Large language models in interdisciplinary research settings

A reflection

Malte Vogl, Alexander von Schwerin, and Sabrina Kirschke

1. Introduction

In the research project "Analysis of narratives in the genetic engineering discourse"[1], an interdisciplinary group of researchers with backgrounds in history (of science), political science, and digital humanities jointly analyzed the narratives of stakeholders in the German discourse on genetic engineering over a period of forty years. The challenge of the research project was to extend the established method of narrative analysis to a long period of time and to compare several points in time. For this aim, a large collection of documents, such as statements, press releases, and position papers, was collected and cataloged. The cataloged documents were analyzed using qualitative methods of empirical social research. In order to complement the qualitative analysis, a topic analysis workflow was developed that included a Large Language Model (LLM) to produce "meaningful" names for the detected topics. In this contribution, we want to reflect our experiences – the necessary translation processes, constructive friction, and critical points from the different backgrounds – in the usage of that method. We see the LLM as a connecting center for our interdisciplinary research, which led to new insights in the corresponding disciplines. However, we also challenge the "black box" that is introduced by the usage of the LLM.

Genetic engineering (GE) is an increasingly prevalent, yet controversial topic. From a social and political science perspective, there is an increasing amount of literature on the discourse of GE in different countries. However, temporal comparison may be key in understanding the role of narratives and respective contextual factors in policy-related discourse and the development of GE policies. Against this background, our study provided a historical narrative approach and analyzed how different stakeholder groups

1 https://www.museumfuernaturkunde.berlin/de/forschung/analyse-von-narrativen-im-gentechnikdiskurs

have used narratives at different times between 1985 and 2024 in Germany. We implemented the narrative analysis using a multi-step qualitative text analysis based on the Narrative Policy Framework, which is common in political science. Based on the current discourse, we have identified eight narratives. Of these, we selected one narrative and examined the extent to which the various stakeholders used this narrative and how it changed or did not change over time, focusing on three different time phases in the overall study period. The results show that certain stakeholder groups used the narrative over the entire period, but that certain features of the narrative changed in the process.[2] We assume that the value and goal framework of the stakeholders is expressed in the continuity and that changes in the context cause changes in the narrative. However, these assumptions have not yet been investigated further.

We developed the topic analysis in parallel to the main qualitative work in order to test the extent to which a computer-based topic analysis can support a conventional text evaluation or can be seen as an alternative research approach. The need to develop alternative methods arises in particular from the fact that the comparison of different points in time and stakeholder groups greatly increase the complexity of conventional text analysis.

Our article focuses on the development and application of the topic analysis in this research context, referring to the interaction between developer and users. By "developer", we refer to MV as the DH specialist in our project; by "users", we refer to AS (historian) and SK (political scientist), who expected topic modeling to assist them in narrative analysis.

2. Data and computation

As stated above, the source material for the research project was heterogeneous. AS collected data from online sources and archives into a Zotero database. Approximately 1200 documents of varying lengths (from several lines to several pages) have been collected in this way. One entry corresponded to one document. Each entry got tagged with a controlled vocabulary concerning the group that authored the material (e.g., governmental, industry and business, science, and non-governmental organization), which general values the group represented (e.g., ecologically aligned, industry-related, technological-innovative etc.) and also the position towards genetic research and its applications. Importantly, each entry also had a date and we had access to the full text of the document. With the help of Aron Marquart from University of Leipzig a pipeline that runs optical character recognition on all collected sources was developed. To get an overview picture of the material, MV developed a pipeline to run topic modelling that made use of the BERTopic package written in Python, specifically the module TextGeneration (Grootendorst 2025).

In the pipeline, several steps are based on machine learning (ML) models, e.g., the embedding of sentences. However, in its standard configuration the pipeline returned topics simply as lists of a fixed number of words. Depending on the selected parameters

2 See (Von Schwerin et al 2025) for details of the research context.

the algorithm would potentially output more than a hundred of such lists. To verify the validity of the result, an expert of the source material would need to look over the lists and give an interpretation of what they would entail. This exact step was then replaced by the LLM in a subsequent version of the software.

The prompt that is provided by the BERTopic tool gathers the mentioned lists of words with a random selection of the source text material together with an example of such a question and the desired answer. With the aim of reproducible research a number of open LLMs were tested with larger models strongly improving the results. For testing, LLAMA 3–8 was sufficient. To produce research output, the pipeline was run with LLAMA 3–70B on a local workstation. The Python code closely follows an example by the BERTopic package (Vogl 2024 , Grootendorst 2025a). The relevant prompt[3] is constructed for each topic by randomly selecting 10 texts from the found topic together with the returned words describing the topic. The LLM is tasked with a system prompt to be an "honest assistant for labeling topics" and the task is described in an example prompt. The final part of the prompt is then the task to create a short label of the topic in a selected language and return only this label. Since the locally available GPU only provides 48GB of RAM, the process is effectively running on CPU and 512GB of RAM. The number of topics that has to be processed by the LLM depends on parameters of the BERTopic pipeline, which greatly influences the runtime. For fine-grained visualizations, the computation can be up to 12 hours and more.

One challenge in the adaptation of this process with our source material lies in the relative length of the material. The prompt provided by the package apparently was designed for rather short text snippets. The material for this research project ranges from 0,5 to more than 30 pages and has diverse levels of structuring. To overcome this, a simple cutting approach was chosen, that clipped all text into pieces of a fixed word length. While this potentially cuts the texts in the middle of an argumentative structure, each text still has potentially several topics assigned to it, as is usually the case in topic modeling. The rationale behind this approach is that the large number of used text parts will still allow a meaningful picture of the evolving debates in the corpus. An additional challenge could have been the language of the corpus which is mostly in German. However, the different ML parts of the BERTopic package provide multilingual versions and the prompt was slightly reformulated to include the information which language the source material has and the returned topics should have.

3. (Limits of) interdisciplinary translation work

The discussion between the developer (MV) and the users (AS, SK) led to a number of modifications to the original computer-based approach. Initially, the changes consisted of the basic adaptation of the computational tools to the research project and required some development work, as described in section 2. In addition, problems with respect to the LLM data processing arose during the joint exchange that were less easy to solve.

3 The prompt for the LLAMA LLM is based on https://maartengr.github.io/BERTopic/getting_started/representation/llm#prompt-engineering

After an initial development phase, the software gave results that seemed convincing to the developer MV. However, the communication about the gathered results opened another challenge to the team. How could one talk about the resulting document classifications and relations in the corpus? Luckily, an additional external software package, datamapplot (McInnes 2025), allowed the use of BERTopic output for interactive visualizations, see Figure 1. In the first adaptation, the text snippets (publications or parts of publications) were visualized in the form of dots with a color that corresponds to the found topic and the LLM-generated topic name was displayed.

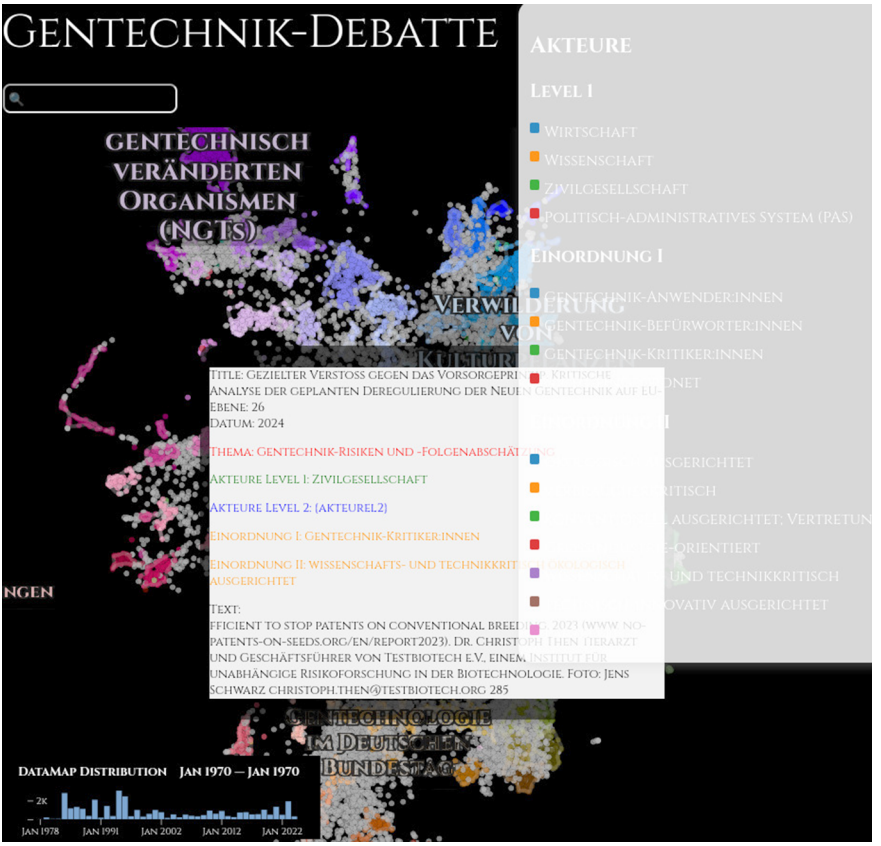
Figure 1: Screenshot of the topic modeling result with LLM topic names.

The figure shows the first adaptation (visualization of topics and topic labels) in a zoomed out view for text parts selected between 2018 and 2024 (color online).

While maybe visually pleasing, the analysis from the user's perspective was still challenging for several reasons: How could one compare the identified topics and labels with the source material? How could one identify the temporal aspects of the topic without checking all text snippets (dots)? What actors in the debate used which topics as part of

their narrative? This feedback led to a second adaptation of the interactive visualization which incorporated a histogram of the times for which documents exist, a hover menu with the classification and text snippet of the text document, a selection menu to limit the visualization to any time period, and a selection menu by actor group and attitudes, see Figure 2.

Figure 2: Screenshots of the topic modeling result with LLM topic labels.



The figure shows the second adaptation (integration of time and document text information and selection modi) displaying information about a certain text, the timeline overview and selection menu and the selection menus for the whole corpus (color online).

Once the software tool had been developed, the users (AS, SK) had further questions. Their wishes were to adapt the results of the topic analysis and the topic labels even more closely to the requirements of the narrative analysis and to manipulate the calculation methods implemented in the LLM accordingly. For example, they noted that the topic names were too general; instead, they wanted the topic names to express the content preferences and values of the actors even more strongly. The developer then explained to

the users the various settings available to him to change the results of the LLM and the prompt in the visualization.

During the joint discussion, it became clear that it was very difficult to explain the LLM calculation methods to the users. However, it also became clear that neither the developer nor the users were able to clarify exactly what changing the adjusting screws meant for the validity of the results. It also remained unclear how far one could go to ultimately make the results fit in the perspective of the users, i.e. where the limit of inadmissible manipulation of the data lay.

4. Discussion of the results

Our research process revealed some mixed experiences regarding the combination of LLMs and traditional qualitative research. On the one hand, our collaboration highlighted well-known challenges of interdisciplinary cooperation. In our case, the focus was on applying the digital humanities (DH) to a historical-political science issue, by translating the requirements of a methodology in the historical and political science (historical policy narrative analysis) into a DH methodology. If the relevant competencies are not combined in a single person, there is automatically an increased need for communication between the researchers in order to establish a mutual understanding of approaches, methods, their possibilities and limitations.

However, given the high amount of trust between the researchers, it was not necessarily a disadvantage that the DH methodology was a black box for the users, i.e., that the users only understood the DH methodology to a limited extent. Also, the researchers dedicated enough time to discuss open questions. In fact, the adaptation of the DH methodology to the historical-political science requirements took place in several steps. And even after the successful joint development of the software tool in its individual elements and processing steps, there was a need for further adaptation on the part of the users.

These post-development adjustments described above can be categorized as a normal procedure when handling statistical procedures that can be classified as a form of calibration. Processes of empirical social research often involve such interactive and, in principle, incomplete adaptation of tools to the research question. One example is the calibration of the calculation methodology for cluster analyses.

On the other hand, we faced limitations in our collaborative research process that go beyond these well-known interdisciplinary challenges. Calibrating the finalized computational tool to the requirements of the historical and political science research question would have required an interdisciplinary understanding of the exact relationship between the DH computational method, outputs, and the respective research questions. In other words, the black box would have had to be opened up. However, a mutual understanding could not be easily established in the exchange because such an understanding required a deeper competence in the other discipline – in our case, an understanding of the users of the structure of LLMs.

Further, and contrary to more established statistical or DH methods, the use of LLMs in interdisciplinary research settings leads to three main LLM-specific challenges.

First, most disciplines in the humanities and social sciences have not yet established what should be considered best-practice in the use of LLMs, leading to a certain lack of trust in such methods. Second, and related to the first point, few if any universities offer introductory courses on the statistical background of LLMs and ML in general. Contrasting this, e.g., with courses such as quantitative sociology or foundations of digital humanities, shows again the lack of established usage patterns. The third and final challenge is the more fundamental question of reproducibility. On the one hand, LLMs are designed to give statistical answers, i.e. not necessarily repeat the same answer to the same question. On the other hand, slight changes in the prompt formulation can give significantly different results. It remains unclear how one should weigh the error of different topic representations or what error bars should follow for different formulations of a prompt. Together, these challenges make using LLMs in an interdisciplinary setting an "extra-dark" black box – for both, the users AND the developers.

5. Conclusion for future cooperation

It is important to manage expectations as to what the tool can and cannot achieve: The identification of topics is exciting, but is limited in its informative value. In our case, the tool was very helpful in identifying which topics are taken up by which actor. However, when it came to questions of attitudes of the actors and the identification of narratives, the informative value was limited. For example, a specific term could be taken up but evaluated very differently by different types of actors.

The use of LLMs is associated with a subject-specific language that is not immediately comprehensible to qualitative researchers. This leads to communicative challenges in the research process, which can also be addressed if sufficient time is provided. This also includes building trust in the methods and in the researchers applying the respective methods.

As an outsider, it is not possible to understand how the tools come to specific results. This is also a problem in qualitative analysis given the subjectivity of the coder. However, in qualitative research, there are clearly defined processes for increasing the intersubjectivity of the results such as double coding of a minimum of two researchers (Mayring 2014), which are not applicable here, although similar standards exist with respect to LLMs such as pair programming or code review through third parties. The question is therefore whether the results can be recognized as scientifically valid and under what conditions this is the case.

In sum, LLMs can be considered as a helpful compass rather than a quick fix in our field of research related to the identification of narratives of different actor groups and in different moments in time, see also (Purificato et al 2025). A sufficient amount of time is needed for expectation management, traceability of results and communication as it is common in interdisciplinary work, more generally. Future research also has to put special emphasis on developing standards for cross-checks that are acceptable for both the LLM and qualitative researchers involved.

References

Grootendorst M (2025) BERTopic. Available at: https://maartengr.github.io/BERTopic/getting_started/representation/llm (accessed 6 January 2025).

Grootendorst M (2025a) Topic modelling with LLAMA. Available at: https://colab.research.google.com/drive/1QCERSMUjqGetGGujdrvv_6_EeoIcd_9M (accessed 16 September 2025).

Mayring P (2014) Qualitative content analysis: Theoretical foundation, basic procedures and software solution. Klagenfurt: Beltz. Available at: https://nbn-resolving.org/urn:nbn:de:0168-ssoar-395173 (accessed 6 November 2025).

McInnes L (2025) DataMapPlot. Available at: https://datamapplot.readthedocs.io/ (accessed 6 November 2025).

META-LLAMA (2025) Llama 3. Available at: https://huggingface.co/meta-llama/Meta-Llama-3-8B (accessed 16 September 2025).

Purificato E, Bili D, Jungnickel R, Ruiz Serra V, Fabiani J et al. (2025) The Role of Artificial Intelligence in Scientific Research – A Science for Policy, European Perspective. *Publications Office of the European Union*. DOI: 10.2760/7217497.

Vogl M (2024) SemanticLayerTools (1.0.0). Zenodo. DOI: 10.5281/zenodo.14190505.

Von Schwerin A, Kirschke S, Marquardt A, and Vogl M (2025) The narrative of (technological) progress in the German genetic engineering discourse. A temporal comparison approach. *(submitted)*.