Contribution ID: **18** Type: **Poster + Lightning Talk**

EncouRAGe: Evaluating RAG local, fast and reliable

Wednesday 16 July 2025 10:18 (3 minutes)

We introduce **EncouRAGe**, a comprehensive Python-based framework designed to streamline the development and evaluation of Retrieval-Augmented Generation (RAG) systems using local Large Language Models (LLMs). Encourage integrates leading tools such as vLLM for efficient inference, Jinja2 for dynamic prompt templating, and MLflow for observability and performance tracking. It supports both in-memory (Chroma) and scalable (Qdrant) vector databases for optimized context retrieval. The framework offers modular RAG methods, customizable inference templates, and detailed evaluation metrics, enabling rapid prototyping and benchmarking of context-aware LLM applications. Encourage aims to democratize LLM-based development with a focus on flexibility, speed, and reproducibility.

I want to give a Lightning Talk

yes

Author: STRICH, Jan (Universität Hamburg)

Presenter: STRICH, Jan (Universität Hamburg)

Session Classification: Lightning Talks