STRIVEWORKS

Custom LLM-RAG Workflow Transforms Information Retrieval for Special Operations

Data science teams blended semantic search and generative AI to source difficult-to-find information and unlock new, rapid analysis.

Special Operations Data Overload Solved With Hybrid RAG System

To comb large volumes of data for specific information, analysts working for a U.S. Special Operations command introduced a hybrid retrieval-augmented generation (RAG) pipeline into their workflow. This system enabled analysts to efficiently locate and understand critical information hidden in the command's archives.

WHAT IS RETRIEVAL-AUGMENTED GENERATION?

Retrieval-augmented generation is a process for supplementing LLM workflows with an authoritative set of external data. It enables LLMs to reference proprietary data for use in responses, providing more accurate and relevant outputs.

Massive Volume of Semi-Structured Data Impeded Efficient Analysis

Analysts working for U.S. Special Operations need to locate high-value information from document repositories to support strategic decisions. These repositories contain tens of thousands of documents that often run hundreds of pages, posing a data management challenge. Open-source and commercially available data sits alongside classified data in a range of formats—some tabular, some free text, some hybrid.

The volume of data and urgency of requests overwhelm the efforts of even expert analysts to find the information they need in a timely fashion. As the command regularly adds new documents and onboards new analysts, the difficulty only multiplies.

Configuring a RAG Pipeline With the Striveworks MLOps Platform

To improve information sourcing from these documents, one Special Operations team experimented with **large language models (LLMs) and RAG**—a technique that optimizes LLMs by having them reference an authoritative data source prior to generating a response.

Working in conjunction with Striveworks, Special Operations Command developed a system that paired a **semantic search** algorithm with an open-source LLM and a RAG pipeline, creating a **hybrid workflow** that rapidly delivered specific responses to search queries.

Scan for digital version

Hybrid Search, RAG, and LLM System Empowers Analysts to Find Critical Data

Once up and running, the system allowed an analyst to submit a query in plain English and rapidly receive a generated summary of relevant information from internal documents. Each summary also included abstracts for the documents and links to source files, enabling analysts to easily verify ground truth.

The hybrid system also offered **capabilities that neither semantic search nor an LLM performed well** alone. By including full-text search capabilities, the system excelled at identifying rare or unusual names—a problem for most RAG workflows using semantic search.

The RAG pipeline also automated data management. New documents in the repository were automatically ingested, cataloged, and made available instantly for search. Modular by design, the RAG infrastructure remains decoupled from the generative LLM, which allows the system to easily upgrade to new and improved models as they become available.

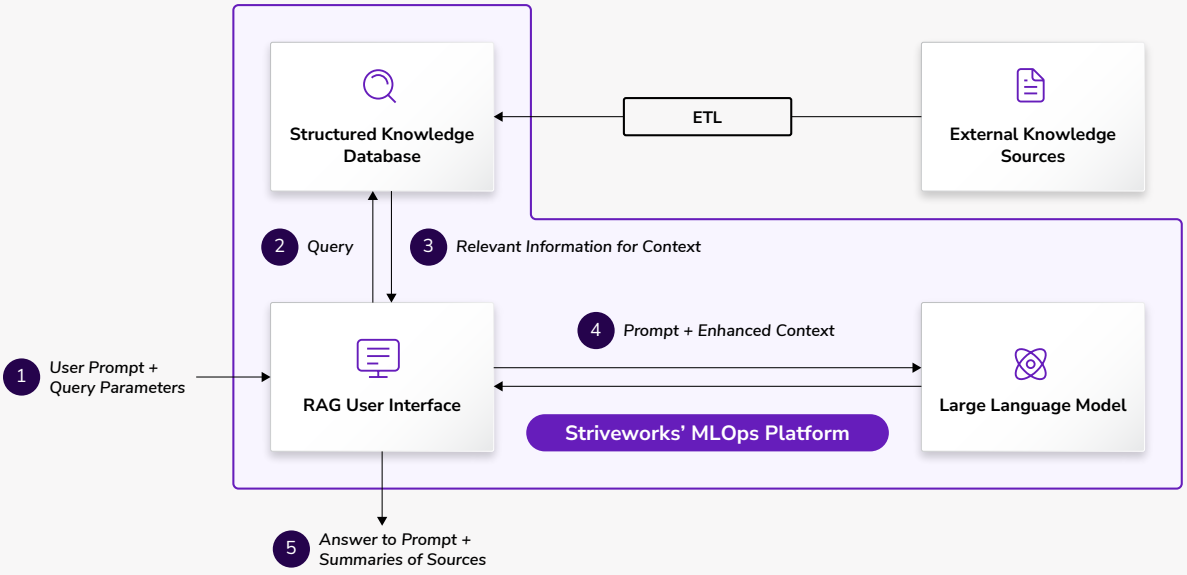
Another challenge for defense customers is **the need to work across multiple networks and classifications**; the ability to provide SOCOM users everywhere with the full capabilities of this LLM-RAG solution was a critical requirement.

Broad Use Cases for Hybrid RAG

Within weeks, the analyst team adopted full control of the Striveworks-powered system. Immediately, analysts were able to submit live queries to the system and retrieve the information they needed from their documents, neatly summarized and packaged with the original source material.

The success of the hybrid RAG approach within the defense sector establishes the potential of this approach to be used successfully across the **intelligence and commercial sectors** as well. Users needing to source private information from caches of unorganized data can incorporate full-text search as part of a RAG pipeline, greatly accelerating information retrieval and analysis.

THE STRIVEWORKS LLM-RAG WORKFLOW



STRIVEWORKS.com
sales@striveworks.com

Scan for digital version