Contribution ID: **40** Type: **not specified**

Exercise: LLMs in Production: RAG pipelines and beyond

Tuesday 25 March 2025 16:45 (1 hour)

Running Large Language Models (LLMs) in production presents lots of complexities extending far beyond your choice in model. Key challenges include:

- How do you address knowledge staleness (i.e. your model being trained on out of date / not relevant information)?
- How do you balance cost optimisation with model latency?
- How do you reduce bias and factual hallucinations?

A widely adopted approach to address these is Retrieval Augmented Generation (RAG).

RAG pipelines implement a two-tiered approach (Retrieval & Generation): allowing models to be given domain-specific information prior to generating their response to a question. Through these techniques, "Off the Shelf" LLMs can be applied to a much wider domain context than what they were originally trained for.

In this lecture, we will explore how to improve the adaptability of LLMs without the need for fine-tuning: covering RAG and related architectures, physics-based approaches like entropix that allow for self-reasoning / context aware sampling, and the challenges with applying these techniques in a production context.

Attended school

tCSC 2023 (Split)

Number of exercise hours

0 (no exercises)

Number of lecture hours

1

Author: MUNDAY, Jack Charlie (CERN)

Presenter: MUNDAY, Jack Charlie (CERN)