Large Language Models integration in Digital Humanities

Giovanni Sullutrone[1,*]

[1]*First Year PhD Student, ICT Doctorate at DBGroup*, University of Modena and Reggio Emilia, UNIMORE*

Abstract

The exponential growth of available data to Digital Humanities (DH) has created an impending need for tools capable of analyzing and extracting information from multi-lingual historical documents. This paper explores the research directions of my PhD project: providing DH scholars with effective, efficient, and explainable tools based on recent advancements in Large Language Models (LLMs). Two are the main directions of investigation: Self-Improving LLMs applied to Text-to-SQL and Topic Modeling, with a focus on interacting with and augmenting existing DBMS; Knowledge Graph (KG) creation and integration to mitigate hallucination, improve transparency and reasoning in question-answering systems. At the heart of my research lies the Digital Maktaba (DM) project which seeks to create a digital library for assisting in the preservation and analysis of multicultural non-latin heritage documents using, among others, cutting edge techniques for Natural Language Processing (NLP) and Data Science. The DM objectives and ideals align with the ultimate goal of the PhD project: the creation of instruments capable of aiding human-data interaction and information extraction while keeping the user at the center of an ever-evolving system. These tools have the potential to revolutionize the way DH scholars interact with historical documents, leading to new insights and discoveries for the field at large.

Keywords

Large Language Models, Cross-Language Document Analysis, Self-Improving Large Language Models, Knowledge Graph Integration

1. Introduction

Digital Humanities (DH) is an interdisciplinary field that merges humanities research with digital technologies with the goal of revolutionizing the analysis of cultural and historical artifacts. The exponential growth of available digital data has led to new challenges in accessing, studying and comparing this vast amounts of information, creating the need for more advanced tools that may assist the user in the analysis and cataloging of documents, be it for the major latin languages or non-latin ones.

To assist in this endeavor new algorithms based on recent developments in Natural Language Processing (NLP) and, more specifically, Large Language Models (LLMs) are gaining traction. These methods harness advancements in the former to enable cross-lingual data mining and language processing, facilitating extraction of information and user interaction with vast databases without the need for extensive technical expertise or costly human intervention.

SEBD 2024: 32nd Symposium on Advanced Database Systems, June 23-26, 2024, Villasimius, Sardinia, Italy

[*] https://dbgroup.unimore.it/site/home.html

[*] Corresponding author.

giovanni.sullutrone@unimore.it (G. Sullutrone)
https://github.com/giosullutrone (G. Sullutrone)
0009-0006-5556-1827 (G. Sullutrone)

In this research project, my primary focus will be on leveraging newly developed methodologies for self-improving LLMs to dynamically adapt to underlying data distributions and improve performances in a self-supervised manner. This will be initially used for improving current techniques of Text-to-SQL and Topic Modeling for the interaction and augmentation of DBMS.

Another area of research will be on transparency, hallucinations and reasoning. By Knowledge Graph (KG) creation, augmentation and integration, I seek to provide, on one side, easily accessible and human readable explanations of the model knowledge and planning steps and, on the other, explicit and reusable thought processes.

In this paper, I first provide, in Section 2, an overview of the Digital Maktaba case study, pointing out its objectives and significance within, and not limited to, the landscape of digital humanities. Following that, in Section 3, I will discuss the related works that will act as the foundation for my research. In Section 4 I will provide a comprehensive overview of my ideas, addressing current limitations and outlining strategies for improvement. Finally, in the concluding section, a summary of all the points explored in this paper will be given.

2. Digital Maktaba Project

In this PhD, the Digital Maktaba (DM), defined as WP5 in the ITSERR project, functions as the primary source and ultimate goal. This work package is dedicated to crafting a digital library that can analyze and extract information from multi-lingual documents, particularly from Arabic scripts (Arabic, Persian and Azerbaijani).

To facilitate this endeavor, multiple books have been provided from the "Giorgio La Pira" library in Palermo, hub of the FSCIRE foundation dedicated to history and doctrines of Islam. This repository of high-quality documents given to the DM project will be invaluable for testing new ideas and techniques, particularly in languages with limited available resources.

My contributions toward realizing the DM's overarching objectives involve providing tools that facilitate the exploration of vast datasets, extraction of pertinent information, and answering complex questions. In Figure 1, we see a representation of the main modules that I identified in this early stage of research.

- **Text-to-SQL**: Retrieval-Augmented Generation (RAG) methods have been shown to reduce hallucinations and improve the quality of responses [1]. In a similar vein, the integration of a Text-to-SQL module will act as an interface to extract relevant information from databases to ground responses for Question-Answering functionalities.
- **Metadata Generation**: one of the most sought-after features for any digital library is the generation of document metadata that can help researchers and even our Text-to-SQL module navigate the vast corpora available. LLMs will be used to fill this role with an initial focus being placed on Topic Modeling.
- **Knowledge Graphs**: hallucination, lack of transparency and poor reasoning are major critiques of current LLM-based Question-Answering systems that make them a difficult proposition for any knowledge-intensive application like digital libraries. KG creation and navigation will be used to provide a clear view of the model knowledge and thought process by conditioning the model into saving important relationships and reasoning steps.

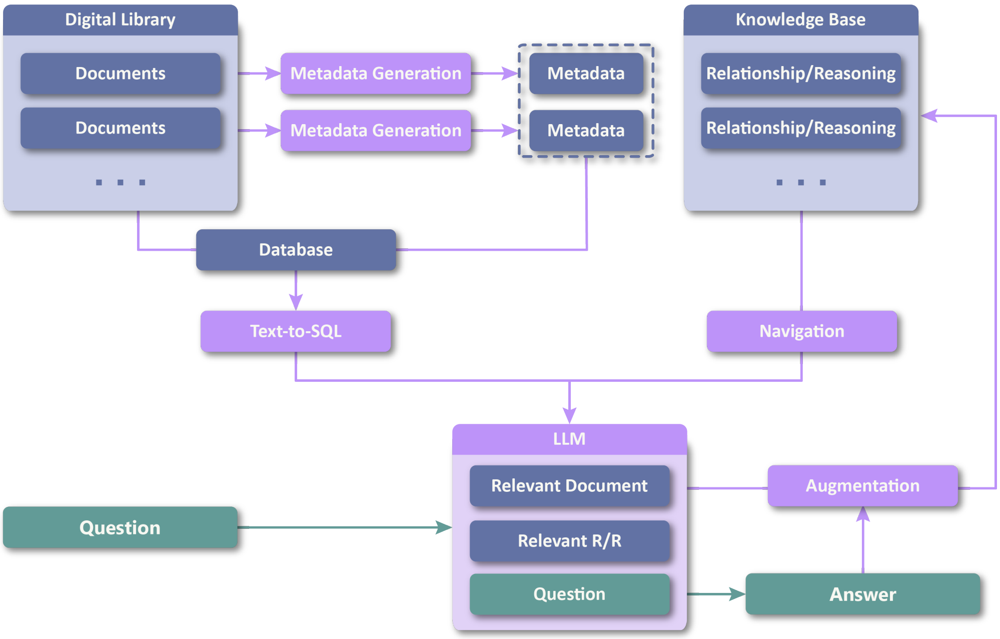

Figure 1: Simplified schema of the modules identified for Digital Maktaba. From the left, a question is presented by the user, Text-to-SQL and KG navigation are then used to get relevant documents, relationships and past reasoning steps. The LLM, given the query and the retrieved context, generates the answer and the information acquired during inference is used to expand the knowledge base.

3. Related Works

3.1. Self-Improvement

The capability of Large Language Models to enhance their own performance through self-training and self-supervision represents a significant advancement in the field. Early methodologies aimed at improving model performance on specific tasks such as code generation [2] or mathematical reasoning [3] were based on human-labeled data, which, while effective, were costly or difficult to obtain.

Another branch of research, instead, focused on using highly performant and expensive models to generate high-quality data to train smaller ones [4, 5], effectively transferring part of their capabilities to a weaker but faster counterpart.

However, the cutting edge of the domain is moving towards self-improving methodologies for creating better-performing models at a low human cost. In [6], a novel fine-tuning method was proposed to refine model capabilities by pitting instances of the model against each other in a adversarial setting, progressively enhancing performance without human intervention.

Concurrently, [7] introduced a self-rewarding approach that uses the model under scrutiny both as predictor and as judge of the response, generating new feedback signals for iterative DPO training [8]. Notably, each iteration of improvement enhances both the judge and the judged capabilities providing interesting possibilities.

3.2. Text-to-SQL

The realm of Text-to-SQL translation has seen considerable advancements, propelled by both traditional and innovative methodologies. Early efforts in this area often centered around templates and heavy human engineering [9]. The advent of machine learning, instead, introduced a wave of data-driven approaches, culminating in recent procedures that make use of Large Language Models to solve the task with satisfactory results. These new up and coming techniques are dominating the leaderboards for the major Text-to-SQL datasets [10, 11]. However, while the usage of open-weight models are garnering interest, they still underperform the best closed models, as of the time of writing, GPT-4 [12].

3.3. Topic Modeling

Traditional Topic Modeling approaches like Latent Dirichlet Allocation (LDA) [13] rely on statistical distributions to uncover latent thematic structures within a text corpus.

In recent years, instead, attention has shifted towards utilizing autoregressive LLMs. In [14] and [15] both closed-source and open-weight models were tested for defining and assigning topics in different datasets and contexts, demonstrating significant improvements compared to baselines, with mixed results for open-weight models compared to closed ones.

3.4. Knowledge Graph and LLMs Synergy

As stated in [16] the integration of Knowledge Graphs and Large Language Models has emerged as a compelling area of research due to their complementary strengths.

LLMs excel in language understanding, generation and processing of extensive textual data. Yet, challenges in interpretability and controllability remain. Conversely, KGs provide clear, structured factual knowledge, enhancing explainability and reasoning.

Therefore synergistic methodology are an intriguing avenue of research. KGs can combat LLMs' propensity to hallucinate facts by injecting real-world knowledge during training [17, 18] or inference [19]. In turn, LLMs can facilitate the construction [20] and enrichment [21] of KGs through their ability to process and interpret vast amounts of unstructured text. Finally, LLMs can act as a powerful tool for interacting with available knowledge bases, enabling users to ask natural language questions to receive structured and informative responses that leverage the depth and interconnectedness of KGs [22].

4. Research Objectives

As described in Section 3, the current best performing techniques for both Topic Modeling and Text-to-SQL are based on closed source models but this is a more broad phenomena as shown in the current ranking for chatbots [23]. This reliance poses several problems for real-world application. These include cost estimation difficulties, dependency on third-party availability, privacy concerns, potential biases and poor transparency. These issues are particularly pronounced in digital humanities research, where the analysis of lengthy historical and multifaceted documents is central.

The objective of this PhD is to address these challenges by improving existing open-source models, thereby providing scholars with effective, efficient and explainable tools for analyzing multilingual documents and interacting with complex data systems. This will be achieved through three primary avenues: self-improvement of Text-to-SQL and Topic Modeling, and KGs integration.

4.1. Self-Improvement

In the domain of Text-to-SQL, current methodologies mainly focus on massive and closed LLMs solutions that are a difficult proposition for a lot of real-world use cases [10, 11].

The subject of research will be the use of self-improving mechanisms to train the model in a self-supervised manner by generating new samples of natural language queries through in-context learning as in [7]. The results will then be judged by the same model, utilizing additional information from SQL execution to provide reinforcement feedback for training. The creation of new examples and the exploration of the database during the iterative improvements could lead to important gains in performance.

Similarly, for Topic Modeling, new documents on the same subjects will be created, the associated topics will be predicted and judged using additional information that can be extracted from more traditional methodologies. This pipeline will provide interesting data for the iterative reinforcement learning procedure. The effectiveness of this specific approach, while no intermediate results are available as of the time of writing, shows promise as it has been shown that Large Language Models are good judges of topic categorization and have a strong correlation with human judgments [24].

4.2. Knowledge Graphs Synergy

The synergy with knowledge graphs has become more and more apparent in recent years leading us to explore their usage for LLMs and with LLMs. This research will focus on leveraging KGs to ground model responses, mitigate hallucinations, and identify conflicting information among documents. In particular, following the work of [25], I plan to use the knowledge retrieved at inference time by RAG systems to generate and augment the available knowledge graphs by self-reflecting on the contextual and reasoning information. By utilizing in conjunction the aforementioned KG alteration with their exploration, as in [22], I foresee explicit and transparent reasoning steps being saved and reused by the LLM for solving progressively more complex tasks. These capabilities will be crucial for digital humanities research, allowing deeper, richer and more transparent analysis of multi-document and multilingual subjects.

5. Conclusions

This paper presented the research objective and foundational works of my studies which aims to create instruments capable of aiding data interaction and information extraction utilizing Self-Improvement methodologies and Knowledge Graphs. The DM project has been shown as an ideal starting point for building a digital library capable of providing scholars with efficient and effective tools for the analysis and cataloguing of multi-lingual, non-latin and lengthy historical documents. Key areas of research were outlined with great emphasis placed on the application of cutting-edge techniques. This project will hopefully contribute a key piece to the puzzle in creating ever-evolving systems built with humans and for humans.

6. Acknowledgments

This work was supported by the PNRR project Italian Strengthening of Esfri RI Resilience (ITSERR) funded by the European Union – NextGenerationEU (CUP:B53C22001770006).

References

[1] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. arXiv:2312.10997.

[2] B. Roziere, J. Gehring, F. Gloeckle, S. Sootla, I. Gat, X. E. Tan, Y. Adi, J. Liu, T. Remez, J. Rapin, et al., Code llama: Open foundation models for code, arXiv preprint arXiv:2308.12950 (2023).

[3] Z. Yuan, H. Yuan, C. Li, G. Dong, C. Tan, C. Zhou, Scaling relationship on learning mathematical reasoning with large language models, arXiv preprint arXiv:2308.01825 (2023).

[4] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, O. Saarikivi, et al., Textbooks are all you need, arXiv preprint arXiv:2306.11644 (2023).

[5] Y. Li, S. Bubeck, R. Eldan, A. D. Giorno, S. Gunasekar, Y. T. Lee, Textbooks are all you need ii: phi-1.5 technical report, 2023. arXiv:2309.05463.

[6] Z. Chen, Y. Deng, H. Yuan, K. Ji, Q. Gu, Self-play fine-tuning converts weak language models to strong language models, arXiv preprint arXiv:2401.01335 (2024).

[7] W. Yuan, R. Y. Pang, K. Cho, S. Sukhbaatar, J. Xu, J. Weston, Self-rewarding language models, arXiv preprint arXiv:2401.10020 (2024).

[8] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, C. Finn, Direct preference optimization: Your language model is secretly a reward model, Advances in Neural Information Processing Systems 36 (2024).

[9] C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the -divergence, Neural computation 23 (2011). doi:10.1162/NECO_a_00168.

[10] T. Yu, R. Zhang, K. Yang, M. Yasunaga, D. Wang, Z. Li, J. Ma, I. Li, Q. Yao, S. Roman, Z. Zhang, D. Radev, Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task, 2019. arXiv:1809.08887.

[11] J. Li, B. Hui, G. Qu, J. Yang, B. Li, B. Li, B. Wang, B. Qin, R. Geng, N. Huo, et al., Can llm already serve as a database interface? a big bench for large-scale database grounded text-to-sqls, Advances in Neural Information Processing Systems 36 (2024).

[12] D. Gao, H. Wang, Y. Li, X. Sun, Y. Qian, B. Ding, J. Zhou, Text-to-sql empowered by large language models: A benchmark evaluation, arXiv preprint arXiv:2308.15363 (2023).

[13] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (2003) 993–1022.

[14] C. M. Pham, A. Hoyle, S. Sun, M. Iyyer, Topicgpt: A prompt-based topic modeling framework, arXiv preprint arXiv:2311.01449 (2023).

[15] H. Wang, N. Prakash, N. K. Hoang, M. S. Hee, U. Naseem, R. K.-W. Lee, Prompting large language models for topic modeling, in: 2023 IEEE International Conference on Big Data (BigData), IEEE, 2023, pp. 1236–1241.

[16] S. Pan, L. Luo, Y. Wang, C. Chen, J. Wang, X. Wu, Unifying large language models and knowledge graphs: A roadmap, IEEE Transactions on Knowledge and Data Engineering (2024).

[17] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, Q. Liu, ERNIE: Enhanced language representation with informative entities, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 1441–1451. URL: https://aclanthology.org/P19-1139. doi:10.18653/v1/P19-1139.

[18] C. Rosset, C. Xiong, M. Phan, X. Song, P. Bennett, S. Tiwary, Knowledge-aware language model pretraining, arXiv preprint arXiv:2007.00655 (2020).

[19] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in Neural Information Processing Systems 33 (2020) 9459–9474.

[20] A. Kumar, A. Pandey, R. Gadia, M. Mishra, Building knowledge graph using pre-trained language model for learning entity-aware relationships, in: 2020 IEEE International Conference on Computing, Power and Communication Technologies (GUCON), IEEE, 2020, pp. 310–315.

[21] Z. Zhang, X. Liu, Y. Zhang, Q. Su, X. Sun, B. He, Pretrain-KGE: Learning knowledge representation from pretrained language models, in: T. Cohn, Y. He, Y. Liu (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 259–266. URL: https://aclanthology.org/2020.findings-emnlp.25. doi:10.18653/v1/2020.findings-emnlp.25.

[22] J. Sun, C. Xu, L. Tang, S. Wang, C. Lin, Y. Gong, H.-Y. Shum, J. Guo, Think-on-graph: Deep and responsible reasoning of large language model with knowledge graph, arXiv preprint arXiv:2307.07697 (2023).

[23] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al., Judging llm-as-a-judge with mt-bench and chatbot arena, Advances in Neural Information Processing Systems 36 (2024).

[24] D. Stammbach, V. Zouhar, A. Hoyle, M. Sachan, E. Ash, Revisiting automated topic model evaluation with large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9348–9357. URL: https://aclanthology.org/2023.emnlp-main.581. doi:10.18653/v1/2023.emnlp-main.581.

[25] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, J. E. Gonzalez, Memgpt: Towards llms as operating systems, 2024. arXiv:2310.08560.