Kernel Memory

Kernel Memory (KM) is a **multi-modal AI Service** specialized in the efficient indexing of documents and information through custom continuous data pipelines, with support for **Retrieval Augmented Generation** (RAG), synthetic memory, prompt engineering, and custom semantic memory processing.

KM supports PDF and Word documents, PowerPoint presentations, Images, Spreadsheets and more, extracting information and generating memories by leveraging Large Language Models (LLMs), Embeddings and Vector storage.

Utilizing advanced embeddings, LLMs and prompt engineering, the system enables Natural Language **querying for obtaining answers** from the information stored, complete with citations and links to the original sources.

Kernel Memory is designed for seamless integration with any programming language, providing a web service that can also be consumed as an OpenAPI endpoint for ChatGPT, web clients ready to use, and a Plugin for Microsoft Copilot and Semantic Kernel.

Kernel Memory (KM) and Semantic Memory (SM)

Semantic Memory (SM) is a library for C#, Python, and Java that wraps direct calls to databases and supports vector search. It was developed as part of the Semantic Kernel (SK) project and serves as the first public iteration of long-term memory. The core library is maintained in three languages, while the list of supported storage engines (known as "connectors") varies across languages.

Kernel Memory (KM) is a service built on the feedback received and lessons learned from developing Semantic Kernel (SK) and Semantic Memory (SM). It provides several features that would otherwise have to be developed manually, such as storing files, extracting text from files, providing a framework to secure users' data, etc. The KM codebase is entirely in .NET, which eliminates the need to write and maintain features in multiple languages. As a service, **KM can be used from any language, tool, or platform, e.g. browser extensions and ChatGPT assistants.**

Here's a few notable differences:

Feature	Semantic Memory	Kernel Memory
Data formats	Text only	Web pages, PDF, Images, Word, PowerPoint, Excel, Markdown, Text, JSON, more being added
Search	Cosine similarity	Cosine similarity, Hybrid search with filters, AND/OR conditions
Language support	C#, Python, Java	Any language, command line tools, browser extensions, low-code/no-code apps, chatbots, assistants, etc.
Storage engines	Azure AI Search, Chroma, DuckDB, Kusto, Milvus, MongoDB, Pinecone, Postgres, Qdrant, Redis, SQLite, Weaviate	Azure AI Search, Elasticsearch, Postgres, Qdrant, Redis, SQL Server, In memory KNN, On disk KNN. In progress: Chroma

and **features available only in Kernel Memory**:

- RAG (Retrieval Augmented Generation)
- RAG sources lookup
- Summarization
- Security Filters (filter memory by users and groups)
- Long running ingestion, large documents, with retry logic and durable queues
- Custom tokenization
- Document storage
- OCR via Azure Document Intelligence
- LLMs (Large Language Models) with dedicated tokenization
- Cloud deployment
- OpenAPI
- Custom storage schema (partially implemented/work in progress)
- Short Term Memory (partially implemented/work in progress)
- Concurrent write to multiple vector DBs