

Easily accessible LLMs in historical studies: opportunities, limitations, pitfalls

Wednesday, 19 March 2025 11:30 (30 minutes)

In the field of Social Sciences, Arts and Humanities (SSAH), researchers have started to explore the possibilities of machine-learning techniques in several directions. With the current and imminent generations of open-source Large Language Models (LLMs) it seems already attainable for individual researchers to speed up onerous but necessary tasks on personal computers, while keeping control of their datasets at all times; does that mean that every SSAH researcher will have a range of useful AI-driven aides on their desktop in the near future? Automation of certain parts of data collection and data processing would certainly enable researchers to skip some of the more painstaking tasks such as qualitative data coding, leaving more time for the actual analysis and opening up the possibility to work with larger data sets. However, the application of LLMs comes with challenges of its own, especially when working with data - such as historical datasets - that are divergent from the data that the model was trained on.

Applying the Phi-3-mini model, this paper explores the ability of open source, low-threshold LLMs to perform specific tasks such as qualitative data coding. It takes the data of the Common Rules Project of research group Social Enterprises & Institutions for Collective Action (SEICA, Erasmus University Rotterdam) as a case study. The interdisciplinary research group studies all kinds of bottom-up organisations in which the participants manage their resources collectively: institutions for collective action (ICAs). Since 2007, members of SEICA have coded rule sets of historical common lands across Europe manually and consolidated these in a database. Presently, there are plans to extend the database with rule sets from other ICAs, e.g. early modern fishery cooperatives, 19th-century consumer cooperatives, and modern-day citizen collectives. The extended database can be used for several ongoing research programmes within the SEICA research group. One is a fairly new programme that assesses the environmental literacy embedded in historical regulations and evaluating their effectiveness in translating that literacy into actionable governance. Another programme that may benefit from the application of LLM's encompasses the comparative analysis of ICAs' regulations across centuries, providing present-day citizen collectives with evidence-based knowledge on more or less successful institutional setups through the knowledge exchange platform CollectieveKracht ("CollectivePower").

By trying to replicate the qualitative coding efforts done manually for the Common Rules Project in the past decades, and applying the resulting *modus operandi* to a newly acquired historical rule set, this paper will assess the opportunities, pitfalls, and limitations to be reckoned with when applying LLMs to historical data, discussing whether an easily available desktop application based on open-source LLMs is already within grasp of SSAH researchers.

Primary author: Dr GROEP-FONCKE, Marianne (Erasmus University Rotterdam)

Co-author: Prof. DE MOOR, Tine (Erasmus University Rotterdam)

Presenter: Dr GROEP-FONCKE, Marianne (Erasmus University Rotterdam)

Session Classification: Social Sciences, art & Humanities

Track Classification: Track 4: Social Sciences, Arts and Humanities (SSAH) Applications