# Intel AI for Enterprise RAG

Unlocking Business Potential with GenAI

# Redefining enterprise with AI

Generative AI (GenAI) has revolutionized the AI landscape, pushing enterprises to rapidly embrace this technology for competitive advantage. Large models trained on vast datasets have demonstrated incredible potential for general-purpose use cases. These advancements have helped organizations enhance productivity, introduce new products, and improve operational efficiency.

TODAY

**AI Assistants**

AI augmented humans

- Code generation
- Enterprise search, Knowledge Discovery
- Defect & Fraud Detection

EMERGENT

**AI Agents**

AI automated domain workflows

- Customer service/support
- Enterprise Assistants
- Security & automation agents

FUTURE

**AI Functions**

Hybrid human & AI custodian enterprising

- Realtime process/product iteration
- Finance & Health diagnostics
- Inventory & supply chain management

## We are at the threshold of three key eras of AI-driven enterprise transformation:

### Today: AI Assistants

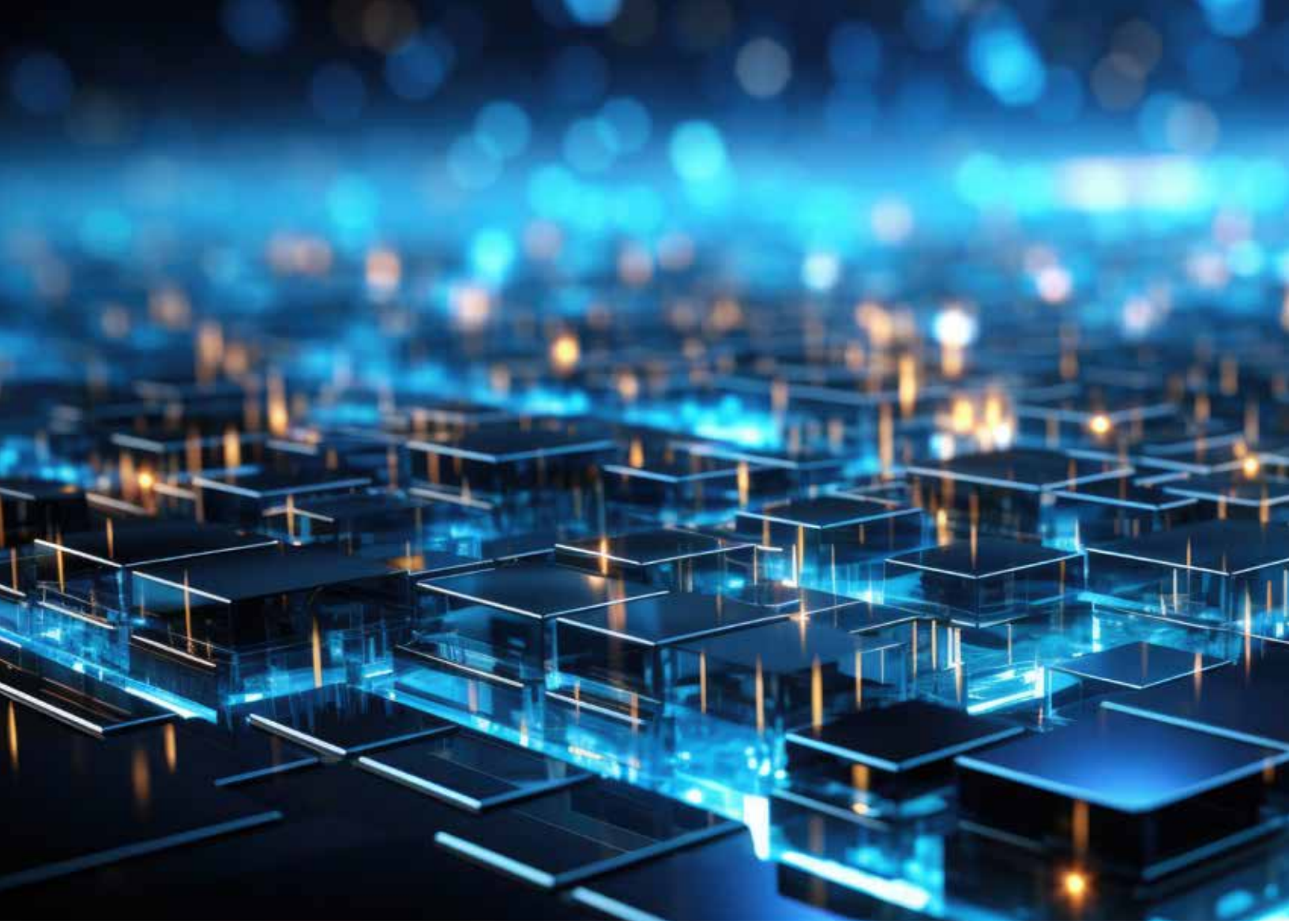
In the current stage, AI is augmenting human efforts through AI Assistants. These assistants provide support in areas like code generation, enterprise search, and defect and fraud detection. Here, AI empowers individuals, allowing them to be more efficient and effective by handling repetitive tasks and offering valuable insights.

### Emergent: AI Agents

The next era involves the rise of AI Agents—automated systems that independently manage domain workflows. AI Agents are starting to revolutionize customer service and support, and can function as enterprise assistants or security agents, streamlining complex processes. They are not merely assisting humans but actively taking over specific tasks, resulting in substantial efficiency gains for businesses.

### Future: AI Functions

Looking ahead, AI Functions will automate complex enterprise-level outcomes. This will lead to a hybrid human and AI enterprise, where intelligent systems work alongside humans to manage high-level tasks such as finance and health diagnostics, inventory, and supply chain management. This transformation will redefine how enterprises operate, leading to more responsive, adaptive, and efficient systems.



## Unlocking potential with GenAI and RAG

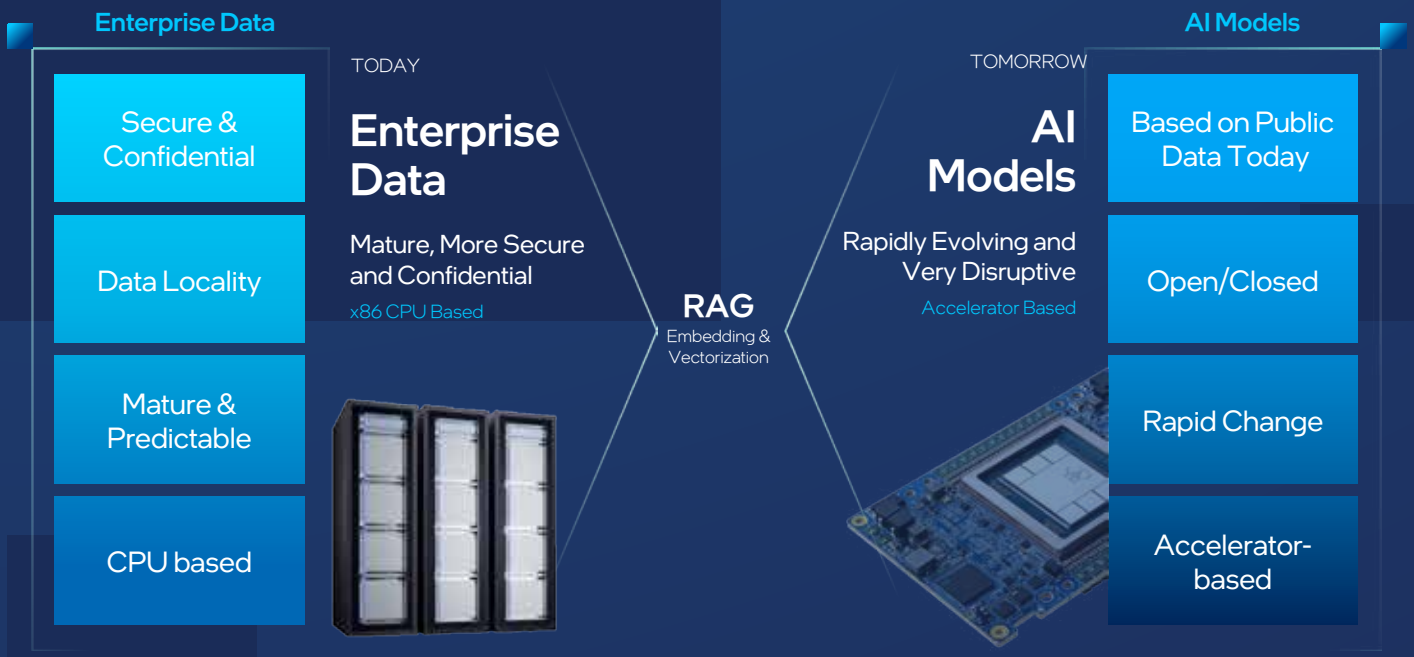
To fully leverage the potential of AI, enterprises need models that are not only powerful but also specific to their needs. While third-party large models are excellent for broad, general-purpose applications, most enterprise use cases require customization. Fine-tuning these models with proprietary data can provide more relevant results, but it also demands considerable resources. Moreover, the integration of proprietary data into these models introduces potential security risks, making it crucial for enterprises to prioritize robust data protection and governance measures throughout the customization process.

This is where retrieval-augmented generation (RAG) comes in. RAG allows enterprises to enhance pre-trained GenAI models using their own data, without the need for costly retraining or the risks of sharing data with external services. By incorporating domain-specific information, RAG creates customized LLMs that can securely generate business-relevant insights, thereby ensuring that the AI outputs are tailored to the enterprise's specific context. With Intel's optimizations and platform support, RAG offers a practical and scalable path for enterprises to achieve meaningful AI-driven outcomes—whether through today's AI assistants, emerging AI agents, or future AI functions.

# What is retrieval-augmented generation?

Foundation models are powerful tools for general-purpose applications. However, without integrating enterprise-specific and domain-specific data, they often fail to provide the most value for business needs. Retrieval-augmented generation (RAG) is an AI framework designed to bridge this gap by enriching Large Language Models (LLMs) with real-time, dynamic data from private sources securely, without requiring retraining or fine-tuning.

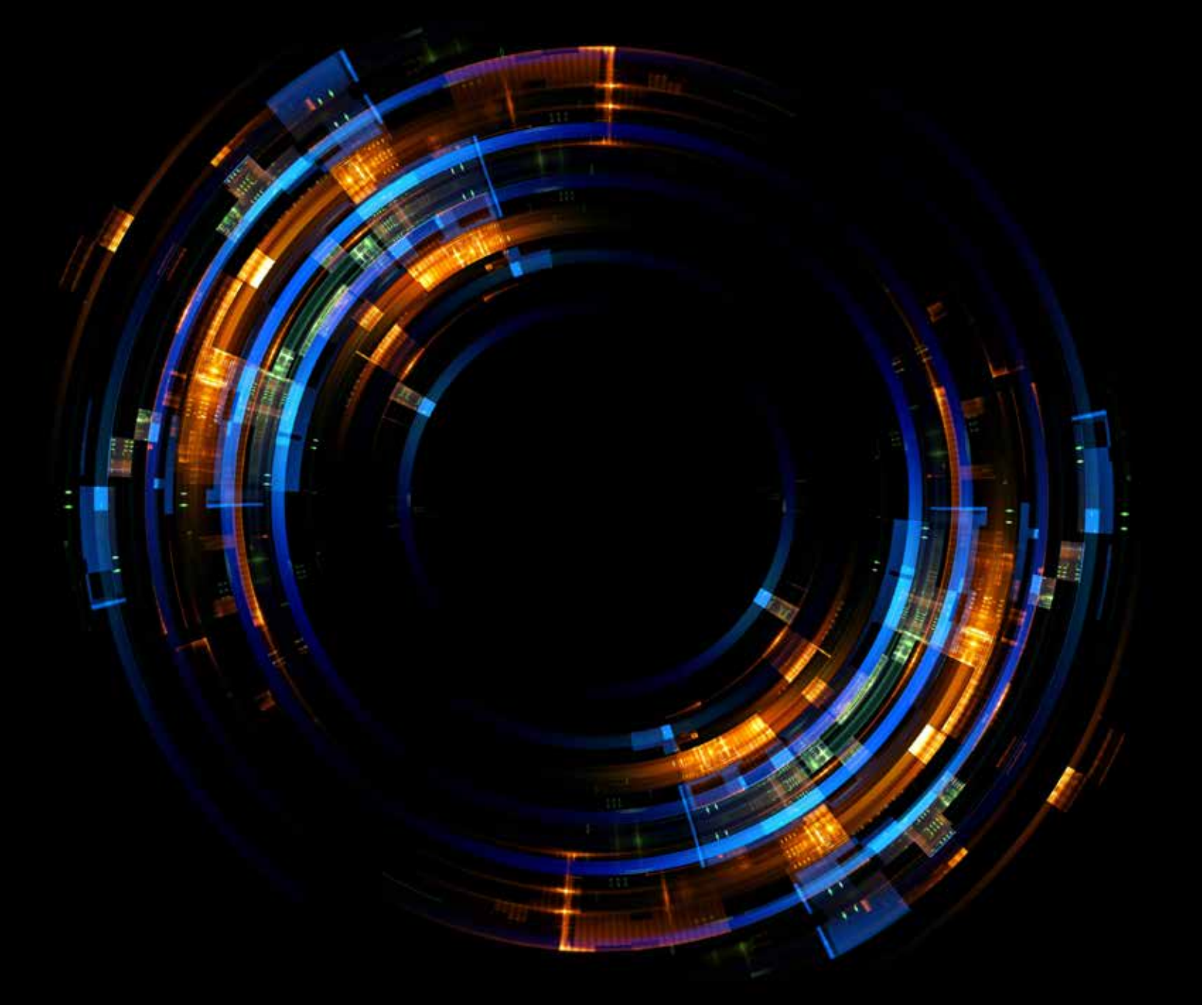
## Bringing Value To Enterprises With RAG




## How does RAG work?

RAG augments the model's prompt stream with data retrieved dynamically based on user queries. When a user provides a query, RAG retrieves relevant, context-rich information from a custom-built knowledge base—often constructed from company specific data, logs, and documents. This retrieved data is combined with the user prompt, enhancing the language model's output to produce more accurate and relevant responses. By adding dynamic, query-specific data to the prompt, RAG allows enterprises to leverage their own proprietary data while keeping it private and secure, as it remains entirely within the organization's systems.

RAG's utility is not limited to text; it can also improve video search, enable more dynamic and interactive document exploration, and even allow chatbots to reference detailed documents. The consistent flow of user query, retrieval, and context incorporation in RAG applications makes them effectively "RAG pipelines."



## Benefits of RAG

- **Custom insights without retraining:** RAG allows enterprises to customize LLMs with proprietary data, providing relevant insights without the need for costly retraining or fine-tuning.
- **Access to real-time context:** By connecting LLMs to a proprietary data knowledge base, RAG offers a practical way to continuously inject fresh data into the model, reducing hallucinations and improving output quality.
- **Secure and private data utilization:** Data remains secure as it is not shared with third-party entities managing the model, ensuring privacy and regulatory compliance.
- **Accelerated AI applications:** With RAG, organizations can launch customized generative AI applications more quickly and cost-effectively by utilizing existing models enhanced with their specific knowledge base.

# Transforming Diverse Use Cases

Enterprise AI solutions have the potential to address a wide range of use cases, and in this section, we'll focus on seven of the most common and impactful applications. RAG plays a key role in enhancing each of these use cases—whether by ensuring alignment with company policies, minimizing the risk of hallucinations, or generating content that reflects the organization's tone of voice. RAG enables AI to deliver more accurate, business-specific insights, providing practical value across a variety of business needs.

Intel AI for Enterprise RAG

## Delivering End-to-End Solutions for the Enterprise

AI solution catalog launching Q4'2024

Easy to Deploy, Fully Validated

- Chat Question & Answer (Q&A)
- Audio Q&A
- Visual Q&A
- Content Summarization
- Frequently Asked Question Generation
- Code Generation
- Code Translation

## Let us take a look at some examples of use cases:

- **Chat Question & Answer (Q&A):** Enable instant responses to common queries across various domains, improving efficiency and access to information.

  **Example:** An HR chatbot using RAG could answer employee questions about benefits or guide them to specific policy documents, reducing the workload on HR personnel.

- **Audio Q&A:** Provide answers using audio input, offering an alternative mode of interaction for employees or customers.

  **Example:** A customer service line for a telecommunications company could use Audio Q&A to provide spoken instructions for troubleshooting common device issues, ensuring better accessibility for users.

- **Visual Q&A:** Enhance operational efficiency by analyzing visual content and identifying relevant sections or insights.

  **Example:** In a media production company, Visual Q&A powered by RAG could be used to search through hours of video footage to find specific segments based on content descriptions, significantly speeding up the editing process by pinpointing the most relevant clips.

- **Content Summarization:** Transcribe and summarize meeting notes or extracts key insights from large documents, enabling faster decision-making.

  **Example:** A legal team could use RAG to summarize hundreds of pages of contract details, providing the key points in a matter of seconds and significantly reducing manual review time.

- **Frequently Asked Question Generation:** Automatically create a set of FAQs to support frontline employees or answer common customer questions.

  **Example:** A retail company could use RAG to generate a comprehensive FAQ database for new product launches, helping support staff handle customer inquiries faster during a high-demand period.

- **Code Generation:** Help developers efficiently generate or modify code scripts, allowing them to focus on higher-level tasks.

  **Example:** A software engineering team working on custom integrations could use RAG-powered code generation to automate repetitive script writing, reducing development time.

- **Code Translation:** Assist in updating legacy systems or migrating old code bases to modern platforms, ensuring compatibility with newer technologies.

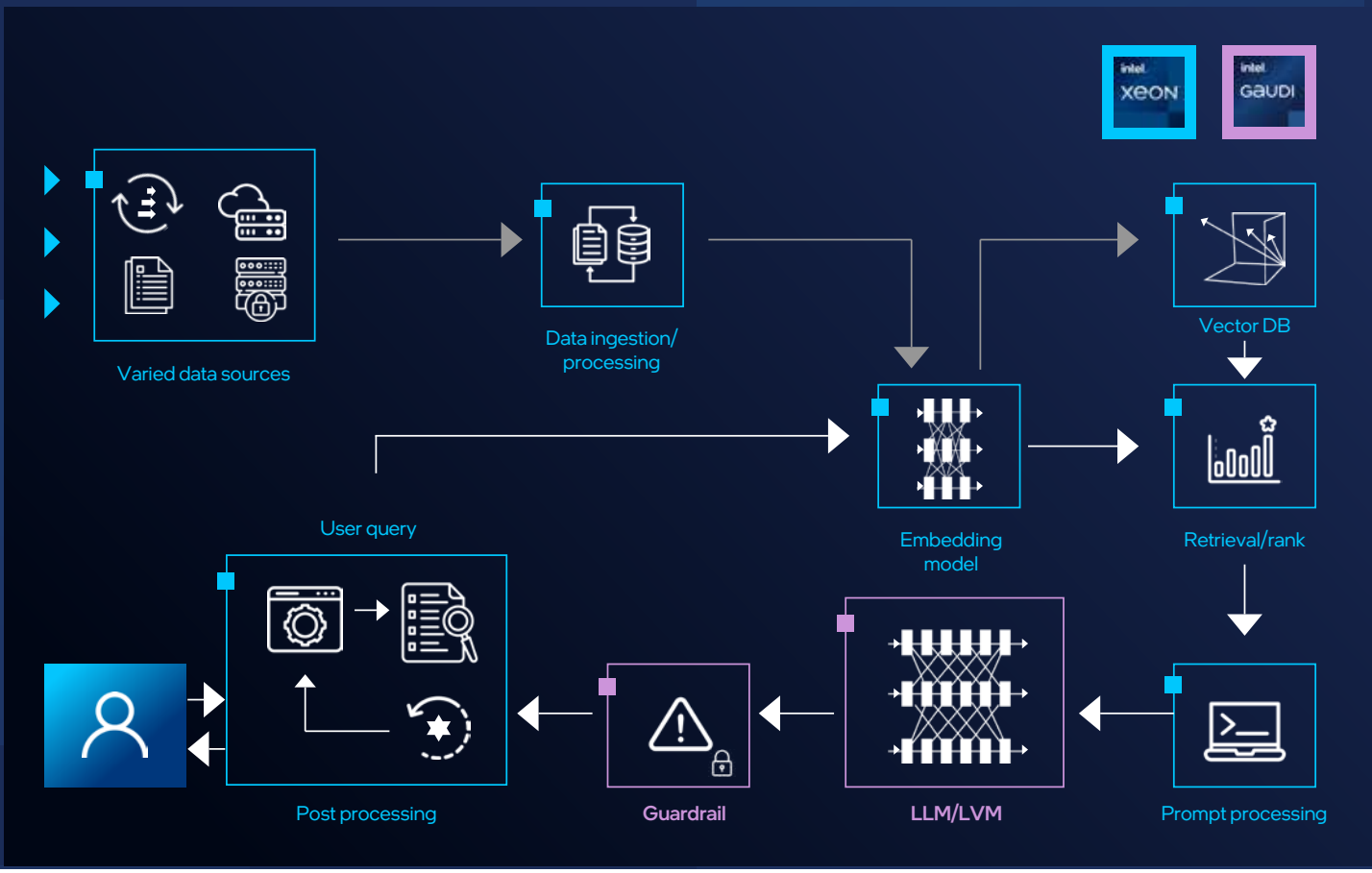
  **Example:** A financial institution could leverage RAG to translate COBOL-based legacy systems into more modern programming languages like Python, ensuring easier maintenance and scalability.

These initial solutions form the core of Intel AI for Enterprise RAG. This end-to-end AI software catalog is designed to be easy to deploy and fully validated, ensuring that enterprises can quickly integrate these RAG-powered applications into their workflows, unlocking new efficiencies and capabilities.

# Intel AI for Enterprise RAG architecture

Intel's enterprise RAG architecture simplifies how organizations leverage retrieval-augmented generation to deliver accurate, context-aware responses by integrating scalable data retrieval, intelligent language models, and secure processing—all optimized for real-world business needs.

This architecture leverages Intel's technology stack, including Intel® Xeon® processors for data processing and storage as well as Intel® Gaudi® AI accelerators for optimized AI inference, ensuring efficiency, scalability, and security throughout the RAG workflow.




The flow starts with varied data sources, which are ingested and processed to build a structured knowledge base. User queries initiate the retrieval process, where relevant vectors are searched and re-ranked. The most relevant context is then used in conjunction with a large model to generate the response, which is delivered back to the user.

Throughout this flow, monitoring components ensure performance, quality, and safety at every step of the pipeline.

The core components of the flow are (1) Building the Knowledge Base, (2) Query and Context Retrieval, and (3) Response Generation. Let's expand on these components.

## 1. Building the Knowledge Base

- **Data collection:** Gather information from varied data sources, such as documents, transcripts, and cloud storage, to build a comprehensive knowledge base.
- **Data processing pipeline:** The collected data is ingested and processed through a pipeline that includes text extraction, formatting, and chunking data into manageable pieces for further processing.
- **Vectorization / Embedding model:** The processed data chunks are passed through an embedding model to convert them into vector representations, which capture semantic relationships.
- **Vector database storage:** The resulting vectorized data is stored in a scalable vector database to enable efficient retrieval for subsequent stages.

## 2. Query and Context Retrieval

- **User query submission:** Users submit queries through a chat interface or via API calls, authenticated by a secure service to ensure appropriate use.
- **Query processing/Embedding model:** Once a query is received, it undergoes input validation for security and compliance. After validation, the query is vectorized to make it suitable for efficient searching.
- **Retrieval/Rank:** A vector search is performed to retrieve relevant vectors from the vector database. The retrieved vectors are then re-ranked using more advanced techniques to ensure the context provided is accurate and relevant to the query.

## 3. Response Generation

- **Prompt processing:** Once the relevant context is retrieved and ranked, it is processed through a prompt processing stage. This combines the user's query with the selected context, preparing it for inference by a large model.
- **LLM/LVM Inference:** The combined prompt and context are fed into a large language model (LLM) or language vision model (LVM). This model generates the response based on its language capabilities and the provided context, ensuring it is coherent and informative.
- **Guardrails**: To ensure the generated output is compliant and safe, guardrails are applied during the response generation. These measures help filter out inappropriate or biased content and align the output with enterprise standards.
- **Post-processing/response delivery:** The final response is delivered back to the user or subsystem through the original interface, providing a coherent and contextually relevant answer that meets user needs. The generated response undergoes post-processing to enhance its quality—refining language, ensuring accuracy, and structuring it for clear communication—before being delivered to the user.

**Intel AI for Enterprise RAG Architecture** ensures that all key components, from knowledge base creation to delivering the final response, operate smoothly and efficiently. The architecture utilizes Intel® Xeon® processors for data processing and vectorization, while Intel® Gaudi® AI accelerators are employed to optimize LLM/LVM inference, enabling effective scaling and secure data handling for enterprise use cases.
In addition, the Intel® Tiber™ AI Cloud plays a vital role in continuously tracking several critical components. **Retrieval performance** is monitored for latency and accuracy to guarantee precise and efficient information delivery, with logs kept for auditing. **Re-ranking efficiency** is observed to maintain context relevance and optimal system speed. **Inference service quality** is tracked to measure latency and response quality, with continuous logging for improvements.

# Four Key Considerations

Accelerating RAG in production environments requires careful attention to several critical factors. RAG pipelines are computationally demanding, and delivering low-latency responses is crucial for end users. When RAG is used with confidential data, ensuring the entire pipeline remains secure becomes even more important.
Intel technologies provide the capabilities to power RAG pipelines effectively, enabling secure, high-performance generative AI solutions tailored to specific industries. Below, we explore these four key considerations and how they contribute to creating high-performance, secure, scalable, and open RAG solutions.

## An Open, Systems-based Approach

### Delivering End-to-End Solutions for the Enterprise

**Scalable**
Gives enterprises flexibility to quickly adapt to changing workload requirements

**Open**
At every layer of the stack, leverages OPEA & partner contributions

**Secure**
Silicon-based security features and trust service - built on the secure technology businesses trust

**TCO**
Leveraging the TCO (Total cost of ownership) advantages of Xeon 6 & Gaudi 3 AI Accelerators

## 1. Computational demand and scalability

LLM inference is typically the most resource-intensive part of the RAG pipeline, especially in real-time production. Similarly, building the initial knowledge base—processing data and generating embeddings—can be computationally heavy, depending on data complexity and volume. Intel's advancements in computing technologies, AI accelerators, and confidential computing help address these challenges, ensuring both high performance and data privacy throughout the pipeline. As transaction demands increase, the compute infrastructure may face latency due to the load created by vector database queries and inference calculations. Therefore, having scalable infrastructure and readily available compute resources is crucial. Additionally, applying key optimizations for embedding generation, vector search, and inference helps boost the system's overall performance.

## 2. Data Privacy and Security

Ensuring data privacy and security is essential in RAG applications, particularly when dealing with confidential or sensitive information. Intel's silicon-based security features help protect data throughout the RAG pipeline—from embedding generation to retrieval and inference. Implementing secure processing technologies and establishing safeguards helps to maintain data integrity and prevent unauthorized access, ensuring that enterprise data remains private and compliant with security standards, even in highly regulated industries like finance and healthcare.

## 3. Open, systems-based approach

Adopting an open ecosystem approach is crucial for building interoperable and scalable RAG solutions. By leveraging open-source tools and optimizations, enterprises can integrate seamlessly across different systems, ensuring flexibility and adaptability. Intel embraces an open ecosystem approach to AI, leveraging OPEA contributions and working with industry partners to create an open and interoperable RAG solution.

## 4. Total Cost of Ownership (TCO)

Total cost of ownership (TCO) encompasses all costs related to deploying, integrating, and maintaining AI systems, including hardware, software, and operational expenditures. Achieving cost efficiency is essential for enterprises looking to maximize the value of their AI investments. By right-sizing their AI resources to align with specific needs, enterprises can avoid over-provisioning while still meeting performance demands.

Intel® Xeon® processors and Intel® Gaudi® 3 AI accelerators provide the flexibility to balance performance and cost, allowing organizations to scale efficiently while maintaining control over TCO. Furthermore, these solutions ease integration challenges and streamline the developer experience, enabling smoother deployments with minimal friction.

Optimizing AI systems from training through to inference not only improves throughput and latency but also reduces overall costs, making it easier for enterprises to manage both high-performance and cost-effective AI solutions.

Unlocking Business Potential with GenAI:

# The Open Platform for Enterprise AI

**The challenge for many businesses is not in accessing AI tools, but in effectively implementing them. The sheer number of AI models, data tools, and deployment methods available today can be overwhelming, leaving companies wondering where to begin.**

This is where the Open Platform for Enterprise AI (OPEA) comes into play. OPEA, an innovative framework introduced by the **Linux Foundation along with Intel and many other leading players from the AI ecosystem**, offers a solution to the complexity of adopting AI by providing a streamlined, cloud-native platform that allows businesses to build, deploy, and scale custom AI solutions with ease. Instead of starting from scratch, companies can leverage preassembled components that integrate seamlessly into their existing infrastructure, making the transition to AI much smoother.


OPEA is not about creating new technology but rather streamlining the integration of existing components and ecosystem players. This collaborative approach aims to accelerate the development of end-to-end generative AI solutions, reducing time to market and addressing fragmentation and complexity issues.

## The Problem: AI Overload

One of the biggest challenges businesses face today is AI overload. With an ever-expanding library of AI models, datasets, and tools, companies are often paralyzed by choice. Not only is it difficult to select the right tools for their needs, but even after choosing, it's hard to ensure that the AI system delivers tangible results. Questions like "Is this saving us money?" or "Is it increasing productivity?" often arise, highlighting the need for a more guided and measured approach to AI.
This challenge underscores the importance of enterprise choice and flexibility in AI adoption. Intel's commitment to an open ecosystem, including integrations like OPEA, empowers businesses to select the right tools and models for their specific needs. This flexibility ensures that AI investments are aligned with business goals, delivering measurable results in productivity and cost efficiency.

## The Solution: A Modular Approach

This approach addresses these concerns by offering a modular, flexible framework that allows businesses to customize their AI systems without needing to develop everything from scratch. Think of it like assembling a model kit—you have all the parts you need, and you simply select the ones that fit your specific requirements. This modularity is enhanced by pluggable components, allowing businesses to upgrade individual parts of their AI systems—such as swapping out an AI model for a more advanced version—without rebuilding the entire system. Whether processing larger datasets or handling more complex tasks, this adaptable framework can adjust to growing business needs with minimal effort.

## Prebuilt Solutions for Common Challenges

Another standout feature is the availability of "reference flows"—pre-built templates designed to handle common AI tasks. Whether incorporating a chatbot into your website to manage customer service or automating document processing for legal teams, these solutions provide ready-made options that businesses can easily deploy. These reference flows act as a guide, helping companies step-by-step to build effective AI solutions.

## The Power of Data Integration and Vector Databases

A key benefit is its ability to integrate with existing data systems. For many companies, legacy infrastructure remains a vital part of their operations, and making these systems work seamlessly with AI is crucial. This solution's focus on interoperability bridges the gap between older systems and cutting-edge AI technologies. Additionally, it supports vector databases, which are particularly useful for handling unstructured data like text and images. By converting complex data points into vectors, these databases make data more accessible and understandable by AI algorithms. This allows businesses to build custom AI solutions that are not only tailored to their needs but also powered by rich, organized data.

## The Future of Enterprise AI: Flexibility, Standardization, and the Road Ahead

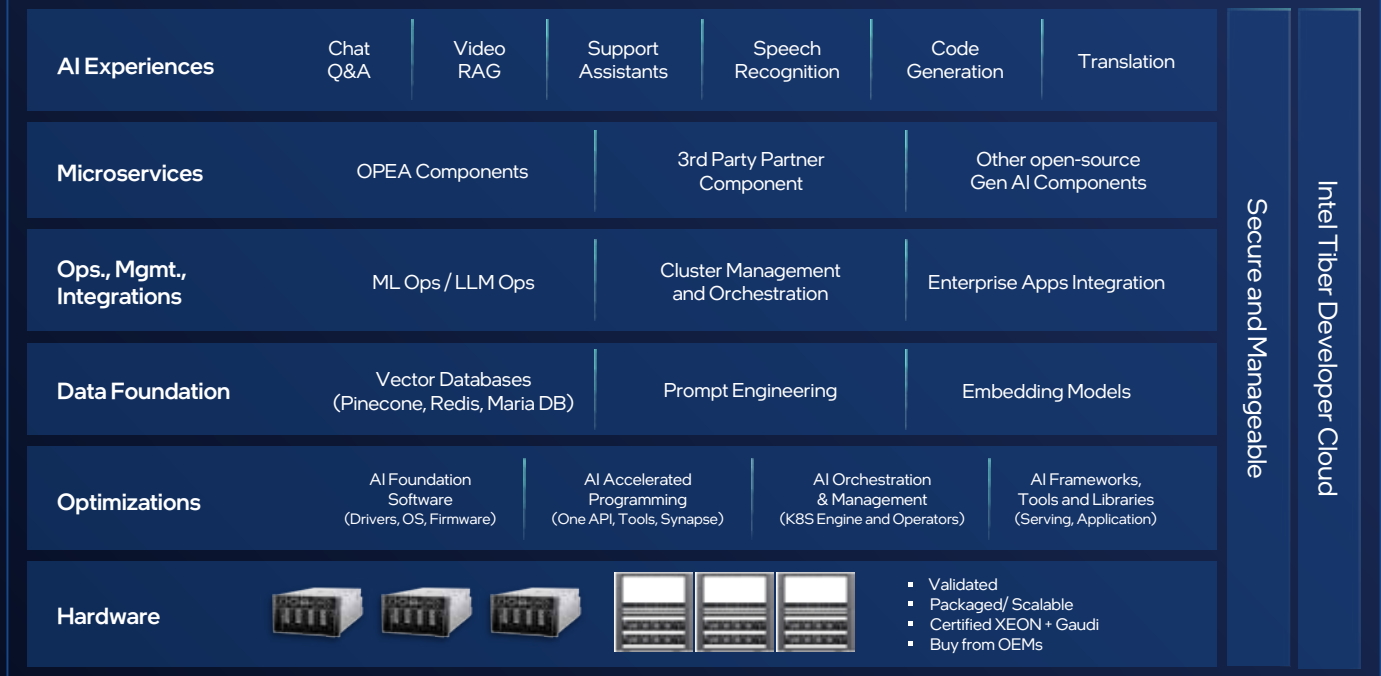
The future of enterprise AI depends on platforms that prioritize flexibility and standardization. Creating a common framework where different AI components—models, data tools, deployment strategies—can work together seamlessly, regardless of the technical expertise of the business, is essential. This standardization leads to faster development, easier collaboration, and ultimately, a more level playing field for businesses of all sizes. The modular design, pre-built solutions, and focus on data integration offer powerful tools for organizations looking to harness the full potential of AI.

# The Open Platform for Enterprise AI

## OPEA's key features include:

- **Open platform:** Enables configurable and composable components from various ecosystem partners, allowing the construction of custom solutions from multi-partner components.
- **Cloud-native deployment:** Components are designed to be cloud-native, facilitating seamless deployment in enterprise environments.
- **Specification for interoperability:** OPEA defines specifications to ensure compatibility between different components, such as vector databases and APIs.
- **Pre-validated components and reference flows:** OPEA offers pre-validated components and reference flows, simplifying development and reducing integration complexities. Users can leverage existing reference flows or customize them to build their solutions.
- **Evaluation and benchmarking:** OPEA provides tools and frameworks for evaluating and benchmarking generative AI solutions, enabling enterprises to measure performance and identify areas for optimization.
- **Microservices for AI pipeline stages:** The platform aims to develop microservices for each stage of the AI pipeline, further simplifying development and deployment for enterprises.

# Full Stack Delivery of Validated RAG Experiences




**Intel AI for Enterprise RAG delivers a streamlined, scalable solution for deploying GenAI applications across enterprises.** Powered by Intel® Gaudi® accelerators and Intel® Xeon® processors, this platform allows businesses to focus more on outcomes and less on operational complexity.

Key benefits include:

- **Streamline Development:** Automate up to 50% of software development tasks using pre-packaged use cases while maintaining accuracy and efficiency.
- **Ease of Use:** Access a catalog of RAG workflows and integrate seamlessly into existing systems, available both on the cloud and on-premises.
- **Engineered by AI Experts:** Built and validated by Intel engineers, enabling rapid deployment and scaling of enterprise-wide GenAI applications.
- **Security:** Protect valuable data with built-in security features
- **Open Ecosystem:** Embrace an open and interoperable approach, leveraging OPEA and third-party integrations for flexible, scalable AI solutions.

Intel AI for Enterprise RAG simplifies AI deployment and provides a cost-efficient path to scaling AI across your organization with confidence.

## Explore and deploy with Intel AI for Enterprise RAG

Choose from a broad range of solutions designed to simplify and accelerate your AI development timeline, from training and optimization to deployment. Our catalog includes specialized tools and frameworks optimized for RAG implementations, ensuring you have the resources needed to build cutting-edge AI applications that leverage the latest in retrieval-augmented generation technology.

Discover our AI software catalog

intel