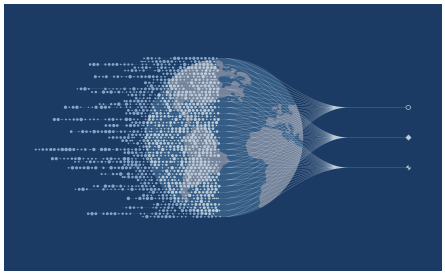

ID: **O4.3-565** Type: **Oral**

framework for developing Large Language Model (LLM) applications

We present a framework for developing Large Language Model (LLM) applications that integrate with various data sources and systems, enabling advanced AI-driven capabilities. Our approach focuses on automating time-consuming/repetitive tasks that require knowledge work through incremental deployment of LLM applications, which can process unstructured information in a common sense manner. We utilize a Retrieval Augmented Generation (RAG) framework to incorporate external knowledge and in-context learning techniques, allowing our LLM applications to learn new skills and adapt to changing contexts. Our framework is built upon open source tools, which provide a scalable and flexible platform for developing and deploying LLM applications locally in our dedicated GPU infrastructure. We demonstrate the capabilities of our framework through various Comprehensive Nuclear-Test-Ban Treaty Organization use cases, including purpose-built AI assistants, coding assistants, and research assistants such as a paper reviewer and a plagiarism detector.

E-mail

evangelos.dellis@ctbto.org

In-person or online preference

Primary author: Mr DELLIS, Evangelos (CTBTO Preparatory Commission)

Co-authors: Mr WIRAWAN, Cahya (CTBTO Preparatory Commission); Mr DEL ROSARIO, Janero (CTBTO Preparatory Commission)

Presenter: Mr DELLIS, Evangelos (CTBTO Preparatory Commission)

Session Classification: O4.3 Use of enabling Information Technologies

Track Classification: Theme 4. Sustainment of Networks, Performance Evaluation, and Optimization: T4.3 Use of enabling Information Technologies