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.

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

```
IT Team → Enterprise Documentations → Embedding Model → Raw Output
          Query → Web Applications → Vector Database
                  ↓                                ↓
                  Enriched Output
```

Try our RAG Demo
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.