AMAX

On-Prem Retrieval Augmented Generation for Enterprise AI

What is RAG?

Retrieval-Augmented Generation (RAG) combines the capabilities of generative AI with advanced embedding techniques to search an external dataset for relevant information. It leverages a pre-trained Large Language Model (LLM) to refine the extracted data into precise, tailored responses.

Businesses can harness the versatility of Retrieval-Augmented Generation (RAG) to customize Large Language Models (LLMs) for various departmental needs, creating a suite of specialized AI assistants. For instance, a tailored GPT for marketing could generate compelling copywriting and content strategies, while a finance-adapted GPT might handle complex economic forecasting and manage accounting data. Similarly, an HR-focused GPT could streamline recruitment processes and enhance employee engagement strategies.

This targeted approach ensures that each department benefits from a personalized AI that speaks its language and meets its unique requirements, leading to increased efficiency and more informed decision-making.

Try our RAG Demo

Key Benefits

Up-to-date Insights

Real-time relevant information retrieval

In-House Data Protection

Enhanced security within your network

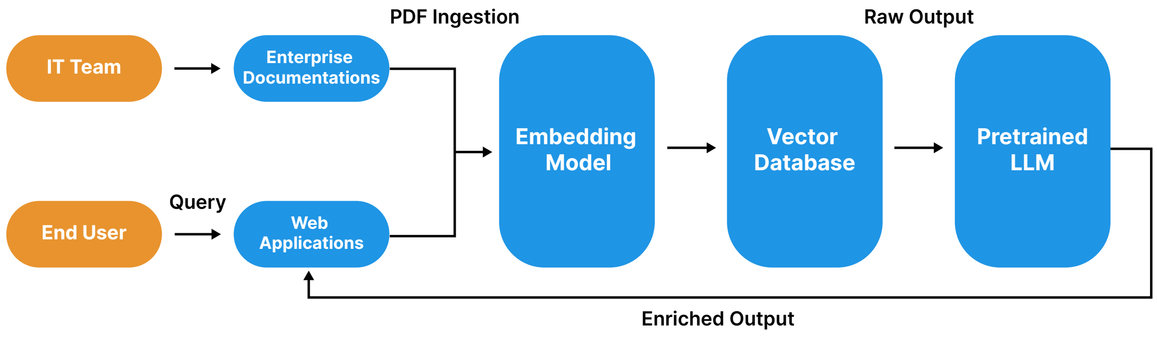
Custom Configuration

Precisely tuned to your business needs

Reliable Performance

Consistent performance, exclusively for your use

RAG Architecture



PDF Ingestion

Raw Output

IT Team

Enterprise Documentations

Embedding Model

Vector Database

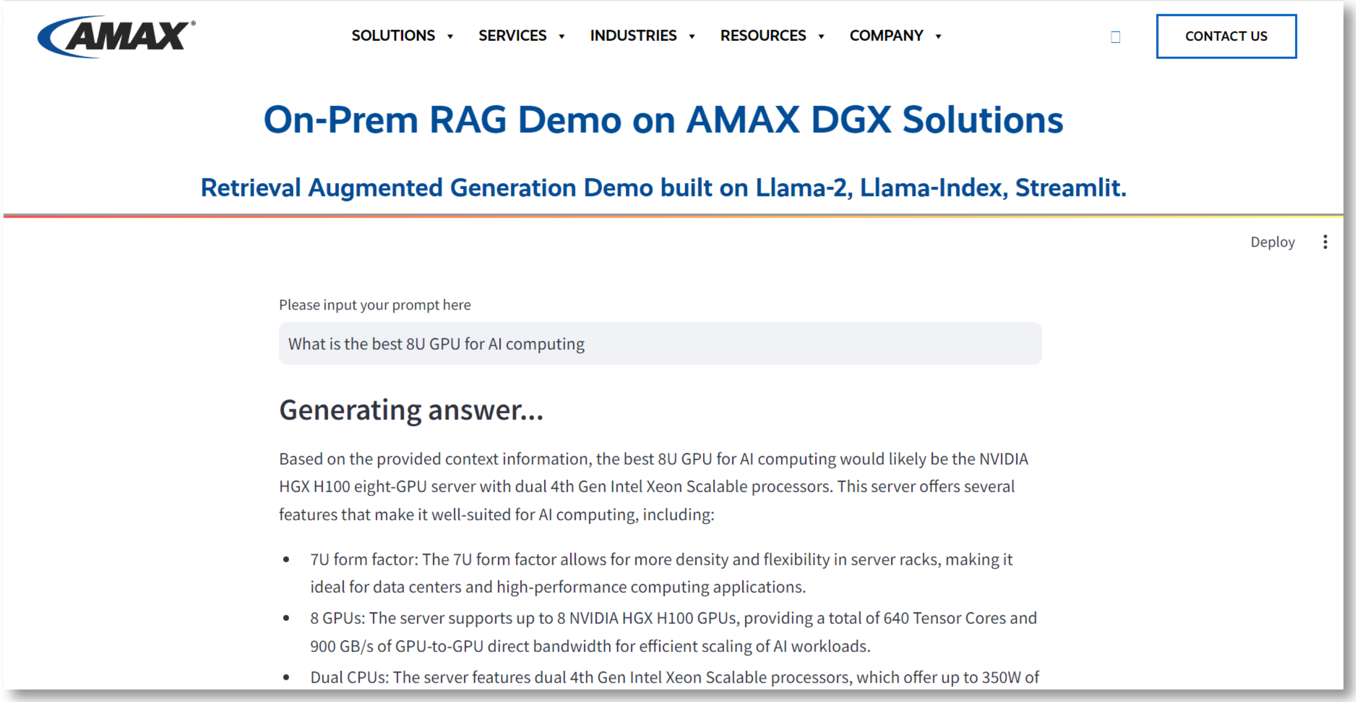
Pretrained LLM

End User

Query

Web Applications

Enriched Output



On-Prem RAG Demo on AMAX DGX Solutions

Retrieval Augmented Generation Demo built on Llama-2, Llama-Index, Streamlit.

Please input your prompt here

What is the best 8U GPU for AI computing

Generating answer...

Based on the provided context information, the best 8U GPU for AI computing would likely be the NVIDIA HGX H100 eight-GPU server with dual 4th Gen Intel Xeon Scalable processors. This server offers several features that make it well-suited for AI computing, including:

- 7U form factor: The 7U form factor allows for more density and flexibility in server racks, making it ideal for data centers and high-performance computing applications.
- 8 GPUs: The server supports up to 8 NVIDIA HGX H100 GPUs, providing a total of 640 Tensor Cores and 900 GB/s of GPU-to-GPU direct bandwidth for efficient scaling of AI workloads.
- Dual CPUs: The server features dual 4th Gen Intel Xeon Scalable processors, which offer up to 350W of

Using RAG to Enhance your LLM Deployment

Essentially, RAG acts as an extension to your LLM, securely fed with your business's internal data. When you pose a query to the RAG system that's specific to your business, products, or services—for instance, "Which AMAX 8U GPU server is best suited for AI computing?"—it draws from those provided documents to deliver accurate and relevant responses that standard pre-trained models cannot offer. With an on-premises RAG solution, your queries and data remain securely within your local infrastructure, ensuring they are not released to the cloud.

Enhancing LLMs with RAG allows for real-time data sourcing, providing current, contextually relevant responses and reduces the chance that an LLM will inadvertently reveal sensitive data or 'hallucinate' incorrect or misleading information.

AMAX's On-Prem Engineered Solutions for AI

Essentially, RAG acts as an extension to your LLM, securely fed with your business's internal data. When you pose a query to the RAG system that's specific to your business, products, or services—for instance, "Which AMAX 8U GPU server is best suited for AI computing?"—it draws from those provided documents to deliver accurate and relevant responses that standard pre-trained models cannot offer. With an on-premises RAG solution, your queries and data remain securely within your local infrastructure, ensuring they are not released to the cloud.

Enhancing LLMs with RAG allows for real-time data sourcing, providing current, contextually relevant responses and reduces the chance that an LLM will inadvertently reveal sensitive data or 'hallucinate' incorrect or misleading information.

Copyright © 2023 AMAX. All rights reserved. All trademarks are the property of their respective owners. Technical information is subject to change without notice. All company and product names are trademarks or registered trademarks of their respective owners. v11102023