Digital Government Authority
هيئة الحكومة الرقمية

Retrieval-Augmented Generation & Enabling Enterprise Innovation

5 January, 2026
Document Type: Brief Study
Document Classification: Public
Issue No.: 1.0

Contents

1. Introduction

Large language models (LLMs), such as OpenAI's ChatGPT, Google Gemini, and Anthropic Claude, are driving new capabilities in AI across industries. But they have a core limitation: these models can only generate answers based on the data they were trained on. They cannot access the latest enterprise knowledge, confidential records, or internal policies without major retraining. This becomes a serious issue in situations where decisions depend on accurate and current information, for example, when responding to regulatory inquiries, checking for compliance with internal policies, or answering staff or citizen questions based on organization-specific data.

Enterprise Retrieval-Augmented Generation (RAG) addresses this problem by retrieving relevant content from internal enterprise resources, such as databases, policy documents, or knowledge repositories, at the time a question is asked. It then combines that information with the generative power of AI to produce fact-based, context-aware responses. This avoids the need to retrain the generative AI model, which is often costly and time-consuming. Importantly, RAG systems are designed with security in mind; enterprise data remains securely hosted on internal servers and is only accessed by the AI system when needed to generate a specific response.

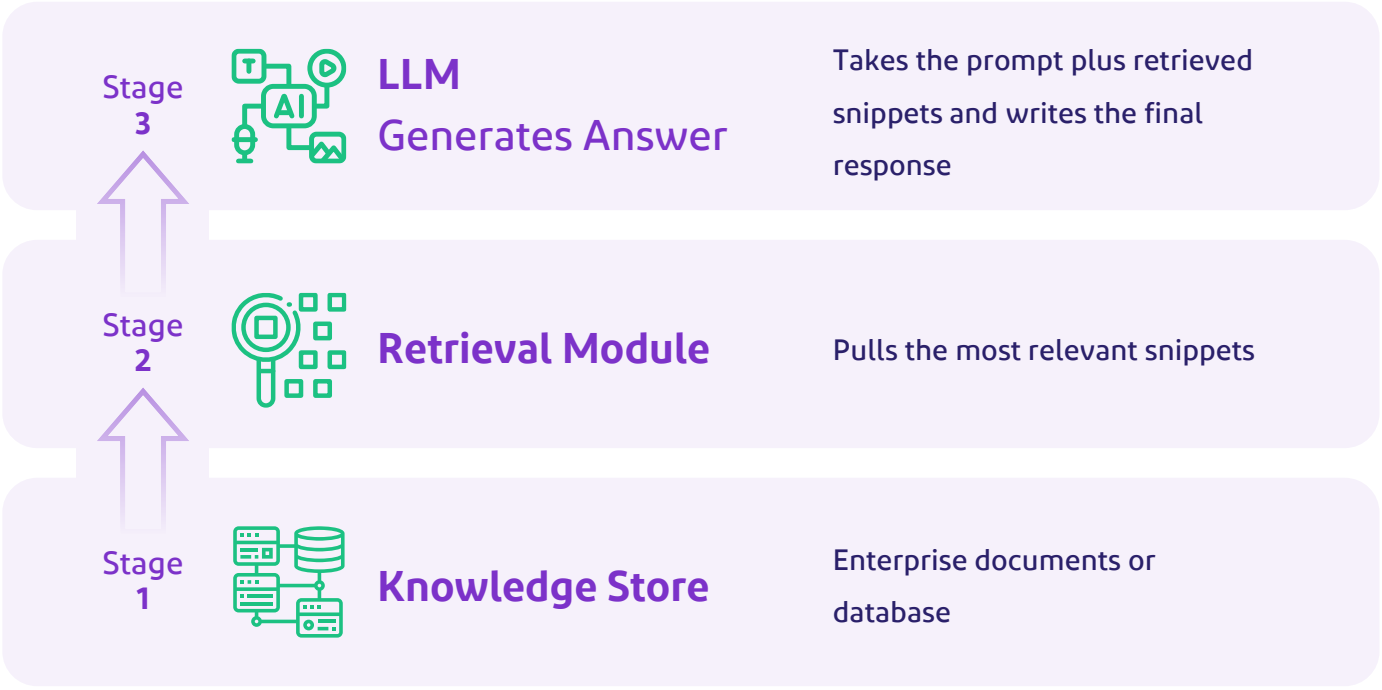

Figure 1: Enterprise Retrieval-Augmented Generation (RAG) Framework

For the public sector, the ability to deliver accurate, timely information is essential. Government agencies manage large volumes of data across laws, regulations, services, and citizen records. Enterprise RAG allows government systems to answer complex queries, assist with service navigation, and support staff with accurate internal information. This builds trust by ensuring that responses are both relevant and aligned with official sources.

In Saudi Arabia, Enterprise RAG aligns closely with national digital transformation priorities. It supports the work of the Digital Government by enabling more responsive, accurate public services. Ministries, authorities, and municipalities will be able to use RAG to provide consistent answers across digital platforms while keeping full control over sensitive information. As internal policies or documents are updated, the system automatically reflects those changes in its responses. This means staff do not need to manually update the AI or reconfigure how it works; the RAG system answers will always be based on the latest available information.

This study is primarily descriptive and conceptual, offering a high-level overview of Enterprise RAG and its relevance to public institutions. It begins by defining how Enterprise RAG works and outlining its key features, then traces the evolution of the technology, highlights global and local examples, and addresses important challenges around implementation. The final sections explore how governments and enterprises can adopt RAG in a secure and effective way, especially in support of Saudi Arabia's national digital transformation goals.

Difference Between RAG and LLMs

Traditional LLMs generate answers based only on what they learned during training, which means their knowledge can become outdated or incomplete.

RAG improves this by retrieving information from trusted external sources at the time of the query, so the answers generated by the LLM are more accurate and up-to-date.

Why Governments Need Enterprise RAG

Governments face a unique challenge: their most important data (laws, policies, and citizen records) is internal and highly sensitive. Enterprise RAG addresses this by securely linking LLMs to government data sources, allowing them to generate accurate answers while protecting confidentiality.

2. Definition of Retrieval-Augmented Generation (RAG)

Retrieval-Augmented Generation (RAG) is a method for enhancing the accuracy and usefulness of generative AI by connecting it to real-time, trusted information. According to Gartner, RAG offers a practical way to overcome one of the biggest limitations of large language models (LLMs). Tools such as ChatGPT, Google Gemini, and Mistral AI are trained on fixed datasets and cannot access private enterprise data or recent information unless they are retrained. This makes them less suitable for tasks that rely on up-to-date or organization-specific knowledge.

RAG solves this problem by retrieving relevant content from internal sources, such as databases or policy documents, at the moment a question is asked. It then uses that information to guide the AI's response. This makes the answers more accurate, more relevant, and better aligned with the needs of the organization. RAG also avoids the need for expensive and time-consuming model retraining, while keeping sensitive data securely stored within the organization.

Gartner provides a four-step model to explain how RAG systems operate:

1

Rewriting the question: The system takes the user's request and improves it so it can find the right information.

2

Searching for answers: It looks through internal knowledge sources to find the most relevant content.

3

Adding the information: The system combines the original request with the retrieved content to form a complete, well-informed prompt.

4

Generating a response: The AI uses the combined prompt to produce a clear and accurate answer.

Core Components of RAG

Retrieval

RAG begins by identifying and extracting relevant information from internal sources such as policy documents, case files, or knowledge bases. This ensures that the AI system works with verified, current content specific to the organization.

Augmented

The retrieved content is then added to the user's original query. This creates a richer prompt that gives the AI the context it needs to produce a meaningful and accurate response.

Generation

With the augmented prompt, the AI system generates a final response. The output reflects not just the model's language capabilities, but also the organization's internal knowledge and priorities.

2.1 Comparison with Related Approaches

It is useful to understand how RAG differs from other common methods used to adapt large language models (LLMs). The following table provides a simplified overview of four main approaches that organizations use to make AI systems more relevant and reliable in practice, highlighting how each one works, when it is most effective, and its main limitation.

Approach	Simplified definition	Best use	Main limitation
Prompt Engineering	Writing clear and detailed instructions to guide how the model responds, without changing how it was trained.	Useful for improving consistency in citizen-facing chatbots, such as ensuring the same greeting tone and format across digital platforms.	Fragile to wording; does not actually improve knowledge or factual accuracy.
In-context Learning	Giving the model examples within the same prompt so it imitates them when answering.	Suitable for short-lived pilots, e.g., teaching an AI assistant how to answer permit-related questions by showing a few sample Q&As.	Large prompts can be costly; cannot access new or secure internal data.
Fine-tuning	Training the model on organization data so it learns the agency's language, policies, or workflow patterns.	Ideal for mature use cases, such as automatically drafting standard reports or summarizing ministry regulations.	Expensive and time-consuming; must be retrained when policies or data change.
Retrieval-Augmented Generation (RAG)	Connecting the model to secure internal data sources (laws, circulars, reports) in real time to generate grounded and current answers.	Best for knowledge-intensive tasks, such as answering citizen inquiries about new regulations or benefits using the latest official documents.	Requires proper data governance, validation, and periodic audits to ensure accuracy.

In short, prompt engineering and in-context learning improve how the model communicates and adapts to examples, while fine-tuning helps it learn organizational tone and workflows. RAG complements all three by giving the model access to trusted, up-to-date information, making it the most practical option for government and enterprise contexts where accuracy and compliance are critical.

3. History of Enterprise RAG

The history of Enterprise RAG reflects the long progression of technologies that made it possible to connect generative AI with real-time, organization-specific information. This timeline highlights key milestones, from early chatbots and expert systems to the rise of vector search, language models, and retrieval frameworks.

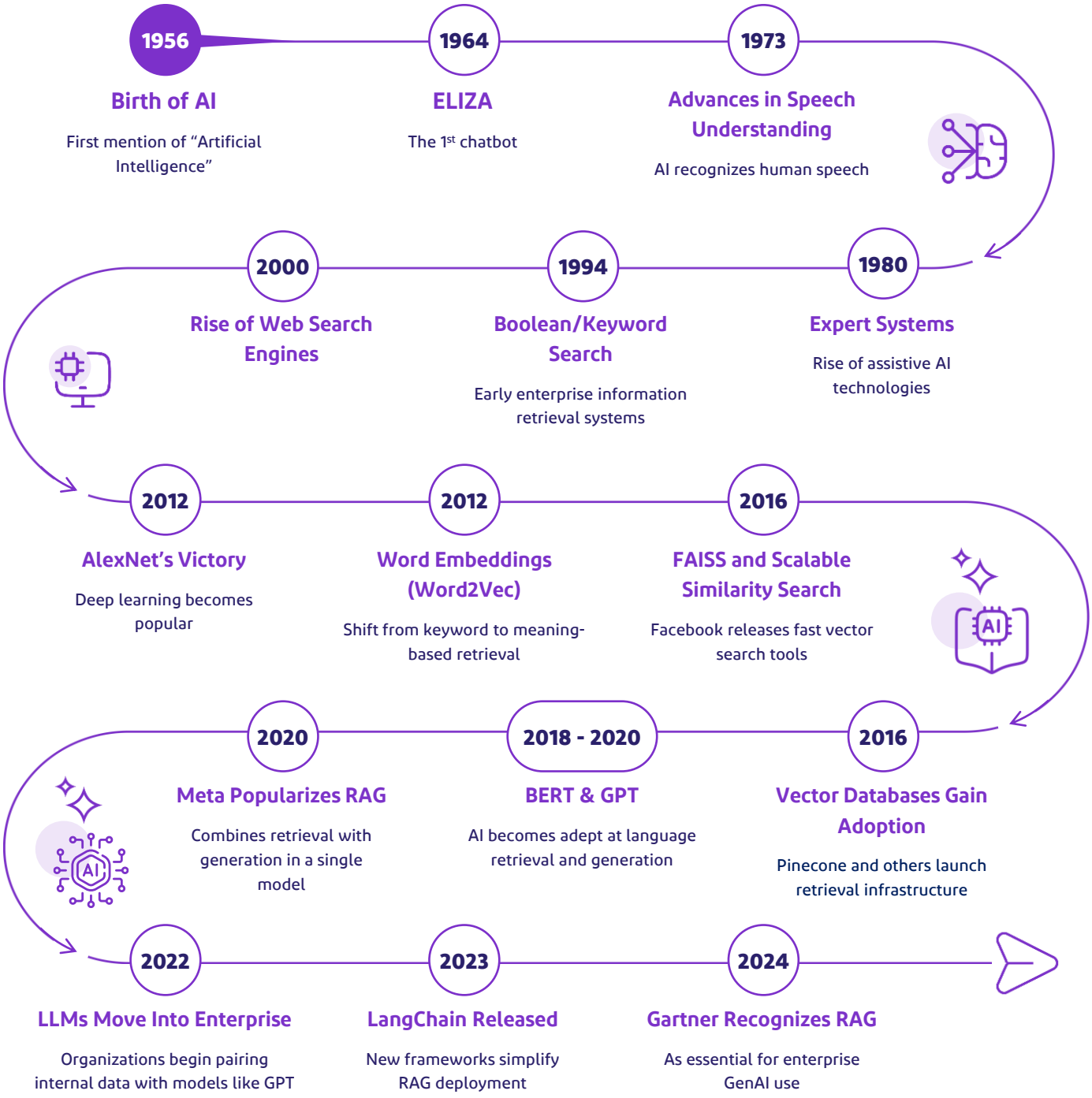

Figure 2: History of Enterprise RAG (Timeline)

The evolution of enterprise RAG reflects the convergence of information retrieval, natural language processing, and generative AI. Starting in 1956, foundational AI research laid the groundwork for computational reasoning, while ELIZA (1964) demonstrated the potential of conversational interfaces. By the 1970s, advances in speech understanding and the development of expert systems enabled machines to process structured data and support complex reasoning. In the 1990s, Boolean and keyword search matured in enterprise settings, giving rise to early information retrieval platforms. The explosion of web search in the 2000s marked a turning point, making large-scale access to digital information a public expectation.

The next major shift came with the introduction of word embeddings in 2012, enabling search systems to match meaning rather than just keywords. That same year, AlexNet's success in computer vision signaled a broader leap in deep learning. In 2014, dense vector search methods emerged, soon followed by scalable tools like Facebook's FAISS and the rise of vector databases such as Pinecone. These developments made it possible to retrieve relevant content with far greater accuracy and speed. Between 2018 and 2020, language models like BERT and GPT brought new fluency to AI systems, setting the stage for Meta's 2020 release of the first popular retrieval-augmented generation architecture.

More recently, RAG has moved from theory to enterprise adoption. In 2022, organizations began pairing large language models with internal knowledge sources to deliver tailored, secure answers. The release of LangChain in 2023 simplified RAG implementation through reusable tools and frameworks. By 2024, Gartner identified RAG as a core capability for enterprise generative AI. Today, RAG is being deployed across government, healthcare, and regulatory systems, offering a scalable solution for delivering real-time, data-grounded intelligence while maintaining full control over sensitive content.

Over the past decades, AI has evolved from static knowledge systems to dynamic, data-aware tools that generate real-time answers grounded in trusted enterprise information.

4. Significance of RAG

Retrieval-Augmented Generation (RAG) is reshaping how artificial intelligence engages with enterprise data, public systems, and global knowledge infrastructure. By combining retrieval with generation, RAG enables AI systems to deliver timely, accurate, and context-specific responses based on trusted information sources. This section explores the significance of RAG from three key perspectives: global, local, and public sector. Each view highlights how RAG is being adopted around the world, how it aligns with Saudi Arabia's digital transformation strategy, and how it is enabling more grounded, secure, and reliable AI use across sectors.

4.1 Global Perspective

RAG is gaining rapid momentum as organizations seek AI systems that are both powerful and grounded in real-world information. Globally, demand for RAG skills is rising sharply. In the United States, job postings referencing Retrieval-Augmented Generation grew by 2,047% between 2023 and 2024, making it one of the fastest-growing generative AI skills. It ranked second only to prompt engineering (+2,238%) and well ahead of tools like ChatGPT (+1,566%). This sharp increase reflects growing interest in AI that can provide accurate, context-specific answers using an organization's own data. Supporting this shift, new tools such as Anthropic's Contextual Retrieval and benchmarks like Ragnarok, CRAG, and FinanceBench are helping standardize RAG performance and extend its use to high-stakes domains like finance.

Enterprise AI adoption is accelerating worldwide. In 2024, AI use in business jumped from 55% to 78%, with 71% of companies reporting use of generative AI in at least one function. As organizations move from pilot projects to full-scale deployment, demand for RAG systems is expected to grow—driven by the need for secure, cost-effective solutions that can deliver accurate responses grounded in internal knowledge. At the same time, governments are ramping up investment and oversight. Saudi Arabia's $100 billion Project Transcendence ranks among the world's largest national AI initiatives, and AI safety institutes have been launched in the U.S., U.K., EU, Japan, and Saudi Arabia. With context windows in leading language models now reaching 2 million tokens, RAG will play a critical role in helping AI systems extract timely, relevant insights from increasingly large and complex sources of information.

These developments show that RAG is no longer a niche capability: it is a global standard for delivering relevant, explainable, and up-to-date AI responses. As both governments and enterprises embed generative AI into daily operations, RAG provides the foundation for accuracy, trust, and responsible use at scale.

4.2 Local Perspective

As Saudi Arabia advances its National Digital Government Strategy (NDGS), Enterprise RAG stands out as a strategic enabler for knowledge-driven government. It enhances the value of generative AI by grounding outputs in internal policies, regulations, and public sector data. These systems can support informed decision-making, reduce response times, and deliver consistent answers across services. When aligned with NDGS priorities, RAG can help accelerate progress on public sector modernization and digital excellence.

Citizens Satisfaction

Enterprise RAG improves citizen services by enabling accurate, real-time responses grounded in official data and policy. Applications include 24/7 chatbots for license renewal, healthcare support, and education services. By retrieving answers directly from trusted government sources, RAG reduces confusion, shortens wait times, and ensures information is consistent, accessible, and up to date.

Enabling Businesses Sector

Enterprise RAG supports business enablement by delivering timely, policy-aligned answers to regulatory, licensing, and service-related queries. It empowers entrepreneurs and SMEs by simplifying access to government resources and compliance information. This helps streamline operations, reduce administrative overhead, and create a more transparent environment for economic participation.

Effective Governance

Enterprise RAG strengthens governance by making internal knowledge quickly accessible to decision-makers and staff. It supports AI assistants that retrieve official procedures, policy history, and legal references on demand. This helps ensure consistency in interpretation and improves the quality of decision-making. In addition, it enhances responsiveness across ministries and authorities.

Effective Investment

Enterprise RAG maximizes the value of national data assets by enabling AI systems to surface insights from government records, research outputs, and strategic documents. By improving access to structured and unstructured information, RAG supports better investment planning, evidence-based policy, and more transparent public-sector project evaluation.

Regulation And Compliance

Enterprise RAG helps government bodies enforce and monitor regulations more effectively by grounding AI tools in up-to-date legal and compliance documents. It supports automated assistants that answer complex regulatory questions with precision, helping staff and stakeholders align with evolving standards. This enables faster and more accurate enforcement while maintaining fairness and clarity.

Expedited Transformation

Enterprise RAG accelerates digital transformation by enabling scalable AI tools that can instantly access and explain internal government knowledge. Whether deployed in service centers, analytics platforms, or internal portals, RAG makes institutional knowledge more usable. This strengthens interoperability, reduces silos, and speeds up the rollout of data-driven public services.

4.3 Advantages of Integrating RAG into the Public Sector

Supporting Equitable and Personalized Public Services

RAG helps tailor service delivery based on specific user needs and regional conditions by retrieving information that reflects local contexts. It also ensures that underserved populations receive accurate guidance, even in remote or complex service environments.

Example: In rural health clinics, RAG can assist with surfacing appropriate guidance for remote consultations, based on current Ministry of Health protocols.

Delivering Trusted and Up-to-Date Public Services

Enterprise RAG enables government platforms to generate accurate responses drawn directly from internal records, regulations, and service documents. This ensures that citizens receive answers based on current policies—not outdated or generalized content. By grounding outputs in trusted data sources, RAG helps increase public confidence in digital government tools.

Example: RAG can be used to enhance platforms like “Absher” or “Tawakkalna”, enabling them to respond to complex citizen queries with personalized, policy-based answers..

Empowering Government Employees with On-Demand Knowledge

Public sector staff often need quick access to detailed internal information, from HR policies to legal protocols. RAG systems allow employees to ask questions and receive responses sourced from organizational documents, without the need to manually search through files. This improves efficiency and consistency in internal decision-making.

Example: Administrative staff in ministries or health centers could use RAG-based assistants to retrieve up-to-date procedures, forms, or compliance rules instantly.

Strengthening Regulatory Compliance and Transparency

By grounding outputs in the latest legal texts and internal controls, RAG helps institutions ensure consistent application of rules and avoid misinterpretation. This supports regulatory bodies in enforcing standards more efficiently, with clear, explainable answers.

Example: A financial regulator could use a RAG-powered assistant to answer questions about updated compliance frameworks or reporting timelines, ensuring consistency across teams.

Accelerating Digital Transformation with Secure, Scalable AI

RAG systems can be deployed securely within existing government IT environments, including sovereign clouds and on-premise servers. This aligns with national goals for data privacy and sovereignty. By avoiding the need to retrain models, RAG also reduces costs and accelerates implementation timelines.

Example: SDAIA's Deem Cloud or the Unified Government Cloud can host RAG-enabled systems that scale across ministries while maintaining full control over sensitive information.

Unlocking the Value of Government Data Assets

Government institutions hold large volumes of internal data across sectors. RAG makes it easier to unlock insights from these data sources by combining them with AI's language capabilities. It helps transform unstructured archives into actionable knowledge, improving planning and analysis.

Example: Urban development teams could use RAG to retrieve zoning regulations, project records, or demographic studies to support smarter infrastructure decisions.

5. Opportunities for Using RAG to Support Digital Transformation in the Kingdom

Building on the significance of RAG, particularly the local perspective outlined previously, this section highlights practical opportunities for integrating RAG into Saudi Arabia's public sector. It identifies key sectors where RAG offers immediate benefits through secure knowledge retrieval, real-time AI support, and improved operational efficiency.

Public Administration

- **Opportunities:**

Enterprise RAG offers a high-impact solution for public administration by enabling secure, real-time access to internal knowledge across ministries, authorities, and agencies. Through systems like SDAIA's DEEM cloud platform, government entities can deploy RAG models locally, ensuring full control over sensitive information while scaling knowledge access across departments. Platforms such as the Saudi Data Portal, which hosts over 7,000 high-quality public datasets, provide a strong foundation for document-grounded AI systems that deliver reliable answers to staff and citizens. With over 100 government entities already pursuing AI adoption, RAG can help unify data usage, eliminate information silos, and reduce time spent navigating complex policy or procedural content.

- **Impact:**

The potential for RAG to improve internal efficiency is substantial. By retrieving answers directly from approved internal sources, RAG systems reduce errors, accelerate response times, and support better-informed decision-making. These capabilities help build a more agile and consistent public sector, especially as Saudi Arabia continues to rank among the global leaders in digital government. With over 320 systems connected to the National Data Bank (NDB) and more than 100 terabytes of government data now accessible, RAG provides a scalable, low-risk way to turn information into action—helping ministries meet national transformation goals while enhancing trust and transparency.

Healthcare

- **Opportunities:**

Enterprise RAG can significantly enhance healthcare delivery by connecting AI systems to up-to-date clinical protocols, internal guidelines, and public health documentation. Platforms such as Sehhaty and the Seha Virtual Hospital already serve millions of users and can be further strengthened with RAG-enabled capabilities that provide grounded, real-time responses to patient and provider queries. Saudi Arabia's Remote Health Program, which facilitated over 2.4 million virtual consultations in 2023, illustrates the growing demand for fast, accurate information at the point of care. With over 90% of public hospitals digitally connected and unified data platforms in place, RAG offers a natural extension to surface verified content during clinical workflows or public interactions—without compromising data privacy.

- **Impact:**

RAG helps healthcare staff retrieve answers instantly from internal protocols, policy updates, and documentation, saving time and improving accuracy in decision-making. This is especially valuable in remote or high-demand settings, where access to centralized information can reduce service delays. In support of the Ministry of Health's focus on equity, personalized medicine, and digital upskilling, RAG enables tailored support for clinicians, administrative staff, and citizens—while ensuring that guidance is always aligned with the latest national health standards. As digital health expands to reach 97% of population clusters, RAG systems will be essential to scale secure, reliable access to knowledge across the Kingdom.

Enterprise RAG also holds promise for sectors like tourism, energy, and public safety, where frontline decisions depend on rapid, context-specific access to official information. It enables AI systems to deliver precise answers grounded in sector-specific data, policies, or infrastructure records. With support from the Digital Government Authority, RAG can help ensure these services are reliable and aligned with national performance standards.

6. Conceptual Framework and Critical Success Factors for RAG Deployment

Enterprise RAG brings real benefits when it is implemented with care. By using RAG, organizations can extract more value from their internal data while maintaining full control over sensitive information. However, to move from experimentation to practical impact, they need structured strategies for deployment. Realizing the full potential of Enterprise RAG requires a clear framework that supports secure implementation, high-quality data management, accurate search performance, and user trust. Successful deployment also depends on tailoring systems to specific use cases, ensuring relevant information is retrieved effectively, and delivering clear, explainable responses. As AI becomes more integrated into decision-making, RAG provides a practical path to scale these tools in a responsible and strategic way.

Conceptual Framework and Critical Success Factors

1 Technology Description

Enterprise RAG connects generative AI with live enterprise data, allowing systems to produce responses based on current, internal information rather than static, or outdated, training data.

2 Adoption & Activation

RAG should be integrated into high-value applications such as public service delivery, compliance checks, and staff support tools. Focus on responsible rollout across departments.

3 Potential Benefits

Improved decision-making, faster access to trusted information, and reduced risk of misinformation. RAG strengthens knowledge management and enhances public sector performance.

4 Expected risks

RAG systems must address challenges related to data security, information accuracy, and relevance. Poorly managed RAG can return outdated or misleading responses.

5 Critical Success Factors

Success depends on curated knowledge bases, effective search optimization, clear prompts, and human oversight to maintain reliability and accountability.

6 Institutional Empowerment

Ongoing investment in AI infrastructure, content preparation, and secure data environments is needed to scale RAG across government and enterprise systems.

This structured roadmap provides a foundation for deploying RAG effectively, with safeguards that protect data and frameworks that support real-world use. By adopting these principles, Saudi Arabia can lead in secure and strategic AI adoption, reinforcing its position as a global innovator in digital government and knowledge-driven transformation.

6.1 KPIs to Track RAG Implementation Success

To ensure that the adoption of RAG delivers measurable value, organizations should track its impact using clear performance indicators. These Key Performance Indicators (KPIs) help decision-makers evaluate whether RAG is improving accuracy, efficiency, and citizen satisfaction in digital services. The following table provides suggested KPIs that can be used as a start and be adapted to each organization's use cases.

KPI	Description	Example Metric
Response Accuracy	Measures how correct and reliable the system's answers are	% of responses validated as accurate by human reviewers
Response Speed	Evaluates how quickly the system provides answers	Average response time in seconds
Citizen Satisfaction	Tracks overall user experience with digital services	% of citizens rating answers as "satisfactory" or higher
Adoption Rate	Monitors how widely the system is used in the organization	# of queries handled through RAG vs. traditional channels
Cost Efficiency	Assesses savings in time and resources	Reduction in staff workload hours

7. Challenges and Considerations for Enterprise RAG Deployment

While Enterprise RAG offers clear benefits for making generative AI more accurate and useful, its successful adoption depends on addressing key operational challenges. From ensuring secure access to internal data to avoiding misinformation, RAG systems require careful design and oversight. Saudi Arabia's government entities already manage over 100 terabytes of data across 320+ connected systems, with 8,700+ datasets available through platforms like the National Data Bank and Saudi Data Portal. As this information becomes more central to AI workflows, deploying RAG in a secure, transparent, and reliable manner is essential. This section highlights common risks and offers practical guidance to help organizations integrate RAG responsibly and in line with national digital goals.

1. Data Source Reliability

Key Concern

RAG systems generate answers based on the documents they retrieve, which makes the accuracy and quality of those sources critical. If a system pulls outdated, inconsistent, or low-quality internal content, the generated response can be misleading. In high-stakes sectors like healthcare or law, this can erode trust and lead to poor decisions. Globally, 71% of organizations using GenAI cite "reliable data access" as a key deployment challenge (AI Index 2025).

Mitigation

Organizations should maintain curated, up-to-date content repositories with version control and archiving policies. Before deployment, RAG systems should undergo content audits to ensure key policies and procedures are indexed properly. In Saudi Arabia, integrating with centralized resources such as the Saudi Data Portal or sector-specific platforms like "Sehhaty" can help ensure source quality. Ongoing validation workflows are also needed to ensure retrieved material remains relevant..

2. Retrieval Accuracy and Hallucination Risk

Key Concern

If a RAG system retrieves irrelevant or weakly related documents, it may generate convincing but incorrect responses—commonly referred to as hallucinations. In sensitive domains such as law or healthcare, this undermines trust, introduces compliance risks, and can damage institutional credibility. According to Anthropic's 2024 benchmark data, retrieval or grounding errors still account for a significant portion of factual inaccuracies in RAG deployments.

Mitigation

Use retrieval quality metrics such as precision@k and human evaluation to test system outputs before rollout. Prioritize implementation of hybrid search (semantic + keyword) and fine-tune ranking algorithms for enterprise context. Responses should always cite source documents to allow verification. Consider deploying internal "RAG moderators"—AI tools or human reviewers—that assess confidence scores before surfacing responses in high-risk settings..

3. Content Drift and Maintenance

Key Concern

RAG systems rely on external documents that may change over time. When internal policies, laws, or procedures are updated, outdated content may continue to be used unless the retrieval index is refreshed. This poses compliance risks and can lead to inconsistencies across departments. In a government context with frequent regulatory updates, this is a growing concern.

Mitigation

Establish automated re-indexing of content sources whenever updates occur. Connect the RAG pipeline to document management systems with change detection, so new material is indexed promptly. Encourage staff to flag outdated content. SDAIA's push for centralized cloud services (e.g., Deem Cloud) creates a strong foundation for coordinated content maintenance across entities.

4. Secure Access and Data Protection

Key Concern

RAG systems interact with internal and sensitive enterprise data. Without security controls, there is a risk of unauthorized access or data leakage through generative outputs. According to the AI Index Report 2024, privacy and security remain the top global concerns in AI deployments, especially in government and healthcare sectors.

Mitigation

Deploy RAG systems within secure government environments such as SDAIA's Deem Cloud. Apply role-based access controls and encryption for document access. Ensure that RAG models do not retain prompts containing sensitive data. Align with global standards like ISO 27001, NCA ECC, and GDPR-equivalent protections for public data handling.

5. User Preferences and Institutional Adoption

Key Concern

Even when technically sound, RAG systems may fail if they do not align with user needs. Complex interfaces, irrelevant results, or a lack of transparency can frustrate employees and reduce adoption. Inconsistent outputs can also lead to resistance or reliance on outdated manual processes. Without usability, the system's impact is limited.

Mitigation

Use human-centered design principles to develop intuitive interfaces with clear prompts, transparent citations, and feedback options. Conduct structured user testing with government staff to refine functionality before deployment. Embed RAG into existing workflows, such as helpdesks, service portals, or internal dashboards.

6. Data Bias and Legal Compliance

Key Concern

RAG systems can raise legal and ethical issues if they misuse sensitive data or generate biased outputs. For example, if a government health portal retrieves outdated medical records, it may disclose private patient data without consent. Similarly, if the system produces biased advice (e.g., prioritizing a group of citizens over another), this can undermine fairness, erode trust, and expose the institution to ethical challenges.

Mitigation

Organizations should set clear rules for how citizen data can be used and ensure all retrieval complies with privacy laws. Regular fairness audits can detect bias in system outputs, and accountability frameworks should define who is responsible when errors occur. Independent oversight committees and staff training help ensure RAG systems are applied responsibly in sensitive areas like healthcare or social services.

8. General Recommendations

To fully realize the potential of Enterprise RAG, strategic action is needed to embed this capability into Saudi Arabia's public sector in a secure and scalable way. As the Kingdom accelerates digital transformation through national initiatives led by the Digital Government Authority (DGA) and Saudi Data & AI Authority (SDAIA), this section outlines key recommendations that provide a clear pathway for integrating RAG into government systems.

Establish National Infrastructure for RAG Deployment

Action: Develop a unified national framework for RAG deployment across government entities, supported by dedicated infrastructure such as SDAIA's Deem Cloud. This includes setting standards for retrieval indexing, content validation, security protocols, and data access management.

Impact: Provides public sector organizations with a structured way to implement RAG, supporting the secure scaling of RAG capabilities across ministries and sectors.

Upskill the Public Sector for RAG-Driven Workflows

Action: Develop RAG-specific training programs in partnership with national academic institutions. Equip staff to understand how RAG systems work, validate their outputs, and integrate them into daily operations. Include modules on source citation, prompt formulation, and ethical use of generative AI tools.

Impact: Builds institutional readiness and ensures responsible use of RAG systems; minimizes risks from misinterpretation or overreliance on automated responses.

Integrate RAG Governance into Digital Government Standards

Action: Embed RAG guidelines into existing digital governance frameworks under the oversight of the Digital Government Authority. This includes defining policies for document versioning, indexing frequency, and cross-entity coordination. Establish audit mechanisms for continuous monitoring and improvement.

Impact: Supports trust, consistency, and excellency by setting clear expectations for RAG use across public institutions..

Institutionalize Human-in-the-Loop (HITL) Evaluation

Action: Introduce a national framework for human review of RAG system outputs. Under this approach, automated answers in critical sectors (such as healthcare, law, or citizen services) are reviewed by trained government staff before publication. Reviewers confirm that responses are accurate, relevant, and phrased appropriately for the public. Clear workflows should define when human review is required, how feedback is recorded, and how recurring errors are corrected in the system.

Impact: Builds public trust by ensuring human oversight where accuracy and sensitivity matter most. Reduces the risk of misinformation or biased content, and helps agencies continuously improve their RAG models through structured and documented feedback.

Establish a National RAG Governance Framework

Action: Create a centralized national framework to govern the use of RAG systems across ministries and public agencies. This should define clear policies for data access, model transparency, privacy protection, and content validation. The framework can be led by the Digital Government Authority (DGA) in coordination with Saudi Data & AI Authority (SDAIA) to ensure consistent standards, certification processes, and compliance monitoring across the government ecosystem.

Impact: Ensures consistency, accountability, and responsible use of RAG technology at the national level. Builds citizen trust by enforcing unified governance rules that protect data integrity, enhance transparency, and promote ethical use of generative AI and TAG in public services.

9. Conclusion

The Enterprise Retrieval-Augmented Generation (RAG) technology is redefining what generative AI can achieve in the government sector. By enabling systems to operate using internal government data in real time, these systems become more effective in regulatory, administrative, and service contexts. As demonstrated in this study, RAG serves as a strategic enabler, helping government entities fully harness the value of their data while reducing the risks of inaccurate or outdated responses.

This study highlighted how RAG technology can enhance data utilization within government agencies without compromising sensitive information privacy. By retrieving content from trusted sources, the technology enables AI systems to deliver more accurate responses. By establishing application standards, these entities can play a pivotal role in accelerating the adoption of retrieval-augmented generation.

The recommendations presented in this study outline a clear path for deploying the technology safely and effectively. For instance, by building shared infrastructure, training government employees, and integrating RAG into digital governance standards, the digital government can launch AI systems capable of quickly retrieving relevant documents and accurately answering agency-specific queries. These recommendations reinforce the objectives of the national digital government strategy and position the technology as a key component of knowledge-based public service delivery.

This focus aligns with the global expansion of generative AI, which has increased the need to ensure that such systems are accurate, transparent, and grounded in up-to-date institutional knowledge. RAG technology directly addresses this need. Global demand for related skills has surged dramatically, with mentions in U.S. job postings increasing by over 2,000% in a single year. The digital government occupies a leading position in this field, supported by major national initiatives such as the $100 billion “Project Transcendence” and large-scale government platforms like the “Deem Cloud.”

By adopting explainable AI and prioritizing dedicated data accessibility in human-centered AI initiatives, the digital government can build on this success, enabling citizens to better understand government data and fostering greater digital engagement.

As governments enter a new phase of AI adoption, the entities that lead the integration of AI with institutional knowledge and updated data will determine the trajectory of public service evolution. By establishing retrieval-augmented generation as a national standard, the digital government can reinforce the Kingdom’s leadership in digital governance and contribute to shaping global best practices. The authority’s role extends beyond driving internal transformation—it influences global thinking about AI reliability and its societal impact.

At its core, retrieval-augmented generation is not merely about improving system performance but about enhancing governments’ ability to serve beneficiaries—citizens, residents, and visitors alike. As the digital government seeks to accelerate the adoption and integration of this technology within its infrastructure, new opportunities arise to deliver more precise and efficient public services. The essence of these opportunities lies in the technology’s ability to support decision-making based on reliable and updated institutional knowledge, in line with the Kingdom’s vision of an effective, world-leading digital government.

10. Definitions

Term	Definition
Enterprise RAG	A method for enhancing generative AI by retrieving internal, real-time content from enterprise sources (such as databases or documents) to improve the accuracy and relevance of AI responses.
Retrieval Module	The component of a RAG system that searches and selects relevant content from knowledge repositories to guide AI responses.
Prompt Augmentation	The process of combining the original user query with retrieved content to create a richer prompt that provides the AI system with context.
Content Drift	The risk that outdated or revised documents remain in a RAG system's index, leading to inaccurate or obsolete AI responses.
Document Grounding	The practice of ensuring AI outputs are based on specific, verifiable documents or data sources, enhancing transparency and trust.
Deem Cloud	A sovereign cloud platform developed by Saudi Data & AI Authority (SDAIA) that hosts secure government AI applications.

11. Bibliography

1. Stanford Institute for Human-Centered Artificial Intelligence (HAI). (2025). Artificial Intelligence Index Report 2025. Stanford University. https://hai.stanford.edu/assets/files/hai_ai_index_report_2025.pdf

2. Stewart, D. (2024, May 8). Getting started with retrieval-augmented generation (ID G00811814). Gartner.

3. Weizenbaum, J. (1966). ELIZA – A computer program for the study of natural language communication between man and machine. Stanford University. https://web.stanford.edu/class/cs124/p36-weizenabaum.pdf

4. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems (pp. 1097-1105). https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

5. Meta AI. (n.d.). FAISS: Facebook AI similarity search. Retrieved from https://ai.meta.com/tools/faiss/

6. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Riedel, S. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems (NeurIPS 2020)

7. Saudi Data & AI Authority (SDAIA). (2024, September). State of AI in Saudi Arabia. Global AI Summit. https://globalaisummit.org/Documents/StateofAIinSaudiArabia.pdf

8. Health Sector Transformation Program. (2024). Health Sector Transformation Report 2024. Vision 2030, Kingdom of Saudi Arabia. https://www.vision2030.gov.sa/media/h0yb5d03/health-sector-transformation-report-2024.pdf

For continuous development purposes, please fill out the following survey

Survey

SCAN HERE

For more digital studies, please follow the link below

Digital Studies

SCAN HERE

هيئة الحكومة الرقمية
Digital Government Authority

DGA.GOV.SA