LLMs as interfaces: Creating custom chatbots to explore Linked Open Data
Author: Massimiliano Carloni (ACDH-CH, ÖAW) - *Type*: Live demonstration

Large language models (LLMs) are increasingly used in the social sciences and humanities (SSH) [1]. However, there are specific limitations to their use. Not only do LLMs hallucinate, but when they are used for information retrieval, they often have little knowledge of specific research domains. For such domains, the SSH community has built up large datasets over the years [2], but these are available to LLMs only in a limited way.
These limitations of LLMs can be partially overcome by using Retrieval Augmented Generation (RAG) [3], which allows LLMs to access external sources such as unstructured text, semi-structured documents, or databases. RAG enables the identification of relevant information from the sources to help LLMs generate better answers.
This live demonstration features a custom LLM-based chatbot that uses knowledge graphs (KGs) to obtain more contextual information. KGs in the SSH domain often take the form of Linked Open Data (LOD) [4], i.e. datasets in which each piece of information is modelled as a triple (subject-predicate-object) and can be linked to other triples by use of common identifiers (URIs) [5]. Such datasets have enormous potential as they represent knowledge in a structured and machine-readable way, and can also define logical rules through ontologies, enabling inference and reasoning [6]. Integrating LLMs with KGs can help reduce the hallucinations of LLMs and allows users to identify the provenance of the information provided by the LLM. This represents one of the possible applications of the emerging field of “neurosymbolic AI” [7], which combines machine learning with the logical reasoning provided by symbolic knowledge representation.
The custom chatbot in this demonstration is based on the LlamaIndex framework [8], which allows for easy setup and has a wide range of integrations [9]. The chatbot uses as examples datasets created with OpenAtlas [10], a database application focused on humanities research data. The datasets are all modelled according to the CIDOC CRM ontology [11], widely used in the SSH community to describe cultural heritage. Therefore, the chatbot can be easily adapted to other datasets that use the same ontology.
Building custom chatbots is not without its challenges. For example, representing the full expressive potential of ontologies in a way that is easily interpreted by an LLM is often a challenging task, requiring explicit declaration of inferences as well as appropriate verbalisation of logical rules. Also, there may be copyright and privacy concerns when using an LLM provided by an external vendor (such as OpenAI).
Nevertheless, such integrations of KGs with LLMs can be a good starting point for providing new interfaces to datasets in the SSH, especially for those users who do not master specific query languages (like SQL or SPARQL [12]), or for those who want to explore semantic relationships that cannot be logically inferred from the data. In future research, this could lead to other possible applications that combine Linked Open Data and LLMs – even beyond RAG – and that could allow for a more advanced use of the expressive power of ontologies.

References

[1] Gu, Peiran, Fuhao Duan, Wenhao Li, Bochen Xu, Ying Cai, Teng Yao, Chenxun Zhuo, Tianming Liu, and Bao Ge. “Bridging Technology and Humanities: Evaluating the Impact of Large Language Models on Social Sciences Research with DeepSeek-R1.” arXiv, April 15, 2025. https://doi.org/10.48550/arXiv.2503.16304.

[2] Chen, Shu-Heng, and Tina Yu. “Big Data in Computational Social Sciences and Humanities: An Introduction.” In *Big Data in Computational Social Science and Humanities*, edited by Shu-Heng Chen, 1–25. Cham: Springer International Publishing, 2018. https://doi.org/10.1007/978-3-319-95465-3_1.

[3] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.” In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 9459–74. NIPS ’20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[4] Berners Lee, Tim. “Linked Data - Design Issues,” June 18, 2009. https://www.w3.org/DesignIssues/LinkedData.html.

[5] W3C Interest Group. “Cool URIs for the Semantic Web,” December 3, 2008. https://www.w3.org/TR/cooluris/.

[6] Allemang, Dean, Jim Hendler, and Fabien Gandon. *Semantic Web for the Working Ontologist: Effective Modeling for Linked Data, RDFS, and OWL*. 3rd ed. Vol. 33. New York, NY, USA: Association for Computing Machinery, 2020.

[7] Hitzler, P., and M. K. Sarker, eds. *Neuro-Symbolic Artificial Intelligence: The State of the Art: 342*. Amsterdam ; Berlin ; Washington, DC: IOS Press Inc, 2021.

[8] “run-llama/llama_index.” Python. 2022. April 26, 2025. https://github.com/run-llama/llama_index.

[9] “Llama Hub.” Accessed April 27, 2025. https://llamahub.ai/.

[10] “craws/OpenAtlas.” Python. 2017. craws.net, April 18, 2025. https://github.com/craws/OpenAtlas.

[11] Bekiari, Chryssoula, George Bruseker, Erin Canning, Martin Doerr, Philippe Michon, Christian-Emil Ore, Stephen Stead, and Athanasios Velios. “Volume A: Definition of the CIDOC Conceptual Reference Model. Version 7.1.3,” February 2024. https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.1.3.pdf.

[12] W3C SPARQL Working Group. “SPARQL 1.1 Overview,” March 21, 2013. https://www.w3.org/TR/sparql11-overview/.