**Difference Between Retrieval-Augmented Generation (RAG) and Semantic Search**

**Retrieval-Augmented Generation (RAG)** and **semantic search** are both used to improve the accuracy and relevance of LLM-generated responses, but they serve different purposes and work in complementary ways.

**Retrieval-Augmented Generation (RAG)**

RAG enhances LLM responses by incorporating external knowledge sources before generating an answer. It retrieves relevant data based on a user’s query and uses this information to augment the prompt, allowing the model to provide more informed and context-aware responses.

However, retrieving the right information at scale can be challenging, especially when dealing with vast amounts of unstructured data across multiple sources. This is where **semantic search** plays a crucial role.

**Semantic Search**

Semantic search improves the retrieval process by understanding the intent and meaning behind a user’s query rather than relying on exact keyword matches. It scans large and diverse datasets—such as manuals, FAQs, research reports, and customer service documents—to extract the most relevant information.

For example, if a user asks, *"How much was spent on machinery repairs last year?"*, a traditional keyword-based search might return a list of documents containing the words “machinery,” “repairs,” and “spent.” In contrast, semantic search understands the context and retrieves specific text passages with the actual expenditure data.

**How Semantic Search Enhances RAG**

Semantic search significantly improves RAG by retrieving the most meaningful and relevant information before it is fed into the LLM. Instead of relying on conventional keyword search, which may miss important context, semantic search ensures that the RAG model receives high-quality data. This leads to more accurate, context-rich, and useful AI-generated responses without requiring developers to manually structure and preprocess data.

Ref:<https://aws.amazon.com/what-is/retrieval-augmented-generation/#:~:text=Retrieval%2DAugmented%20Generation%20(RAG),sources%20before%20generating%20a%20response>