Getting Started with Azure AI + RAG

April 2025

By: Kam Nazridoust, Ph.D.

Getting Started with Azure AI + RAG

Table of Contents

Introduction

Azure AI provides powerful cloud-based services for building intelligent applications. Combined with Retrieval-Augmented Generation (RAG), it enables you to create AI solutions that can process and understand unstructured data from PDFs and other documents, enhancing the capabilities of large language models.

This guide will walk you through the essentials of setting up Azure AI services, understanding RAG architecture, and implementing PDF processing solutions in both Python and JavaScript.

Understanding RAG

Retrieval-Augmented Generation (RAG) is a technique that enhances language models by retrieving relevant information from external knowledge sources before generating responses. Instead of relying solely on the model's pre-trained knowledge, RAG allows the model to "look up" information as needed.

How RAG Works

1. **Document Processing**: Convert documents (PDFs in our case) into a format suitable for AI processing
2. **Chunking**: Break down documents into smaller, manageable pieces
3. **Embedding**: Convert text chunks into vector representations (embeddings)
4. **Indexing**: Store embeddings in a vector database for efficient retrieval
5. **Retrieval**: When a query is received, find the most relevant document chunks
6. **Generation**: Combine retrieved information with the language model to generate accurate responses

Benefits of RAG with Azure AI

- More accurate and up-to-date responses
- Ability to reference specific documents and data sources
- Reduced hallucinations (fabricated information)
- Control over the knowledge used by the model
- Transparency in AI-generated responses