VentureBeat

s3: The new RAG framework that trains search agents with minimal data

Ben Dickson

@BenDee983

May 28, 2025 3:51 PM

Image credit: VentureBeat with Gemini

Researchers at University of Illinois Urbana-Champaign have introduced s3, an open-source framework designed to build retrieval-augmented generation (RAG) systems more efficiently than current methods.

s3 can benefit developers creating real-world large language model (LLM) applications, as it simplifies and reduces the cost of creating retriever models within RAG architectures.

RAG retrieval

The effectiveness of any RAG system hinges on the quality of its retrieval component. In their paper, the researchers categorize the evolution of RAG approaches into three distinct phases.

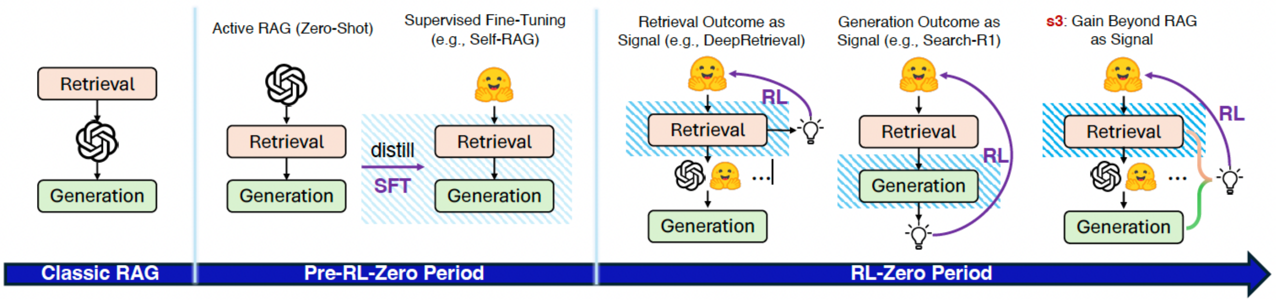
1. “Classic RAG” systems rely on static retrieval methods with fixed queries, where retrieval quality is disconnected from the ultimate generation performance. These architectures struggle with queries requiring contextual or multi-hop reasoning.

2. A subsequent phase, dubbed “Pre-RL-Zero,” introduces more active LLM participation during inference. These techniques involved multi-turn interactions, interleaving query generation, retrieval, and reasoning. However, they typically depend on zero-shot prompting and lack trainable components to optimize retrieval through direct outcome signals.

3. The most recent phase, “RL-Zero,” leverages reinforcement learning (RL) to train

models to act as search agents, improving through outcome-based feedback like answer correctness. An example is Search-R1, which trains the model to interleave reasoning with search queries and retrieved context.

Despite their advancements, existing RL-Zero approaches often optimize retrieval using search-centric metrics that ignore downstream utility. Moreover, they require fine-tuning the LLM, which is costly and error-prone. By entangling retrieval with generation, they limit real search utility and compatibility with frozen or proprietary models.



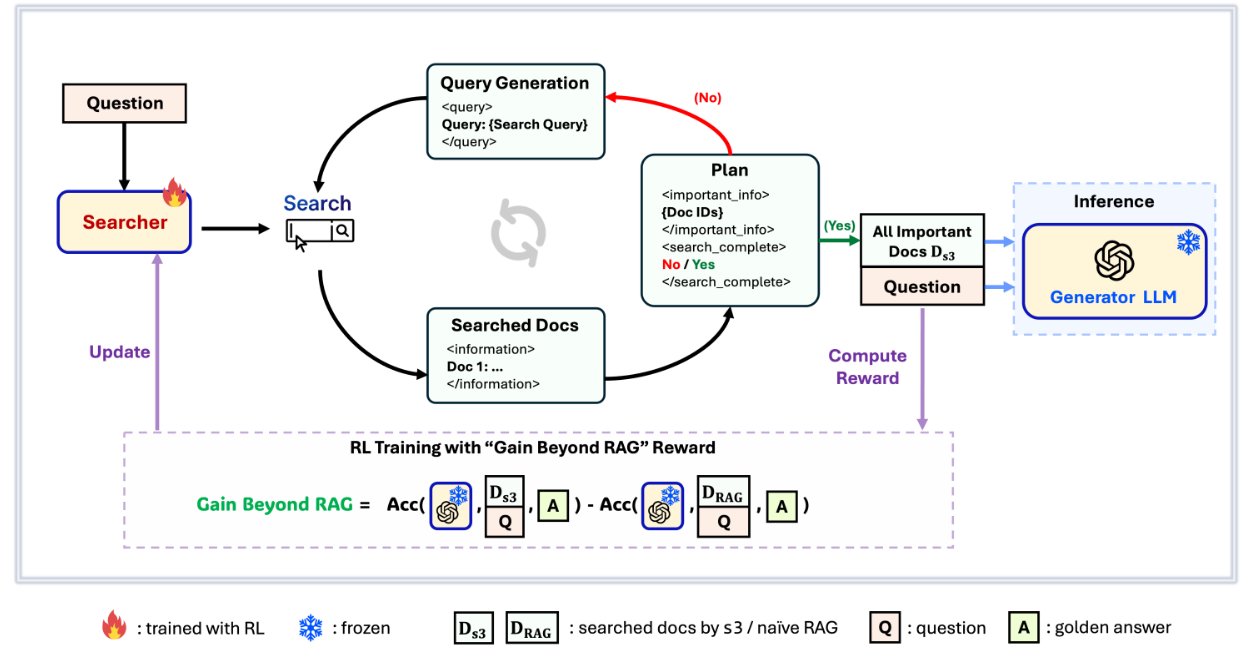
Different types of RAG Source: arXiv

As the researchers put it, “This motivates a shift toward a modular framework where search and generation are cleanly separated, and optimization focuses purely on search quality with respect to downstream utility.”

s3

The s3 framework addresses this challenge with a model-agnostic approach. The main idea is to train a search agent with structured, multi-turn access to external knowledge. This search agent improves the quality of the retrieval stage without affecting the LLM that generates the final answer.

In s3, a dedicated searcher LLM iteratively interacts with a search engine. It generates queries based on the prompt, retrieves relevant documents, selects a useful subset of evidence, and decides whether to continue searching for more information. Once the search concludes, a separate, frozen generator LLM consumes this accumulated evidence to produce the final answer.



s3 framework Source: arXiv

A core innovation of s3 is its reward signal, Gain Beyond RAG (GBR). GBR quantifies the improvement in the generator's accuracy when conditioned on documents retrieved by s3, compared to a baseline that retrieves the top documents matching the query. This reward incentivizes the searcher to find documents that truly enhance the generator's output quality.

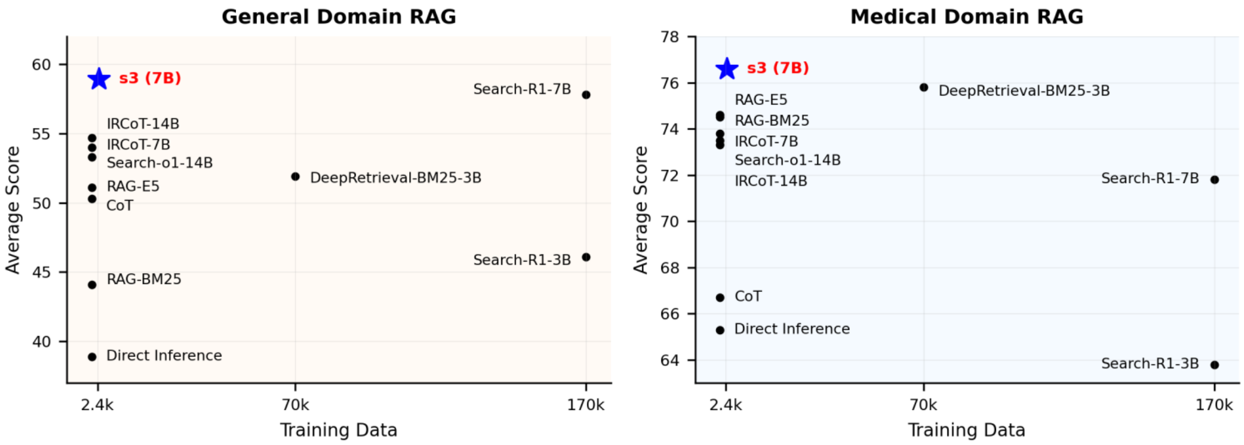
"s3 decouples the retriever (searcher) from the generator. This lets companies plug in any off-the-shelf or proprietary LLM—whether GPT-4, Claude, or an internal model—without having to fine-tune it," Patrick (Pengcheng) Jiang, lead author of the paper and doctoral student at UIUC, told VentureBeat. "For enterprises with regulatory or contractual constraints on model modification, or those that rely on closed-source LLM APIs, this modularity makes s3 highly practical. It allows them to enhance search quality without touching their generation infrastructure."

s3 in action

The researchers tested s3 across six general-domain question-answering benchmarks, comparing it against three categories of RAG systems: End-to-end fine-tuning (e.g., Search-R1), static retrieval with frozen generators (such as classic RAG pipelines) and active retrieval with frozen generators (e.g., combining documents obtained by Search-R1 with a frozen LLM). In their experiments, they used Qwen2.5-

7B-Instruct as the base model for the searcher and Qwen2.5-14B-Instruct and Claude 3 Haiku as the frozen generator LLMs.

s3 surpassed static, zero-shot and end-to-end tuned baselines on most benchmarks and achieved an average score. Its data efficiency is particularly noteworthy: s3 achieved strong gains with only 2.4k training examples, significantly less than the 70k examples required by DeepRetrieval (a static retrieval framework) or the 170k needed by Search-R1, while outperforming both in context quality and final answer performance.



s3 vs other RAG techniques Source: GitHub

“Many enterprises lack large-scale annotated QA datasets or the GPU infrastructure to fine-tune end-to-end LLM systems. s3 lowers the barrier by enabling strong retrieval performance with minimal supervision and compute,” Jiang said. “This means faster prototyping, reduced costs and quicker time-to-deployment for AIpowered search applications."

The findings suggest a fundamental shift in optimization strategy. As the researchers note in the paper, most of the performance gain in RAG stems from "improving the search capability instead of aligning generation outputs," which implies that focusing RL on search strategy rather than combined generation alignment yields better results.

Another crucial finding for enterprise applications is s3's ability to generalize to domains it has not been trained on. s3 showed zero-shot success on medical QA despite training only on general QA, suggesting that "reinforcement-learned search skills generalize more reliably than generation-tuned approaches," according to the researchers.

This cross-domain adaptability makes s3 well-suited for specialized enterprise applications that often deal with proprietary or bespoke datasets without requiring extensive domain-specific training data. This means that a single trained searcher could serve different departments (e.g., legal, HR, customer support) or adapt to evolving content such as new product documents.

"We see immediate potential in healthcare, enterprise knowledge management, and scientific research support, where high retrieval quality is critical and labeled data is often scarce," Jiang said.