiv.org/abs/2409.14924>.
 688 *arXiv:2409.14924* [cs].

689

690 Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Livia Sun, Jeff Huang, Cody Hao Yu, Shiyi
 691 Cao, Christos Kozyrakis, Ion Stoica, Joseph E Gonzalez, et al. Sqlang: Efficient execution of
 692 structured language model programs. *Advances in Neural Information Processing Systems*, 37:
 693 62557–62583, 2024.

694

695

696

697

698

699

700

701

APPENDIX

A AN ILLUSTRATION OF LLM-BASED SEARCH AGENTS

Figure 7 shows an example of a search agent process. Faced with a complex query ("Curious is a women's fragrance by a singer born in what city and state?"), the search agent first engages in preliminary reasoning ("I need to find out which city and state a singer..."). Recognizing a knowledge gap regarding the "Curious fragrance," the model proactively decides to initiate a search ("search Curious fragrance information"). Upon receiving the crucial information ("Curious is a women's fragrance by Britney Spears"), the model doesn't conclude its process. Instead, it integrates this new knowledge into its subsequent thought process and reasoning. This triggers further searches, of which the retrieval result is concatenated with previously generated tokens and re-injected into LLMs. Through this dynamic cycle of "think-search-rethink," the model progressively assembles the necessary pieces of the knowledge puzzle required to answer the question fully. This culminates in a high-quality answer that addresses all aspects of the initial query ("McComb, Mississippi"). This ability to autonomously plan retrieval actions and iteratively incorporate new information into its reasoning process allows the search agent to tackle more complex questions and deliver better responses, moving beyond reliance solely on pre-trained knowledge or a single retrieval.

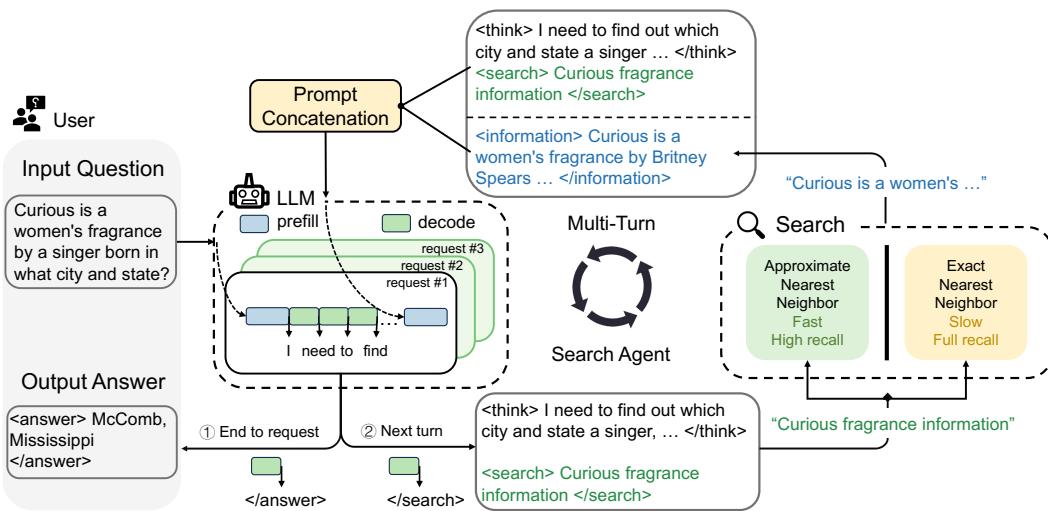


Figure 7: An illustration of reasoning and search interleaved LLM-based search agents.

B IMPLEMENTATION DETAILS

B.1 SEARCHAGENT-X EXECUTION

This section outlines the high-level execution flow of the SearchAgent-X system, as depicted in Algorithm 1, complementing the conceptual component descriptions in Section 3 of the main paper. SearchAgent-X orchestrates LLM inference (with prefix caching, Section 3.1), dynamic high-recall approximate retrieval, Priority Scheduling (Section 3.2), and Non-Stall Retrieval (Section 3.3) to achieve efficient search agents. The system initializes an LLM inference engine and manages incoming requests, active asynchronous search tasks, and their results.

The main execution loop begins by ingesting new user requests into the LLM engine's pool (line 5-7). Concurrently, if Non-Stall Retrieval is active, SearchAgent-X consults an external signal to identify and expedite the completion of any ongoing retrieval tasks that have reached sufficient maturity or for which LLM engine readiness dictates early termination (line 10-11), thus preventing pipeline stalls. Upon completion of a search (either normally or via early termination), retrieved documents are concatenated with the original context, and the augmented request is resubmitted to the LLM engine (line 15-20).

756 The core of the processing loop involves LLM generation and the agentic control flow. Before
 757 each LLM generation step, SearchAgent-X’s Priority Scheduling policy is applied to the queue
 758 of waiting requests, reordering them to optimize system throughput and KV-cache utilization (line
 759 23). Following token generation by the LLM, the output for each active sequence is parsed (line
 760 25-28). If a `<search>` tag is detected, indicating a need for external knowledge, SearchAgent-X
 761 halts further generation for that sequence and launches an asynchronous high-recall ANN retrieval
 762 task (line 29-34). Conversely, if a `<answer>` tag is identified or the sequence naturally concludes,
 763 the request is finalized (line 35-38). This iterative and asynchronous process enables the dynamic
 764 interleaving of LLM reasoning, external knowledge retrieval, and intelligent scheduling, which is
 765 fundamental to SearchAgent-X’s efficient handling of complex search agent workloads.
 766

Algorithm 1 SearchAgent-X Main Execution Loop

```

 1: Initialize LLM_Engine, ArrivalQueue, ActiveSearchTasks, FinishedOutputs
 2: Configure PriorityScheduling (enabled/type), NonStallRetrieval (enabled)
 3: while LLM_Engine has unfinished requests or not ActiveSearchTasks is empty or not Ar-
 4:   rrivalQueue is empty do
 5:     // Step 1: Ingest new requests
 6:     for each request  $R_{new}$  in ArrivalQueue ready for processing do
 7:       Add  $R_{new}$  to LLM_Engine’s request pool
 8:       Remove  $R_{new}$  from ArrivalQueue
 9:     end for
10:    // Step 2: Non-Stall Retrieval Check (if enabled)
11:    if NonStallRetrieval is enabled and ActiveSearchTasks is not empty and LLM_Engine has
12:      waiting requests then
13:        TerminatedSearchIDs  $\leftarrow$  CheckExternalNonStallSignal()
14:        // Queries for searches to terminate early
15:      end if
16:      // Step 3: Process completed search tasks
17:      for each search task  $S_i$  in ActiveSearchTasks that has completed do
18:         $R_{orig}, retrieved\_docs, search\_finish\_time \leftarrow S_i.getResult()$ 
19:        new_context  $\leftarrow$  Concatenate( $R_{orig}.context, retrieved\_docs$ )
20:        AddResumedRequest( $R_{orig}, new\_context, search\_finish\_time$ ) to LLM_Engine
21:        Remove  $S_i$  from ActiveSearchTasks
22:      end for
23:      // Step 4: LLM Generation Step
24:      if LLM_Engine has unfinished requests then
25:        // Section 3.2
26:        ApplyPriorityScheduling(LLM_Engine.waiting_requests)
27:        LLM_Outputs, Scheduler_Info  $\leftarrow$  LLM_Engine.step()
28:        RecordTokenTimingsAndPrefixCacheStats(Scheduler_Info)
29:        for each output  $O_j$  in LLM_Outputs do
30:          current_text  $\leftarrow O_j.getGeneratedText()$ 
31:          if DetectSearchTag(current_text) then
32:            query  $\leftarrow$  ExtractSearchQuery(current_text)
33:            LLM_Engine.abortRequest( $O_j.request\_id$ )
34:             $S_{new} \leftarrow$  LaunchAsyncRetrievalTask( $O_j.request\_id, current\_text, query$ )
35:            // High-recall ANN
36:            Add  $S_{new}$  to ActiveSearchTasks
37:          else if DetectAnswerTag(current_text) or  $O_j.isFinished()$  then
38:            Add  $O_j$  to FinishedOutputs
39:            LLM_Engine.abortRequest( $O_j.request\_id$ )
40:          end if
41:        end for
42:      end if
43:    end while
44:  CleanupRemainingTasks()
  
```

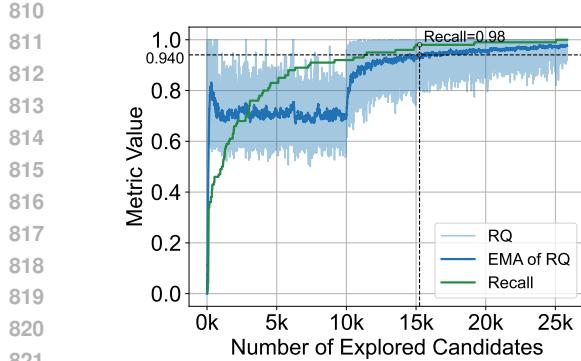


Figure 8: **EMA Signal of Retrieval Maturity.** RQ means relative quality of newly explored candidates. EMA represents smoothed RQ. The vertical line marks maturity, where the improvement of recall and EMA decreases to about 0.

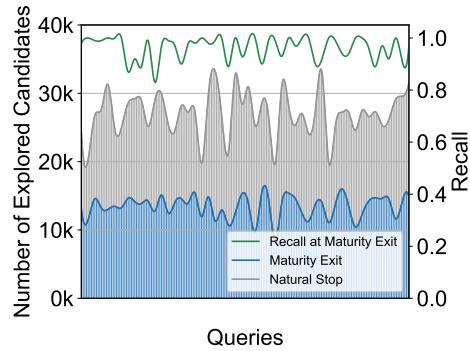


Figure 9: **Comparison of Maturity Exit and Natural Stop.** The shadows represent the number of candidates explored by two methods (showing a similar trend). The curve represents the recall of maturity exit (consistently high).

B.2 RETRIEVAL MATURITY ESTIMATION

In practice, the raw RQ_t signal described in Section 3.3 exhibits short-term fluctuations that may hinder robust maturity estimation. To address this, `SearchAgent-X` applies an exponential moving average (EMA) (Holt, 2004), with a window size of 500, to smooth the signal.

Selecting an appropriate threshold τ for the EMA is critical to balancing retrieval quality and latency. To determine a suitable value, we sample queries from the Musique dataset and record the evolution of the EMA curve as the number of explored candidates increases. For each query, we identify the point where the EMA curve flattens—i.e., where marginal improvements approach zero—indicating that newly explored candidates are far from the query and contribute little to quality. This point reflects the onset of retrieval maturity. As shown in Figure 8, the recall at this stage also stabilizes and reaches a high level (around 0.98). We adopt the corresponding EMA value ($\tau = 0.9$) at this “knee” as the practical threshold τ in `SearchAgent-X` to reliably trigger maturity exit. We provide more analysis of this design to verify its effectiveness in Appendix C.3.

B.3 DETAILED EXPERIMENTAL SETUP

We implement `SearchAgent-X` by building upon vLLM (Kwon et al., 2023), a state-of-the-art LLM inference engine to use its efficient PagedAttention mechanism. For retrieval component, we use a knowledge base constructed from a chunked Wikipedia dataset, containing approximately 21 million text chunks. Each chunk is embedded into a 384-dimensional vector using the `all-MiniLM-L6-v2` model (Reimers & Gurevych, 2019). An Approximate Nearest Neighbor (ANN) index is built offline over these embeddings using HNSWlib (Malkov & Yashunin), configured with parameters such as up to 32 neighbors per node and an `efConstruction` (candidate list size during build) of 500. This index serves as the foundation for all ANN-based retrieval methods in our experiments. For these ANN searches (employed by `SearchAgent-X` and approximate retrieval baselines), we generally set the search range (e.g., `efSearch`) to 10,000 to achieve high recall with acceptable computational overhead, based on empirical analysis. These HNSW ANN searches leverage inter-query parallelism with 4 threads to optimize throughput while managing memory access contention.

Specific configurations for the different systems are as follows. For the exact retrieval baseline (`vLLM_ENN`), we adapt HNSWlib to perform exhaustive search more efficiently by enabling intra-query parallelism, utilizing 6 threads. For `SearchAgent-X`, beyond using the aforementioned high-recall ANN setup, we set its unique parameters: the priority scheduling level G is configured to 6 (we note that `SearchAgent-X` exhibits low sensitivity to this specific value, as shown in our ablation study in Appendix C.4). The threshold τ_{EMA} for estimating retrieval maturity in the non-stall mechanism is set to 0.9, determined via offline profiling detailed in Appendix B.2. The approximate retrieval

864 baselines (vLLM_ANN, CachevLLM_ANN) also utilize the general ANN search settings described above,
 865 including the search range of 10,000 and 4 threads for inter-query parallelism.
 866

868 C MORE RESULTS

870 C.1 DETAILED OVERALL EFFICIENCY

872 We note that SearchAgent-X outperforms all baselines in different scenarios including model sizes,
 873 deployment methods (single GPU or distributed GPUs), top- k values, and query datasets.
 874

875 **Different Model Sizes and Top-k Values** Table 2 presents detailed results of overall efficiency
 876 across different methods for Search-R1 under 7B (on a single L40 GPU) and 14B models (on two
 877 40GB A100 GPUs). The advantage of SearchAgent-X is bigger in the 7B model and small top- k
 878 values of the 14B model. This is due to two reasons. First, although the 14B model distributes model
 879 weights across two A100s, its KV-cache space is still smaller than the 7B model (because the model
 880 weights are larger, resulting in a larger KV-cache for a single token). The 7B model has a larger
 881 available KV-cache space, thus yielding greater benefits from managing the prefix cache. Second, the
 882 14B model we test calls retrievals more cautiously, while the 7B model calls more retrievals (e.g., 4.9
 883 of the 7B model vs. 3.3 of the 14B model when top- k = 3 for Musique dataset), resulting in a greater
 884 distinction in request priority.
 885

886 **Different Datasets** Moreover, as shown in Table 3, on the HotpotQA dataset, SearchAgent-X also
 887 outperforms all baselines, achieving 1.47x to 2.55x higher throughput and 1.62x to 2.81x lower
 888 latency compared to the strongest baseline, CachevLLM_ANN.
 889

890 **Different Search Agent Models** To further demonstrate the generality of our method, we apply it
 891 to a different search agent architecture, ReCall (Chen et al., 2025). We observe that SearchAgent-X
 892 consistently delivers best efficiency. As shown in Table 4 below, when applied to ReCall’s 7B model,
 893 SearchAgent-X attains 1.12x to 1.74x higher throughput, 1.31x to 1.81x lower latency, and 1.49x
 894 to 2.86x higher cache hit rate than baselines. The results demonstrate that SearchAgent-X still
 895 outperforms the most competitive baseline in this different search agent architecture, confirming the
 896 generality of SearchAgent-X.
 897

898 C.2 DETAILED GENERATION QUALITY

900 Table 5 further shows the generation details beyond Table 1 of SearchAgent-X and exact retrieval
 901 (vLLM_ENN) across different datasets. We note that SearchAgent-X achieves similar generation
 902 length as exact retrieval across all datasets (6822 tokens vs. 6826 tokens), indicating that our non-stall
 903 retrieval does not lead to unusual search agent model/retriever behaviours, making SearchAgent-X
 904 possible for maintaining perfect generation quality.
 905

906 C.3 ANALYSIS OF MATURITY EXIT MECHANISM

907 **The maturity exit mechanism effectively halts unnecessary searches without compromising**
 908 **retrieval quality.** To validate the effectiveness of non-stall retrieval, we analyze whether the maturity-
 909 based termination reliably halts unnecessary ANN iterations. We compare the retrieval traces of
 910 representative queries under two settings: maturity-based early stop and standard natural stop. As
 911 shown in Figure 9, we make several observations. First, query difficulty varies significantly across
 912 the dataset, resulting in different numbers of candidate nodes explored before natural convergence.
 913 This highlights the need for an adaptive termination strategy rather than relying on a fixed exploration
 914 budget. Second, for queries of varying difficulty, the number of candidates explored by the maturity-
 915 based strategy closely matches the natural termination point of standard search, indicating our maturity
 916 exit accurately captures query difficulties. More importantly, the recall achieved by maturity-stopped
 917 queries remains consistently high (0.963 on average). These results confirm that our non-stall retrieval
 effectively terminates redundant search iterations while preserving retrieval quality.

918
 919
 920 Table 2: Comparison across seven key metrics and top- k values for different methods. Throughput
 921 and efficiency gains are marked by \times multipliers. Lower values are better for metrics marked with \downarrow .

Metric	Top-1	Top-2	Top-3	Top-4	Top-5	Top-1	Top-2	Top-3
	Search-R1-7B					Search-R1-14B		
Throughput								
vLLM_ENN	0.44	0.69	0.62	0.63	0.61	0.46	0.47	0.43
vLLM_ANN	0.62	1.04	0.76	0.73	0.68	0.94	0.77	0.61
CachevLLM_ANN	1.18	1.25	0.89	0.77	0.69	1.08	0.89	0.70
SearchAgent-X	1.59	2.36	1.64	1.39	1.01	1.40	1.09	0.76
Max Ratio	3.61 \times	3.42 \times	2.65 \times	2.20 \times	1.66 \times	3.04 \times	2.32 \times	1.77 \times
Min Ratio	1.35 \times	1.89 \times	1.84 \times	1.81 \times	1.46 \times	1.30 \times	1.22 \times	1.09 \times
Token Throughput								
vLLM_ENN	69.90	90.85	86.12	97.79	84.26	156.35	127.28	111.21
vLLM_ANN	101.21	136.81	106.09	114.28	94.65	320.76	206.46	159.70
CachevLLM_ANN	191.65	164.46	124.59	119.17	95.64	366.85	239.01	182.60
SearchAgent-X	259.94	309.73	229.96	216.21	139.25	472.76	292.13	199.14
Max Ratio	3.72 \times	3.41 \times	2.67 \times	2.21 \times	1.65 \times	3.02 \times	2.30 \times	1.79 \times
Min Ratio	1.36 \times	1.88 \times	1.85 \times	1.81 \times	1.46 \times	1.29 \times	1.22 \times	1.09 \times
Latency \downarrow								
vLLM_ENN	1300.56	1066.05	1089.37	1154.62	1205.16	1642.86	1567.85	1614.71
vLLM_ANN	571.36	625.33	790.68	930.29	1020.46	767.46	923.74	1052.74
CachevLLM_ANN	429.60	562.47	759.57	916.58	1026.91	673.35	805.60	980.35
SearchAgent-X	238.00	266.50	347.14	466.78	620.07	502.10	690.20	939.42
Max Ratio	0.18 \times	0.25 \times	0.32 \times	0.40 \times	0.51 \times	0.31 \times	0.44 \times	0.58 \times
Min Ratio	0.55 \times	0.47 \times	0.46 \times	0.51 \times	0.60 \times	0.75 \times	0.86 \times	0.96 \times
P99 Latency \downarrow								
vLLM_ENN	1758.64	1441.18	1462.45	1569.80	1641.90	2205.26	2136.75	2237.55
vLLM_ANN	915.04	956.46	1160.86	1348.87	1459.13	1066.63	1334.67	1555.47
CachevLLM_ANN	609.70	797.46	1073.03	1296.03	1446.88	930.06	1159.02	1454.61
SearchAgent-X	373.20	421.32	566.28	716.88	993.98	732.16	958.78	1362.24
Max Ratio	0.21 \times	0.29 \times	0.39 \times	0.46 \times	0.61 \times	0.33 \times	0.45 \times	0.61 \times
Min Ratio	0.61 \times	0.53 \times	0.53 \times	0.55 \times	0.69 \times	0.79 \times	0.83 \times	0.94 \times

954
 955
 956
 957 Table 3: Comparison of system efficiency under the HotpotQA dataset.

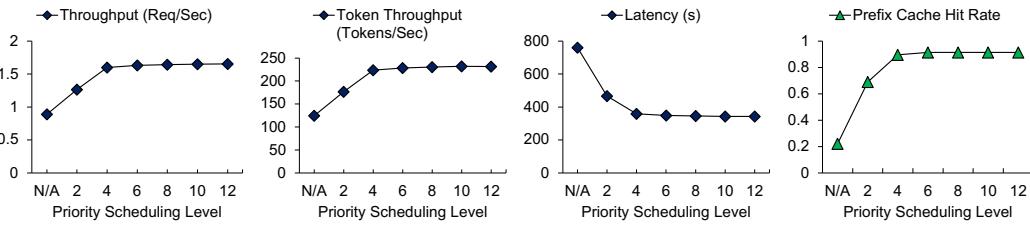
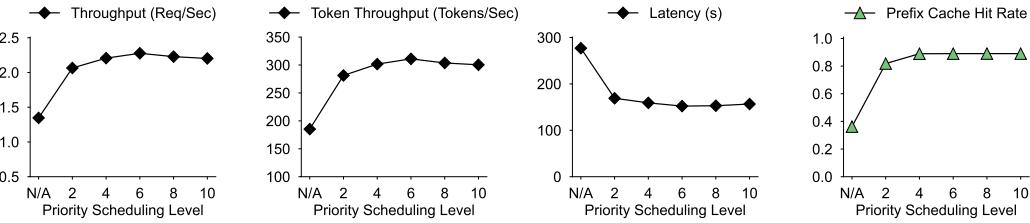
TopK	Method	Throughput	Latency	Token Throughput	Cache Hit Rate
3	vLLM_ENN	0.78143	483.4094	106.0500	0.0000
	CachevLLM_ANN	1.30926	282.9210	179.9000	0.3890
	SearchAgent-X	1.99253	171.9478	271.8200	0.8890
4	vLLM_ENN	0.74530	505.7417	102.6500	0.0000
	CachevLLM_ANN	0.95950	375.9803	133.7800	0.3040
	SearchAgent-X	1.57188	210.8670	217.4600	0.9100
5	vLLM_ENN	0.68400	543.9000	87.5970	0.0000
	CachevLLM_ANN	0.81700	443.1300	105.2410	0.1706
	SearchAgent-X	1.20400	273.6900	154.8331	0.7340

972
973 Table 4: Comparison of system efficiency under the ReCall search agent model.
974

TopK	Method	Throughput	Latency	Token Throughput	Cache Hit Rate
3	vLLM_ENN	0.6370	376.375	224.550	0.000
	CachevLLM ANN	0.9861	273.360	342.490	0.570
	SearchAgent-X	1.1035	207.920	391.898	0.849
4	vLLM_ENN	0.6200	423.070	205.630	0.000
	CachevLLM ANN	0.8242	342.870	276.810	0.383
	SearchAgent-X	1.0780	234.390	364.066	0.804
5	vLLM_ENN	0.7180	393.390	220.810	0.000
	CachevLLM ANN	0.7566	383.490	236.660	0.250
	SearchAgent-X	0.9570	282.085	302.049	0.716

988
989 Table 5: Generation Quality of SearchAgent-X and Exact Retrieval.
990
991

Dataset	Musique	NQ	2Wiki	HotpotQA	Bamboogle	StrategyQA	Avg.
Retrieval Counts							
Exact Retrieval	3.247	2.288	3.126	2.702	2.440	2.496	2.717
SearchAgent-X	3.251	2.292	3.138	2.699	2.448	2.476	2.717
Output Length							
Exact Retrieval	8125	5839	7575	6840	6152	6402	6822
SearchAgent-X	8134	5847	7600	6839	6151	6382	6826

1010
1011 Figure 10: **Comparison of Different Priority Levels G on Musique dataset.** The numbers on the
1012 numbers on the X-axis represent different priority scheduling levels; N/A indicates that priority scheduling is not
1013 used.1024
1025 Figure 11: **Comparison of Different Priority Levels G on HotpotQA dataset.** The numbers on the
1026 numbers on the X-axis represent different priority scheduling levels; N/A indicates that priority scheduling is not
1027 used.

1026

1027

Table 6: System performance under different concurrency for vLLM and SearchAgent-X.

1028

1029

1030

1031

1032

1033

1034

1035

1036

1037

1038

1039

1040

1041

Cache Hit Rate

Concurrency	1	10	50	100	200	300	400	500
Cache Hit Rate								
vLLM_ANN	0.131	0.065	0.088	0.135	0.168	0.173	0.173	0.177
SearchAgent-X	0.999	0.920	0.920	0.920	0.920	0.589	0.555	0.495
End-to-End Latency								
vLLM_ANN	2181.75	633.37	477.88	448.64	437.12	425.89	415.44	411.67
SearchAgent-X	1378.82	323.21	212.05	199.29	208.07	295.39	314.80	327.87
Throughput								
vLLM_ANN	0.190	0.604	0.782	0.843	0.853	0.865	0.862	0.865
SearchAgent-X	0.203	0.848	1.348	1.517	1.582	1.137	1.100	1.060

C.4 COMPARISON OF DIFFERENT PRIORITY LEVELS G

The performance of SearchAgent-X is insensitive to its priority level setting. The priority level G mentioned in Section 3.2 is an important hyperparameter of our method. In this section, we conduct an ablation study to evaluate the performance of SearchAgent-X with different priority levels G . The evaluation is performed on the Musique and HotpotQA datasets using the Search-R1 7B model in offline scenarios. We set $G = 2, 4, 6, 8, 10$, and 12, and compare them with the baseline without priority scheduling (N/A). The results are shown in Figure 10. We note that the performance of SearchAgent-X is not sensitive to the choice of G , and all efficiency metrics (including throughput, token throughput, latency, and prefix cache hit rate) first improve and then stabilize after $G = 6$. This is expected because the average retrieval number of the 7B model is within 4 and 6, while the primary objective of priority scheduling is to distinguish requests with different retrieval numbers for effective management.

We further validate this observation on the HotpotQA dataset. As shown in Figure 11, the efficiency consistently improves as G increases (e.g., up to a $1.62 \times$ throughput gain), and stabilizes once $G \geq 6$. This again confirms that the performance of SearchAgent-X is not sensitive to the choice of G , making it straightforward to tune across different settings.

C.5 COMPARISON OF DIFFERENT CONCURRENCY SETTINGS

SearchAgent-X consistently outperforms baselines across different concurrency levels. To explore the impact of concurrency levels on system performance, we conduct an analysis to validate the effectiveness of SearchAgent-X under explicitly controlled request concurrency by tuning key parameters: iteration-level concurrency (*max_num_seq*) in vLLM (Kwon et al., 2023). The evaluation is conducted on the Musique dataset using the Search-R1 7B model in offline scenarios.

As detailed in Table 6, when increasing concurrency from 1 to 500, SearchAgent-X maintains a high cache hit rate up to around 200. Beyond that point, the cache hit rate drops more significantly. Meanwhile, system efficiency first increases then declines, showing a turning point: latency is minimized around concurrency = 100, and throughput peaks at concurrency = 200. This suggests that the GPU becomes saturated or experiences resource contention when concurrency exceeds 300. In practice, we typically choose the default value 256, which yields performance close to the optimal. The results demonstrate that SearchAgent-X consistently outperforms vLLM in both throughput and latency while exhibiting a tradeoff between cache hit rate and concurrency. We also observe that vLLM consistently suffers from low cache hit rates and lower system efficiency compared to SearchAgent-X across all concurrency levels. This is because suboptimal scheduling and retrieval under interleaved reasoning-retrieval workloads, where cache for previous tokens is frequently evicted—even with relatively small concurrency—causing inefficient reuse.

C.6 COMPARISON WITH A NAIVE EARLY-STOPPING BASELINE

SearchAgent-X aligns stopping decisions with the LLM’s prefill/decode progress, making the stop decision more timely and effective. In contrast, naive early-stopping approaches make halting

1080

1081 Table 7: Performance comparison of vanilla early stop and SearchAgent-X’s Non-Stall Retrieval.

1082

Method	Throughput (req/sec)	Latency (sec)	Cache Hit Rate	Accuracy
Vanilla HNSW	0.7872	661.85	0.412	0.15
Early Stop	0.8050	634.77	0.564	0.15
SearchAgent-X	0.9502	620.07	0.645	0.15

1083

1084

1085

1086

1087

1088

1089

Table 8: Comparison of system efficiency under the NSW search engine.

1090

1091

1092

1093

1094

1095

1096

1097

1098

1099

1100

1101

1102

1103

TopK	Method	Throughput	Latency	Token Throughput	Cache Hit Rate
3	vLLM_ENN	0.7880	395.040	97.640	0.000
	CachevLLM_ANN	1.7401	173.026	216.970	0.666
	SearchAgent-X	2.1270	129.862	265.690	0.891
4	vLLM_ENN	0.7196	578.940	99.487	0.000
	CachevLLM_ANN	1.0706	285.390	147.590	0.350
	SearchAgent-X	1.6560	172.340	227.730	0.864
5	vLLM_ENN	0.6940	428.642	87.531	0.000
	CachevLLM_ANN	0.8959	333.206	115.081	0.281
	SearchAgent-X	1.1402	241.234	146.360	0.690

1104

1105

1106

1107

1108

1109

1110

1111

1112

1113

1114

decisions based exclusively on retrieval-side signals, such as a machine learning model that predicts the “maturity” of retrieval, while disregarding the state of LLM inference. To further validate the effectiveness of SearchAgent-X’s non-stall retrieval, we conduct a comparison against a vanilla early-stop baseline that only relies on retrieval status to decide the stopping point. The evaluation is conducted on the Musique dataset using the Search-R1 7B model with Top-5 retrieval in offline scenarios. Table 7 shows SearchAgent-X outperforms vanilla HNSW and the early-stop baseline. Early-stop baseline (Li et al., 2020) boosts throughput by 2.26% and cache hit rate by 0.152, cutting latency by 27.08 sec while keeping accuracy at 0.15. SearchAgent-X further increases throughput by 18.04% and cache hit rate by 0.081, reducing latency by 14.70 sec, with accuracy unchanged at 0.15. This confirms the efficiency gains of SearchAgent-X’s non-stall retrieval via aligned LLM scheduling and retrieval.

1115

1116

C.7 SCALABILITY TO OTHER RETRIEVAL METHODS

1117

1118

1119

1120

1121

1122

1123

1124

To further validate the generalizability of our method, we implemented and evaluated it on an alternative retrieval backend, NSW (Malkov et al., 2014), which is another widely-used approximate search method. The experiments are conducted on the Musique dataset in offline scenarios using the Search-R1 7B model. As shown in Table 8, our method boosts the throughput of the NSW-based system by up to 2.7× and the cache hit rate by 2.46×, while reducing latency by 3.04×. These results confirm that our method is robust across retrieval strategies and our design is compatible with any iterative retriever, including cluster-based (e.g., IVF (Johnson et al., 2021)) and graph-based (e.g., HNSW (Malkov & Yashunin, 2018)) methods.

1125

1126

D DISCUSSION

1127

1128

D.1 APPLICABILITY

1129

1130

1131

1132

1133

Other Inference Engines Our implementation of SearchAgent-X is a direct modification of vLLM. We therefore compare against vLLM-based baselines so that all systems share the same serving stack (PagedAttention, batching logic, etc.), avoiding interference with the performance observations from unrelated optimizations. However, The proposed optimization techniques are not specific to vLLM. SearchAgent-X only assumes: 1) token-level generation scheduling, and 2) an FCFS (or

1134 FCFS-like) base policy that can be overridden. SGLang also adopts token-level generation scheduling
 1135 and an FCFS-style policy, so SearchAgent-X can be integrated into SGLang with minimal changes
 1136 for system improvement.

1137

1138 **Beyond Graph-Based ANN** For non-graph indices such as IVF or PQ-based methods, we treat
 1139 probing additional clusters or codebook entries as units of search effort and apply the same maturity-
 1140 based stopping rule; in large-scale industrial deployments it is common to tune these parameters to
 1141 trade recall against latency (Wang et al., 2021; Ren et al., 2025), making SearchAgent-X readily
 1142 applicable in such environments.

1143

1144 **Search APIs** While non-stall retrieval assumes controllable local searches, it is still possible to
 1145 initiate concurrent requests to search APIs with multiple different search parameters, and do non-stall
 1146 retrieval with an asynchronous manner as our compatible extension.

1147

1148 D.2 FUTURE EXTENSIONS

1149 **Multi-Agents** Our priority-aware scheduling and non-stall retrieval mechanisms offer a natural
 1150 starting point for coordinating across agents, for example by: 1) introducing agent-level priorities or
 1151 budgets, and 2) propagating these priorities to the shared SearchAgent-X scheduler.

1152

1153 **Other Tools** The core ideas of SearchAgent-X, including balancing retrieval accuracy and system
 1154 efficiency, priority-aware scheduling, and allocating “just enough” tool-use effort, can be applied
 1155 broadly to any scenario with multi-round, dynamically autonomous tool invocations.

1156

1157 E THE USE OF LARGE LANGUAGE MODELS

1158

1159 We use the large language model for proofreading and language refinement for this paper. The LLM’s
 1160 contribution is confined to improving grammar and prose; all research concepts and results are the
 1161 original work of the authors.

1162

1163

1164

1165

1166

1167

1168

1169

1170

1171

1172

1173

1174

1175

1176

1177

1178

1179

1180

1181

1182

1183

1184

1185

1186

1187