Case Study

Artificial Intelligence
Intel® Xeon® Processor
Intel® Advanced Matrix Extensions (Intel® AMX)

Private Enterprise-Grade RAG Chatbot: Intel and VMware Cloud Foundation Private AI Services

A Blueprint for an Enterprise-Grade RAG Chatbot

VMware Cloud Foundation Private AI Services provide a secure, hybrid platform to deploy Generative AI applications. These services facilitate model governance, RAG (Retrieval-Augmented Generation) application development, and vector databases, allowing enterprises to securely manage private data and models with improved TCO.

Too often, AI initiatives fail because they start at the wrong end of the problem they're trying to solve. Instead of taking the most common approach, hardware procurement, a better starting point is the business intent and its desired result. VMware Cloud Foundation Private AI Services and Intel have flipped the script, starting with the business challenge and ending with the best technology to get the job done.

Background

As GenAI has moved beyond experimentation to its position as a strategic imperative for most businesses, the path to adoption has been slowed by the limitations inherent in LLMs:

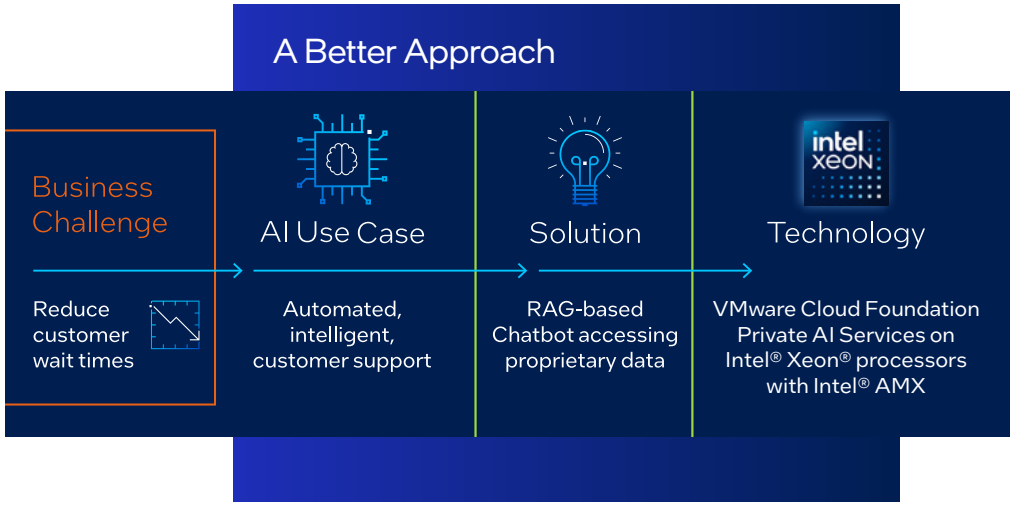
- Training data that is time-bound, providing answers that are quickly out-of-date.
- A lack of proprietary data or internal documents reduces the accuracy and impact of your results.
- Data security and privacy concerns that give C-suite decision-makers reason to pause.

RAG (Retrieval-Augmented Generation) addresses these limitations to provide accurate, timely, contextual, and customized answers tailored to your specific business needs. RAG utilizes up-to-date information, enhances factual accuracy by querying and augmenting proprietary data, and provides access to secure information.

RAG allows you to plug your proprietary information directly into a pre-trained model. It significantly improves the output accuracy and relevance to your queries and reduces model hallucinations. RAG allows you to interact with your documents and data without needing to retrain or fine-tune a model. RAG is the fastest way to turn "a general LLM" into "an expert on your data."

Solution Overview

Many organizations become paralyzed by a "technology-first" mindset that focuses on factors such as GPU availability, token speeds, and model sizes, while losing sight of the actual business challenges they set out to solve.



VMware Cloud Foundation Private AI Services and Intel have partnered to change this approach, focusing first on the business challenge, providing a turnkey, secure, and scalable platform for Enterprise Chatbot Solutions. By combining the simplicity of VMware Cloud Foundation Private AI Services with the ubiquitous power of Intel® Xeon® processors and Intel® Advanced Matrix Extensions (Intel® AMX) acceleration, we enable enterprises to run AI inference on the trusted infrastructure they already own. This is not just about technology; it is about aligning AI directly with business outcomes without rearchitecting your infrastructure.

How it Works

The Enterprise Chatbot Solution, powered by VMware Cloud Foundation Private AI Services and Intel, delivers a complete, full-stack AI infrastructure platform designed to deploy production-grade Retrieval Augmented Generation (RAG) applications. This validated solution combines industry-leading Intel Xeon processors with built-in AI acceleration and VMware Cloud Foundation Private AI Services platform to provide a secure and simple-to-deploy alternative to complex, GPU-dependent architectures.

We are finally moving past the misconception that AI requires a GPU-only environment. For many inference workloads, especially those involving RAG and chatbots, Intel Xeon CPUs offer ample performance and are highly cost-efficient. In addition, the solution is highly flexible. This heterogeneous approach uses GPUs when needed, then leverages your Intel Xeon CPU infrastructure when not needed. This approach helps avoid the risks of overinvestment in GPU infrastructure and of losing control over costs. Let your CPU and GPU scale AI across your entire datacenter footprint without creating dedicated silos.

The Components

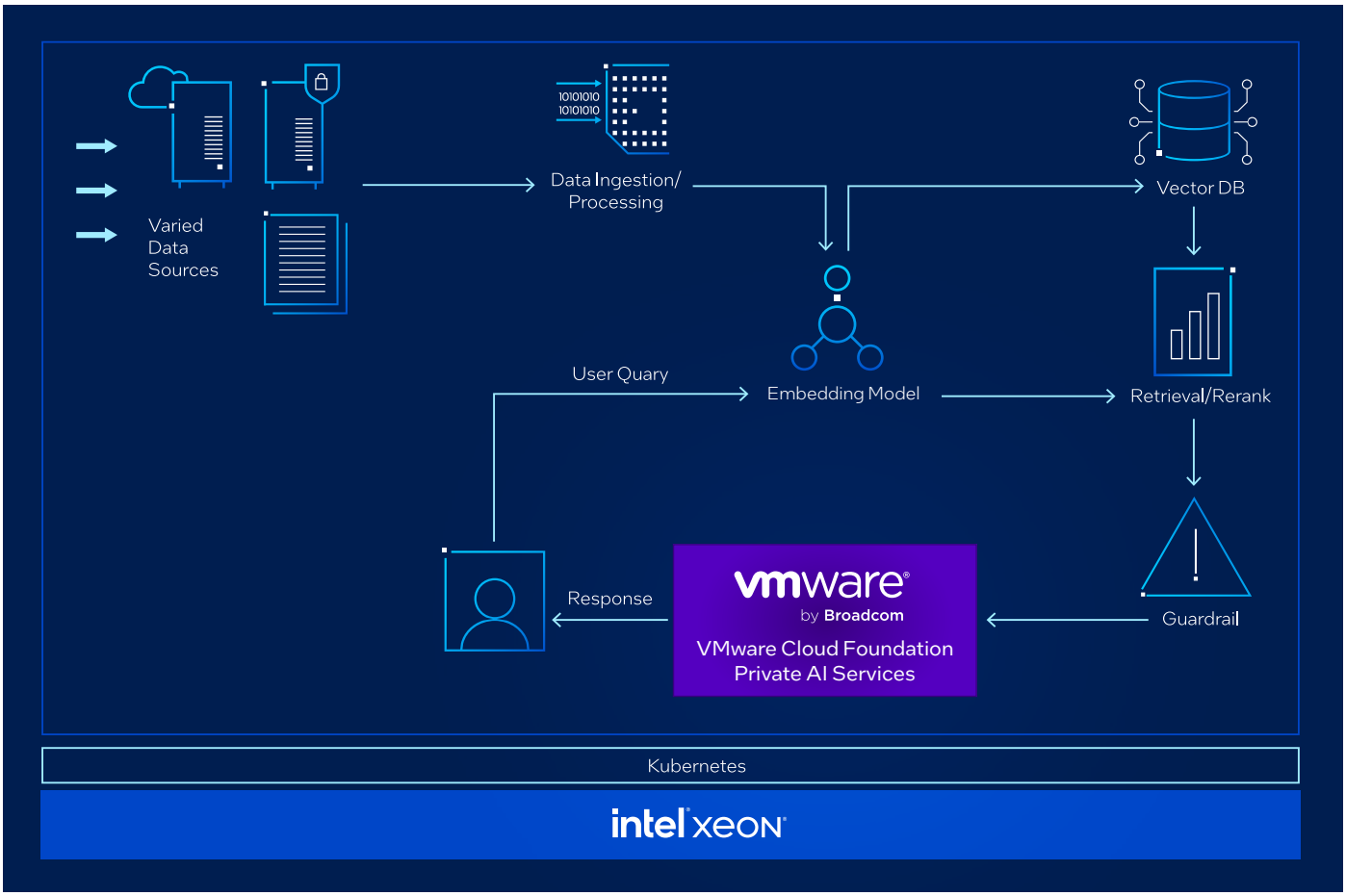
VMware Cloud Foundation Private AI Services provides end-to-end lifecycle management for AI applications with integrated security, role-based access controls (RBAC), and multi-environment deployment capabilities across data center, edge, and cloud.

Intel® AI for Enterprise RAG is an opensource stack that simplifies the RAG implementation, transforming your enterprise data into actionable insights. Powered by Intel Xeon processors, it integrates components from industry partners to offer a streamlined approach to deploying enterprise solutions.

Intel Xeon Processors (4th Gen and newer) with **Intel AMX**, a built-in hardware accelerator optimized for AI inference and matrix operations critical to RAG workloads.

The solution architecture leverages Intel AMX to accelerate the computationally intensive embedding generation, vector similarity operations, and model serving that are at the heart of RAG pipelines. At the same time, VMware Cloud Foundation Private AI Services provides secure model endpoints, manages the inference pipeline, and supports scaling when traffic patterns change. **Intel AI for Enterprise RAG** enables intelligent AI experiences that understand your business context.

Architecture Overview



1. **Ingest:** Private documents (PDFs, databases, internal wikis) are ingested safely within your perimeter.
2. **Embed/Vectorize:** Data is converted into vector embeddings using Intel-optimized models.
3. **Retrieve:** When a user asks a question, the system retrieves relevant context from your private data.
4. **Reranker:** Retrieved documents are reordered using advanced algorithms to prioritize the most relevant content for the query.
5. **Generate:** The LLM, running on Intel Xeon processors with Intel AMX, generates a precise answer using that context.
6. **Deliver:** The insight is delivered to the employee or customer via a secure chatbot interface.

Key Benefits

Faster Time to Production:

Deploy enterprise RAG applications in days, not months. VMware Cloud Foundation Private AI Services validated platform, Intel AI for Enterprise RAG enterprise-grade application, and Intel's built-in acceleration, Intel AMX, eliminate the complexity of GPU procurement, custom infrastructure builds, and multi-vendor integration challenges. Organizations can prototype, validate, and scale RAG use cases on the same platform without infrastructure redesign.

Enterprise-Grade Security and Compliance:

Keep sensitive corporate data on-premises with private RAG endpoints that never expose proprietary information to public AI services. VMware Cloud Foundation Private AI Services provides RBAC, comprehensive audit trails, content filtering via safety models, and encryption at rest and in transit.

Simplified Operations at Scale:

Manage the entire AI lifecycle from model deployment to endpoint monitoring through the solution. Deploy RAG applications consistently across data center, edge, and cloud environments without replatforming. Support multiple concurrent RAG endpoints and general workloads on the same infrastructure with intelligent resource scheduling. Scale your solution easily by adding more platforms to increase the number of parallel users.

Use Cases by Industry

	Business Challenge	AI Solution	Benefit
Finance & Banking 	Compliance teams are often overwhelmed by the rapid changes in regulations.	**Regulatory RAG Bot:** Instantly searches thousands of local/state/national regulations to answer compliance queries.	Reduces risk and speeds up audit responses; data never leaves the secure perimeter.
Retail 	Customer support costs are rising, while personalization remains low.	**Shopping Assistant:** A Generative agent that suggests products based on inventory and user history.	Runs efficiently on edge servers (Intel Xeon processors) at store locations; improves conversion rates.
Government 	Agencies have vast archives of unsearchable data.	**Policy Analyst:** Summarizes and retrieves specific clauses from decades of records.	High security (FedRAMP ready potential); utilizes existing trusted hardware.

Technical Highlights – A Look Under the Hood

This solution utilizes a full-stack approach, optimizing every layer of the infrastructure—from silicon to software—to deliver performant AI without the complexity of proprietary hardware silos.

1. The Compute Engine: Intel Xeon Processors with Intel AMX

The "secret sauce" of this solution is Intel AMX, a built-in accelerator engine. (Available on 4th Gen Intel Xeon processors and newer.)

What it does: Intel AMX introduces a new architectural component called Tiles (2D registers) and TMUL (Tile Matrix Multiply), which are purpose-built to accelerate the matrix math required by Transformers within the model.

- **The Benefit**: It allows you to run highly performant BF16 and INT8 inference directly on the CPU.
- **Performance Metrics:** Experience low Time-to-First Token (TTFT) for popular SLM/LLM models under 15B parameters (e.g., Llama 3 1B/3B/8B, Mistral 7B), ensuring a "real-time" feel for chatbot users without needing a GPU.

2. The Software Platform: VMware Cloud Foundation Private AI Services

VMware Cloud Foundation Private AI Services Enterprise AI provides a consistent, cloud-like operating model for AI workloads.

- **Simplified AI Operations:** Empower scalable GenAI app development with user-friendly metrics, dashboards, and enterprise-grade API delivery. Simplify AI operations with an intuitive interface for deploying, monitoring, and securing LLMs and APIs.
- **Accelerate GenAI adoption:** Accelerate GenAI through seamless integration, validated LLMs, and custom model options (including private uploads).
- **AI Model Management:** Streamline AI adoption by simplifying LLM deployment and securing API workflows, reducing learning curves.
- **Private Control and Governance:** VMware Cloud Foundation Private AI Services helps your Enterprise adhere to compliance and governance with control for your data, agents, and infrastructure, creating unique AI boundaries.
- **Zero-Trust and Dark Site Capable:** The platform supports air-gapped deployments, meaning your AI models and vector databases can run entirely disconnected from the internet, a critical requirement for government and defense.

3. The Chatbot Platform: Intel AI Enterprise RAG

- **Domain-Specific Intelligence:** Enrich conversations and document processing with your organizational knowledge without training or fine-tuning models.
- **Multiple Use Cases:** Support for Chat Q&A for conversational AI and Document Summarization for extracting key insights from documents.
- **Rapid Deployment:** Transform enterprise documents into AI-powered experiences in minutes, not months.
- **Enterprise-Ready Scale:** Deploy secure, compliant AI solutions that grow with your business needs.
- **One-Click Enterprise Deployment:** Fully automated Kubernetes cluster provisioning with Ansible playbooks, supporting both single-node and multi-node configurations with comprehensive infrastructure setup.
- **Optimized AI Hardware Support:** Native support for Intel Xeon processors and Intel® Gaudi® AI accelerators with Horizontal Pod Autoscaling (HPA), balloons policy for CPU pinning on NUMA architectures, and performance-tuned configurations.
- **Enterprise-Grade Security and Compliance:** Integrated Identity and Access Management (IAM) with Keycloak, programmable guardrails for fine-grained control, Pod Security Standards (PSS) enforcement for secure enterprise operations, role-based access control for vector databases, and Intel® Trust Domain Extensions (Intel® TDX) support for confidential computing.
- **Comprehensive Monitoring and Observability:** Integrated telemetry stack with Prometheus, Grafana dashboards, distributed tracing with Tempo, and centralized logging with Loki for full pipeline visibility.

Why VMware and Intel?

Security and Privacy

Unlike public API-based models, where data leaves your control, the VMware Cloud Foundation Private AI Services and Intel partnership provide enterprise-grade security, keeping all inference and data retrieval on-premises or in your private cloud, with no external LLM endpoints.

Fast, Turnkey Deployment

A full-stack platform that deploys in days, not months. Benefit from a validated, integrated lifecycle-managed RAG pipeline with a consistent cloud-operating model.

Simple to Manage and Start

This solution leverages trusted VMware Cloud Foundation Private AI Services and Intel Xeon processors' proven infrastructure, enabling organizations to manage AI workloads the same way they manage other enterprise applications. Teams can leverage familiar tools and processes to maintain, monitor, and scale AI infrastructure, minimizing the creation of AI-specific silos. This unified approach ensures IT teams remain in control while accelerating AI adoption across the business.

Cost Efficiency and Sustainability

By utilizing Intel AMX built into Intel Xeon processors, you can immediately deploy on existing infrastructure without additional hardware costs. It also avoids the so-called "GPU Tax," the high cost, power consumption, and scarcity associated with specialized hardware for workloads that don't require it.

Simplicity and Scalability

VMware Cloud Foundation Private AI Services has the "easy button" for infrastructure. Agentic AI requires a platform that hosts both the AI and application logic side by side, allowing the AI agents to interact with other software autonomously. Eliminate infrastructure silos to enable AI as an integral part of your application stack.

Conclusion

GenAI is no longer the future...it's here today. Intel and VMware Cloud Foundation Private AI Services provide a smart, powerful, and cost-efficient enterprise RAG solution that leverages the best hardware and software, ready to deploy on your existing infrastructure. Don't let the "speeds and feeds" conversation slow down your progress to achieving an AI-ready enterprise today. "Private AI" and "fast AI" are no longer mutually exclusive. Intel AMX and the VMware Cloud Foundation platform have changed that equation. You don't have to choose between performance and data sovereignty. You can have both, on hardware you already own. Let Intel and VMware help you flip the script for faster business insights and results.

Get Started

Here is some initial deployment guidance to help you get started. What follows are some tested SLAs based on the provided system requirements for on-prem or cloud deployments. Note that these are provided as a starting point. These configurations can easily be scaled in VMware Cloud Foundation Private AI Services and Intel AI Enterprise RAG to support customer environment needs.

Capacity Requirements

Resource Type	Guidance
VMs	5 VMs: 3 control plane + 2 worker
Logical Cores	152 vCPUs total
RAM Memory	352 GB total
Disk Space	784 GB total
Operating System	Ubuntu 24.04 Server
Hypervisor	ESXi 7.0 Update 3 or later (8.x preferred)

Node Reference

Node	Role	vCPUs	RAM	Storage
controlVM0–2 (x3)	Control Plane	8 each	32 GB each	128 GB each
workerVM3-4 (x2)	Worker	64 each	128 GB each	200 GB each

Control plane nodes handle Kubernetes orchestration only. All AI inference workloads run on the two worker nodes.

Intel AMX Prerequisites (ESXi)

Confirm the following before deployment — a mismatch on any of these will silently mask AMX acceleration, even on capable hardware.

Requirement	Specification	Notes
BIOS Configuration	Intel AMX must be Enabled	Check "Intel AMX" or "Advanced Matrix Extensions" in BIOS
Minimum ESXi Version	8.0 or later	Required for VM Hardware Version 20, which is needed for AMX instruction support
VM Hardware Compatibility	Version 20 (introduced with vSphere 8.0)	Older versions mask AMX features
EVC Mode	Sapphire Rapids or higher, or Disabled	Older baselines (Ice Lake, Cascade Lake) will mask AMX

Note: Users can introduce other model sizes, but that could impact compatibility and performance. Carefully evaluate your requirements and test thoroughly.

Are you already a VMware Cloud Foundation Private AI Services customer? You're invited to try Intel AI for Enterprise RAG today.

Get more information on VMware Cloud Foundation Private AI Services.

Contact your VMware Cloud Foundation Private AI Services or Intel representative today to schedule a demo of the Enterprise RAG Solution.

vmware by Broadcom | intel

Performance varies by use, configuration, and other factors. Learn more on the Performance Index site. Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software, or service activation.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

0726/BP/HBD/PDF