How Retrieval-Augmented Generation gets results in the public sector

What does AI-enabled efficiency look like in the public sector?

Contents

01
The RAG advantage

What is RAG?

Retrieval-Augmented Generation (RAG) is an architecture that augments the capabilities of LLMs by adding an information retrieval system that provides grounding data, resulting in answers that are anchored in facts rather than solely relying on the model's internal training data. This grounding is crucial for providing factual, context-aware answers and reducing "hallucinations."

How does RAG work?

When a query is made or a prompt provided, a retrieval process finds the contextually relevant information from a specified collection of texts, such as a website or a document database and ranks these documents or passages. This is the retrieval step.

The retrieved documents are combined with the original query to form an enriched context. This is the "augmentation" step of RAG.

Then, a generative language model such as GPT uses the augmented context to generate a coherent and informed response.

RAG is having a moment – and it couldn't happen at a better time for public sector leaders who are being asked to do even more with less. RAG combines **traditional AI** and **generative AI (GenAI)** to provide better and more useful outputs than either can provide on their own. It provides GenAI models with the transparency, accuracy and reliability required to deliver smarter, more trustworthy answers even faster.

In practice, this means that where general-purpose AI tools draw from broad swaths of data sources to deliver insights that are helpful but not necessarily conclusive, RAG focuses on data that is vetted and often proprietary – the same types of data that are managed by public sector organizations. So, when users execute RAG-enabled queries, they can be confident that the responses they receive from the systems are based on reliable sources.

If users still have reasons to doubt or question RAG responses, every query is supported by thorough sourcing – they can simply click through to examine the source data for every finding. Ultimately, this saves countless hours spent by agency professionals who are otherwise only able to use AI results as a jumping-off point into further research and decision making.

The potential applications of RAG in the public sector are as varied as public sector organizations themselves. In this e-book, we'll share several applications in the public sector to offer a practical look at what's possible and what's already being accomplished in real-world scenarios that have a direct impact on citizen service delivery.

02 Urgent care coordination

In one European country, a national health agency deploys a central call center to respond to tens of thousands of inquiries each year from doctors and other care providers on their patients' medical histories – everything from vaccination records to medications, health charts and much more. Several thousand of these queries are urgent and involve potentially life-threatening scenarios – the faster the medical team can get the information they need, the greater the likelihood of achieving the desired patient outcomes.

Manual reckoning

In the past, call center staffers relied on a heavily manual process: A call comes in, a ticket is opened, then staffers hunt down the relevant records and respond accordingly. Much of this work was coordinated through manual spreadsheets. The shortcomings of this approach were laid bare during a historic power outage across the country, which sparked more than 1,000 calls involving critical incidents in a single day. The call center was swamped in the face of this nightmare scenario – and agency leaders resolved to find a better way.

The 40-source challenge

One of the most challenging aspects of the call center's response strategy was the large number of disparate clinical records applications – 40 in total – that staffers needed to comb through to find relevant patient information. Manually open an application, search for the information, capture it, repeat as needed across other applications, report out – that was the process.

Agency leaders needed a system to automate as much of the end-to-end research process as possible, giving call center staffers more effective tools for accessing useful, relevant data across all those applications – all to reduce response times and improve accuracy and reliability.

Enter RAG and agentic AI

LLMs had a role to play in the agency's new strategy for tapping into expansive clinical records databases – but LLMs don't solve problems on their own. In this case, they needed to be supported with other tools that could better operationalize them through automation.

That's where a combination of RAG and agentic AI tools came into play. With RAG, the agency organized and prepared the data contained in clinical records applications, making it easily accessible for queries from call center staff. Agentic AI capabilities added a critical layer of decision intelligence, automatically setting queries in motion based on a range of agency-defined parameters.

Today, an incoming medical query is automatically analyzed by an AI agent for keywords such as the patient's name, condition and specific question: "I need the full medication history for Jane Doe for the past five years." Then, another **AI agent** immediately queries the relevant databases to find the information, using RAG-enabled systems. The agent then provides the call center staffer with its findings – all sourced with clickable reference information. No manual queries required – staffers are instead able to apply their human insights to double-check that the information is reliable and present it to medical staff in ways that best meet their needs. Just as important, they can respond faster and more accurately to many more queries than ever before.

This RAG-enabled agentic AI system is already resulting in a 20% improvement in the number of tickets solved by the call center and an accompanying 8% – 15% reduction in costs.

The agency's RAG- and agentic AI-enabled system resulted in:

- 20% improvement in the number of tickets solved
- 8% – 15% reduction in costs

03
Citizen responsiveness

How can government organizations make themselves more responsive to citizen needs and input? It's not a new question – this has been a core challenge for most agencies since their inception. But it has grown more pressing as citizen expectations continue increasing, accelerated by their experiences as consumers. They want Amazon-level speed and responsiveness from the government that serves them – and public agencies have access to both the data and technology tools that they need to deliver on these expectations.

That was the problem one state's Port Authority was grappling with recently. Responsible for managing everything from airports to train stations and shipping ports, this Port Authority was on the receiving end of a steady stream of citizen feedback, arriving in the form of large volumes of unstructured data. A broken door in a train station, flooding at a bus station, a pothole at the airport, staffing problems, persistent delays – any of these issues at any port could lead to numerous emails, calls, website submissions and survey data demanding a fix.

Data deluge

While there are many aspects of successfully responding to citizen needs, data is the core driver of the entire system – and for this Port Authority, data was a problem. The agency had too much data, from too many systems, in too many formats, to be responsive in a timely manner using the data management systems it had in place. It needed a new strategy for gathering, accessing and analyzing this incoming data in short order to set a response in motion quickly.

Finding focus with RAG

Using LLMs, the Port Authority was able to better manage and access incoming citizen feedback covering a range of issues across the state's ports. But some strategic issues were more important than others for the agency – the challenge was to find a way to isolate the most relevant data. SAS' natural language processing capabilities helped identify this information, filtering for contextually relevant documents.

Once those documents were generated, the agency used a RAG-based approach to summarize the data and generate a range of possible responses. For example, staff attitude was a top issue for the Port Authority. Using RAG, the agency was able to query the data on staff attitudes at a specific port from many different angles: What were the specific complaints about staff attitude? At what time of day were these issues most acute? And so on. From there, the agency was able to cross-reference adjacent data to determine that staffing levels were insufficient at a handful of critical times – and was able to base staffing decisions on this data to improve citizen feedback on staff attitudes.

04
IoT-based predictive machine maintenance

Machine maintenance can be a significant driver of costs for organizations at all levels of government. IoT-based, AI-enabled predictive maintenance strategies already in use in the private sector can have big implications for organizations looking to reduce such costs.

For any one of the hundreds or thousands of machines operated by public organizations, there are three core data inputs that inform any maintenance intervention:

- **Current operating conditions:** How is the machine operating in real time, and are there any anomalies detected that indicate underperformance or potential failure?
- **Product information and troubleshooting guides:** If a problem is encountered, what information do we need to diagnose and solve the problem?
- **Organizational experience:** What is our experience with this machine and other identical machines?

This wide-ranging, large-volume data is typically contained in various structured and unstructured formats in different parts of the organization – real-time performance data, PDFs, vendor manuals, field service bulletins, charts, tables, emails, handwritten field notes and more.

Increasingly, it is also captured and provided by IoT sensors. Like companies in the private sector, public sector organizations can use RAG tools to pull in all this data, storing it in vector databases to make it easily accessible to AI agents tasked with managing machine learning (ML)-generated alerts from existing maintenance systems.

In practice, this means that as soon as an alert is generated, AI agents instantly swarm on it, drawing from massive volumes of raw maintenance data to present the alert in full, well-sourced context to maintenance professionals. The resulting knowledge service generates a response at the moment it is required – and the scope of the response is limited to the range of documents. For government organizations, this can result in a much more efficient, effective and proactive approach to maintaining the large numbers of machines and vehicles operating throughout their infrastructures.

Activate the knowledge system with AI agents

If the maintenance knowledge system (ML-driven alert systems using machine data, combined with LLM- and RAG-powered, encyclopedic information on the machines and their maintenance histories) is the beehive of maintenance data, AI agents are the worker bees that carry out all the tasks that keep the hive functioning.

These agents are assigned highly specific tasks that take alerts and present them to human operators with all the relevant information available to inform decision making.

How AI agents work with RAG in practice

When RAG is deployed in tandem with AI agents, government organizations gain an army of automated maintenance assistants to aid human experts. Here's a high-level overview of how it works:

1. An anomaly is detected, triggering a "call" to an agent.
2. The agent receives basic information: "Here's the anomaly, and here's some basic information about it."
3. The agent taps into the knowledge system using targeted prompts – this is a RAG query. Through the prompts, the agent can search specific collections of information, or all of them – "search product manuals," "search maintenance history" or "search all," for example.
4. The agent queries the relevant data sources and produces a response. The response is curated in a way that is meaningful, useful and understandable to technicians.
5. Once the agent is created, it can be redeployed into any workflow.

These agents are modular components that are reusable across a range of applications whenever and wherever they are needed. As a result, organizations can extract the maximum value from the time and resources required to develop these valuable agents.

05
Avian influenza monitoring

Avian influenza (also known as bird flu) is one of the most devastating risks facing the poultry industry – it spreads extremely fast and can kill entire flocks within days. All it takes is one infected chicken to set this chain of events in motion. And the impact can be felt far beyond a single poultry facility – avian influenza is a threat to human health and can have a devastating impact on the broader economy and food supply. Eggs and poultry products cost more in the wake of an outbreak, supply chains are disrupted and trade imbalances become even more pronounced.

The US has experienced outbreaks of avian influenza with increasing frequency and severity in recent years, but inspection resource allocations have not kept up. Meanwhile, there is a tremendous range of data being generated at poultry facilities across the US every hour – data that could be used to identify potential outbreaks in time to better manage them, to determine underlying factors contributing to the outbreaks and to avoid the downstream impact of an outbreak on consumers.

Working in the context of a comprehensive decision management workflow, RAG capabilities are ideally suited to this challenge. Here's how they could be put to work in support of more efficient, effective inspections.

Tap into facility-level source data

What is the current state of health of the flock? How many birds are showing signs of physical stress? What does their diet consist of, and where did the feed originate? These are the same types of questions inspectors ask when they're examining a facility, but the data could be fed directly into RAG-accessible LLMs, allowing a combination of AI, analytics and data management tools to work together to identify patterns that signal the potential for an outbreak – as well as its likely cause.

Automate the inspections decision tree

AI agents can carry out many of the routine tasks that keep the inspection process running smoothly and efficiently in a RAG environment. These agents are assigned highly specific tasks that take alerts and present them to inspectors with all the relevant information available to inform decision making.

Once developed, these AI agents can be reused or modified to extract the most value from the time and resources invested in their development. For example, they can be repurposed to assess animal type, facility type and whether action is required, as well as used to score derogatory findings and suggest biosafety actions. As a result, users can uncover issues such as contaminated feed, damaged housing that allows wild birds to enter, unclear recordkeeping and unreported euthanizations – all insights that can help prevent disease outbreaks and protect the food supply.

06 Practical and proven applications of AI

Much has been made of the gulf between AI promise and AI reality. Given the hype behind AI, this is not only unsurprising, but valid. Especially for public sector leaders, it's important to distinguish between practical, proven applications of AI and those that cannot yet be relied upon to deliver on their agencies' missions.

Each of the applications of AI included in this book uses capabilities that have already been proven in both the private and public sectors. In light of mounting pressure to deliver services even more efficiently, this is exactly the moment in which public sector IT and agency leaders should be actively seeking ways to put AI tools to work in support of advancing their missions.

SAS has a long history of helping **public sector organizations** achieve their goals, and we are widely recognized as a world leader in data and AI solutions.

Key point

In light of mounting pressure to deliver services even more efficiently, this is exactly the moment in which public sector IT and agency leaders should be actively seeking ways to put AI tools to work in support of advancing their missions.

Take your RAG capabilities to the next level with SAS Retrieval Agent Manager.

No coding, just insights. Get started at **sas.com/ram**.

To contact your local SAS office, please visit: **sas.com/offices**