Retrieval-Augmented Generation (RAG) Fundamentals and Large-Scale Deployment

Dalmo Cirne, Workday, Inc., USA

Large Language Models (LLMs) are mostly trained with public data, which limits their ability to accurately answer questions that require domain-specific knowledge. In this lecture, we will talk about how we can bring contextual information to an LLM using Retrieval-Augmented Generation (RAG), this way reducing hallucinations and significantly improving the quality of responses.

We will look at examples of an LLM "hallucinating" when trying to answer questions it knows little to nothing about, then apply RAG techniques to introduce domain-specific context that will significantly reduce or eliminate incorrect inferences. Next, we will look at improving the results even further by introducing semantic chunking.

As we explore LLMs and RAG, we will discuss embeddings, determining relevance in a vector space, vector databases, temperature, and other related topics.

The lecture will conclude with us looking at the details, including architectural diagram, of deploying a large-scale system capable of handling high volumes of data ingestion and servicing inferences.