Contribution ID: **1** Type: **not specified**

Verbalisation Process of a RAG-Based Chatbot to Support Tabular Data Evaluation for Humanities Researchers

Thursday 18 September 2025 12:00 (20 minutes)

Scholars have access to large amounts of data and publications stored in RDRs (Research Data Repositories). LLMs (Large Language Models) can efficiently work with textual data. But, since LLMs are pretrained and have a limited context window, they cannot work with large amounts of text. For this, the standard approach is to use RAG (Retrieval Augmented Generation), where an embedding space is built for the text corpus. During answering, the nearest suitable texts are extracted and provided to the context of the LLM. However, data in tables is not evaluated correctly because the embedding treats the tabular data as textual and thus fails to correctly model the semantics, which represents the context, of the tabular data. In this article, we show how tabular data can be used in a RAG-like approach: The key steps are i) a static cloze text is generated and then modified once by an LLM and ii) presented to the scholar for possible modifications. Afterwards, iii) the whole data set is verbalised according to the cloze text and, therefore, iv) usable for RAG. In particular, step iii) is crucial for our system as we add the missing context to the data.
Our feasibility study shows how to efficiently generate a chatbot with a large amount of structured data.

Authors: ASSELBORN, Thomas (Universität Hamburg); BENDER, Magnus (Aarhus University); MARWITZ, Florian (Universität Hamburg); MÖLLER, Ralf (Universität Hamburg); MELZER, Sylvia (Universität Hamburg)

Presenters: ASSELBORN, Thomas (Universität Hamburg); BENDER, Magnus (Aarhus University)