AI-NEPI Conference Proceedings - Enhanced Edition

AI-NEPI Project Team

2025-01-01

Table of contents

Chapter 1

AI-NEPI Conference Proceedings - Enhanced Edition

Chapter 2

Preface

This enhanced edition of the AI-NEPI Conference Proceedings contains presentations on Large Language Models for History, Philosophy and Sociology of Science, held in 2025.

The chapters in this book have been generated from comprehensive XML content reports, providing structured access to the rich discussions and insights shared during the conference.

Each chapter includes:

- Structured presentation content with key sections
- Slide images synchronized with the presentation flow
- Complete speaker abstracts and overviews
- Detailed transcriptions of the presentations

This enhanced format allows readers to follow both the visual presentation materials and the detailed content in an integrated manner.

Chapter 3

Large Language Models for the History, Philosophy and Sociology of Science (Workshop)

The workshop titled “Large Language Models for the History, Philosophy and Sociology of Science,” conducted from April 2-4, 2025, at TU Berlin (Room H2005) and online, was organized by Gerd Graßhoff, Arno Simons, Adrian Wüthrich, and Michael Zichert. The event garnered substantial interest, evidenced by over 50 submissions to the call for papers, from which 16 presentations were selected. Approximately 220 participants registered, with on-site attendance fully booked and a significant onl…

3.1 Overview

The workshop, titled “Large Language Models for the History, Philosophy and Sociology of Science,” was held from April 2-4, 2025, at TU Berlin (Room H2005) and online. It was organized by Gerd Graßhoff, Arno Simons, Adrian Wüthrich, and Michael Zichert. The event attracted substantial interest, receiving over 50 submissions to the call for papers, from which 16 presentations were selected.

Approximately 220 participants registered, with on-site attendance fully booked and a significant online audience. The workshop’s conceptual framework emerged from two key initiatives. The first was the “Network Epistemology in Practice” (NEPI) project, where researchers Arno Simons and Michael Zichert utilized Large Language Models (LLMs) for analyzing physics texts and conceptual issues in physics, respectively.

The second initiative stemmed from Gerd Graßhoff’s long-standing advocacy for employing AI in History and Philosophy of Science (HPS), particularly for investigating scientific discovery processes. Funding for the workshop was provided by the NEPI project’s European Research Council (ERC) Consolidator Grant (Nr. 101044932). The NEPI project itself investigates internal communication within the Atlas collaboration at CERN to understand collective knowledge generation processes.

Methodologies employed by NEPI include network analysis for examining communication structures and semantic tools, incorporating LLMs, for tracing the flow of ideas. Workshop sessions were recorded, including presenter video and full audio, with plans to make these recordings available on NEPI’s YouTube channel, subject to presenter consent. Question and answer sessions were structured to gather multiple questions before presenters provided collective responses.

Interactive engagement was supported through *Etherpad*/*Cryptpad* for asynchronous discussions and *Zoom* chat for real-time interactions. Logistical provisions included on-site coffee breaks and refreshments, with lunch and a reception held in Room H2051.

Keynote presentations featured Nina Tahmasebi (University of Gothenburg) and Pierluigi Cassotti, discussing "Large-scale text analysis for the study of cultural and societal change," with a focus on semantic change detection and data science applications in the humanities. Iryna Gurevych (Ubiquitous Knowledge Processing Lab, Technical University Darmstadt) delivered a keynote titled "How to InterText? Elevating NLP to the cross-document level," addressing information extraction, semantic text processing, machine learning, and NLP applications in the social sciences and humanities. The official workshop website is https://www.tu.berlin/hps-mod-sci/workshop-llms-for-hpss. Supporting entities included nepi (Network Epistemology in Practice), the European Research Council (ERC), and the European Union.

3.2 Workshop Scope and Participation

The workshop, titled "Large Language Models for the History, Philosophy and Sociology of Science," was held from April 2-4, 2025. The venue was TU Berlin, Room H2005, with provisions for online participation. The organizing committee comprised Gerd Graßhoff, Arno Simons, Adrian Wüthrich, and Michael Zichert.

Information and registration for the workshop were accessible via the URL https://www.tu.berlin/hps-mod-sci/workshop-llms-for-hpss and a QR code labeled "Register here." The workshop received support from several organizations, including nepi (Network Epistemology in Practice), the European Research Council (ERC), and the European Union.

The call for papers generated significant interest, receiving over 50 submissions. From these, 16 papers were selected for presentation, a process that involved difficult choices due to the high volume. The workshop attracted a substantial number of participants, with on-site places quickly becoming fully booked.

A large online audience also registered, bringing the total number of registered participants to approximately 220 at the commencement of the workshop, with registrations ongoing. The stated objective of the two-and-a-half-day workshop was to foster an inclusive conversation on the application of Large Language Models in the History, Philosophy, and Sociology of Science.

3.3 Workshop Genesis and NEPI Funding

The conceptualization of the workshop originated from two distinct but complementary initiatives. The first arose from the "Network Epistemology in Practice" (NEPI) project. Within this project, Arno Simons, a project member, trained one of the initial large language models using physics texts, reflecting a core interest of NEPI. Simons proposed discussing such research in a wider academic forum.

Michael Zichert, also a member of the NEPI project team, had been working with large language models to analyze conceptual problems in physics. He concurred that a workshop on this subject would be a valuable endeavor.

The second initiative came from Gerd Graßhoff, a cooperation partner of the NEPI project. Graßhoff has long advocated for employing Artificial Intelligence (AI) in the fields of history and philosophy of science, particularly focusing on AI's role in analyzing processes of scientific discovery. He independently conceived the idea of organizing a workshop centered on new AI-assisted methodologies for these disciplines. Consequently, these parallel interests led to a collaborative effort, resulting in the current workshop.

The workshop is funded by the European Research Council (ERC) through Consolidator Grant number 101044932. This grant specifically supports the "Network Epistemology in Practice" (NEPI) project.

3.4 NEPI Project: Objectives and Methodologies

The "Network Epistemology in Practice" (NEPI) project focuses its research on the internal communication dynamics of the Atlas collaboration. This major particle physics experiment is located at CERN, the European Organization for Nuclear Research. The Atlas collaboration is recognized as one of the largest and most significant international research collaborations.

The overarching goal of the NEPI project is to gain deeper insights into the mechanisms by which such extensive research entities collectively produce new scientific knowledge. To achieve its objectives, the NEPI project employs a dual methodological approach.

Firstly, network analysis techniques are applied to map and understand the communication structure prevalent within the Atlas collaboration. Secondly, semantic analysis tools are utilized to track and analyze the propagation and evolution of ideas through these established network structures.

Large Language Models (LLMs) play a crucial role within this semantic toolkit, specifically in the analysis of idea flow. This particular application of LLMs is highlighted as a significant current interest. The workshop aims to provide a platform for discussing a wide array of other applications for LLMs in related fields.

3.5 Workshop Operations: Recording, Q&A, and Communication

The workshop operates under specific logistical and procedural guidelines. Recording of the proceedings is in progress; participants were notified of this and provided consent during registration. The technical setup for recording includes a single camera focused on the presenter. Audio is captured using four microphones, with an iPhone serving as a backup recording device.

Subject to the consent of individual presenters, videos of the talks—including discussion segments (audio from the discussion and video footage of the presenter only)—are planned for upload to the NEPI project's YouTube Channel following the workshop. The recording setup ensures audience privacy, as the camera is not directed at attendees during discussion periods. Participants requiring more information or wishing to withhold consent for recording are encouraged to approach the organizers.

To facilitate efficient discussions within the large group, a structured question and answer format is implemented. After each presentation, approximately four questions or comments will be collected from the audience. The presenter will then address these points collectively. This approach is designed to optimize time and streamline the interaction process.

Multiple tools are provided to support participant interaction. For asynchronous communication outside of active sessions, an *Etherpad* or *Cryptpad* platform is available (accessible via a QR code). This allows attendees to post comments and questions, and presenters can review and respond to them later. For real-time interaction during sessions, the *Zoom* chat feature is available for both online and in-person attendees to submit questions or comments at any point.

Opportunities for informal networking are integrated into the workshop schedule. Coffee breaks and refreshments are provided at the main workshop venue. Lunch and the workshop reception will take place in room H2051, located down the hall, at the far end of the building, and one floor down. Attendees may walk together to this location. Participation in the workshop dinner is subject to limited seating and prior confirmation.

3.6 Keynote Speakers and Contributions

The workshop features two keynote presentations from distinguished researchers in relevant fields.

The first keynote, titled "Large-scale text analysis for the study of cultural and societal change," was delivered by Nina Tahmasebi and Pierluigi Cassotti. Nina Tahmasebi is the Principal Investigator of the "Change is Key!" research program based in Gothenburg and is affiliated with the University of Gothenburg. Pierluigi Cassotti is a researcher participating in the "Change is Key!" project.

Their collective expertise encompasses semantic change detection, the development of benchmarks for assessing such detection methods, and the broader application of data science methodologies to research questions within the humanities. This focus makes their contribution highly pertinent to the workshop's objectives.

The second keynote speaker was Iryna Gurevych, who presented her talk, "How to InterText? Elevating NLP to the cross-document level," on the following day in the late afternoon. Iryna Gurevych leads the Ubiquitous Knowledge Processing (UKP) Lab at the Technical University Darmstadt.

Her research specializations include information extraction, semantic text processing, and machine learning, with a strong emphasis on applying Natural Language Processing (NLP) techniques to research in the social sciences and humanities. Her work was identified as aligning perfectly with the themes of the workshop.

Chapter 4

A Primer on Large Language Models

The presentation provides a primer on Large Language Models (LLMs), beginning with the Transformer architecture, its encoder-decoder structure, and the concept of contextualized word embeddings. It details the development of pre-trained language models, distinguishing between encoder-based models like BERT (Bidirectional Encoder Representations from Transformers) for full context understanding and decoder-based models like GPT (Generative Pre-trained Transformers) for text generation. The...

4.1 Overview

This presentation provides a primer on Large Language Models (LLMs), beginning with the *Transformer* architecture, its encoder-decoder structure, and the concept of contextualized word embeddings. It details the development of pre-trained language models, distinguishing between encoder-based models like *BERT* (Bidirectional Encoder Representations from Transformers) for full context understanding and decoder-based models like *GPT* (Generative Pre-trained Transformers) for text generation.

The evolution of LLMs in science domains is discussed, highlighting models such as *BioBERT*, *Specter*, and *SciBERT*, and various adaptation methods including continued pre-training, fine-tuning, contrastive learning (e.g., *SentenceBERT*), and Retrieval Augmented Generation (*RAG*). Key distinctions in LLM concepts cover architectures, fine-tuning strategies, and word versus sentence embeddings.

Applications of LLMs in History and Philosophy of Science and Technology Studies (HPSS) research are categorized into: dealing with data/sources, analyzing knowledge structures (e.g., entity extraction, mapping discourses), studying dynamics (e.g., conceptual histories), and examining knowledge practices (e.g., citation context analysis). Trends in HPSS LLM usage include accelerating interest, publication in diverse journals, varied customization levels, and increased accessibility (e.g., *BERTopic*).

Significant concerns are computational resources, model opaqueness, and lack of training data/benchmarks. HPSS-specific challenges involve the historical evolution of language, the need for critical reconstructive perspectives, and issues with sparse, multilingual, or old script data. Recommendations emphasize building LLM literacy, developing shared resources, and maintaining HPSS methodological integrity while leveraging LLMs to bridge qualitative/quantitative approaches and reflect on the field's intellectual history (e.g., co-word analysis).

4.2 The Transformer Architecture: Foundation of LLMs

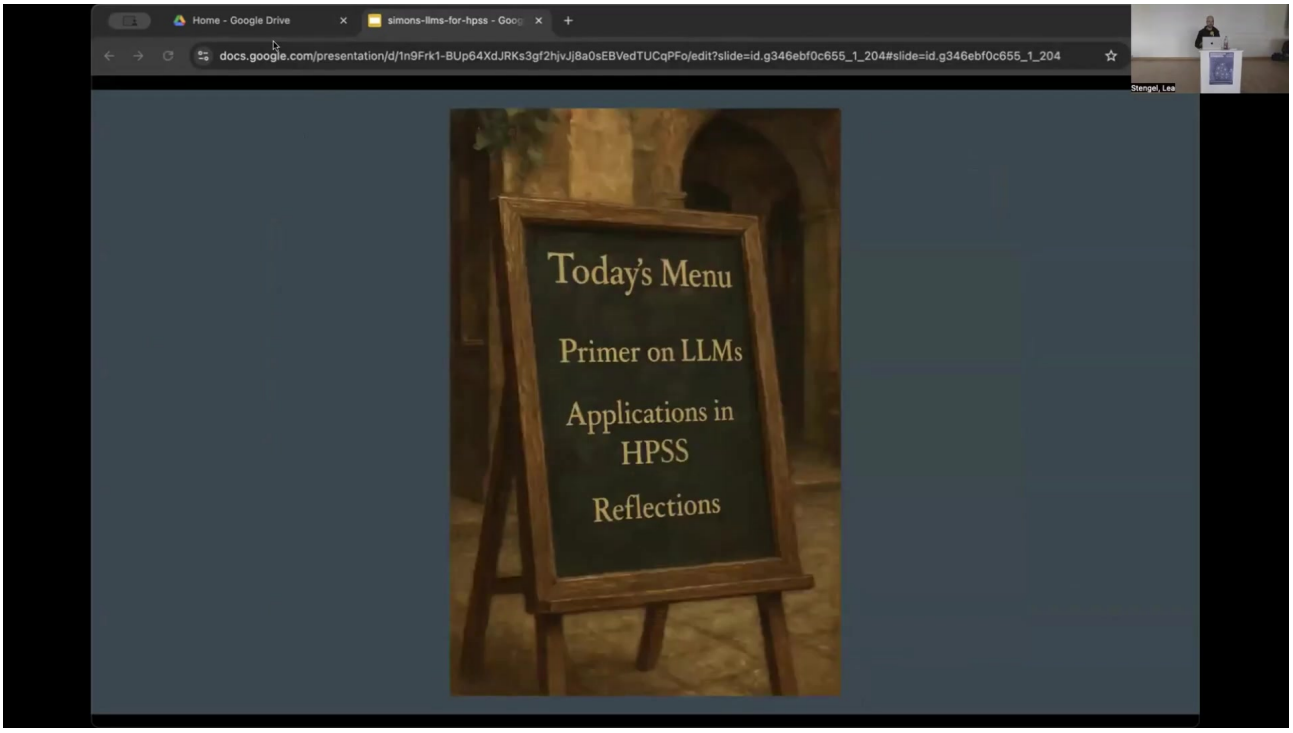

Figure 4.1: Slide 01

The *Transformer* architecture, introduced in 2017, serves as the fundamental framework for virtually all contemporary Large Language Models (LLMs). Originally conceived for machine translation tasks, such as converting German text to English, the *Transformer* model features a distinctive dual-stream structure. These two streams, an encoder on the left and a decoder on the right, are interconnected.

The encoder processes the input source language sentence. For instance, words from a German sentence are fed into the encoder, where they are transformed into numerical representations. These numerical data undergo processing through multiple layers, within which contextualized word embeddings are progressively refined. A key characteristic of the encoder is its ability to process the entire input sentence simultaneously. Each word in the source sentence can interact with, or "attend to," every other word, enabling the model to construct a comprehensive representation of the sentence's overall meaning.

The numerical representation generated by the encoder is then passed to the decoder stream. The decoder's function is to generate the output sentence in the target language, for example, English words. As each English word is produced, it is fed back into the decoder as input for generating the subsequent word. This iterative process continues until the complete target sentence is formed.

Unlike the encoder, the decoder operates with a unidirectional attention mechanism, meaning that generated words can only consider preceding words in the sequence. They cannot access future words, a constraint inherent to the next-word prediction task. However, they can look back at the words already generated in the sequence. Both streams utilize layers to increasingly contextualize word embeddings.

4.3 Pre-trained Language Models: BERT and GPT

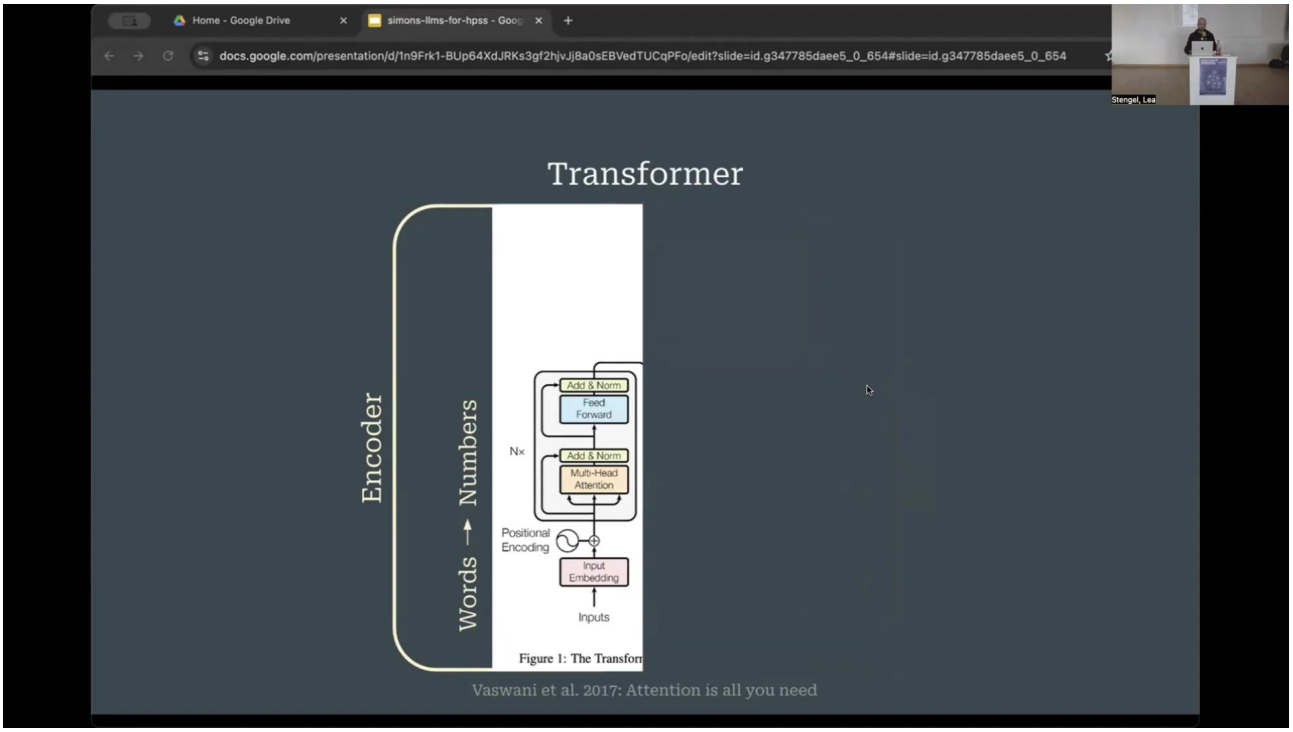

Figure 4.2: Slide 03

Following the introduction of the *Transformer* architecture, researchers began re-engineering its encoder and decoder streams independently to develop pre-trained language models. These models are designed to achieve a strong general understanding or generation capability in a language, which can then be adapted through further, often minor, training for specific Natural Language Processing (NLP) tasks.

On the encoder side, this re-engineering led to the development of models like *BERT*, which stands for Bidirectional Encoder Representations from Transformers. The *BERT* family of models remains highly influential. *BERT*'s operational principle allows every word in an input sequence to attend to every other word, facilitating a comprehensive, full-context understanding of the entire input simultaneously. The term "bidirectional" in its name refers to this ability of words to consider context from both preceding and succeeding words, while "encoder-based" indicates its derivation from the *Transformer*'s original encoder stream.

Conversely, the decoder stream gave rise to models like *GPT* (Generative Pre-trained Transformers), which form the basis of well-known applications such as *ChatGPT*. Due to their architectural constraint of only looking at predecessor words, *GPT* models excel at generating new words and, consequently, new text. This generative capability is a primary differentiator from *BERT* models, which are not inherently designed for extensive text generation.

The core difference lies in their primary functions: *GPT* models are generative, designed to produce language, whereas *BERT*-like models are geared towards a coherent, full-context understanding of sentences. Beyond these two primary types, other model architectures exist, including those that combine encoder and decoder components. Furthermore, there are advanced techniques for utilizing decoders in ways that enable them to perform more like encoders, achieving bidirectional context understanding. An example of such an approach is *XLM*, which is based on *XLNet*. Understanding the distinction between generative models and full-context understanding models is crucial.

4.4 Evolution and Adaptation of LLMs for Scientific Domains

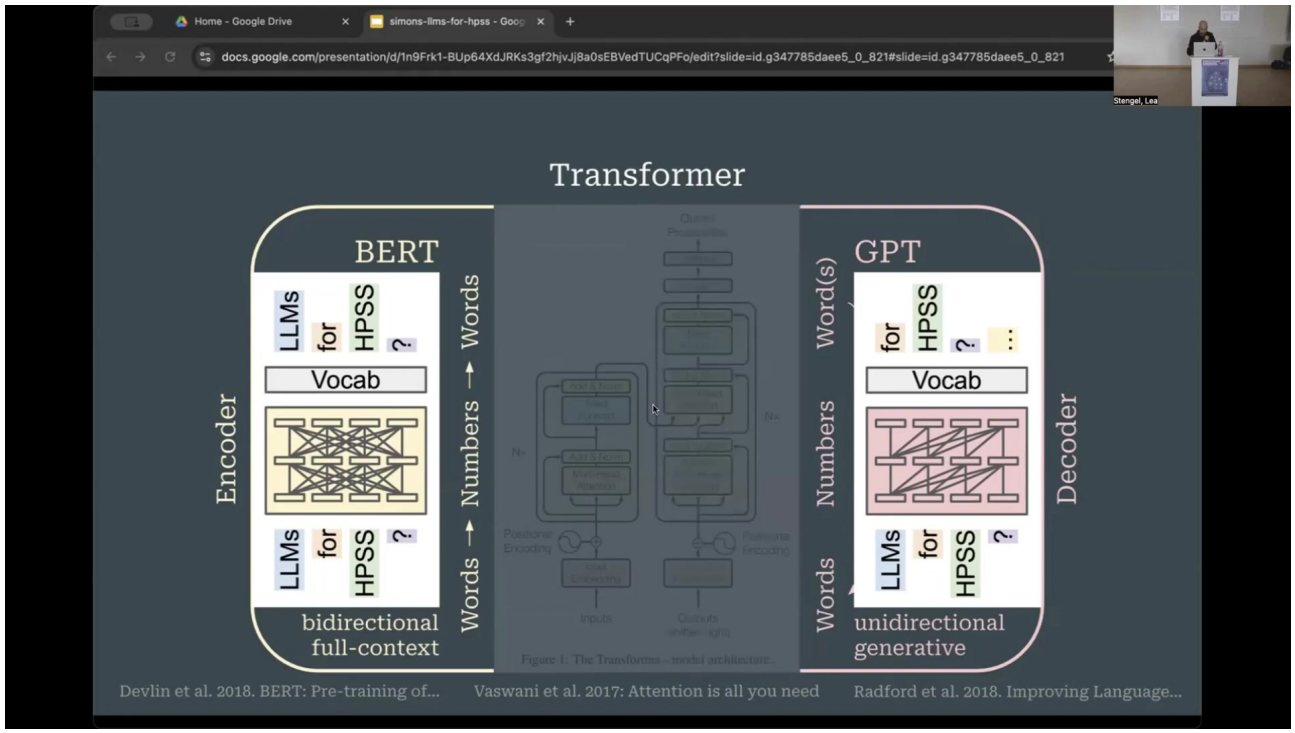

Figure 4.3: Slide 06

The evolution of Large Language Models (LLMs) has seen a significant focus on applications within various scientific domains and tasks. An overview of this development reveals a diverse landscape of models. Notably, encoder-type models, similar to *BERT*, are more commonly developed and applied in scientific contexts than decoder-type models. Early influential models in this space include *BioBERT*, *Specter*, and *SciBERT*. These, along with newer models, cater to a wide array of scientific fields such as biomedicine, chemistry, material science, climate science, mathematics, physics, and the social sciences.

Adapting LLMs to the specific language of scientific domains involves several methods. The foundational method is pre-training, where a model initially learns language patterns. This occurs either by predicting the next token, typical for *GPT*-style models, or by predicting randomly masked words, characteristic of *BERT*-style models. However, full pre-training demands extensive computational resources and vast datasets, making it impractical for many research groups. A more feasible approach is continued pre-training, where an already pre-trained model (e.g., a general *BERT* model) is further trained on a corpus of domain-specific text, such as physics literature.

Another common adaptation strategy is fine-tuning for downstream tasks. This involves adding new layers on top of a pre-trained model and training these specific layers to perform tasks like classification, for example, to determine sentiment or identify named entities. Prompt-based methods were mentioned as another adaptation technique, though not elaborated upon.

Contrastive learning is a key method for generating sentence or document embeddings that reside in the same vector space as word embeddings. *SentenceBERT* is a prominent and widely adopted technique for achieving this. This is particularly relevant as Iryna Gurevych, a keynote speaker, is known for her work in this area and might discuss *SentenceBERT*.

Retrieval Augmented Generation (*RAG*) represents a significant adaptation approach. *RAG* is not a single model but rather a pipeline or system where at least two, and often more, models work together. This technique allows for the adaptation of an LLM to a specific domain, such as a scientific field, without the need for extensive retraining of the core generative model. The *RAG* process typically involves a user submitting a query (e.g., "What are LLMs?"). A model, often *BERT*-like, then encodes this query into a sentence embedding. This embedding is used to search a database of relevant documents, retrieving the most similar passages. These retrieved passages are then integrated into the prompt provided to a generative LLM, which uses this augmented context to produce an answer. This mechanism is commonly used by systems like *ChatGPT* when they access external information, such as searching the internet.

Finally, reasoning models or agents are emerging as complex systems. These are not monolithic LLMs but rather integrated ensembles of multiple LLMs combined with a variety of other software tools.

4.5 Key LLM Concepts and Application Categories in HPSS Research

Several key distinctions are important to retain as a primer on Large Language Models. These include the existence of different model architectures (such as encoder-only, decoder-only, and encoder-decoder structures), a variety of fine-tuning strategies for adapting models to specific tasks, the crucial difference between word embeddings and sentence embeddings (which represent fundamentally different levels of semantic representation), and the varying levels of abstraction at which these models can be applied.

An ongoing survey is being conducted on the use of LLMs as tools in History and Philosophy of Science and Technology Studies (HPSS) research. Preliminary findings from this survey have led to the identification of four main categories, or bins, for classifying these applications:

- Dealing with Data and Sources: This category encompasses uses of LLMs for interacting with research data and primary/secondary sources. Specific tasks include facilitating data discovery, parsing complex or large-scale textual data, and improving the overall management of source materials.
- Knowledge Structures Analysis: LLMs are employed to analyze and extract knowledge structures embedded in texts. This includes the identification and extraction of specific entities relevant to HPSS, such as scientific instruments, celestial bodies, or chemical substances. It also involves mapping complex conceptual landscapes, a traditional area of HPSS inquiry, like analyzing science policy discourses or tracing the formation and interaction of interdisciplinary fields.
- Dynamics: This category focuses on using LLMs to study the evolution and change of concepts and language over time. A prime example is the analysis of conceptual histories of words, tracking how their meanings and usages shift within scientific communities, an approach similar to that demonstrated by researchers like Nina Janich and Pelin Doğan.
- Knowledge Practices: LLMs can be applied to investigate various knowledge practices. One specific example is citation context analysis. This method has an established tradition within HPSS, though in recent times it has often been predominantly used for evaluative purposes (e.g., research assessment). However, it holds potential for a broader range of HPSS-specific analytical tasks.

4.6 Trends, Concerns, and Accessibility in HPSS LLM Usage

Several trends and concerns are evident regarding the use of Large Language Models (LLMs) in History and Philosophy of Science and Technology Studies (HPSS). There is an accelerating interest in LLMs within the HPSS community. Publications detailing LLM applications are found predominantly in information science-oriented journals such as *Scientometrics* and *JASIST* (Journal of the Association for Information Science and Technology). However, an increasing number of papers are also appearing in journals traditionally less focused on computational methods. This suggests

that the enhanced semantic capabilities of modern LLMs are making them more attractive and relevant to qualitative researchers, philosophers, and other scholars within HPSS.

The degree of customization in LLM use varies widely across the field. Some researchers utilize readily available, off-the-shelf tools like *ChatGPT*, while others at the other end of the spectrum are engaged in developing entirely new LLM architectures tailored to specific HPSS needs.

Despite the growing adoption, several repeating concerns are frequently voiced. The substantial computational resources required to train or even fine-tune large models pose a significant barrier. The opaqueness of these models, often referred to as their "black box" nature, makes it difficult to understand their internal decision-making processes, which is a concern for interpretability and trustworthiness. There is also a perceived lack of sufficient training data, particularly for specialized historical contexts or languages relevant to HPSS research. Furthermore, the absence of established benchmarks specifically designed for evaluating LLM performance on HPSS-relevant tasks makes it challenging to compare different models or approaches systematically. Finally, researchers face a trade-off between different types of LLMs, as no single model is universally optimal; the most adequate model must be chosen based on the specific research question and purpose.

On a positive note, there is a discernible trend towards increased accessibility of LLM-related tools. For example, *BERTopic*, a popular library for topic modeling, is noted for its ease of use, largely due to robust maintenance and active development, making advanced techniques more approachable for a wider range of researchers.

4.7 HPSS-Specific Challenges and Methodological Considerations for LLM Adoption

The adoption of Large Language Models (LLMs) in History and Philosophy of Science and Technology Studies (HPSS) necessitates acknowledging several challenges specific to the discipline. A primary challenge is the historical evolution of concepts and language. Most LLMs are trained on contemporary language, which may not align with the historical texts and linguistic conventions central to much HPSS research. This requires strategies such as training custom models on historical corpora or using existing models with a keen awareness of their inherent biases and limitations when applied to historical material.

Another significant challenge stems from the reconstructive and critically reflective perspective characteristic of HPSS. Scholars in this field typically do not take scientific texts at face value but instead engage in critical interpretation, reading "between the lines" to understand the authors' situated contexts, motivations, and subtle discursive strategies, such as boundary work. Current LLMs are generally not trained to detect or analyze these nuanced aspects of texts. Therefore, methods need to be developed to enable models to approximate this type of critical reading. Furthermore, HPSS research often contends with practical data issues like sparse datasets, the presence of multiple languages (often historical variants), and archaic scripts, all of which pose difficulties for standard LLM application.

To address these challenges and effectively leverage LLMs, several recommendations are proposed for the HPSS community. Firstly, there is a need to build LLM literacy. This involves understanding the underlying theory of these models, their capabilities, limitations, and the broader implications of their use. While natural language interfaces for coding may become more common, acquiring some coding skills can be beneficial. A crucial aspect of literacy is moving beyond the superficial use of tools that might produce appealing visualizations or graphs without a deep comprehension of the underlying processes or the significance of the results.

Secondly, the development of shared datasets and benchmarks specifically tailored to HPSS research questions and materials is essential for robust and comparable evaluations. Thirdly, it is vital to stay true to core HPSS methodologies. While HPSS research problems need to be translated into tractable NLP tasks (such as classification, generation, or summarization), care must be taken to ensure that these technical tasks do not overshadow or "hijack" the fundamental research purpose and critical inquiries of HPSS.

Despite these challenges, LLMs also present new opportunities. They offer promising avenues for bridging qualitative

and quantitative research approaches, potentially fostering more integrated and multifaceted analyses. Moreover, the rise of LLMs provides an occasion for HPSS to reflect on its own intellectual history and pre-existing analytical frameworks. For instance, current LLM techniques resonate with earlier methods developed within HPSS, such as co-word analysis, pioneered by scholars like *Michel Callon* and *Arie Rip* in the 1980s, often in conjunction with *Actor-Network Theory* (*ANT*).

Chapter 5

OpenAlex Mapper: Transdisciplinary Investigations

The work presented introduces OpenAlex Mapper, a tool for investigating transdisciplinary applications of scientific models and concepts. The development was a collaboration between Max Neuchel, Andrea Loettgers, and Taya Knuuttila, funded by an ERC grant on "Possible Life." OpenAlex Mapper leverages a fine-tuned Specter 2 language model, the OpenAlex database, and UMAP dimensionality reduction. The Specter 2 model was fine-tuned to enhance its ability to recognize disciplinary boundaries…

5.1 Overview

The work presented introduces *OpenAlex Mapper*, a tool for investigating transdisciplinary applications of scientific models and concepts. Its development was a collaboration between Max Neuchel, Andrea Loettgers, and Taya Knuuttila, funded by an ERC grant on "Possible Life."

OpenAlex Mapper leverages a fine-tuned *Specter 2* language model, the *OpenAlex* database, and *UMAP* dimensionality reduction. The *Specter 2* model was fine-tuned to enhance its ability to recognize disciplinary boundaries using a dataset of articles from similar disciplinary backgrounds.

The core of the system involves a base map created by embedding 300,000 random English-language articles with abstracts from *OpenAlex* using the fine-tuned *Specter 2* model. These embeddings are then reduced to two dimensions with *UMAP*, and the trained *UMAP* model is retained.

OpenAlex Mapper allows users to input arbitrary *OpenAlex* search queries. The tool downloads the results, embeds their abstracts using the same *Specter 2* model, and projects these new embeddings onto the pre-existing 2D *UMAP* base map. This visualization helps users understand the disciplinary distribution and context of search terms, authors, or concepts.

The tool is interactive, allowing users to explore data points on the map and link back to the original source papers. Key features include options for visualizing temporal distributions and citation graphs.

The primary purpose of *OpenAlex Mapper* is to assist researchers in History and Philosophy of Science and Scholarship (HPSS) in addressing challenges related to small sample sizes and generalizing case study findings, particularly in the context of large-scale, contemporary science. It aims to support qualitative heuristic investigations by providing a quantitative, large-scale perspective.

Example applications include tracking the diffusion of model templates (e.g., *Hopfield model*), mapping the interdisciplinary presence of concepts (e.g., "phase transition" vs. "emergence"), and analyzing the distribution of scientific methods (e.g., *random forest* vs. *logistic regression*).

Limitations of the tool include its dependency on the *OpenAlex* database's data quality and coverage, the current use of an English-only language model, the requirement for abstracts or good titles for embedding, and the inherent imperfections of the *UMAP* algorithm (stochasticity and information loss during dimensionality reduction). A working paper with further technical details is available.

5.2 Architecture and Core Methodology

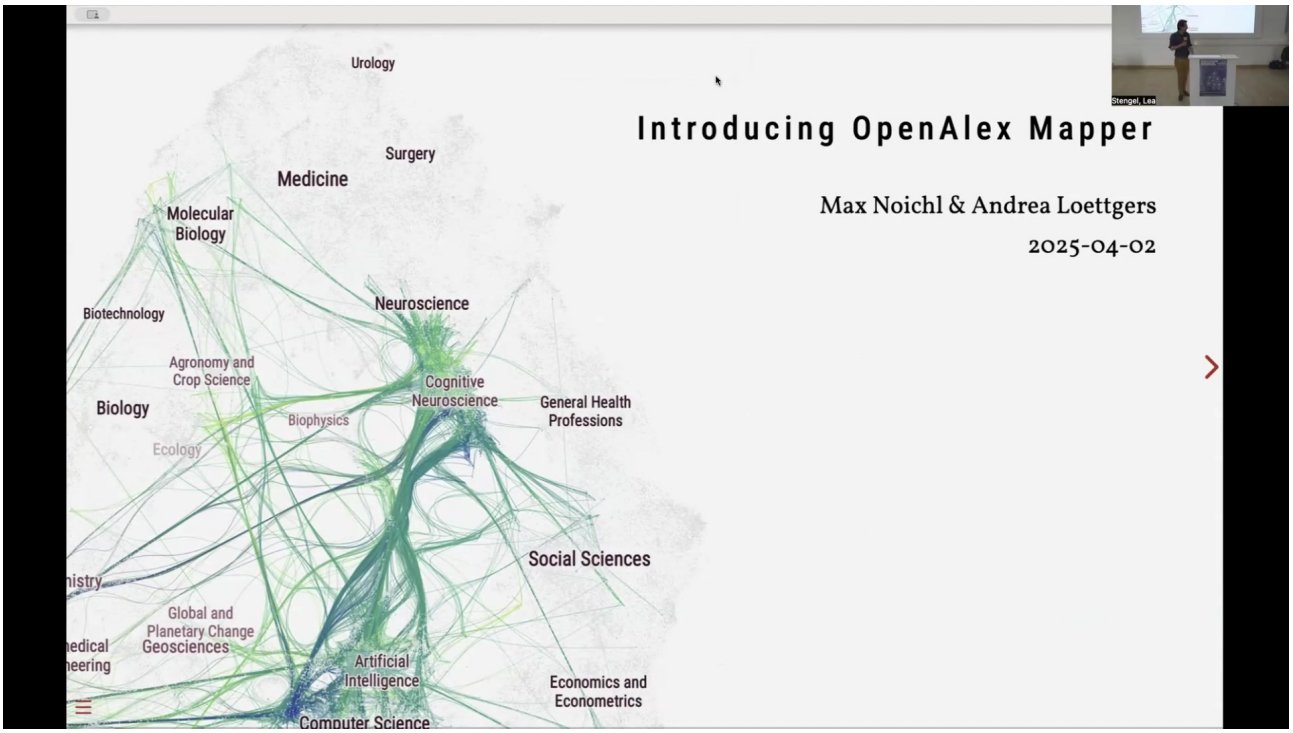

Figure 5.1: Slide 01

OpenAlex Mapper is a tool developed by Max Neuchel, in collaboration with Andrea Loettgers and Taya Knuuttila at the philosophy department in Vienna, with funding from an ERC grant focused on "Possible Life." Max Neuchel is currently a PhD student at the theoretical philosophy department at Utrecht University. The presentation slides, which include interactive elements, are accessible at maxnoichl.eu/talk.

The core methodology behind *OpenAlex Mapper* involves several key steps:

- First, a *Specter 2* language model was fine-tuned. This fine-tuning aimed to enhance the model's ability to recognize and respect disciplinary boundaries. The training process utilized a dataset composed of articles from very similar disciplinary backgrounds, and the model was trained to effectively distinguish between them. This fine-tuning represented minor modifications to the *Specter 2* model, rather than a comprehensive retraining, resulting in a "discipline improved *Specter 2*." The training process itself was visualized using *UMAP* dimensionality reduction.
- Second, the *OpenAlex* database, a vast and inclusive repository of scholarly material, was leveraged. *OpenAlex* is

noted for being larger and more inclusive than *Web of Science* or *Scopus*, though likely smaller than *Google Scholar*. Its significant advantages include being fully open data, easily queryable in batches, and freely accessible.

- Third, a base map was created. This involved sampling 300,000 random articles from *OpenAlex*, with the criteria that they be in English and possess reasonably well-formed abstracts. The abstracts of these 300,000 articles were then embedded using the fine-tuned *Specter 2* model, which produces embeddings with 768 dimensions. Subsequently, *Uniform Manifold Approximation and Projection* (*UMAP*) was employed to reduce these high-dimensional embeddings to a two-dimensional representation. The *UMAP* model trained during this stage is retained for future use.
- Finally, the *OpenAlex Mapper* tool enables users to interact with this system. Users can input arbitrary queries into the *OpenAlex* database through the tool. The tool then downloads the relevant search results. The abstracts of these newly retrieved articles are embedded using the same fine-tuned *Specter 2* model. These new embeddings are then processed through the previously trained *UMAP* model, projecting them onto the 2D base map. This allows the new articles to be positioned on the map as if they had been part of the original layout process, a feature facilitated by *UMAP*'s capability to project new data into an existing learned manifold.

5.3 Demonstration and Interactive Features

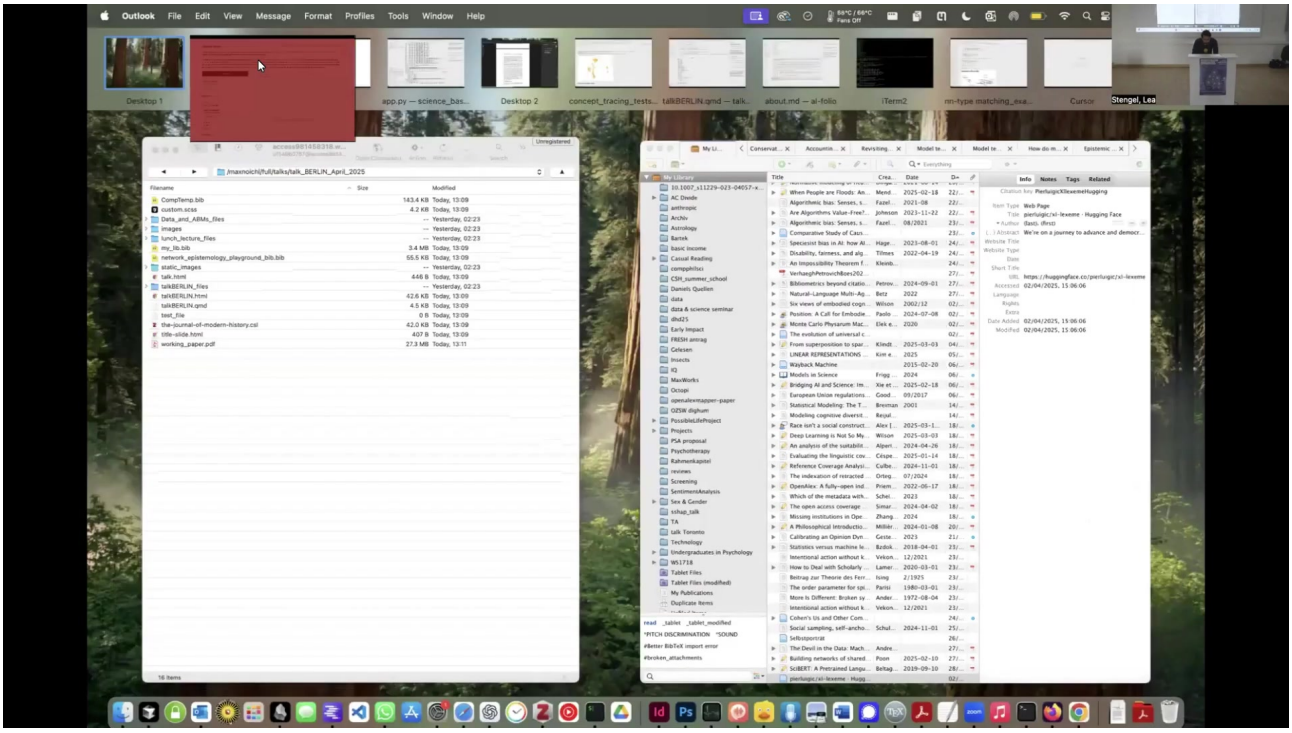

Figure 5.2: Slide 08

Access to the *OpenAlex Mapper* tool is provided through the presenter's website, maxnoichl.eu/talk, which also hosts the presentation slides, or via a direct URL to the tool itself. A live demonstration showcased its operation.

The workflow begins with navigating to the *OpenAlex* search interface to find articles related to a topic, for instance, "scale-free network models." The URL generated by this *OpenAlex* search query is then copied. This URL is subsequently pasted into the *OpenAlex Mapper* interface. Users can adjust various settings before running the query. During the live

demonstration, a minor technical issue involving a red screen overlay, possibly related to *Zoom*, briefly interrupted the display but was resolved by reloading.

Once a query is run, *OpenAlex Mapper* performs several backend processes. It downloads a specified number of records from the *OpenAlex* search results; for the demonstration, this was limited to the first 1000 records to save time. The tool then embeds the abstracts of these downloaded articles. If the user has enabled the option, it also processes the citation graph among these results. Finally, it generates a map visualizing the queried data.

The output is a projection of the search results onto a pre-existing gray base map. This visualization effectively shows where scholarly items that match the query—whether by term usage, citation of a specific author, or other criteria—appear within the broader scientific landscape. An example map for "scale-free network models" was generated live, and a previous example using the search term "coriander" (a standard *OpenAlex* example) was mentioned to illustrate the tool's output.

A key aspect of *OpenAlex Mapper* is its interactivity. Users can explore the map in detail, investigating why certain papers or concepts appear in specific regions. For example, one could examine why a term like "coriander" might surface in publications related to epidemiology or public health. Furthermore, clicking on any individual paper represented on the map will link the user directly to its corresponding website or original source document.

The tool offers several configurable settings. Users can opt to visualize the temporal distributions of their search results or to display the citation graph overlaid on the map. Additionally, an alternative version of *OpenAlex Mapper* was mentioned as being available for a few hours after the presentation. This version runs on a more powerful, higher-latency GPU setup, designed to handle larger and more computationally intensive queries.

5.4 Rationale and Utility for HPSS

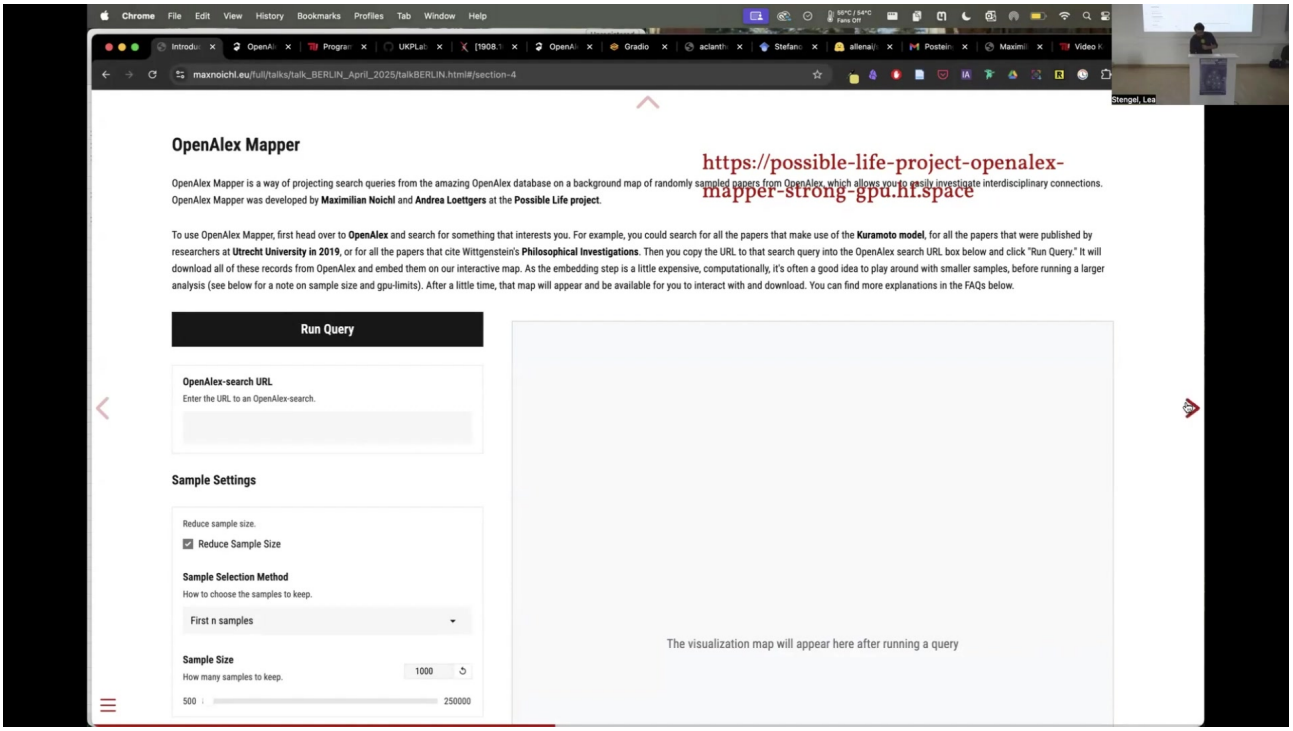

Figure 5.3: Slide 14

OpenAlex Mapper is particularly relevant for the field of History and Philosophy of Science and Scholarship (HPSS). A common challenge within HPSS is the reliance on small samples and in-depth case studies. While these methods provide rich, detailed insights—often derived from close readings of scientific papers, interactions with scientists, or analyses by individuals trained as both scientists and philosophers—generalizing these findings to the vast scale of contemporary science presents a significant hurdle. There is a concern about how to validate such qualitative approaches in the context of what is described as "global big rapid discovery contemporary science."

The *OpenAlex Mapper* tool is designed to help address these issues by offering a broader, large-scale perspective. It assists researchers in answering questions about the actual prevalence, context, and impact of scientific models, concepts, or methods. For instance, concerning a model like the *Hopfield model*, which originated in a particular scientific domain and was subsequently adopted in various other fields, *OpenAlex Mapper* can help investigate where this model truly established a foothold, where it became a significant tool, where it is still actively used, and where it continues to be referenced. The tool facilitates tracking such transportation and adoption of models across diverse scientific disciplines.

The methodological approach of *OpenAlex Mapper* involves sophisticated quantitative methods operating in the background. However, its primary aim is to support what are, in essence, qualitative heuristic investigations. The tool is designed with HPSS users in mind, enabling a fluid transition between a macroscopic view provided by the "big map" and a microscopic examination of specific results. Users can zoom in on particular clusters or individual papers to understand the nuances of what is occurring at a granular level. Crucially, all explorations conducted with the tool can be directly linked back to the original textual sources, as each data point on the map provides access to the corresponding paper.

5.5 Illustrative Applications

Several examples, all representing ongoing work and work in progress, illustrate the application of *OpenAlex Mapper*:

- The first use case involves the investigation of model templates, which was an original motivation for developing the tool. In the philosophy of science, "model templates" refer to a way of conceptualizing how models possessing very similar underlying structures can arise independently in different scientific disciplines. This concept also explores how such shared structures might organize scientific knowledge in a manner that cuts across, or is orthogonal to, traditional disciplinary boundaries. Using *OpenAlex Mapper*, the mapping of three distinct model templates showed that they appear in specific, sometimes non-continuous, locations on the base map of scientific literature.

- A second application is the mapping of scientific concepts. An example provided was the visualization of the concept "phase transition" contrasted with the concept "emergence," with the latter depicted in orange on the map. This type of analysis has been conducted previously, but *OpenAlex Mapper* offers the benefit of broadening such investigations into interdisciplinary contexts. This is particularly advantageous because obtaining specific and comprehensive datasets for interdisciplinary concept analysis can often be problematic.

- The third example focuses on analyzing the distribution of scientific methods across disciplines, particularly in interdisciplinary settings. This application is relevant to a current debate in the philosophy of science concerning the role of machine learning techniques versus more classical statistical methods in scientific research. To explore this, a specific machine learning technique, the *random forest* model, was compared with a somewhat analogous classical method, *logistic regression*. The analysis involved examining how these two methods are distributed across various disciplines. The results indicated "quite distinguishable patterns" in their usage. This observation, in turn, raises interesting philosophical questions, such as why researchers in a field like neuroscience might frequently employ *random forest* models, while those in closely related fields such as psychiatry or mental health research often opt for *logistic regressions*. Such findings can spur deeper inquiry into the underlying reasons for these differential methodological adoptions.

5.6 Technical Considerations and Limitations

Several qualifications and limitations are associated with *OpenAlex Mapper*:

- Firstly, the tool's effectiveness stands and falls with the *OpenAlex* database. While the data quality within *OpenAlex* is considered reasonable, especially when compared to other major scholarly databases, it is not perfect. This inherent imperfection is a constant consideration. Furthermore, there are known coverage issues, with certain disciplines like law and some areas of the humanities potentially being undersampled in the database.
- Secondly, the language model currently employed is an English-only version of *Specter 2*. This naturally limits the scope of analyses that can be performed, although this limitation might be less critical for research focusing on the recent history of science, where English-language publications are dominant. In principle, this could be remedied by incorporating multilingual language models, but there is a current scarcity of high-quality, science-trained multilingual models.
- Thirdly, the tool's embedding process is dependent on the availability of textual data. It is limited to sources that possess abstracts or, at a minimum, well-constructed and informative titles from which embeddings can be generated.
- A fourth significant dependency is on the *Uniform Manifold Approximation and Projection* (*UMAP*) algorithm. The entire method is heavily reliant on *UMAP* for dimensionality reduction. *UMAP*, while powerful, has several imperfections. It is a stochastic algorithm, which means that the specific 2D map generated is one of many possible valid outputs from the same input data. Repeated runs might produce slightly different layouts. More fundamentally, the process of reducing high-dimensional data—such as the 768-dimension embeddings from the *Specter* model—to a mere two dimensions necessitates significant tradeoffs. This reduction involves a degree of "pushing and pulling and misaligning" of the data points to fit them into the lower-dimensional space, which can lead to some loss of information or distortion of relationships.

For those interested in more in-depth technical details, a working paper has been prepared and is available online. This document offers a more comprehensive explanation of the technical components and methodologies underlying *OpenAlex Mapper*.

Chapter 6

Genre Classification for Historical Medical Periodicals

The ActDisease project at Uppsala University focuses on the history of patient organizations in 20th century Europe, using their periodicals as primary source material. The project aims to classify textual genres within these historical medical periodicals to facilitate nuanced historical analysis, study the evolution of communicative strategies, and enhance the accuracy of text mining techniques. The digitization of these periodicals involved Optical Character Recognition (OCR) using ABB…

6.1 Overview

The ActDisease project at Uppsala University focuses on the history of patient organizations in 20th century Europe, using their periodicals as primary source material. The project aims to classify textual genres within these historical medical periodicals to facilitate nuanced historical analysis, study the evolution of communicative strategies, and enhance the accuracy of text mining techniques.

The digitization of these periodicals involved Optical Character Recognition (OCR) using *ABBYY FineReader Server 14*. This process presented challenges with complex layouts, varied fonts, and scan quality, leading to OCR errors and disrupted reading order. Experiments were conducted for post-OCR correction of German texts using instruction-tuned generative models.

A key methodological component of the project is genre classification. Nine distinct genres were defined (Academic, Administrative, Advertisement, Guide, Fiction, Legal, News, Nonfiction Prose, QA) based on communicative purpose, under the supervision of a historian. Annotation was performed at the paragraph level on Swedish and German periodicals, achieving a Krippendorff's alpha of 0.95.

The research explored both zero-shot and few-shot learning approaches. Zero-shot experiments involved mapping ActDisease genres to labels from publicly available datasets (*CORE*, *FTD*, *UDM*) and fine-tuning multilingual encoder models (*XLM-Roberta*, *mBERT*, historical *mBERT* - *hmBERT*). Results indicated that models fine-tuned on *FTD* with custom mapping performed well, and specific models showed aptitude for certain genres (e.g., *hmBERT* for 'Administrative').

Few-shot learning experiments assessed performance with varying training data sizes, with and without prior Masked Language Model (MLM) fine-tuning. MLM fine-tuning significantly boosted performance, with *hmBERT-MLM* showing

the best results, particularly in distinguishing between fiction and nonfiction. Further experiments involved few-shot prompting of *Llama-3.1 8b Instruct*, which demonstrated decent quality but required more examples for complex genres.

The findings underscore that genre classification is vital for analyzing the diverse content of popular historical magazines. While zero-shot methods offer a viable starting point, few-shot learning with prior MLM fine-tuning of multilingual encoders, especially historical models like *hmBERT*, provides superior results. Future work includes refining annotation schemes, synthetic data generation, and active learning to improve classification quality. The tools and resources utilized include the ActDisease dataset, *ABBYY FineReader Server 14*, the *CORE*, *FTD*, and *UDM* datasets, and various language models (*XLM-Roberta*, *mBERT*, *hmBERT*, *Llama-3.1 8b Instruct*).

6.2 ActDisease Project and Dataset

Figure 6.1: Slide 01

The ActDisease project, titled "Acting out Disease - How Patient Organizations Shaped Modern Medicine," is an ERC-funded research initiative (ERC-2021-STG 101040999). It is based at the Department of History of Science and Ideas, Uppsala University, Sweden. Its primary purpose is to investigate the historical role of patient organizations in shaping disease concepts, illness experiences, and medical practices throughout 20th century Europe.

The project focuses on approximately 10 European patient organizations located in Sweden, Germany, France, and Great Britain, covering a period from around 1890 to 1990. The main source material for this research comprises periodicals, predominantly magazines, published by these patient organizations. This collection forms the ActDisease dataset, a private and recently digitized compilation totaling 96,186 pages.

The dataset includes materials related to various diseases and countries:

- Germany: Allergy/Asthma, Diabetes, Multiple Sclerosis.

- Sweden: Allergy/Asthma, Diabetes, Lung Diseases.
- France: Diabetes, Rheumatism/Paralysis.
- UK: Diabetes, Rheumatism.

Illustrative examples of these periodicals include “BRA Review” and “Allergia.” An example of the historical context is Heligoland, Germany, which was the founding place of the Hay Fever Association of Heligoland in 1897.

6.3 Digitization and OCR Challenges

Figure 6.2: Slide 01

The digitization of the ActDisease dataset was performed using Optical Character Recognition (OCR). The software employed for this task was *ABBYY FineReader Server 14*. While this OCR model performed well in recognizing most common layouts and fonts, several challenges persisted. These included difficulties with complex page layouts, slanted text, rare or unusual fonts, and inconsistencies in scan or photograph quality.

Consequently, some issues remain in the digitized collection. These include OCR errors, which are especially prevalent in German and French texts, and instances of disrupted reading order. To address some of these problems, experiments were conducted on post-OCR correction, specifically for German texts, utilizing instruction-tuned generative models.

The findings of these experiments are detailed in a publication by Danilova and Aangenendt (*RESOURCEFUL-2025, ACL*). It was also observed that OCR errors occur frequently in texts with creative formatting, such as advertisements, humor pages, and poems.

6.4 Motivation for Genre Classification

Figure 6.3: Slide 01

Exploration of the ActDisease materials revealed a high diversity of text types, which, however, exhibited similarities across the different magazines. A significant challenge identified is that various text types often appear concurrently on a single page—for instance, an administrative report might be placed next to an an advertisement and a humor section. This heterogeneity poses a problem for standard text mining approaches, as yearly and decade-based topic models and term counts typically do not differentiate between these text types, leading to a likely bias towards the most frequent genres.

To address these challenges, genre emerged as a useful concept. In language technology, genre is often defined as a class of documents that share a common communicative purpose, as described by Petrenz (2004) and Kessler (1997). The primary objective of implementing genre classification is to enable the exploration of the source material from multiple perspectives, thereby facilitating the formulation of historical arguments.

Specifically, genre classification allows for the study of how communicative strategies evolved over time, comparing these strategies across different countries, diseases, and publications. Furthermore, it permits a more fine-grained analysis of term distributions and topic models by examining them within distinct genre categories.

6.5 Genre Definitions and Examples

Figure 6.4: Slide 03

The ActDisease periodicals contain a variety of textual genres. Examples identified include poetry, academic reports (such as studies on the pancreas), legal documents (like deeds of covenant), advertisements (for instance, for diabetic chocolate), instructive or guidance texts (including recipes or dietary advice), patient organization reports detailing meetings and activities, and narratives about patient experiences and lives.

The genre labels used for classification were defined under the supervision of the project's main historian, who specializes in patient organizations. The labels were designed to be useful for segregating content within the materials to support further historical analysis, while also being as general-purpose as possible to allow for potential application to similar datasets.

The following genres, along with their definitions and communicative purposes, were established:

- Academic: Consists of research-based reports or explanations of scientific ideas, such as research articles or reports. Its purpose is to convey information from the scientific and medical communities to the magazine's audience.
- Administrative: Includes documents pertaining to organizational activities, like meeting minutes, reports, and announcements. It aims to report on and inform about the events and activities of patient organizations.
- Advertisement: Features content that promotes products or services for commercial purposes.
- Guide: Provides step-by-step instructions, such as health tips, legal advice, or recipes.
- Fiction: Encompasses texts designed to entertain and emotionally engage readers, including stories, poems, humor, and myths.
- Legal: Contains texts that explain legal terms and conditions, such as contracts, rules, or amendments.

- News: Comprises reports on recent events and developments.
- Nonfiction Prose: Includes narratives of real events or descriptions of cultural or historical topics, such as memoirs, essays, or documentaries.
- QA: Refers to sections structured as questions paired with expert answers, commonly found in the periodicals.

6.6 Annotation Process and Dataset Preparation

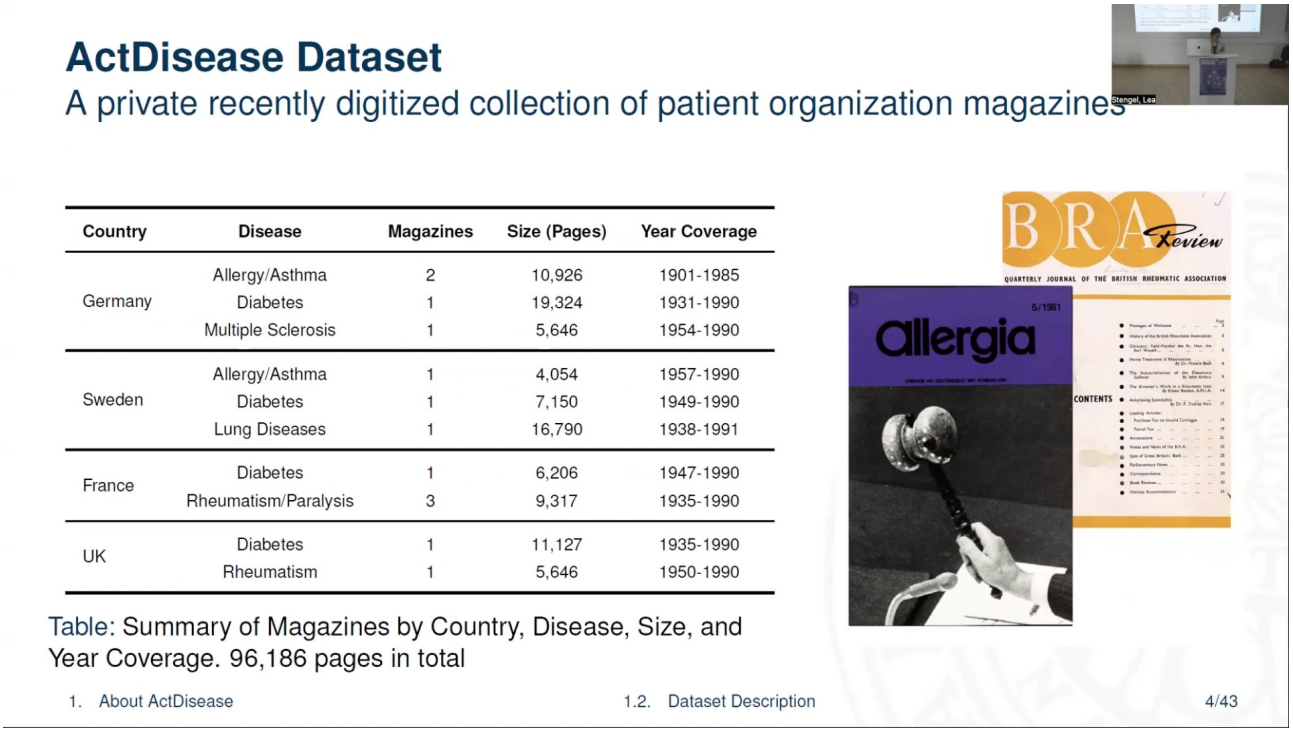

Figure 6.5: Slide 07

The annotation unit for genre classification was defined as paragraphs. These paragraphs were derived from the *ABBYY OCR* output and subsequently merged based on font patterns (type, size, bold, italic attributes) at the page level. For the annotation task, samples were drawn from two specific periodicals: the Swedish magazine "Diabetes" and the German magazine "Diabetiker Journal." The selection comprised the first and mid-year issues from each year of these publications.

A team of six project members, consisting of four historians and two computational linguists, all either native speakers or proficient in Swedish and German, performed the annotations. Two independent annotations were collected for each paragraph. The inter-annotator agreement achieved was 0.95, measured by Krippendorff's alpha, indicating a high level of consistency. An example of the annotation file shows a tabular structure with metadata (Year, Volume, Issue, etc.), the paragraph text, and columns for each genre where annotators made hard assignments.

The annotated dataset was split into a training set of 1182 paragraphs and a held-out set of 552 paragraphs (approximately 30% of the data), with stratification by label. For few-shot learning experiments, six different training set sizes were created (100, 200, 300, 400, 500, and 1182 paragraphs), randomly sampled from the main training set and balanced by label. The held-out set was further divided equally into validation and test sets, also balanced by label. The 'legal' and 'news' genres were excluded from these few-shot experiments due to insufficient training instances. For zero-shot experiments,

the entire test portion of the held-out set was utilized. Analysis of genre distribution in the training and held-out samples revealed a strong imbalance for the 'advertisement' and 'nonfictional prose' genres across the German and Swedish languages.

6.7 Zero-Shot Genre Classification

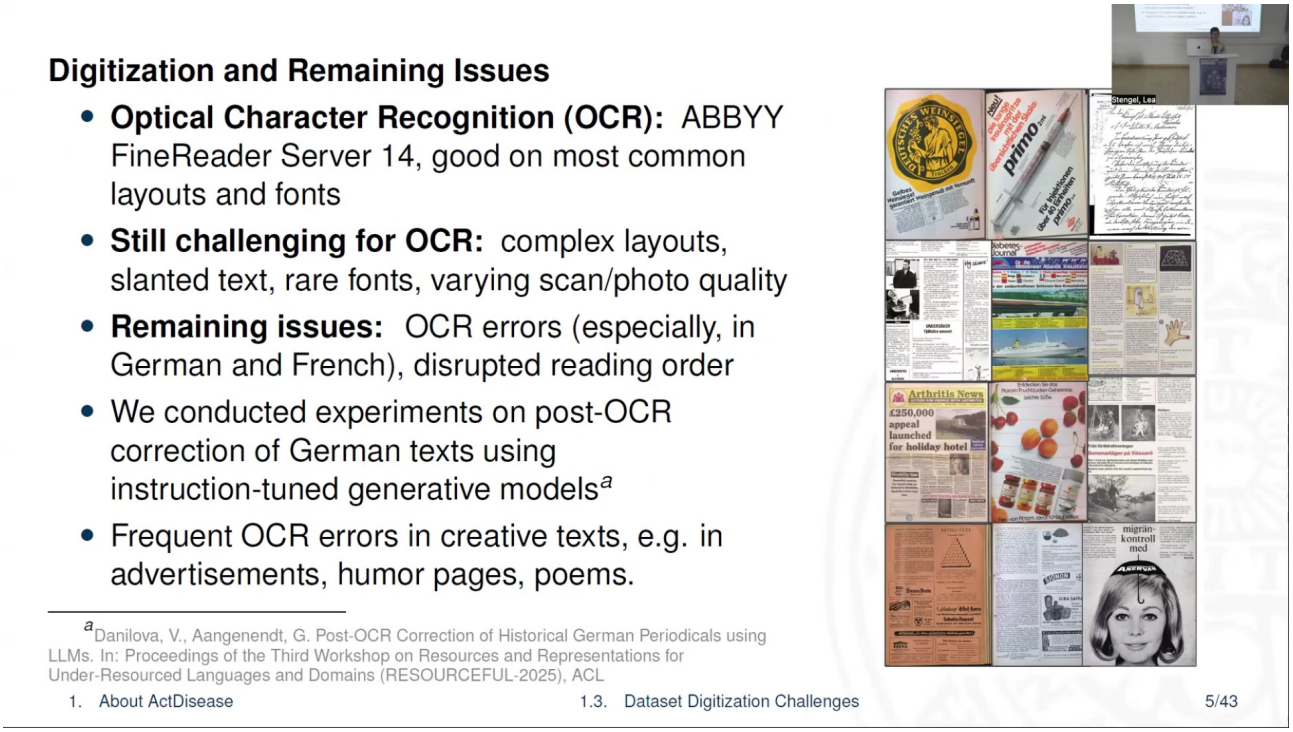

Figure 6.6: Slide 09

For zero-shot genre classification, the research addressed two main questions: whether genre labels from publicly available datasets could be efficiently mapped to the custom ActDisease labels, and how classification performance would vary across different datasets and models. Three publicly available datasets were utilized: the *Corpus of Online Registers of English (CORE)* by Egbert et al. (2015), annotated at the document level primarily in English; the *Functional Text Dimensions (FTD)* dataset of web genres by Sharoff (2018), also document-level, covering English and Russian; and *UD-MULTIGENRE (UDM)*, a subset of *Universal Dependencies* with recovered sentence-level genre annotations in 38 languages (de Marneffe et al., 2021; Danilova and Stymne, 2023).

Genre mapping between ActDisease labels and those in *CORE*, *UDM*, and *FTD* was performed independently by two annotators, with final mappings chosen based on full agreement. For example, the ActDisease 'Academic' genre was mapped to 'research article' (RA) in *CORE*, 'academic' in *UDM*, and 'academic' (A14) in *FTD*. Some ActDisease genres, like 'Administrative' in *CORE* and *FTD*, lacked direct suitable mappings.

The training data creation pipeline involved this mapping, followed by preprocessing (removing web addresses, emails, XML tags, and emojis), chunking, and sampling. Four sampling configurations were applied: using only Germanic languages [G+], using all language families [G-], balancing by ActDisease labels [B1], and balancing by both ActDisease and original dataset labels [B2]. This process generated four distinct training samples for each of the *FTD*, *CORE*, and *UDM* datasets.

Three multilingual encoder models were employed: *XLM-Roberta* (Conneau et al., 2020), *mBERT* (Devlin et al., 2019), and *historical mBERT (hmBERT)* (Schweter et al., 2022). *hmBERT* was of particular interest due to its pretraining on multilingual historical newspapers, including German and Swedish. Each model was fine-tuned on all dataset samples and configurations, resulting in 48 fine-tuned models. Reported metrics are averages across these configurations.

Evaluation of zero-shot predictions was challenging due to imperfect label set overlaps. Therefore, performance was analyzed for each genre separately, supplemented by confusion matrix analysis. The *X-GENRE* web genre classifier (Kuzman et al., 2023) served as a baseline. The evaluation scenario was cross-lingual for *FTD* and *X-GENRE* (which lacked German or Swedish training data for the mapped labels) and partially cross-lingual for *UDM* and *CORE*.

Overall results for zero-shot learning indicated that models fine-tuned on the *FTD* dataset with the custom ActDisease mapping performed better. Models trained on *UDM* and *CORE* exhibited some class-specific biases; for instance, *UDM*-trained models showed a bias towards 'news,' while *CORE*-trained models leaned towards 'guide.' Certain models demonstrated strengths for specific genres: *XLM-Roberta* fine-tuned on *UDM* achieved, on average, 32% more correct predictions for 'QA' compared to *mBERT* and *hmBERT*. Conversely, *hmBERT* fine-tuned on *UDM* yielded 16% more correct predictions for 'Administrative' than *XLM-Roberta* and *mBERT*. *CORE*-based models were effective at predicting the 'legal' genre. Confusion matrices for configurations like *hmBERT_UDM_True_True* illustrated these behaviors. Detailed per-category F1 scores, averaged across data configurations, were presented, with highlighted values indicating robust performance not attributable to systematic biases. For instance, *hmBERT* (in its non-MLM version for zero-shot context) showed strong F1 scores for 'administrative' and 'advertisement'. The different data sampling configurations (B1, B2, G+, G-) had varied impacts: for *FTD*, B2 and G+ decreased performance, while for *UDM*, including other language families and applying balancing generally improved macro F1 scores.

6.8 Few-Shot Classification with Encoder Models

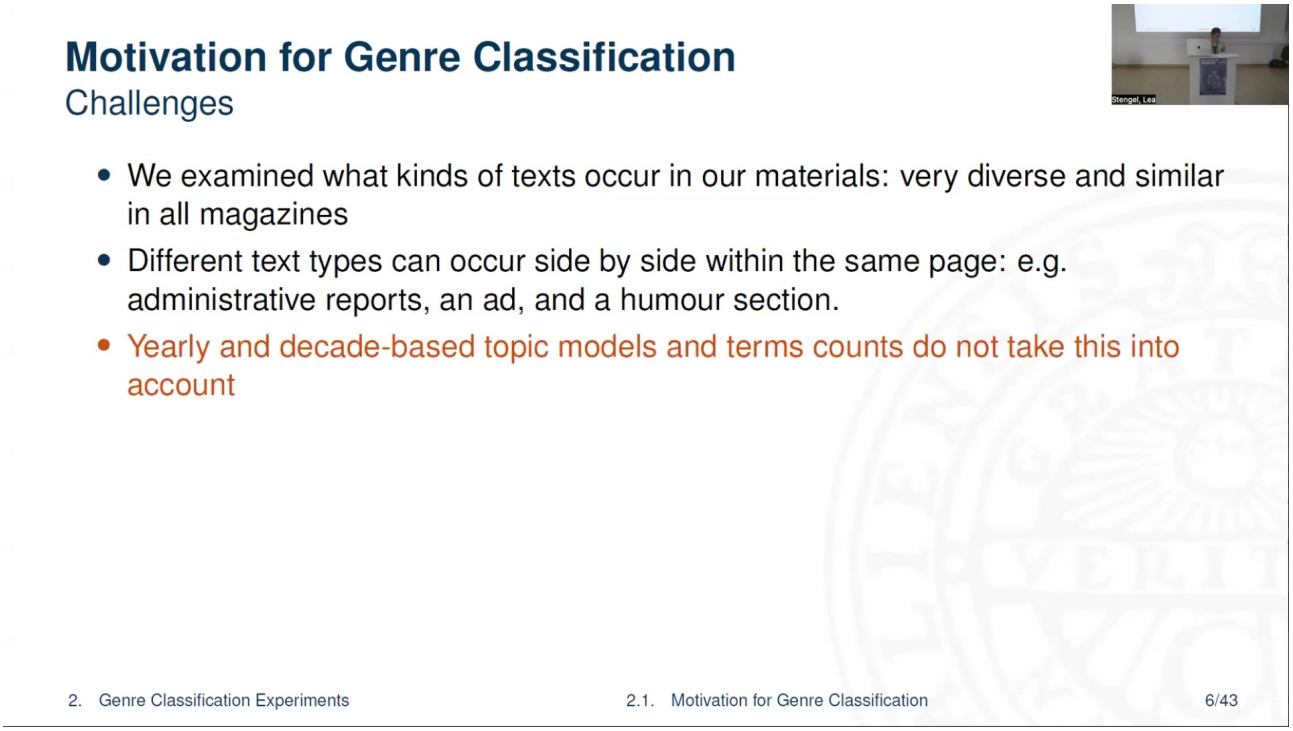

Figure 6.7: Slide 13

Few-shot learning experiments aimed to understand how classification performance changes with varying training set sizes across different models, and whether prior fine-tuning on the full dataset (referred to as MLM, in the context of Masked Language Model pre-training benefits) significantly enhances performance. The models tested included *hmbert*, *mbert*, and *xlmr*, each with a version that had undergone prior MLM fine-tuning on the ActDisease dataset (*hmbert-mlm*, *mbert-mlm*, *xlmr-mlm*). Training dataset sizes ranged from 100 to 1182 instances.

The results demonstrated that prior MLM fine-tuning provided a clear advantage. F1 scores consistently increased with larger training set sizes for all models. However, even with the maximum training size of 1182 instances, F1 scores generally remained below 0.8. The *hmBERT-MLM* model slightly outperformed other models. For example, with 1182 training instances, *hmBERT-MLM* achieved a macro F1 of 0.77 and an accuracy of 0.82, showing strong performance in categories like 'administrative' (0.86 F1), 'advertisement' (0.93 F1), and 'nonfiction prose' (0.82 F1). In comparison, *XLMR-MLM* achieved a macro F1 of 0.76 and accuracy of 0.84, with high scores in 'QA' (0.84 F1) and 'legal' (0.89 F1), but lower for 'nonfiction prose' (0.56 F1).

The superior performance of *hmBERT-MLM* was partly attributed to its sustained ability to differentiate between 'fiction' and 'nonfiction prose' even with the full dataset, a task where other models, particularly *XLM-Roberta*, experienced a significant performance decline. An analysis of the *XLM-Roberta-MLM* confusion matrix using the full-sized training dataset revealed that 'nonfictional prose' was frequently misclassified as 'fiction.' This suggests that within the specific domain of diabetes-focused patient organization magazines, 'fiction' and 'nonfictional prose' might share many thematic and narrative elements, especially since both often revolve around patient experiences. This similarity could become more pronounced with larger datasets, indicating that further data or alternative methods might be necessary to better distinguish these genres.

6.9 Few-Shot Prompting with *Llama-3.1 8b Instruct*

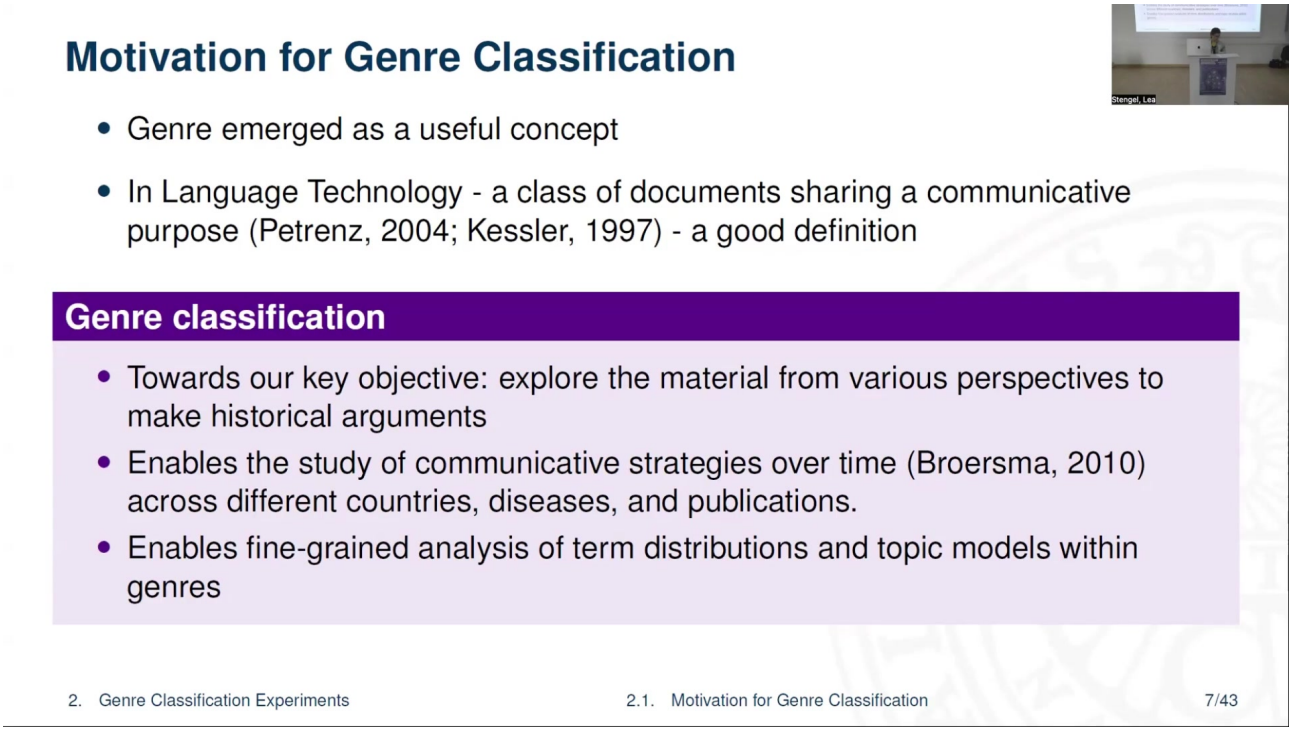

Figure 6.8: Slide 16

Due to the limited availability of annotated data for comprehensive instruction tuning, few-shot prompting experiments were conducted using *Llama-3.1 8b Instruct*. This model is a popular, multilingual, generative language model with open weights. The method involved constructing a prompt that included an instruction section with definitions for each genre, followed by an examples section containing two to three carefully selected examples per genre. The input text from the test set was then provided, and the model was prompted to output the predicted genre.

The evaluation was performed on the zero-shot test set, which is the entire held-out portion of the annotated data. The results indicated that *Llama-3.1 8b Instruct* could handle certain genre labels reasonably well; for instance, it achieved an F1-score of 0.84 for the 'legal' genre. However, the small number of examples (two or three per genre) proved insufficient for the model to adequately learn and represent more complex or internally diverse genres, such as 'nonfictional prose,' 'advertisement,' and 'administrative.' The confusion matrix of its predictions showed correct classifications along the diagonal for several genres but also highlighted areas of confusion.

6.10 Conclusions and Future Work

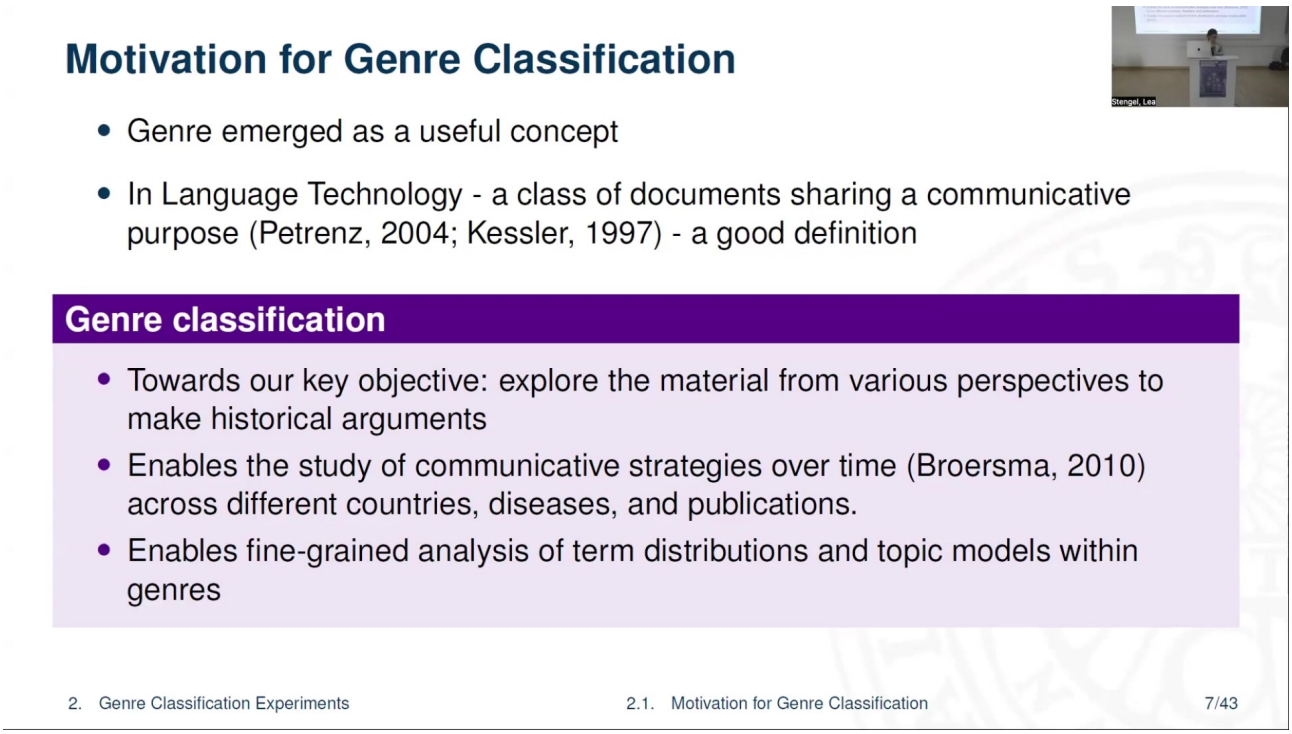

Figure 6.9: Slide 17

The research concludes that popular magazines, rich in varied content, present significant text mining challenges due to their multitude of genres, unlike more homogenous sources like scientific journals and books. These genres are indicative of chosen communicative strategies, and their consideration is vital for an accurate and detailed interpretation of text mining outputs. Genre classification serves as a key method to make these complex historical sources more accessible for computational analysis.

For scenarios with no training data, two potential approaches are suggested: leveraging existing modern datasets if their genre categories are sufficiently general-purpose and align with the target material, or employing few-shot prompting with a capable generative model. However, if some annotated data is available, few-shot learning using multilingual encoders such as *XLM-Roberta* or, particularly, *historical multilingual BERT (hmBERT)*, especially when combined with

prior MLM fine-tuning, proves to be a more effective strategy. The most significant performance improvements from MLM fine-tuning were observed for *hmBERT*, which showed a 24% gain, compared to 14.5% for *mBERT-MLM* and 16.9% for *XLM-RoBERTa-MLM*.

Ongoing and future work aims to further enhance the quality of this research. This includes applying the classification to investigate specific historical hypotheses, developing and implementing a new annotation scheme with more fine-grained genre distinctions, undertaking a new annotation project funded by Swe-CLARIN, exploring synthetic data generation techniques to augment training sets, and employing active learning strategies to optimize the annotation process. These efforts are directed at improving genre classification quality for both the ActDisease project's internal research needs and for the benefit of the wider research community.

6.11 Acknowledgements

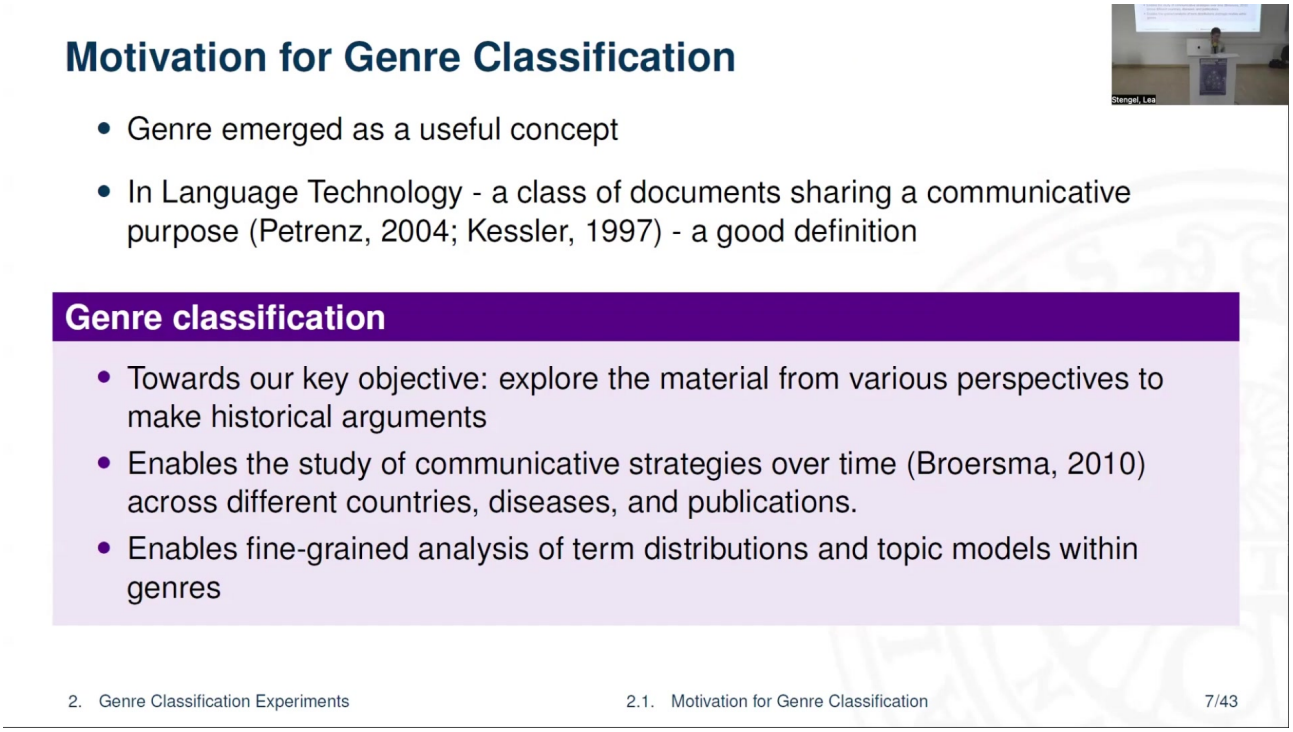

Figure 6.10: Slide 19

Acknowledgements are extended to the annotation team, composed of project members Ylva Söderfeldt, Julia Reed, Andrew Burchell, Maria Skeppstedt, and Gijs Aangenendt. The project received funding from the European Research Council (ERC) under grant ERC-2021-STG 101040999. Support in the form of GPU resources and data storage was provided by the Centre for Digital Humanities and Social Sciences. The contributions of reviewers, including Dr Maria Skeppstedt and anonymous reviewers, are also acknowledged. The project website can be accessed via a QR code presented on the final slide.

Chapter 7

Computational HPSS: Tracing Ancient Wisdom's Influence with VERITRACE

The VERITRACE project (2023-2028), an ERC Starting Grant initiative (101076836) at Vrije Universiteit Brussel (VUB) and accessible via HTTPS://VERITRACE.EU, aims to trace the influence of the early modern 'ancient wisdom' or Prisca Sapientia tradition on the development of natural philosophy and science. This tradition is found in texts such as the Chaldean Oracles, Sibylline Oracles, Orphic Hymns, and the Corpus Hermeticum. The project involves large-scale multilingual exploration of app…

7.1 Overview

The *VERITRACE* project (2023-2028), an ERC Starting Grant initiative (101076836) at Vrije Universiteit Brussel (VUB) and accessible via HTTPS://VERITRACE.EU, aims to trace the influence of the early modern 'ancient wisdom' or *Prisca Sapientia* tradition on the development of natural philosophy and science. This tradition is found in texts such as the *Chaldean Oracles*, *Sibylline Oracles*, *Orphic Hymns*, and the *Corpus Hermeticum*.

The project involves large-scale multilingual exploration of approximately 430,000 printed texts published between 1540 and 1728. These texts are sourced from *Early English Books Online* (*EEBO*), *Gallica* (French National Library), and the *Bavarian State Library*, covering at least six languages.

VERITRACE employs computational History, Philosophy, and Sociology of Science (*HPSS*) methods, including keyword search and textual reuse detection (both lexical and semantic). This approach functions similarly to an "early modern plagiarism detector" and seeks to uncover previously ignored networks of texts, passages, themes, topics, authors, and new patterns in intellectual history.

Core challenges for the project include variable Optical Character Recognition (OCR) quality from raw library-provided texts (in XML, HOCR, HTML formats) without ground truth page images. Further complexities arise from early modern typography and semantics across multiple languages, alongside the sheer volume of data.

Large Language Models (*LLMs*) are utilized in two main capacities:

- *GPT*-based *LLMs* function as 'LLMs-as-Judges' for enriching and cleaning metadata. This involves matching *VERITRACE* records with high-quality metadata from the *Universal Short Title Catalogue* (*USTC*). This approach,

utilizing a panel of *LLMs* (Primary, Secondary, Tiebreaker, Expert), currently faces challenges with hallucinations when using open-source models like *Llama*.

- *BERT*-based *LLMs*, specifically *Language-agnostic BERT Sentence Embeddings* (*LaBSE*), are used for generating vector embeddings. These embeddings encode semantic meaning for text matching and are implemented within the *VERITRACE* web application.

A 15-stage data processing pipeline prepares the textual data, which is then indexed in an *Elasticsearch* backend. The alpha-stage *VERITRACE* web application, currently hosted locally and not publicly available, includes several modules:

- The Explore module provides corpus statistics (e.g., language distribution, documents by decade, sourced from *MongoDB*) and allows detailed metadata viewing. This includes features for multilingual content identification within texts and experimental page-by-page OCR quality assessment.

- The Search module, powered by *Elasticsearch*, enables keyword-based searches with support for complex queries (AND, OR, nested) and proximity queries.

- The Analyse module is planned to include tools for topic modeling, Latent Semantic Analysis (LSA), and diachronic analysis.

- The Read module integrates a *Mirador* viewer for accessing digital facsimiles (PDFs) of the historical texts alongside their metadata.

- The Match module is designed to find textual similarities. It supports lexical (keyword-based), semantic (vector embedding-based), and hybrid matching techniques. Users can customize parameters and select matching modes (Standard, Comprehensive).

A case study involving matching Newton's Latin *Optice* (1719) and English *Opticks* (1718) demonstrated the Match module's capabilities. Lexical matching, in standard mode, correctly found no significant cross-language matches. Semantic matching using *LaBSE* produced conceptually reasonable matches (e.g., similarity scores of 90-92%) but revealed issues with coverage scores and the overall adequacy of the *LaBSE* model. This inadequacy is potentially due to out-of-domain model collapse when processing historical, multilingual, and OCR-affected text.

Future work will address challenges such as selecting or fine-tuning more suitable multilingual embedding models (alternatives include *XLM-Roberta*, *intfloat/multilingual-e5-large*, historical *mBERT*). Other challenges include managing semantic change over time in historical texts, mitigating the impact of poor OCR quality (through selective re-OCR or sourcing higher-quality text versions), and ensuring the scalability and performance of the system for the full corpus of 430,000 texts.

7.2 Project Foundation and Objectives

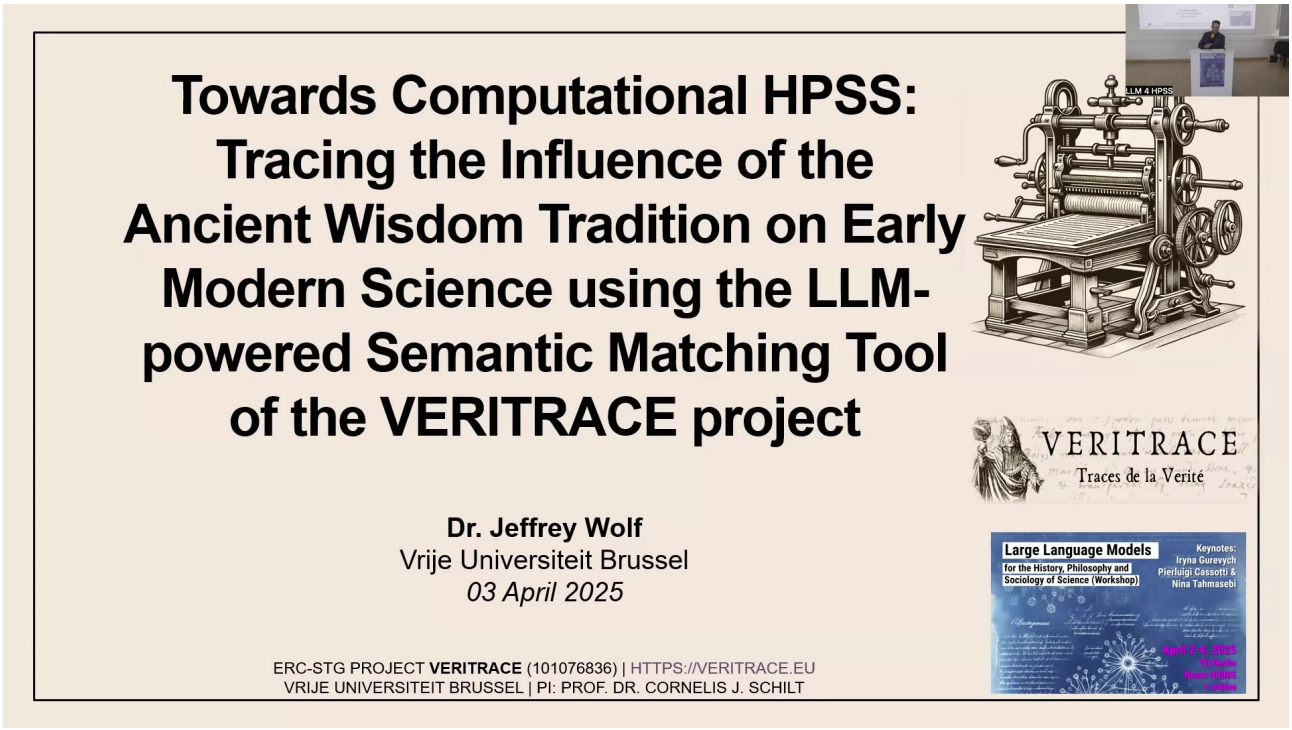

Figure 7.1: Slide 01

The *VERITRACE* project, subtitled "Traces de la Verité," is a five-year initiative funded by an *European Research Council* (*ERC*) Starting Grant (No. 101076836), running from 2023 to 2028. The project is based at the Vrije Universiteit Brussel (*VUB*) under the leadership of Principal Investigator (PI) Prof. Dr. Cornelis J. Schilt.

The research team comprises five members: the PI, Dr. Eszter Kovács (a classicist), Dr. Jeffrey Wolf (a historian of science and medicine, serving as the digital humanities lead for this project), Niccolò Cantoni (a historian), and Demetrios Paraschos (a historian). While the team is primarily based in Brussels, Dr. Wolf operates from Berlin. The project's official website is HTTPS://VERITRACE.EU.

The central research objective of *VERITRACE* is to trace the influence of what is termed an early modern 'ancient wisdom' or *Prisca Sapientia* tradition on the development of natural philosophy and science during the early modern period. This tradition is found in a range of texts, including the *Chaldean Oracles*, the *Sibylline Oracles*, the *Orphic Hymns*, and the *Corpus Hermeticum*, the last of which is particularly noted for its relevance to the history of chemistry. A core collection of 140 works has been identified as representing this 'ancient wisdom' tradition, forming a close reading corpus for the project.

While some connections between prominent scientific figures and these texts are known—for instance, Isaac Newton's engagement with the *Sibylline Oracles* and Johannes Kepler's familiarity with the *Corpus Hermeticum*—*VERITRACE* aims to delve deeper. The project seeks to uncover a much broader network of texts and authors who engaged with this tradition, many of whom may be lesser-known and constitute what one scholar has termed the 'great Unread.' Dr. Wolf, whose primary historical research focuses on the 18th century (a period that begins as the *VERITRACE* project's timeline concludes), is responsible for the digital humanities components of this endeavor.

7.3 Computational HPSS Framework

Figure 7.2: Slide 02

The *VERITRACE* project adopts a *Computational History, Philosophy, and Sociology of Science* (*HPSS*) framework to address its research questions. The core methodology involves large-scale multilingual exploration of its textual corpus. This exploration is facilitated by several computational tools and techniques.

Keyword search capabilities allow for targeted queries within the dataset. A significant focus is placed on identifying textual reuse across the corpus. This includes detecting both direct (lexical) reuse, such as explicit quotations (whether cited or not), and indirect (semantic) reuse, which encompasses paraphrases or subtle allusions that contemporary readers would have recognized (for example, an indirect reference to the *Corpus Hermeticum*).

Given the very large and multilingual nature of the corpus, these tools are designed to operate effectively across diverse linguistic and textual materials. The project aims, in effect, to build an "Early Modern Plagiarism Detector." The primary objective of this computational approach is to uncover networks of texts, passages, themes, topics, and authors that may have been overlooked by traditional historical methods. A secondary objective is the potential discovery of new patterns and insights within the intellectual history and philosophy of science.

7.4 Data Set: Composition and Sources

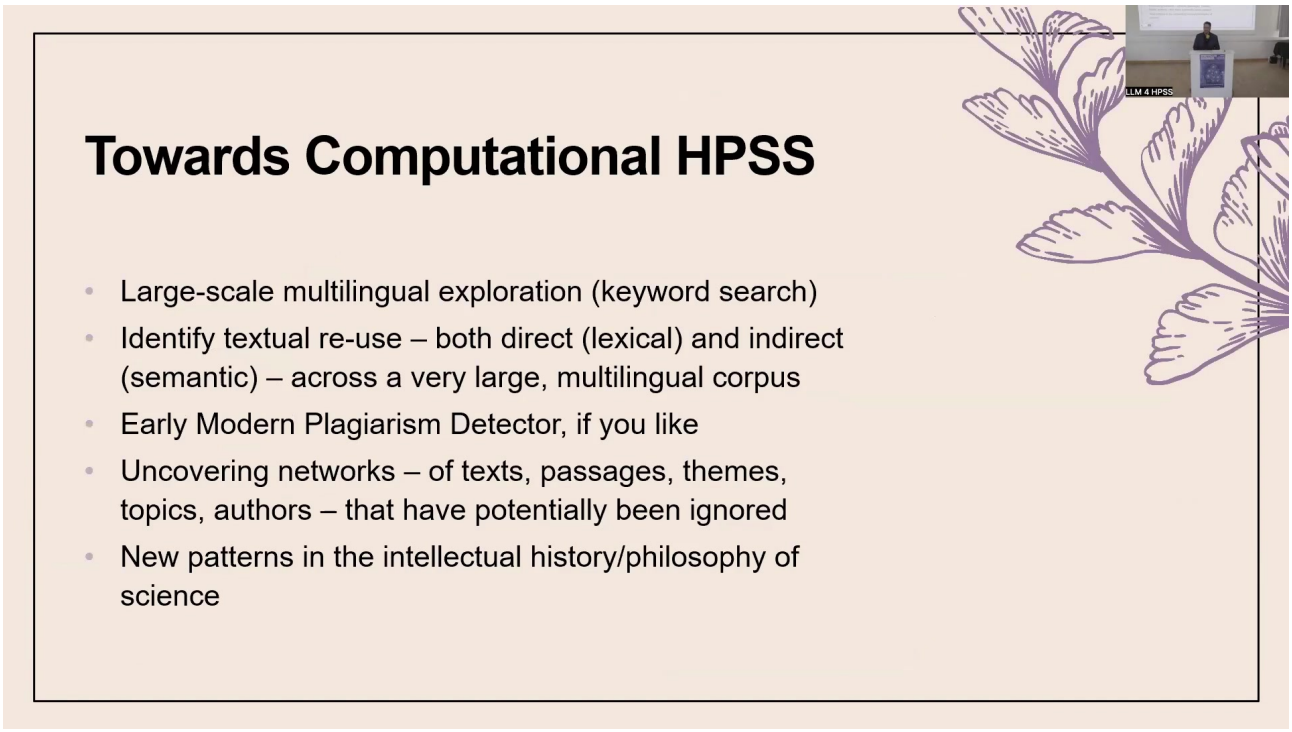

Figure 7.3: Slide 04

The *VERITRACE* project utilizes a large, diverse, and multilingual data set composed exclusively of digital texts derived from printed works; handwritten materials are intentionally excluded from its scope. The chronological span of the corpus covers approximately 200 years, beginning in 1540 (a starting point chosen for various, though unspecified, reasons) and concluding in 1728, a date selected because it is shortly after the death of Isaac Newton. The texts within the corpus are in at least six different languages.

The data is aggregated from three primary multilingual sources:

- *Early English Books Online* (*EEBO*), which is noted as being freely downloadable.
- *Gallica*, the digital library of the French National Library, from which sources have been downloaded.
- The *Bavarian State Library*, which constitutes the largest single data source for the project.

In total, these sources contribute to a corpus of approximately 430,000 texts. The project plans to employ a range of state-of-the-art digital techniques for analysis, including Keyword Search, Text Matching, Topic Modelling, and Sentiment Analysis, among others.

7.5 Challenges and LLM Strategy

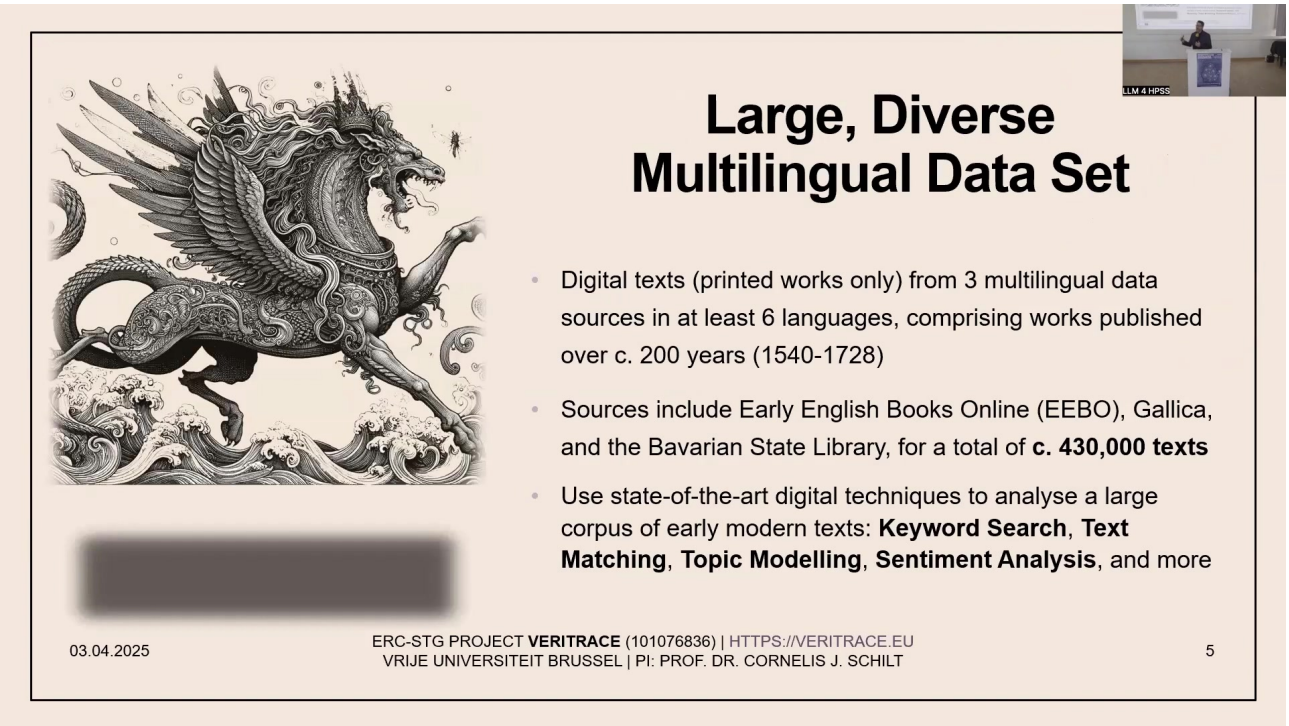

Figure 7.4: Slide 05

The *VERITRACE* project confronts several core challenges inherent in working with large historical textual datasets. A primary issue is the variable quality of Optical Character Recognition (OCR) in the texts, which are supplied by libraries in raw digital formats such as XML, HOCR, or even HTML files. Critically, these raw texts are provided without corresponding ground truth page images, making OCR error correction difficult and impacting all subsequent data processing stages.

A second significant challenge lies in handling early modern typography and semantics across at least six different languages, each with its own historical linguistic complexities. Thirdly, the sheer scale of the data—hundreds of thousands of texts printed across Europe over a span of roughly 200 years—presents substantial data management and processing hurdles.

To address some of these challenges, particularly in text analysis, the project employs Large Language Models (*LLMs*) in two distinct roles. On the "decoder-side," *GPT*-based *LLMs* are utilized to help enrich and clean the metadata associated with the texts. This involves a methodology referred to as "LLMs-as-Judges." While this application is part of the project, it is not the primary focus of this particular presentation. The presentation concentrates on the "encoder-side" application of *LLMs*, where *BERT*-based models are used to generate embeddings. These embeddings aim to encode the semantic meaning of sentences and passages (defined as short groups of sentences) within the textual corpus, primarily to facilitate text matching tasks.

7.6 Metadata Enrichment with LLMs-as-Judges

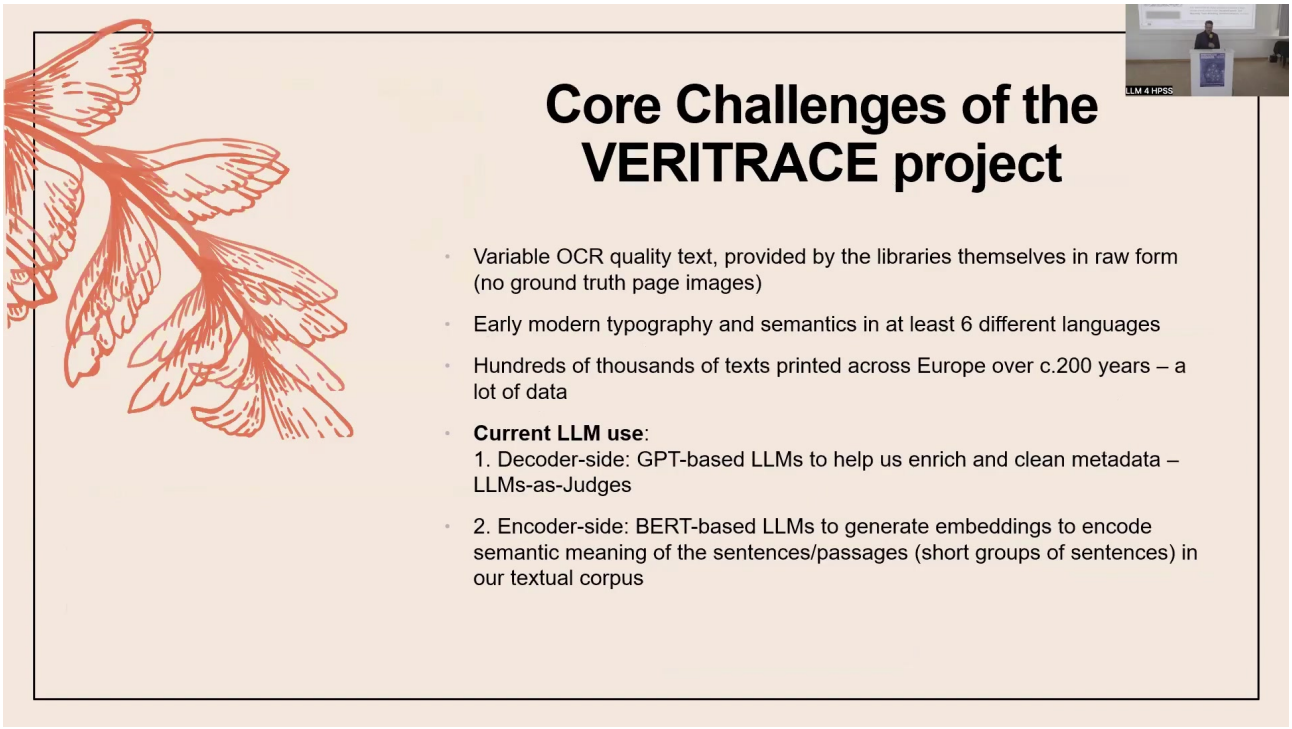

Figure 7.5: Slide 06

A specific application of *LLMs* within the *VERITRACE* project, though not detailed extensively in this presentation, involves their use as "LLMs-as-Judges" for metadata enrichment. The primary motivation for this sub-project is to improve the quality of *VERITRACE*'s metadata by mapping its records to corresponding entries in the *Universal Short Title Catalogue* (*USTC*), a recognized high-quality metadata source available at https://www.ustc.ac.uk.

Successfully matched records result in "enriched" *VERITRACE* metadata that is less likely to require extensive manual cleaning. The challenge stems from the fact that while some record mapping can be automated using external identifiers, a majority of records cannot be matched this way, especially since the initial *VERITRACE* metadata is uncleaned. The alternative, manual comparison of bibliographic details for tens of thousands of record pairs (with each team member initially assigned 10,000 such pairs), is an extremely tedious task.

To automate this, the project is developing an "LLM Bench," a panel of multiple *LLMs*, to evaluate whether a given pair of bibliographic records—one from a low-quality source (*VERITRACE*) and one from a high-quality source (*USTC*)—represent the same underlying printed text. This bench is configured with a Primary *LLM*, a Secondary *LLM*, a Tiebreaker *LLM*, and an Expert *LLM* for handling difficult edge cases.

The process requires these *LLMs* not only to judge whether a match exists but also to provide detailed reasoning for their decisions and a confidence level for each judgment. These *LLM*-generated assessments are then compared against ground truth data, with the *VERITRACE* team conducting a final review. This system relies on extensive prompt guidelines to direct the *LLMs* regarding matching criteria and the desired output format, which includes fields like the ground truth status, the *LLM*'s final decision, a confidence score (e.g., "HIGH (87.7%)"), the decisions of individual models in the bench, key factors considered (such as title similarity, author match, date match, and place match), and a narrative reasoning.

Currently, this *LLM*-as-Judges system is a work in progress and is not yet fully functional. A major challenge encountered

is the occurrence of hallucinations in the *LLM* output, where the models (primarily open-source models like *Llama*, not frontier models) generate information about records that were not part of the input. Attempts to mitigate these hallucinations by requesting more structured output have had mixed results; while structured output can reduce hallucinations, it often leads to the *LLMs* providing more generic and less helpful responses, especially in their reasoning. The process of finding the optimal balance in prompting and model configuration is described as being "more art than science." Despite these ongoing challenges, the potential for this system to save significant time is considered substantial, and the project remains open to advice on improving this aspect.

7.7 Web Application and Data Infrastructure

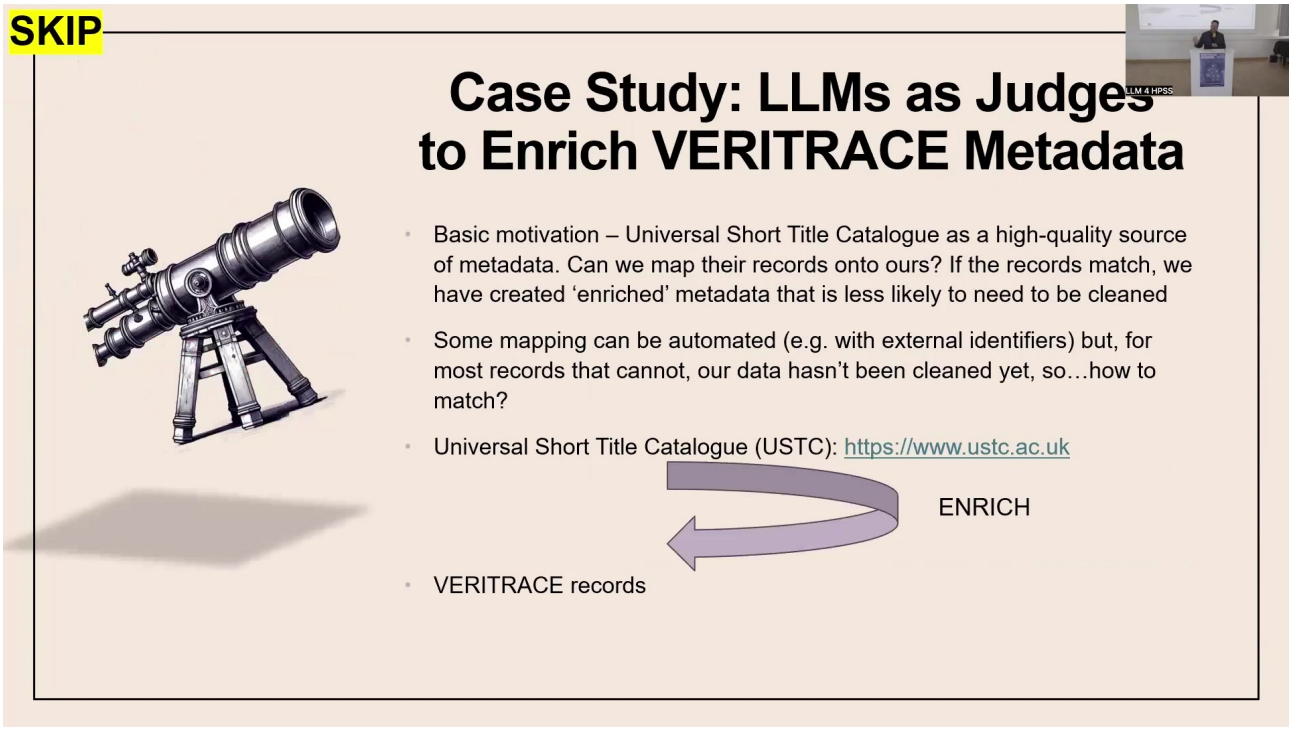

Figure 7.6: Slide 08

The *VERITRACE* project is developing a web application to provide access to its data and analytical tools. This application is currently in an alpha version and is described as extremely new, to the extent that it had not yet been shared with the full project team prior to this presentation. It is not publicly available and runs on the presenter's local computer. The demonstration of this application is intended more as a "promise of what we want to do" rather than a showcase of a finished product.

For semantic analysis within this application, the project is currently testing a *BERT*-based Large Language Model, specifically *LaBSE* (*Language-agnostic BERT Sentence Embeddings*). This model is used to generate vector embeddings intended to represent every passage within the corpus texts. However, a preliminary assessment suggests that *LaBSE* is "probably not good enough" for the project's ultimate requirements, even though it demonstrates functionality in some instances.

Underpinning the web application is a substantial data processing pipeline. This pipeline takes raw text files provided by libraries in formats such as XML, HOCR, and HTML, and processes them for storage and indexing in an *Elasticsearch*

database, which serves as the backend for the website. The pipeline is complex, consisting of 15 distinct stages. Examples of these stages include extracting text into plain text files, generating mappings of all character positions, segmenting the text into meaningful units, and assessing OCR quality. Each of these stages requires careful optimization. The generation of vector embeddings using models like *LaBSE* occurs towards the end of this multi-stage pipeline, after the initial text processing and cleaning steps.

7.8 Explore Module: Corpus Overview

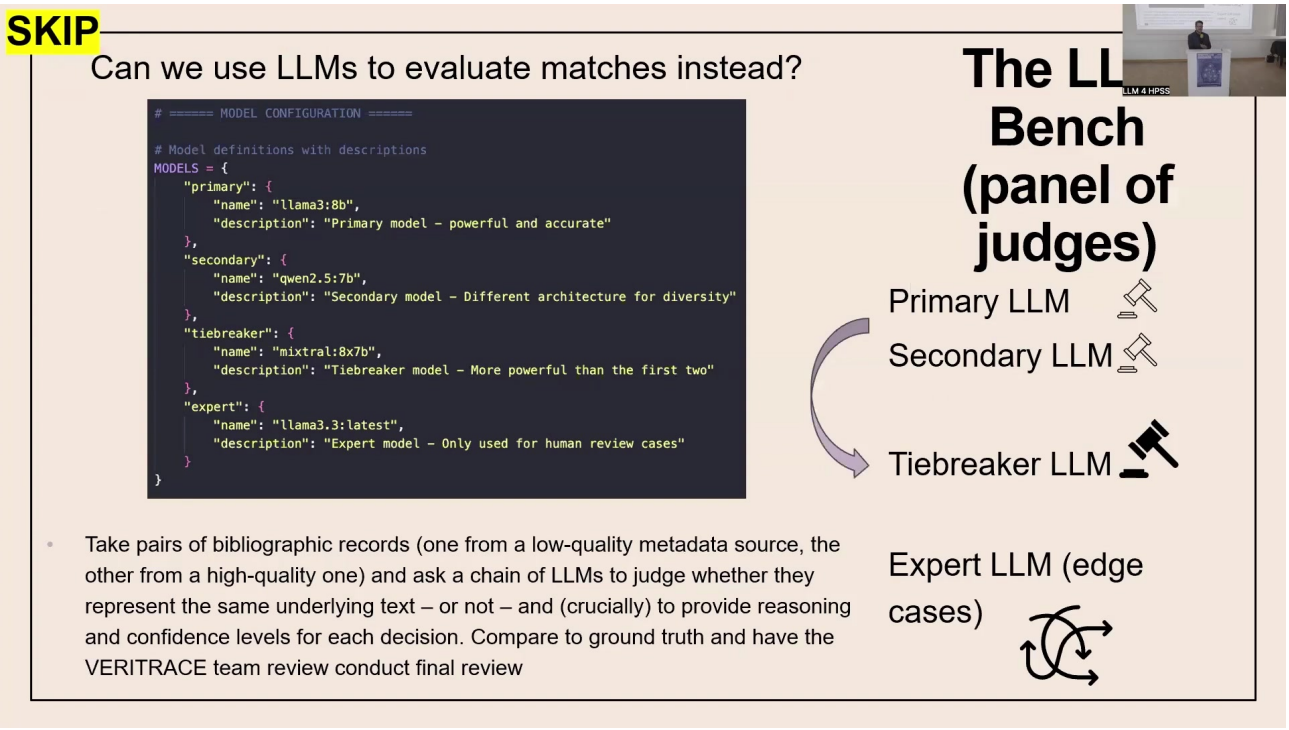

Figure 7.7: Slide 10

The *VERITRACE* web application is structured into approximately five main sections: Explore, Search, Match, Analyse, and Read. The "Explore" section is designed to offer users an overview of the corpus through various statistics and to enable detailed inspection of the metadata associated with the texts.

For corpus-level statistics, this section draws data from a *MongoDB* database. At the time of presentation, the system contained 427,395 metadata records. These statistics are visualized through several charts, including:

- A pie chart for Language Distribution.
- A pie chart showing Documents by Data Source.
- A bar chart illustrating Documents by Decade.
- A bar chart depicting Publication Places.

Within the Explore section, an "*Elasticsearch* Metadata Explorer" provides functionality for users to browse sample records and examine the rich metadata compiled or generated for each text. Two features are particularly highlighted. First, detailed "Language Information" is provided for each document. This is derived from a language identification

process run on every text, capable of analyzing segments as small as approximately 50 characters. This granular approach is crucial for accurately capturing the linguistic makeup of multilingual texts, which might be inadequately described by basic bibliographic metadata (e.g., a book simply labeled "Latin" might contain significant portions in Greek). An example given is a text identified as being 15% Greek and 85% Latin, which can then be classified as "substantively multilingual."

Second, the system attempts an "OCR Quality Assessment." This is a challenging task as it is performed on the raw text without access to ground truth page images. The assessment aims for page-by-page granularity, rather than assigning a single quality score to an entire book. This feature is acknowledged as difficult and experimental. Other metadata fields available for inspection include Document ID, Filename, File Path, Bibliographic Title, Author, Printer, Publication Place, Date, Format, Subject, comprehensive Language Information (including details on multilingual content, secondary languages and confidence scores, and detailed language distribution percentages), OCR Information (including a quality distribution chart), and Document Statistics (such as number of pages, segments, and character length).

7.9 Search Module Functionality

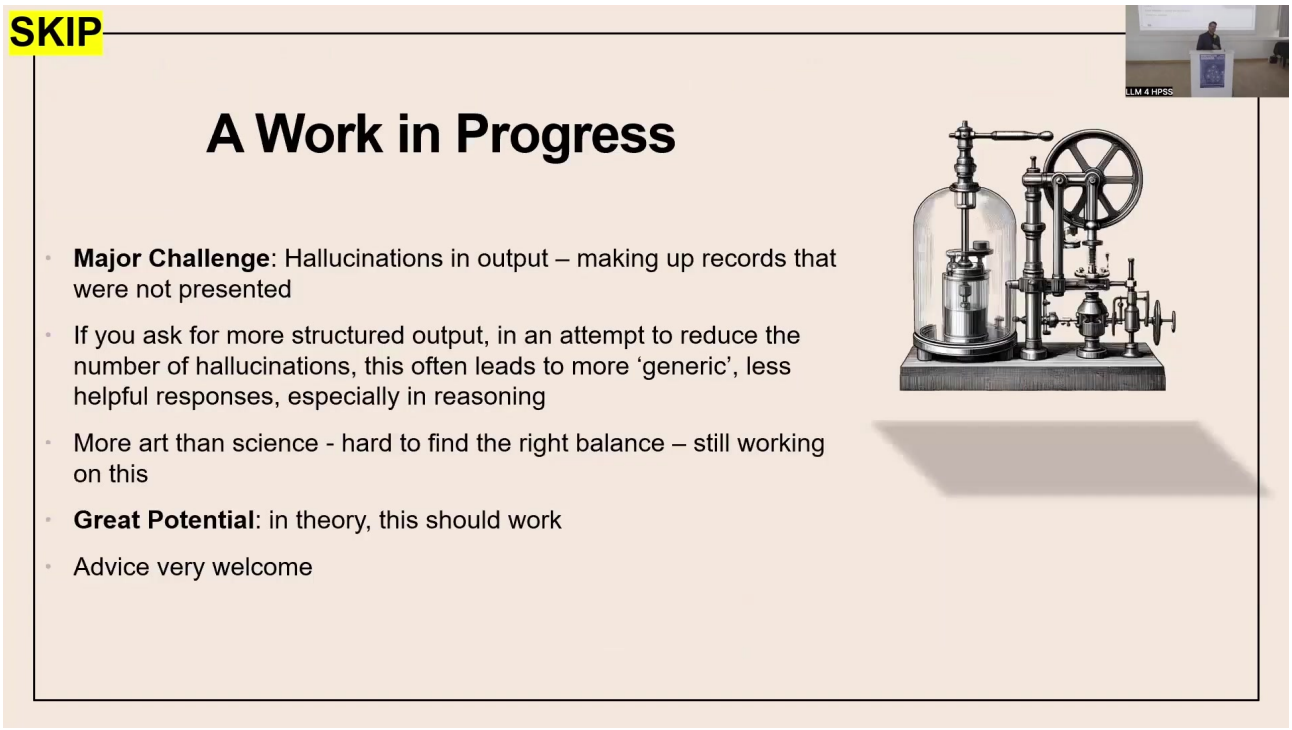

Figure 7.8: Slide 12

The "Search" section of the *VERITRACE* web application is anticipated to be a primary entry point for scholarly users, offering standard keyword search functionalities. This module is powered by an *Elasticsearch* backend. The version demonstrated during the presentation operates on a prototype corpus comprising 132 files, which represent a small fraction of the total 430,000+ texts. Even for this subset, the index contains 16,991,177 segments and occupies 15.37 GB of storage. It is projected that the index for the full corpus will scale into the terabyte range.

The search module supports a variety of query types. Users can perform basic keyword searches; for example, a search for "hermes" within the prototype corpus returned 22 documents with a total of 332 matches. More advanced fielded queries allow users to target specific metadata fields, such as searching for "author:kepler 'hermes'," which in the prototype

identified one document with two matches.

The system also handles complex queries incorporating Boolean operators (AND, OR) and nested query structures. Furthermore, proximity queries are supported, enabling users to find texts where specified terms appear within a certain word distance of each other, for instance, locating instances where “Hermes” and “Plato” are mentioned within 10 words of one another. The search results are displayed in a table format, providing details such as Filename, Title, Author, Date, Language, a relevance score, the number of segments with hits in the document, and the total number of matches for the query terms. The basic search functionalities are reported to be already operational.

7.10 Analyse (Planned) and Read Modules

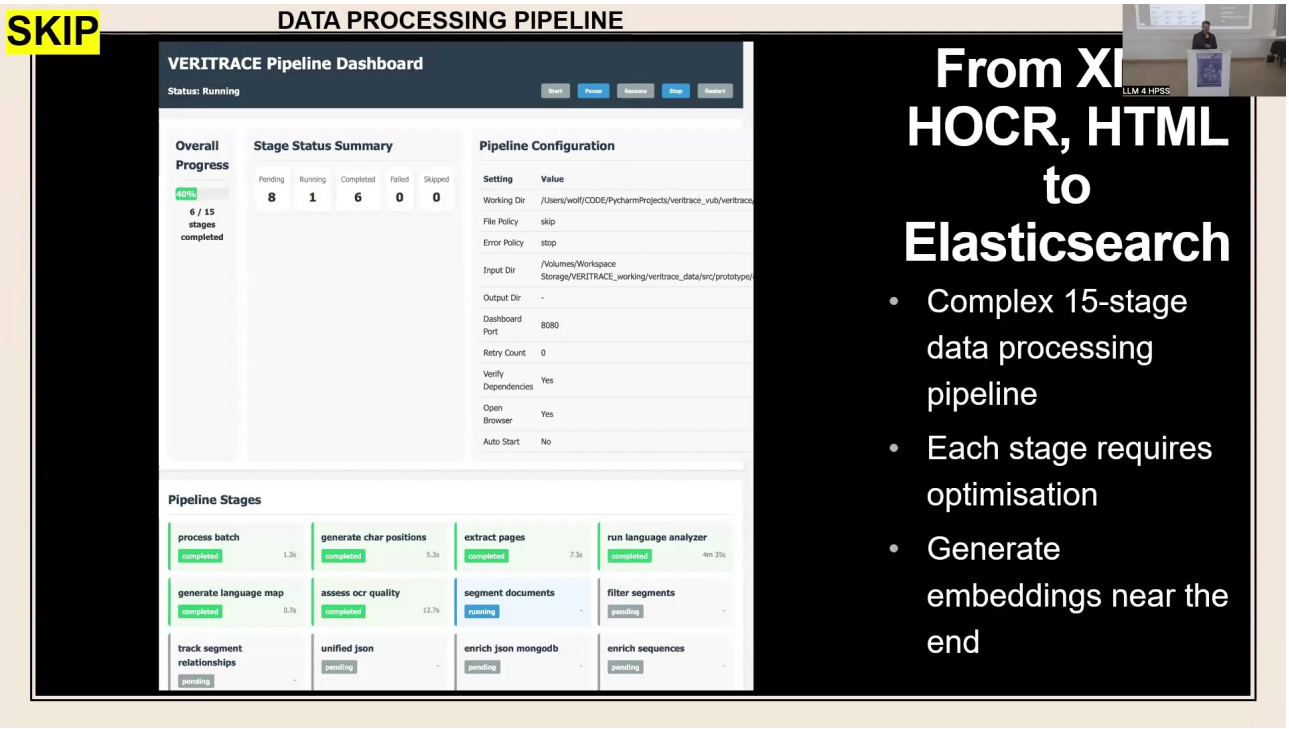

Figure 7.9: Slide 14

The *VERITRACE* web application includes an “Analyse” section, which is currently planned and not yet implemented. This section is intended to house various advanced analytical tools. The planned features include:

- Topic Modeling capabilities.
- Latent Semantic Analysis (LSA).
- Tools for Diachronic Analysis, with the acknowledgment that methods for this are being actively learned and considered, partly based on input from workshop attendees.

In contrast, the “Read” section of the application is already implemented. Its purpose is to provide scholars with the ability to view digital facsimiles of the historical texts within the corpus. This is achieved by offering access to PDF versions of every text. An integrated *Mirador* viewer is used to display these documents, aiming for a user experience comparable to reading texts on a standard library website. Alongside the document image, users can also access relevant metadata for

the text being viewed. The interface for this section shows an image of a historical document page, accompanied by document information fields such as Source Document, Citation, Preferred title of work, Creator, and Title.

7.11 Match Module: Textual Similarity

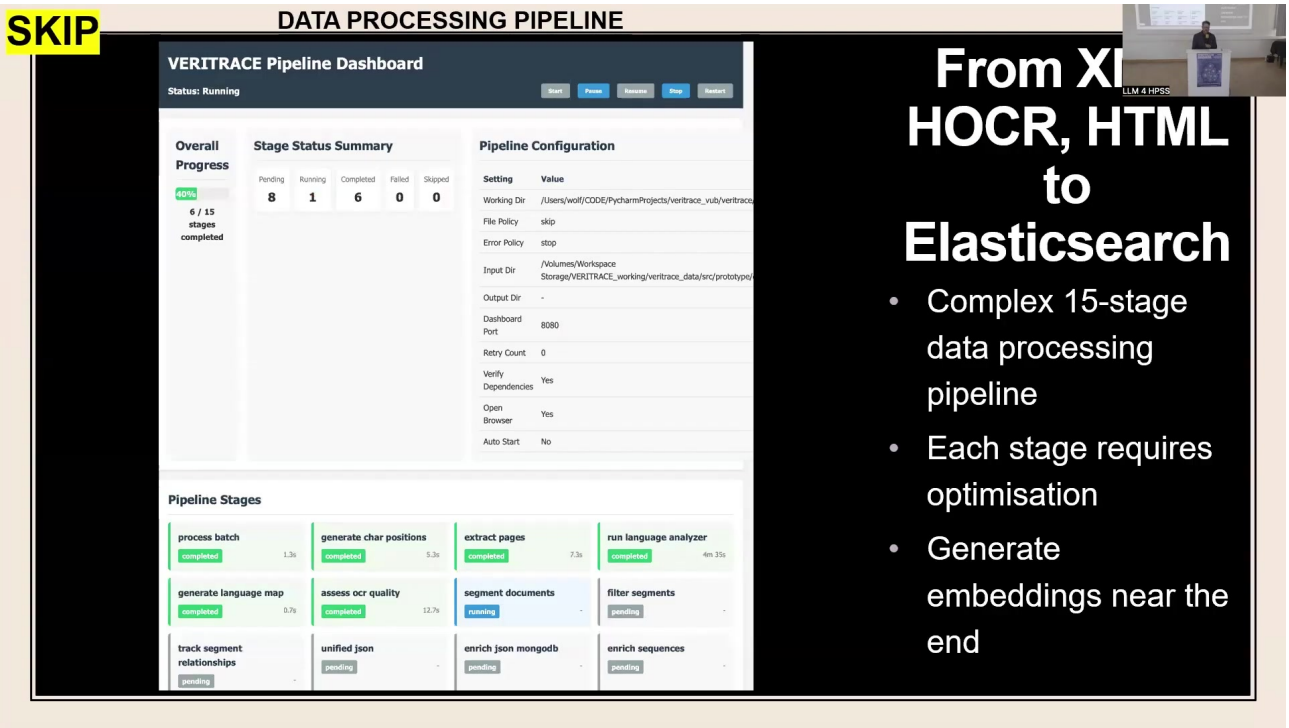

Figure 7.10: Slide 14

The "Match" section of the *VERITRACE* web application is dedicated to identifying textual reuse and similarities between documents, primarily leveraging vector embeddings. Users can input a "Query Text" and a "Comparison Text" through designated text areas. The module supports various matching scopes: users can compare a single document against another single document, perform multi-document comparisons (for instance, comparing Newton's Latin *Opticks* against all of Kepler's works available in the database), or attempt a corpus-wide match, where one text is compared against the entire *VERITRACE* corpus. This last option is acknowledged as computationally intensive, with potential challenges in delivering results within an acceptable timeframe for users.

A "Matching Options Panel" allows users to customize the matching process by adjusting various parameters. The design philosophy is to expose these parameters, such as the minimum similarity score threshold, so that advanced users can fine-tune the search, although default settings will also be available. The module offers three primary match types:

- Lexical Matching: This method relies on keyword matching and is effective when texts share similar vocabulary.
- Semantic Matching: This approach uses vector embeddings to identify conceptually similar passages. It is designed to function across different languages and can detect similarities that are not based on shared vocabulary (e.g., paraphrases).
- Hybrid Matching: This combines both lexical and semantic techniques, potentially allowing users to assign different weights to each approach.

Additionally, different "Matching Modes" are available. A "Standard Mode" uses default settings. A "Comprehensive Mode" employs more computing power and may take longer to execute, but aims to find a more exhaustive set of matches. The existence of a comprehensive mode implies a faster, possibly more selective, mode for quicker analyses.

7.12 Case Study: Newton's Opticks Matching

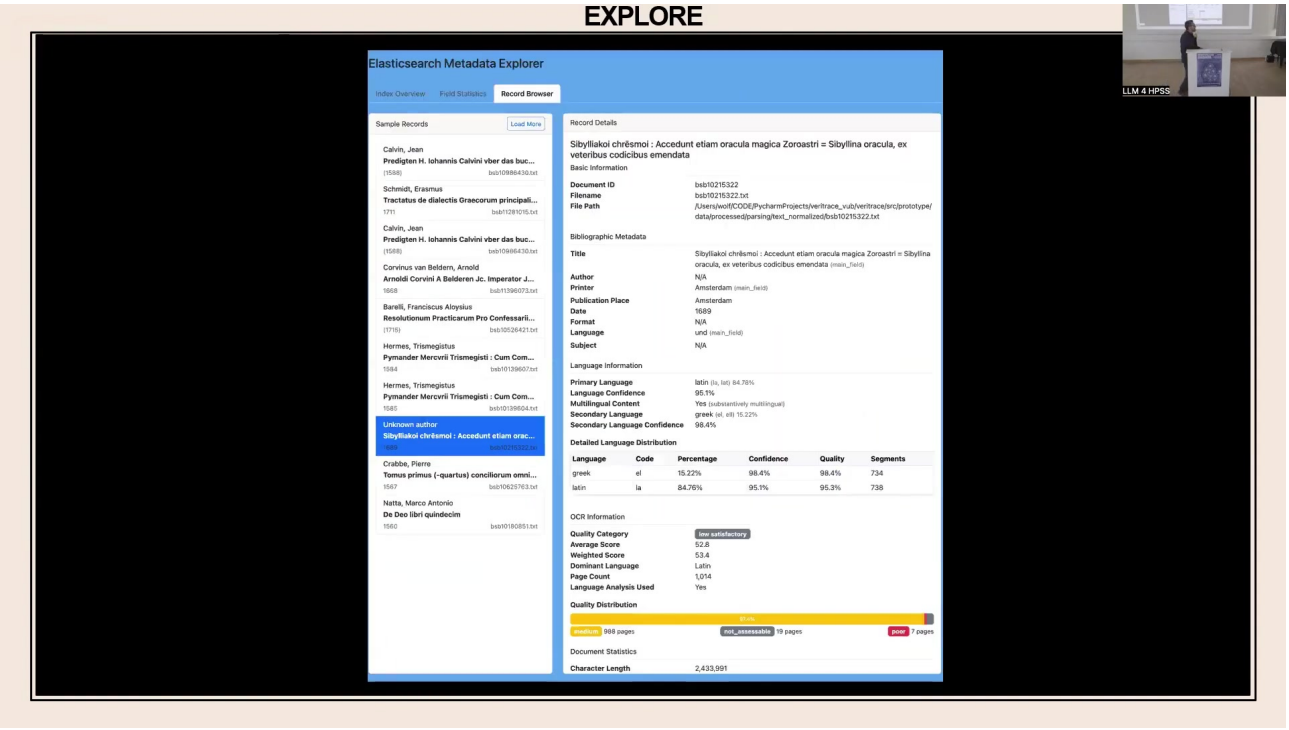

Figure 7.11: Slide 16

To test and illustrate the "Match" module's capabilities, a case study involving two editions of Isaac Newton's *Opticks* was conducted, serving as sanity checks. The query document was the Latin 1719 edition (*Optice: sive de reflexionibus, refractionibus, inflexionibus et coloribus lucis…*), and the comparison document was the English 1718 edition (*Opticks: Or, A Treatise Of The Reflections, Refractions, Inflections and Colours Of Light*).

The first sanity check involved a lexical match (keyword-based) between the Latin and English editions using the standard matching mode. As expected, due to the language difference, no significant matches were found. However, when using the "Comprehensive Mode," three matches were identified; these were in English, suggesting that the Latin edition likely contains some English text, possibly in prefatory material.

To illustrate the output of a lexical match, an example of matching the English *Opticks* against itself was shown, yielding a normalized match score of 100%, a coverage score of 99.7%, and a quality score of 100.0%. The match details view for such a scenario provides side-by-side displays of source and comparison passages with highlighted terms and similarity scores (100% for identical passages). The summary also includes information like the number of comparisons performed (almost 1.3 million in one instance).

The second sanity check performed a semantic match between the Latin and English *Opticks* editions, using the *LaBSE* model for vector embeddings and the standard matching mode. The expectation was to find significant conceptual similarities, as one text is largely a translation of the other. The results indicated that the matches found "seem reasonable,"

even with underlying OCR issues. For example, passages discussing colors in the Latin text were successfully aligned with corresponding passages about colors in the English text, with similarity scores in the example range of 90.35% to 91.77%.

The summary statistics for this semantic match showed a normalized match score of 58%, a coverage score of 36.9%, and a quality score of 91.2%. The quality score was deemed reasonable and high. The low coverage score was noted as requiring further investigation, though it was posited that this might partially reflect genuine textual differences, as the Latin edition is reportedly longer and potentially quite different from the English one. Despite some reasonable matches, other queries run with the *LaBSE* model have led to the assessment that it is generally "inadequate for the task." The hypothesized reason for this inadequacy is an "out-of-domain model collapse," where the model struggles with the historical language, typography, poor OCR quality, and mixed multilingual content of the *VERITRACE* corpus, which significantly deviates from its modern training data.

7.13 Future Challenges

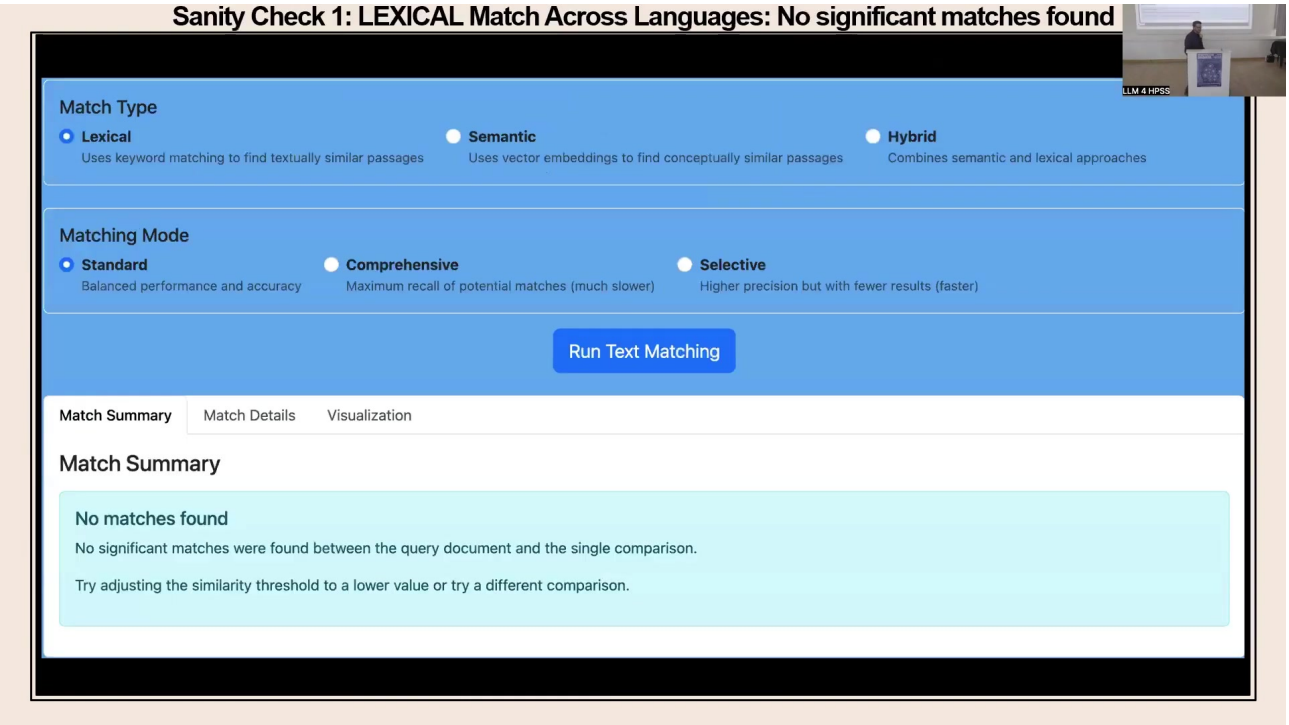

Figure 7.12: Slide 21

Several significant challenges and open questions lie on the horizon for the *VERITRACE* project. A primary concern is the choice of an appropriate embedding model. While *LaBSE* was used as a starting point, it is likely insufficient for the project's needs. Alternative pre-trained multilingual models such as *XLM-Roberta*, *intfloat/multilingual-e5-large*, or various historical *mBERT* models are being considered. However, all these models come with trade-offs concerning accuracy, storage footprint, and inference speed. An alternative strategy under consideration is whether it would be more effective to fine-tune a base embedding model specifically on the *VERITRACE* historical corpus, given its unique linguistic and material characteristics. This is posed as an open question about the viability of using off-the-shelf models versus the necessity of custom fine-tuning.

Another complex issue is handling semantic change over time. The project must grapple with how an embedding model can account for the evolution of word meanings and concepts across several centuries (e.g., comparing a text from 1540 with one from 1700) and across different languages, all while attempting to represent them within a unified vector space.

The persistent problem of poor OCR quality continues to be a major hurdle, as it adversely affects all downstream processing tasks, including fundamental steps like accurately segmenting text into sentences and passages. Re-OCRing the entire corpus of 430,000 texts is not a feasible solution due to resource constraints. Potential mitigation strategies include selectively re-OCRing only the texts identified as having very poor quality, or investing effort in locating and integrating existing, higher-quality digitized versions of these texts from other sources.

Finally, scaling and performance will become increasingly critical issues. Current queries on a small subset of 132 texts take approximately 15 seconds to complete. Extrapolating this to the full corpus of 430,000 texts raises significant challenges for maintaining acceptable query performance and overall system responsiveness. The project actively welcomes advice and input on addressing these multifaceted challenges.

Chapter 8

Explainable AI and Scientific Insights in Humanities

The presentation details the application of Explainable AI (XAI) methods to understand Large Language Models (LLMs) and their application in generating scientific insights within the humanities, specifically focusing on historical texts and images. The work addresses the challenges of interpreting complex "black box" AI systems, particularly the shift from classification models to multi-task generative foundation models. Part A focuses on XAI techniques. It introduces XAI 1.0, prim...

8.1 Overview

The presentation details the application of Explainable AI (*XAI*) methods to understand Large Language Models (*LLMs*) and their application in generating scientific insights within the humanities, specifically focusing on historical texts and images. The work addresses the challenges of interpreting complex "black box" AI systems, particularly the shift from classification models to multi-task generative foundation models.

Part A focuses on *XAI* techniques. It introduces *XAI* 1.0, primarily based on feature attributions like heatmaps for classification models (e.g., image and table classification). It highlights the limitations of these methods for generative AI and proposes *XAI* 2.0, focusing on structured interpretability, feature interactions, and mechanistic views. This includes exploring second-order (pairwise relationships) and higher-order (graph structures, walks) attributions.

Examples include analyzing biases in sentiment prediction based on names using first-order attributions and investigating long-range dependencies in text summarization, finding a bias towards later parts of the input context. Second and higher-order methods are applied to understand similarity predictions in text embeddings, revealing reliance on simple strategies like noun matching, and to analyze complex language structures using Graph Neural Networks (*GNNs*) and walk-based explanations, demonstrating the ability to capture hierarchical relationships like negation.

Part B applies AI methods to humanities research. An initial application involved extracting visual definitions from a corpus of mathematical instruments using class-specific heatmap explanations to identify relevant visual features (e.g., fine-grained scales). A major project focuses on corpus-level analysis of early modern astronomical tables from the *Sacrobosco Corpus* (1472-1650), comprising 76,000 pages of university textbooks.

This project addresses challenges posed by heterogeneous data, limited annotations, and the failure of standard *OCR* and foundation models on this out-of-domain historical data. A workflow named *XAI-Historian* is developed to aid

historians in gaining insights at scale and generating data-driven hypotheses. The method involves data collection, atomization-recomposition (representing tables using bag of bigrams and histograms), and corpus-level analysis through embedding and clustering.

A small, custom-trained model is used to detect bigrams, verified using *XAI* to ensure it correctly identifies matching features. The resulting table representations are used for distance-based clustering. Cluster entropy analysis is applied to investigate innovation spread across European publication locations, revealing differences in print program diversity. Specific case studies in Frankfurt/Main (center for reprinting) and Wittenberg (political control limiting diversity) are identified and validated against historical knowledge.

Key challenges identified include the difficulty of automated analysis of heterogeneous historical corpora with few labels, the limitations of current foundation models for complex research questions despite their utility for intermediate tasks (labeling, curation, error correction), the roadblock of low-resource data for scaling *ML* methods, and the need for thorough evaluation of out-of-domain transfer, especially for historical and small-scale data, as *LLMs* are primarily trained on modern natural language and code. The work emphasizes the necessity of close cooperation between *ML* experts and domain experts (historians) for validation and meaningful interpretation of AI results in humanities research.

8.2 Presentation Structure

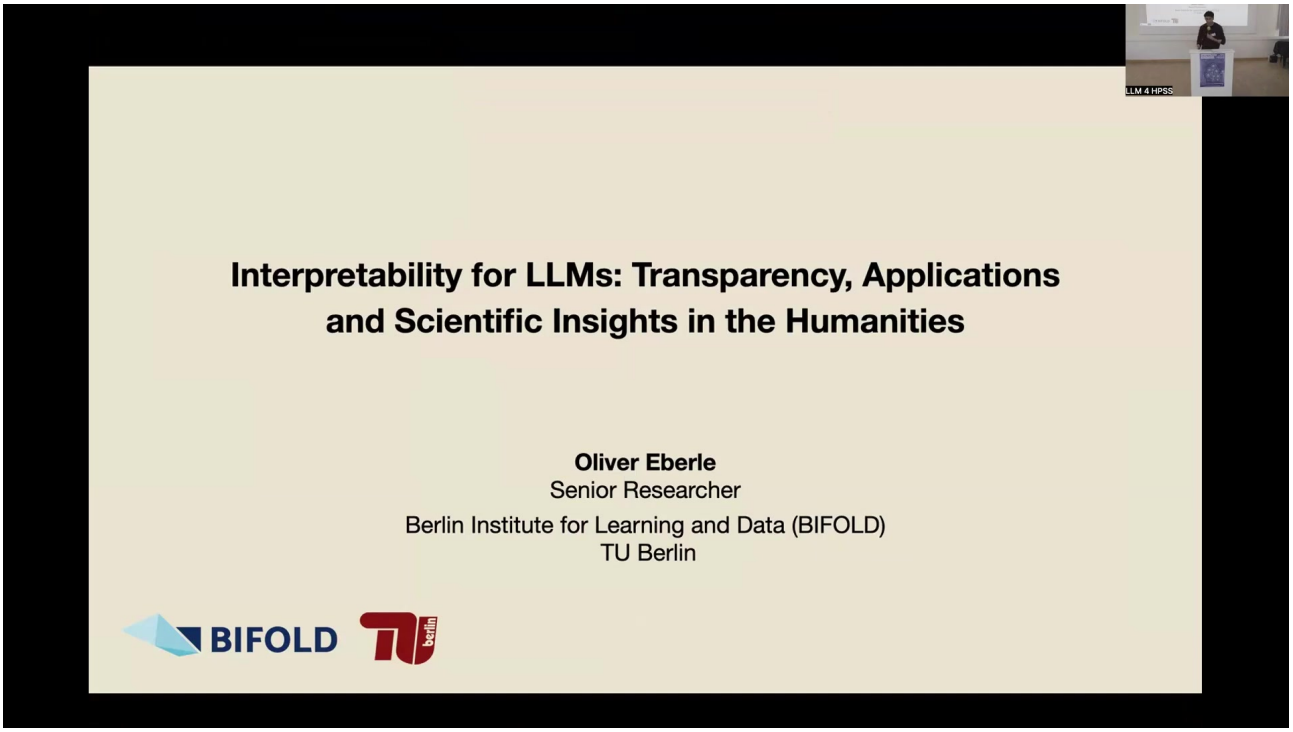

Figure 8.1: Slide 01

The presentation is structured into two primary sections. The first part, designated A, focuses on Explainable AI (*XAI*) and methods for understanding Large Language Models (*LLMs*). This involves developing techniques and approaches to gain insight into the internal workings of these highly complex models.

The second part, designated B, explores the application of AI to generate scientific insights, specifically highlighting applications within the humanities.

8.3 XAI 1.0: Feature Attributions

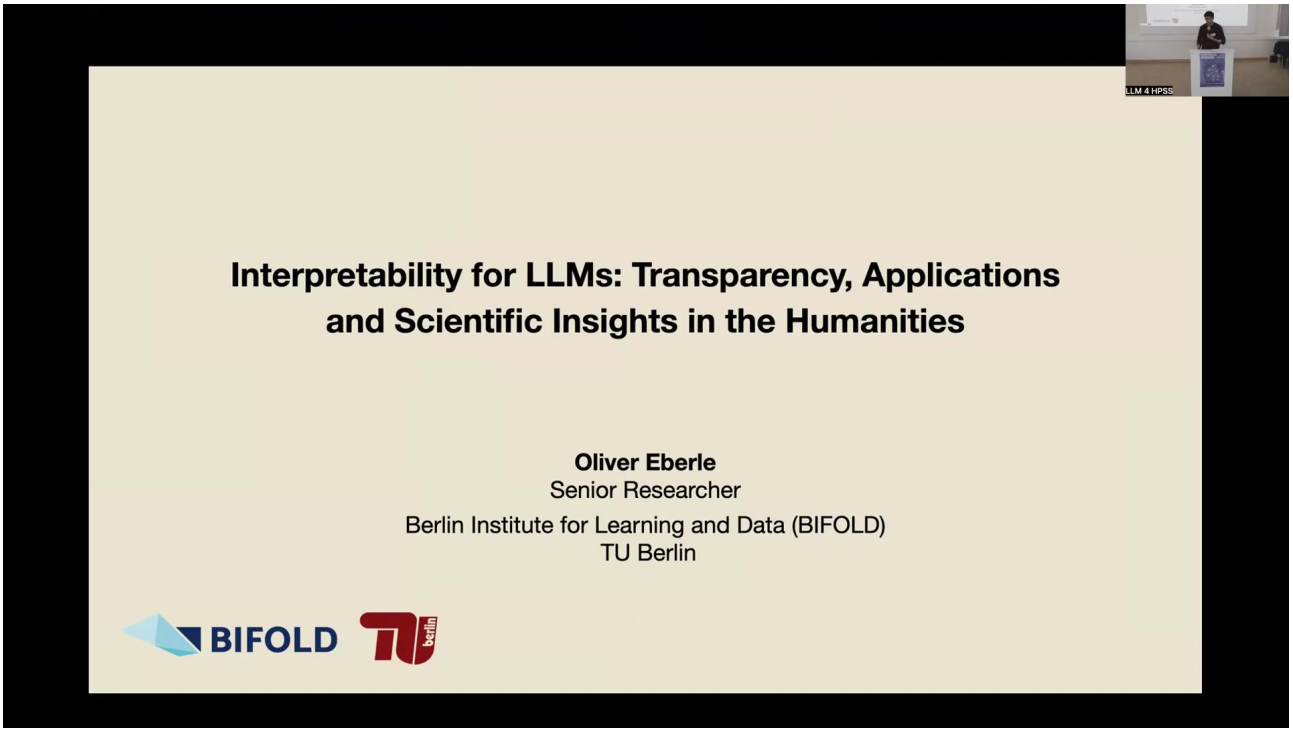

Figure 8.2: Slide 01

Explainable AI (*XAI*) 1.0 represents the initial phase of research in this field, primarily centered on feature attributions. This area of machine learning historically focused on visual data, such as images, with significant advancements in language processing emerging more recently. The core problem addressed was understanding the decision-making process within "black box" machine learning models, particularly classification systems.

In a standard scenario, an input image is fed into a black box AI system, which produces a prediction, such as identifying a "Rooster". However, the user typically has no understanding of which input features led to this specific prediction.

Post-Hoc Explainability was developed as a solution approach, applying explanation methods after the model has generated its prediction. A common output of these methods is a heatmap representation. The heatmap indicates which specific input features, such as pixels in an image, were most responsible for the model's prediction. For instance, a heatmap might highlight the head and neck of a rooster image, demonstrating that these pixels were key to the model's classification decision.

The broader purposes of explainability include:

- Verifying that model predictions are reasonable.
- Identifying flaws and biases to understand how models make mistakes.
- Learning about the underlying problem domain by observing surprising solutions discovered by models.
- Ensuring compliance with regulations such as the European AI Act.

This approach characterized the standard *XAI* scenario until approximately five years ago, as documented by *Samek et al. (2017)*.

8.4 Shift to Generative AI and Foundation Models

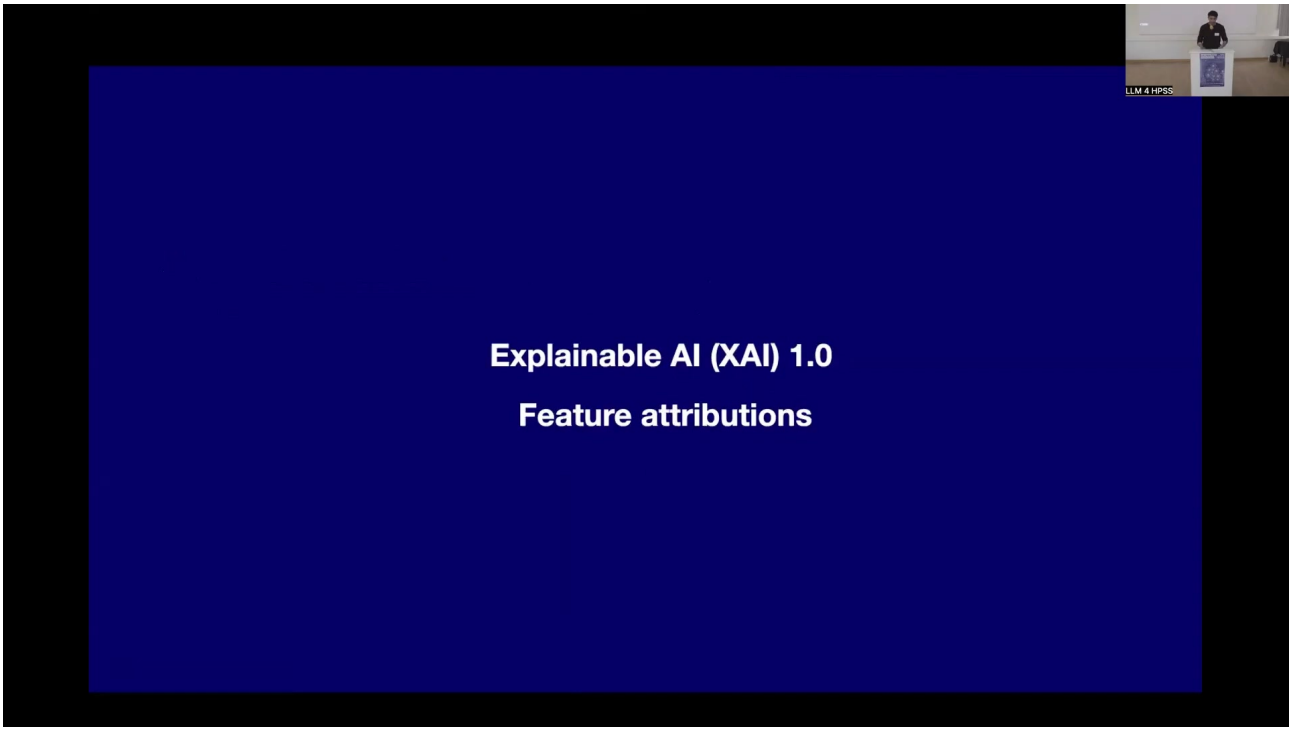

Figure 8.3: Slide 03

The current landscape is dominated by Generative AI (*Gen AI*), marking a significant shift from models primarily focused on classification. Today's models possess multi-task capabilities, extending beyond simple classification to include functions such as finding similar images, generating new images, and answering diverse questions across numerous topics. This expanded functionality presents a challenge: it becomes significantly more difficult to trace and ground a specific prediction or generated answer back to particular input features, unlike the more straightforward case of classification.

To address this, there is a need to develop explanation methods that go beyond simple heatmap representations. Proposed directions include considering feature interactions and adopting more mechanistic perspectives to understand model behavior.

Contemporary foundation models are characterized by their multi-task nature and their capacity to function as "world models," encoding broad knowledge. This characteristic makes them relevant for fields like the humanities, as they can potentially offer insights into societal aspects, the evolution of text over time, and specific features within textual data. The diagram presented, adapted from *Samek et al. (2017)*, illustrates this shift by showing multiple potential outputs from a black box AI system.

8.5 Model Mistakes and Limitations

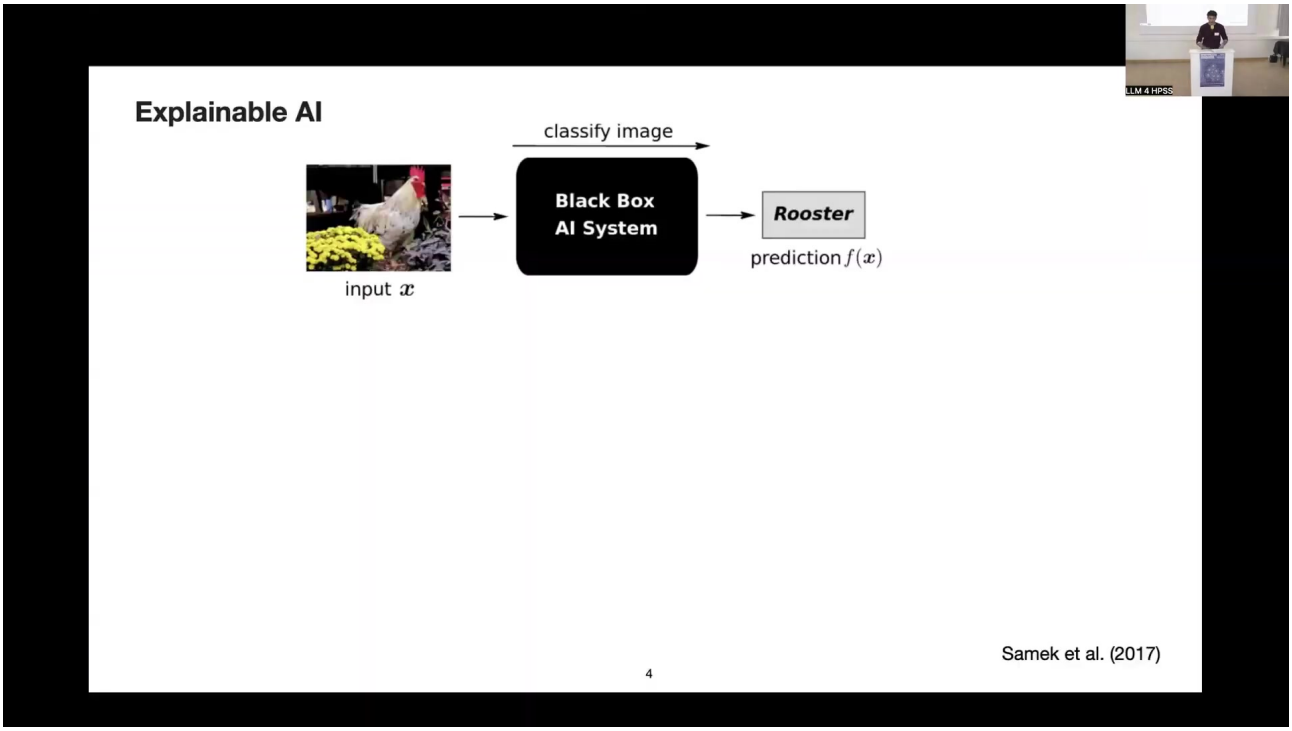

Figure 8.4: Slide 04

AI models, including contemporary *LLMs*, are capable of making surprising mistakes. Two well-known examples illustrate this. The first involves a standard object classifier tasked with identifying a sailboat. The model incorrectly bases its prediction on the surrounding water rather than the boat itself. This error occurs because water is a feature correlated with boats and its texture is easier for the model to detect. This example is documented by *Lapuschkin et al. (Nature Communications, 2019)*.

The second example demonstrates a multi-step planning mistake observed in standard *LLMs*, such as a *Llama 3.something* model. When asked to predict the next step in the Tower of Hanoi puzzle, the model attempts an invalid move: directly moving the largest disk, which is inaccessible due to smaller disks on top, to the final right peg. This indicates that the model failed to understand the physical constraints governing the puzzle. This type of error in reasoning is highlighted by *Mondal & Webb (arXiv, 2024)*.

While more recent reasoning models might perform better, these examples underscore the importance of understanding model limitations.

8.6 XAI 2.0: Structured Interpretability

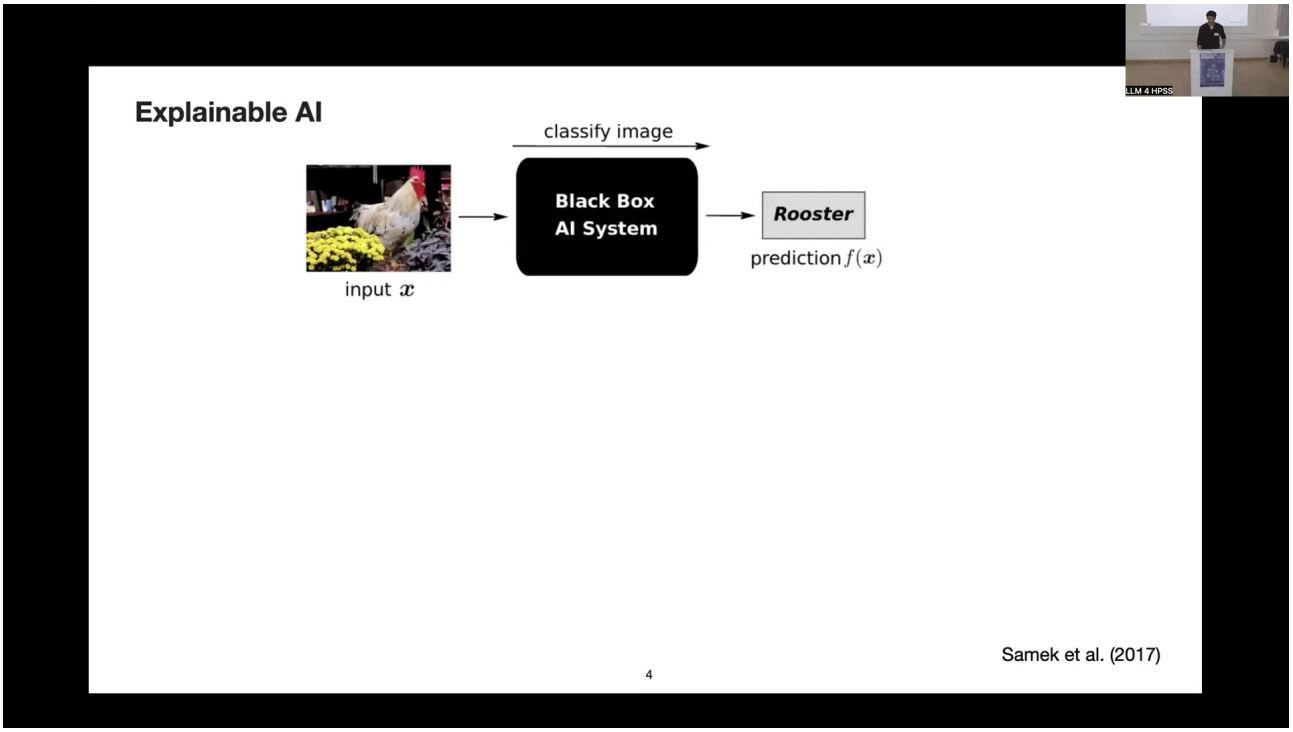

Figure 8.5: Slide 05

XAI 2.0 introduces the concept of Structured Interpretability, aiming to advance beyond the limitations of heatmap-based explanations. This approach focuses on identifying relevant features and understanding their interactions to provide deeper insights into model behavior.

First-order explanations concentrate on the importance of individual features, such as highlighting a single feature (x1) within a set (x1-x4). These are particularly useful for explaining classifier predictions. An example involves a classifier trained on historical table data. Using heatmaps, it was verified that the model correctly focused on the numerical content of the tables, which serves as a good proxy for detecting numerical tables.

Second-order explanations delve into pairwise relationships between features, examining interactions between pairs like x1 and x2 or x1 and x3. This is crucial for explaining similarity predictions, such as those derived from the dot product of embeddings. The method involves computing interaction scores between tokens. In an application explaining the similarity between two historical tables, interaction scores highlighted matching digits, like '38' in both tables, confirming that the model was functioning as intended by identifying identical numerical content.

Higher-order explanations explore more complex structures, including graph structures, feature subgraphs, or feature walks, which represent sets of features that are relevant together. This is depicted as connections between multiple features, potentially forming complex patterns like triangles. These methods are employed to gain more intricate insights into models and move towards a circuit-level understanding of their operations. An example shows a complex network diagram with highlighted elements representing these relevant feature sets.

8.7 First-Order Attributions in LLMs

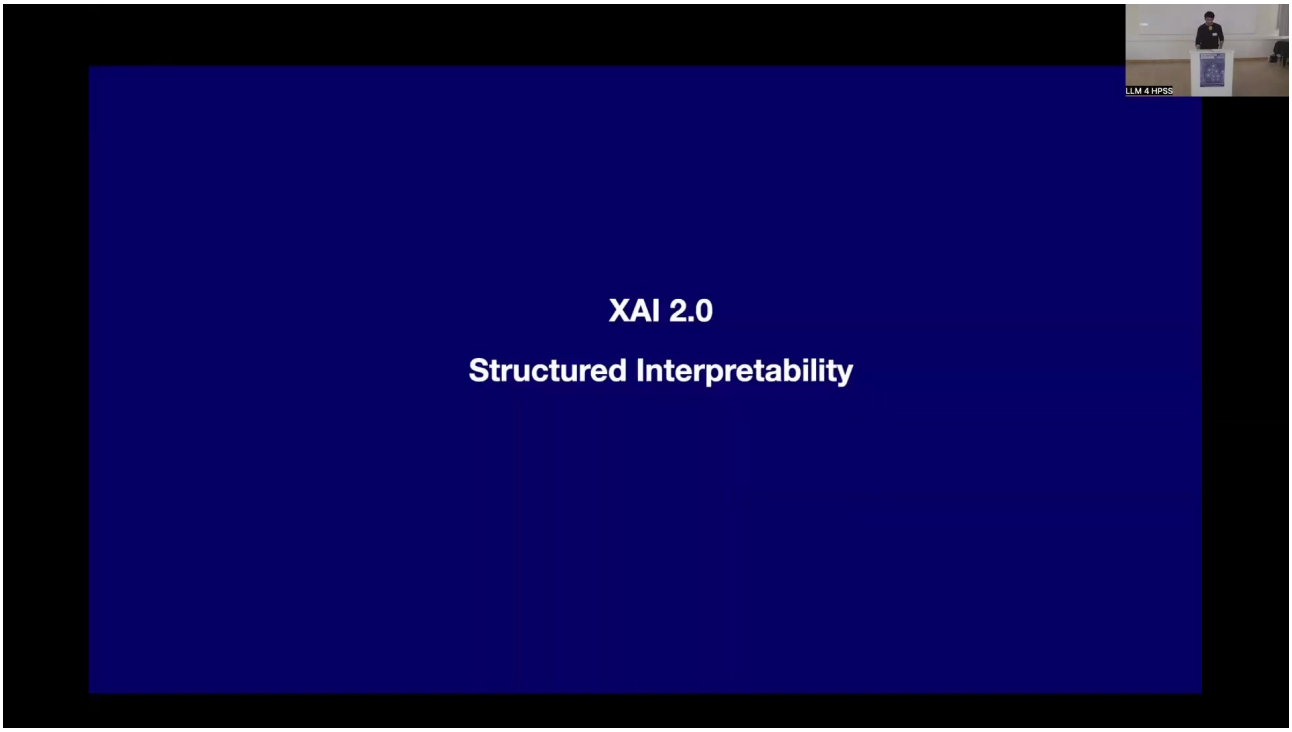

Figure 8.6: Slide 08

First-order attributions have been applied to language data, including examples relevant to the humanities. Example 1a investigates biased sentiment predictions in *Transformer LLMs*. The setup involves using Explainable AI to understand feature importance in these models, specifically analyzing how names influence the prediction of a positive or negative review. The task uses a standard sentiment prediction scenario on movie reviews.

A method proposed for *transformers* is used to compute heatmaps and rank sentences based on name relevance. The results indicate that positive sentiment predictions are more likely when associated with male Western names such as Lee, Barry, Raphael, or the Coen Brothers. Conversely, negative sentiment scores are more likely with names perceived as foreign-sounding, like Saddam, Castro, or Chan. This demonstrates the utility of *XAI* in detecting fine-grained biases within models, a phenomenon now widely recognized in the community. This work is referenced as *Ali et al., XAI for Transformers (ICML, 2022).*

Example 1b explores first-order attributions for long-range dependencies in *LLMs*. The setup involves generating text summaries for long inputs, specifically up to an 8k context window using Wikipedia articles, and analyzing the extent of token dependencies. The task is to provide a long text input and ask the model to generate a summary. The method involves analyzing the origin of the information used in the generated summary within the input context to determine if the model utilizes long-range information.

The results show that the model predominantly focuses on the later parts of the context, prioritizing information presented closer to the prompt. While the model is capable of incorporating long-range information from the beginning of the context, it is significantly less likely to do so, as illustrated by a log scale graph showing counts versus position difference. The implication is that *LLM*-generated summaries may not provide a balanced representation of the entire input text, tending to emphasize recently presented data. This research is referenced as *Jafari et al., SambaLRP (NeurIPS, 2024).*

8.8 Second & Higher-Order Interactions in Text

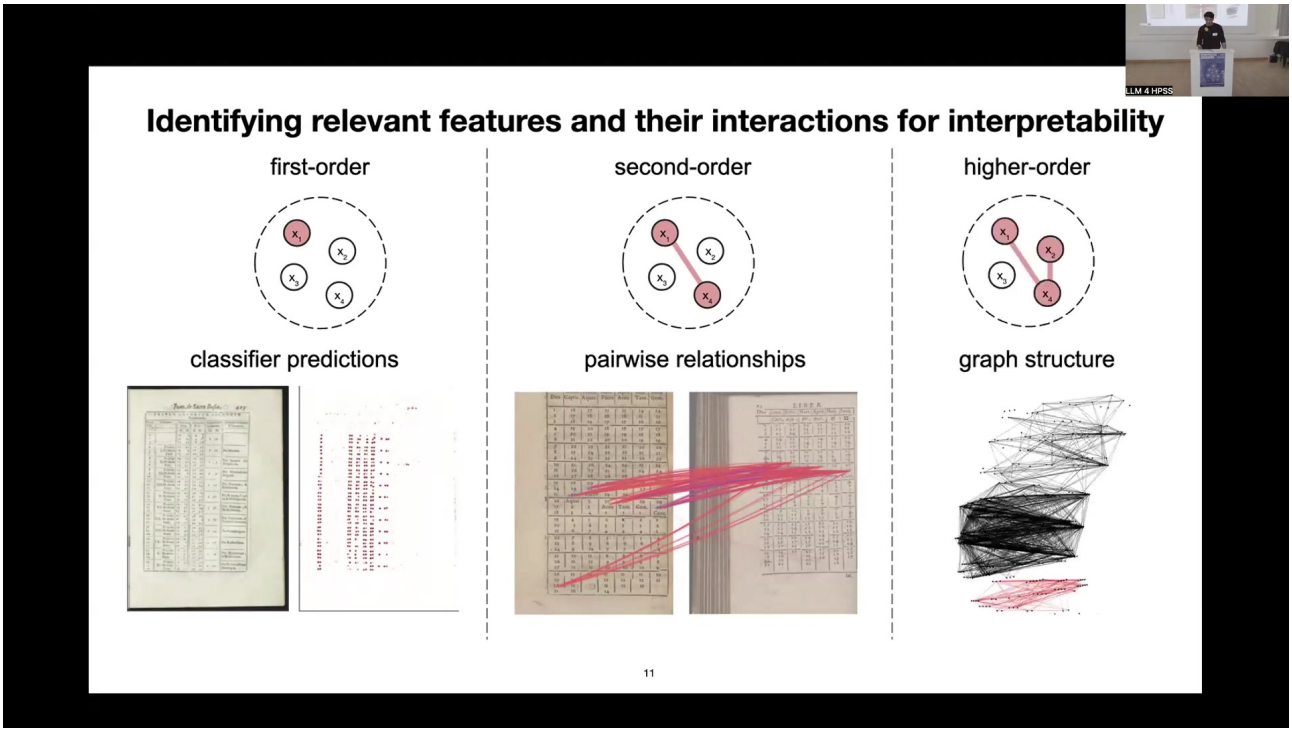

Figure 8.7: Slide 11

The research extends to investigating second and higher-order interactions within text data, particularly in the context of text embeddings and similarity. A standard scenario involves taking two sentences, such as "A cat I really like" and "it is a great cat," obtaining their embeddings from a model like *BERT* or a *sentence BERT* model, and computing a similarity score, typically using a dot product. The challenge lies in understanding the reasons behind a specific similarity score value.

Second-order explanations provide a solution by yielding interaction scores between tokens. These scores help to understand why the model considers the sentences to have high similarity. In a toy example, it was found that noun matching strategies, involving synonyms or identical nouns, are frequently matched and significantly contribute to high similarity predictions.

Analysis at the corpus level, using review data, revealed that models employ quite simplistic strategies to produce high similarity scores. Common patterns include matches between noun tokens (even identical ones), some noun-verb matches, and interactions involving separator tokens. The conclusion drawn is that models, when forced to compress large amounts of information, tend to rely on relatively simplistic strategies, which might not be immediately obvious or intuitive. This implies that when using *LLMs* for embedding data and subsequently computing rankings based on similarity, the underlying features driving high scores could be very simple.

8.9 Graph Neural Networks and Walk-Based Explanations

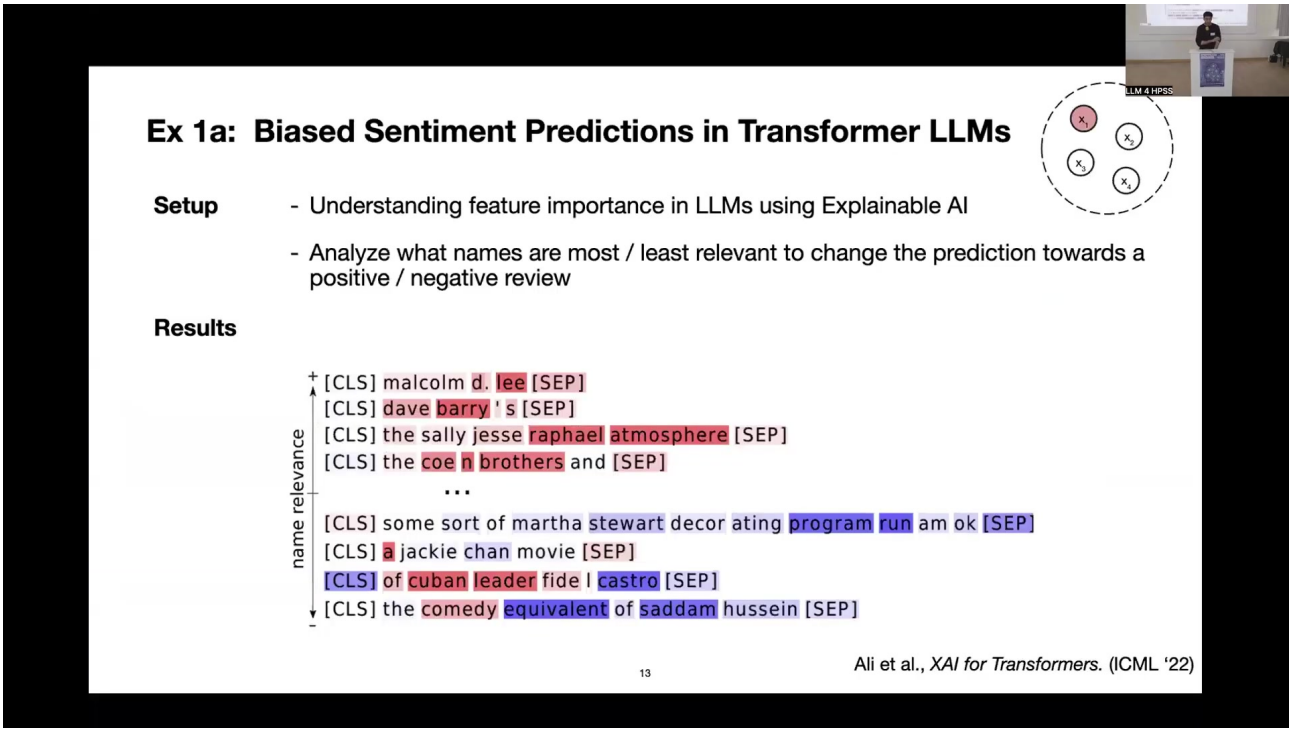

Figure 8.8: Slide 13

Graph Neural Networks (*GNNs*) are utilized for structured predictions, leveraging their ability to encode structural information. A connection is drawn between *GNNs* and *LLMs*: the attention mechanism in *LLMs* can be conceptualized similarly to *GNNs*, indicating which tokens are permitted to exchange information through message passing.

A method called walk-based relevance is employed to explain predictions made by *GNNs* and, by extension, *LLMs* when framed in this manner. This method provides attributions in terms of "walks," which represent interactions between features along paths within the graph structure. The process involves feeding an input graph into the model, which processes information through multiple layers (H0 through HL) involving interactions between nodes. The final prediction is then explained by identifying specific walks within the graph structure that are particularly relevant to that prediction. This approach is detailed in *Schnake et al., Higher-Order Explanations of Graph Neural Networks via Relevant Walks (TPAMI, 2022).*

8.10 Higher-Order Interactions for Complex Language Structure

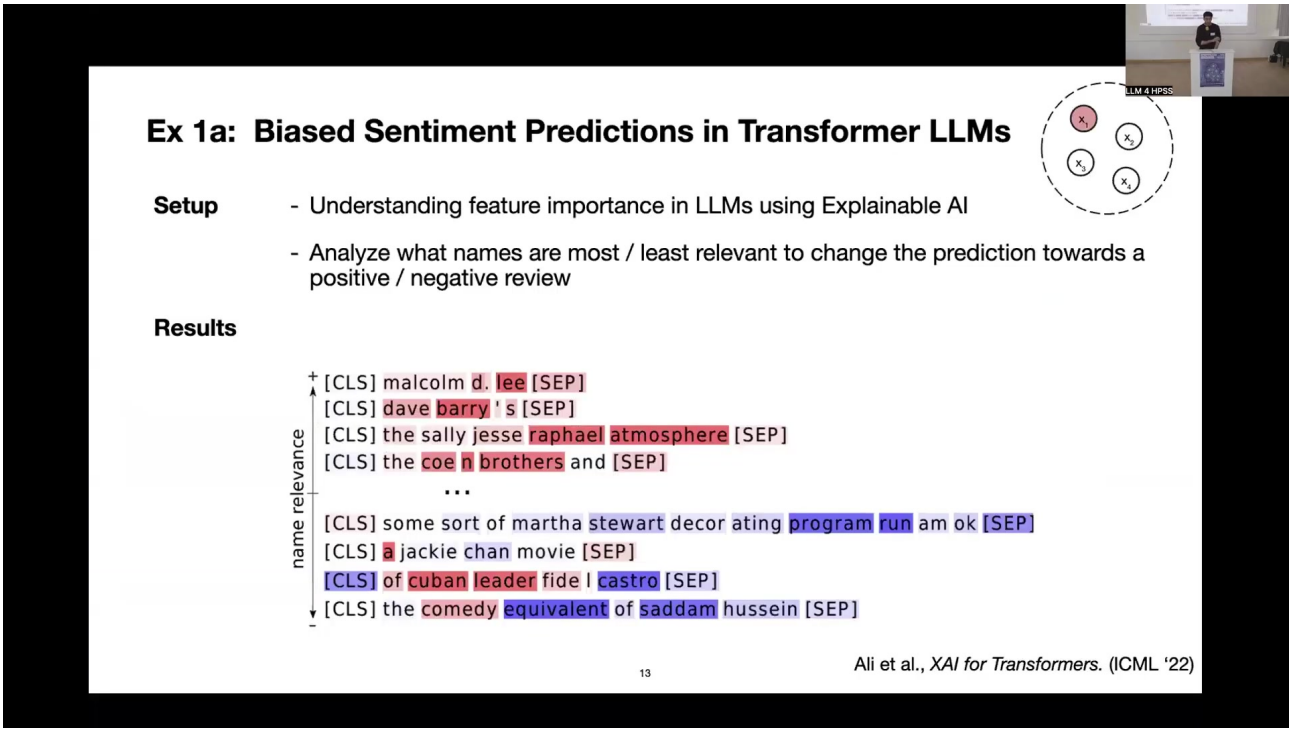

Figure 8.9: Slide 13

Walk-based explanations are applied to analyze complex language structure, leveraging the fact that the hierarchical nature of natural language is well-suited to representation as graph structures. The setup involves training a *GNN* (or an *LLM* framed as a *GNN*) on a movie review sentiment task and then extracting relevant walks to explain predictions.

A comparison is made between high-order interactions and standard explanation methods, such as *Bag of Words (BoW)*. Using the example sentence "First I didn't like the boring pictures, but it is certainly one of the best movies I have ever seen," a standard explanation method like *BoW* fails to capture the complexity. It might assign a high score based on the presence of words like "like" or "in it," completely missing the crucial negation "didn't like."

In contrast, high-order interactions, represented through a tree-like graph structure where nodes are words and edges represent relationships, successfully capture this complexity. The method correctly identifies the first part of the sentence, "First I didn't like the boring pictures," as having a negative sentiment score despite containing potentially positive words, because it understands the negation. It also correctly assigns a positive score to the second part, reflecting the overall sentiment and the hierarchical structure of the sentence. This demonstrates the ability of higher-order methods to understand more intricate linguistic phenomena. This work is also referenced in *Schnake et al., Higher-Order Explanations of Graph Neural Networks via Relevant Walks (TPAMI, 2022).*

8.11 AI Insights in Humanities: Visual Definitions

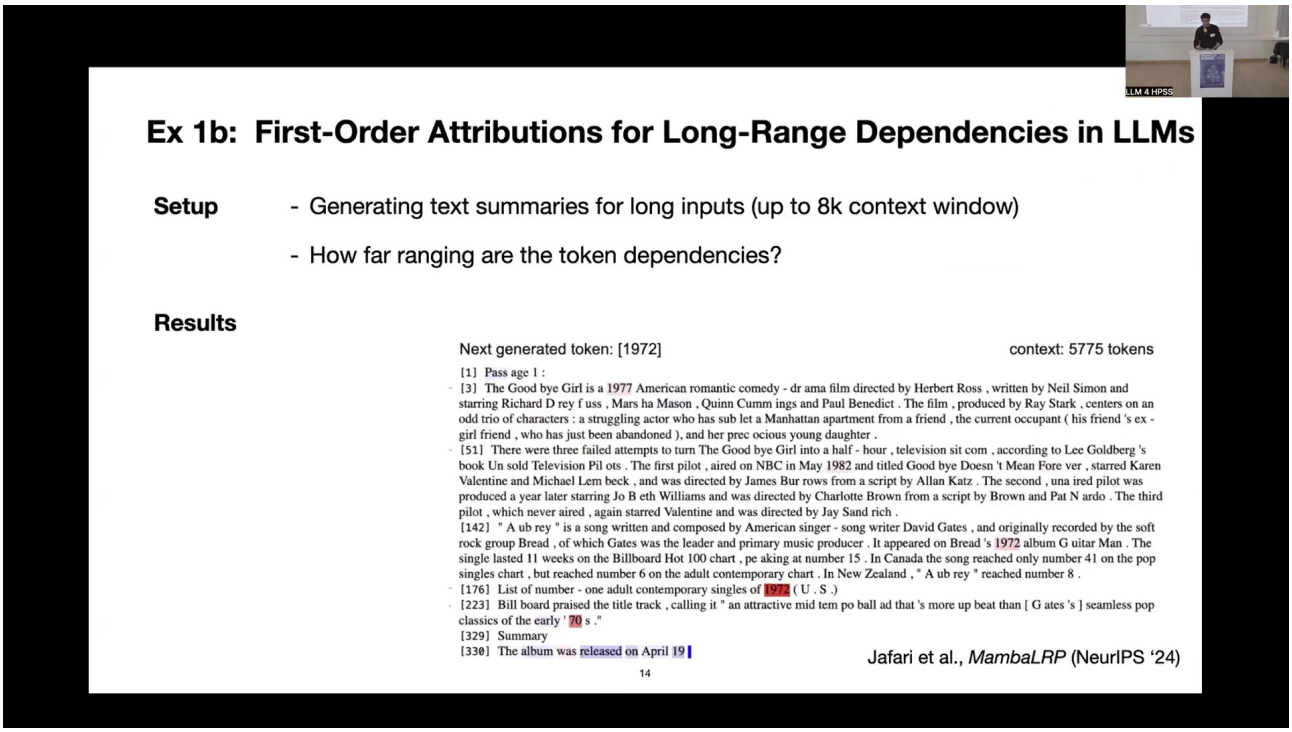

Figure 8.10: Slide 14

Part B of the presentation shifts focus to AI-based scientific insights within the humanities. Example 4 demonstrates extracting visual definitions from corpora. The data corpus used consists of images of mathematical instruments from the *Sphaera Corpus*, as compiled by Valleriani and colleagues in 2019.

An initial approach utilized heatmap-based methods. The task involved building a classifier capable of categorizing these images into specific classes, such as distinguishing between a "machine" and a "mathematical instrument." Class-specific heatmap explanations were employed as the method. The purpose was to assist historians in establishing potentially more objective criteria for defining visual categories within the corpus.

Validation was crucial and involved close cooperation with domain experts, specifically historians like Matteo Valleriani and Jochen Büttner, to verify the meaningfulness of the definitions derived from the AI analysis. The findings indicated that fine-grained scales present on the mathematical instruments were highly relevant features for the models when making classification decisions. This work is documented in *El-Haij & Eberle+, Explainability and transparency in the realm of DH (International Journal of Digital Humanities, 2023).*

8.12 Corpus-Level Analysis of Early Modern Astronomical Tables

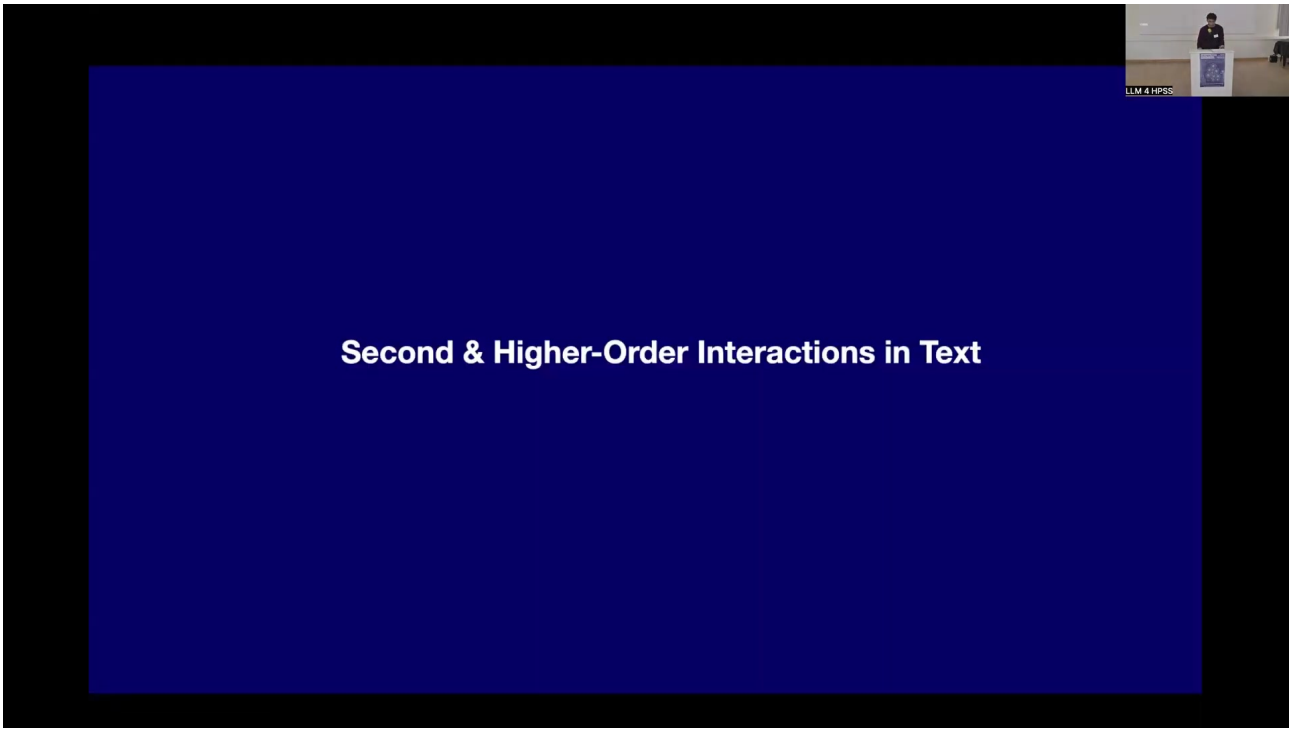

Figure 8.11: Slide 15

Example 5 presents a major project involving corpus-level analysis of early modern astronomical tables. The data corpus is the *Sphaera Corpus*, covering the period from 1472 to 1650. This corpus consists of early modern texts, specifically university textbooks, and comprises approximately 76,000 pages.

The problem addressed was the historians' interest in automatically matching and identifying tables with similar semantics. Manual analysis of this corpus at scale was not feasible. The project faced significant challenges due to the nature of the data: the corpus is highly heterogeneous, and very limited annotations are available. Furthermore, standard *Optical Character Recognition (OCR)* and contemporary *Foundation Models* proved ineffective on this historical, out-of-domain data. The corpus is referenced as *Sphaera Corpus (1472-1650) (Valleriani+ '19)* and *Sacrobosco Table Corpus (1472-1650) (Eberle+ '24)*.

8.13 XAI-Historian Workflow for Historical Insights at Scale

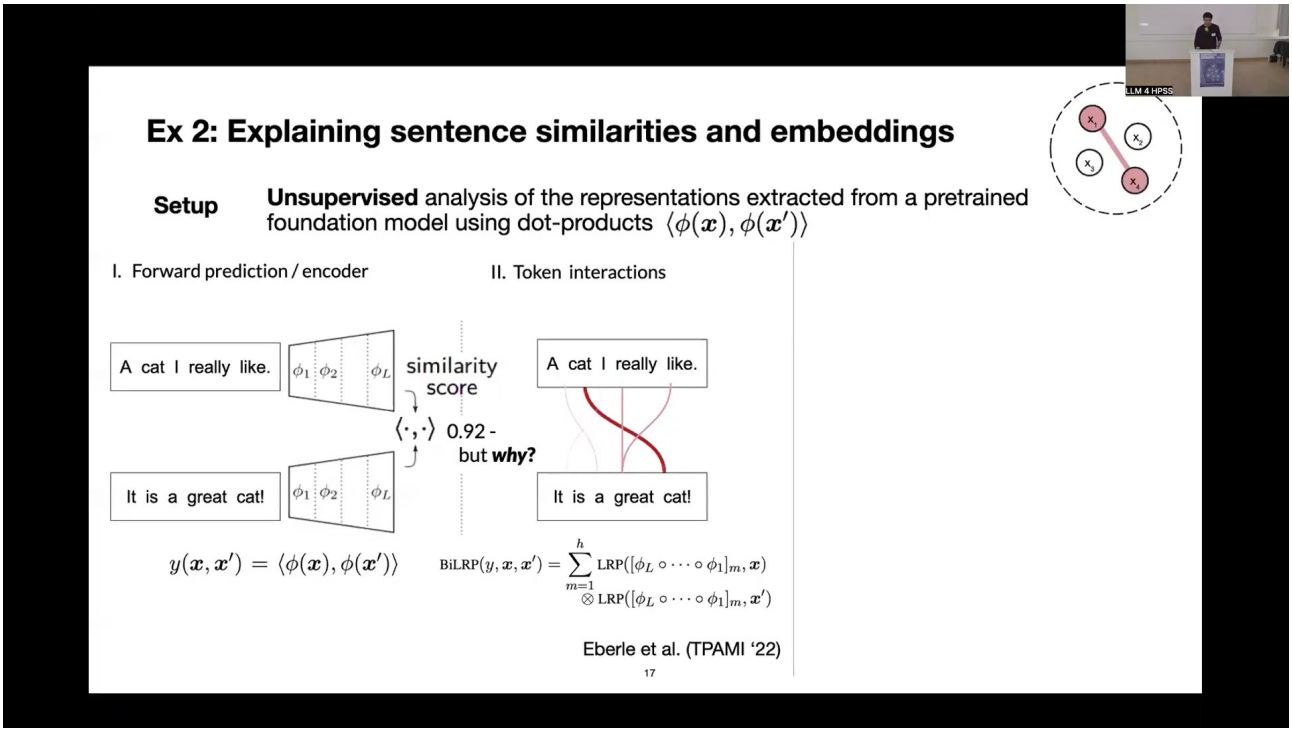

Figure 8.12: Slide 16

To address the challenges of analyzing the historical table corpus at scale, a workflow was developed in collaboration with historians. This workflow is conceptualized as the *XAI-Historian* approach, enabling historians to utilize AI and explainable AI for data-driven hypothesis generation and the discovery of case studies.

The workflow comprises three main steps:

- Data Collections: Starting with the *Sacrobosco* corpus of historical books.
- Atomization-Recomposition: Processes the input tables. Instead of attempting to process the entire table directly with standard foundation models, which are ineffective on this out-of-domain data, the tables are represented using a "bag of bigrams" approach. This involves identifying sequences of two characters, such as '01' or '21'. A custom, small model is trained specifically for the task of detecting these bigrams. Explainable AI methods, such as heatmaps or interaction maps, are then used to verify that this custom model functions correctly, for instance, by checking if it consistently detects matching bigrams like '38' on different input tables. This verification process is crucial for building trust in the model's decisions. The outputs of this step include bigram maps and histograms.
- Corpus-Level Analysis: Takes the historical table embeddings, which are derived from the bigram representations, and applies distance-based clustering. The output is a representation of data similarity, often visualized as a scatter plot showing distinct clusters of tables.

This comprehensive workflow is detailed in *Eberle et al., Historical insights at scale (Science Advances, 2024).*

8.14 Cluster Entropy Analysis for Investigating Innovation Spread

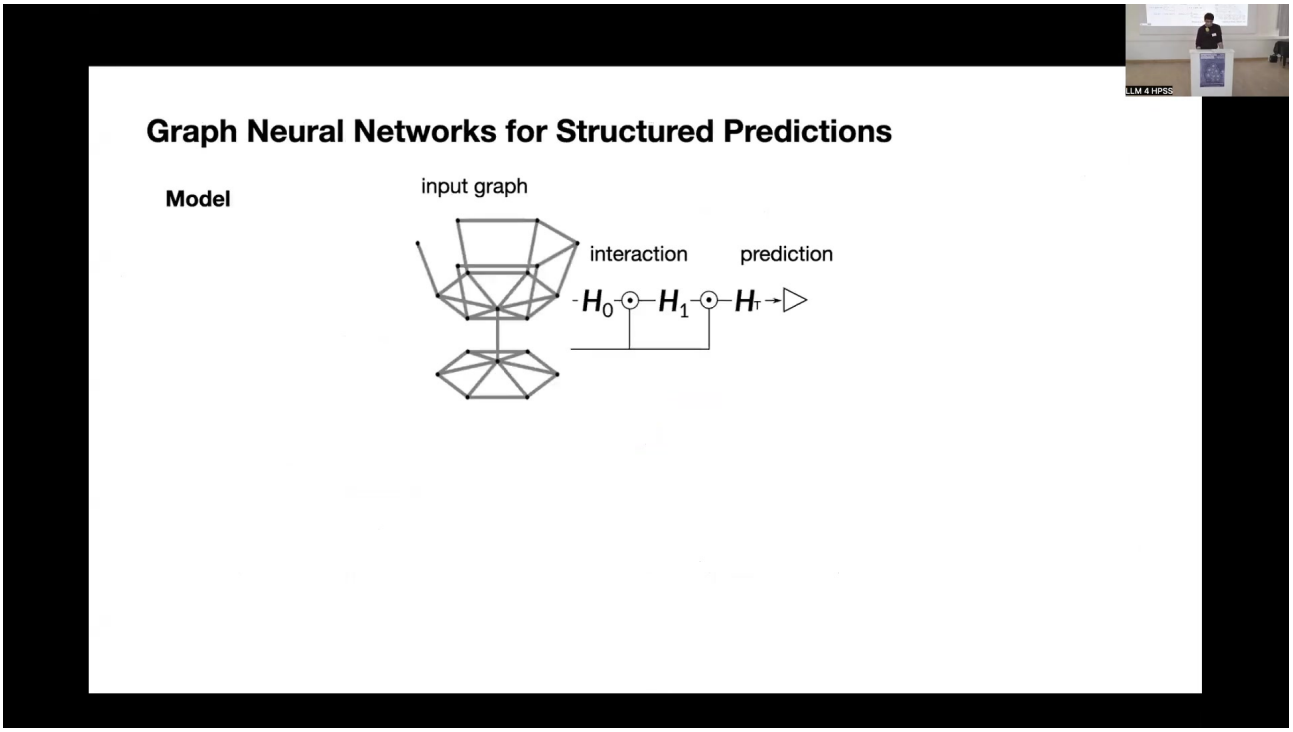

Figure 8.13: Slide 17

Building upon the *XAI-Historian* workflow, cluster entropy analysis was applied to investigate the spread of innovation across Europe during the period of *Sphaera* publication (1472-1650). The method utilizes the output of the clustering approach, which groups historical tables based on their representations derived from the custom bigram model.

The process involves using the table representations obtained from the model, performing distance-based clustering on these representations, and then, for each publication city, determining the diversity of table types produced by counting how many different clusters are represented in that city's output. Entropy is calculated for each city's print program as a measure of this diversity. Low entropy indicates that a city primarily reproduces the same content, signifying a less diverse print program. Conversely, higher entropy suggests a more diverse range of publications. The specific metric used is the difference between the observed cluster entropy H(p) and the maximum attainable entropy H(p_max) at that print location, where lower values indicate lower diversity relative to what is theoretically possible.

This analysis identified two interesting cases with the lowest entropy scores. Frankfurt am Main was found to have low diversity, which aligns with its historical reputation as a center known for reprinting editions repeatedly. A more historically significant finding concerned Wittenberg, which also exhibited an unusually low diversity score. This finding supports the historical understanding that political control exerted by the Protestant reformers, particularly Melanchthon, actively limited the print program by dictating the curriculum. The analysis revealed this historically anomalous low diversity, matching existing historical intuition and supported knowledge. This application of the method is detailed in *Eberle et al. (Science Advances, 2024).*

8.15 Conclusion and Challenges

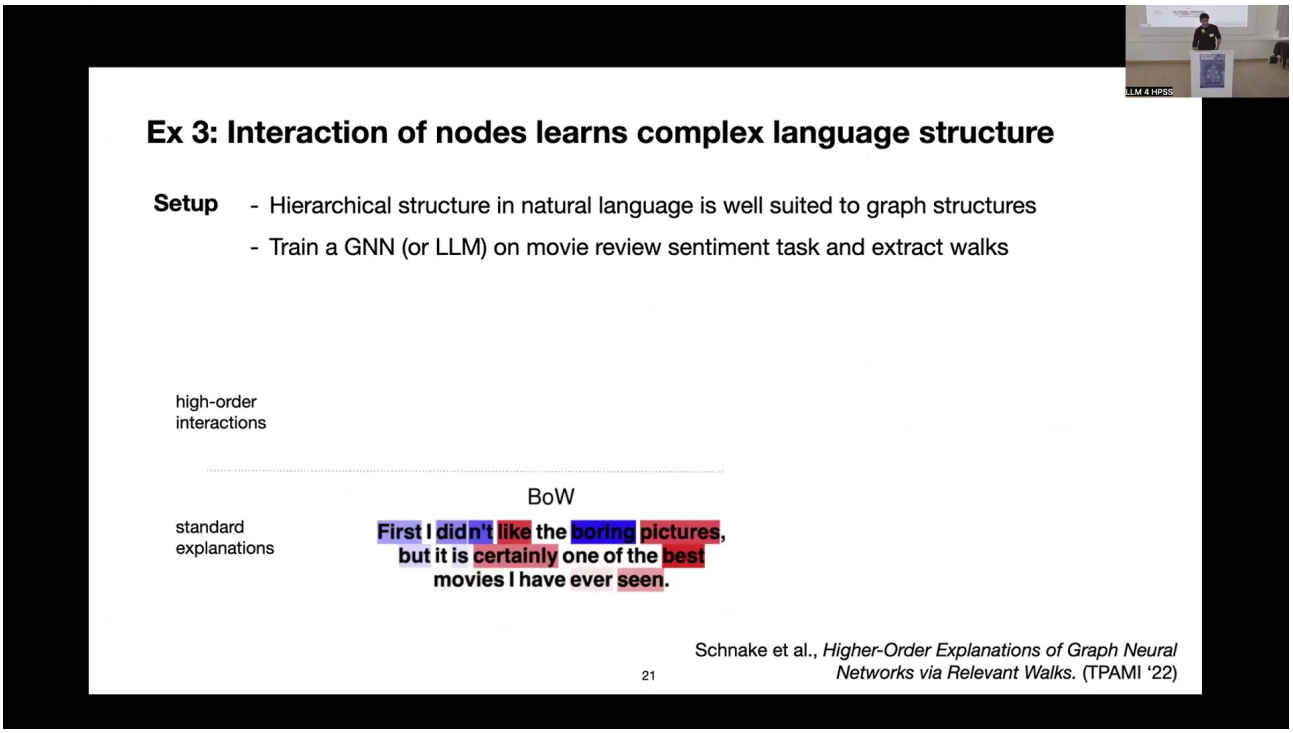

Figure 8.14: Slide 19

In conclusion, researchers in the Humanities and Digital Humanities have primarily focused on the digitization of source material. However, automated analysis of these digitized corpora presents significant challenges due to their inherent heterogeneity and the scarcity of available labels. Multimodality is identified as a relevant aspect for future research in this domain.

The integration of *Machine Learning (ML)* with *Explainable AI (XAI)* holds potential to scale humanities research efforts and facilitate the development of novel research directions. While *Foundation Models* and *Large Language Models (LLMs)*, along with prompting techniques, can provide automated results for intermediate tasks such as labeling, data curation, and error correction, they remain limited when addressing more complex research questions.

Significant challenges persist, including the roadblock posed by low-resource data for applying *ML* methods effectively, particularly in the context of scaling laws. Furthermore, out-of-domain transfer requires thorough evaluation, especially when dealing with historical and small-scale datasets. This challenge arises because current *LLMs* are primarily trained and aligned for tasks involving modern natural language and code generation, making their direct application to historical or highly specialized data problematic without careful adaptation and validation.

Chapter 9

Modeling Science: LLM for the History, Philosophy and Sociology of Science

The presentation addresses the limitations of current Large Language Models (LLMs) for scholarly inquiry, particularly in the history, philosophy, and sociology of science (HPSS). It identifies key missing capabilities in LLMs, including the ability to counter hallucination, understand meaning beyond embedding vectors, formulate justified true statements, avoid repeating unreliable media content, seek best justification, and plan scientific inquiry. The core proposed solution is the conce…

9.1 Overview

The presentation addresses the limitations of current Large Language Models (LLMs) for scholarly inquiry, particularly in the history, philosophy, and sociology of science (HPSS). It identifies key missing capabilities in LLMs, including the ability to counter hallucination, understand meaning beyond embedding vectors, formulate justified true statements, avoid repeating unreliable media content, seek best justification, and plan scientific inquiry.

The core proposed solution is the concept of "Validation is all you need," which involves providing reasons, arguments, and evidence for propositions and actions. This capability is framed within a new proposed discipline called Computational Epistemology, requiring epistemic agency to identify propositions, analyze argumentation, and understand historical actors' intentions, plans, and actions based on documented traces.

A comprehensive research infrastructure is presented to achieve these goals, comprising five key components:

- *Scholarium* (Evidence): Curated scholarly sources validated by editorial boards, including extensive historical collections like the *Opera Omnia Euler*, *Kepler Gesammelte Werke*, and *Brahe Opera Omnia*.
- *Scholarium* (Registry): A structured database serving as an alternative to embedding-based approaches, containing detailed, historically validated content items such as personal chronologies, communication acts, statements, arguments, and records of the use of language, tools, methods, data, and sources. Access is provided via an AI API and the *Model Context Protocol* (*MCP*) API.
- User Interface (*AI Cockpit*): A working environment, specifically the *LettreAI* platform on *Cursor*, integrating multimodal LLMs (*Claude*, *Gemini*, *Llama*, including *Gemini 2.5* for multimodal capability) and featuring a specialized AI agent named "*Bernoulli*" for historical queries.

- *FAIR Infrastructure*: Utilizing *Zenodo*, hosted by CERN, for long-term storage and publication of data, ensuring Findability, Accessibility, Interoperability, and Reusability.

- Technical Support: Provided by the startup *OpenScienceTechnology*, focusing on running the infrastructure, including an *MCP API Server*, and adhering to principles of Open Source, Open Access, Open Data, and Open Collaboration to standardize AI access to knowledge.

The system aims to provide validated, complete answers to complex historical queries by leveraging curated data and instructing LLMs with explicit reasoning rules formulated in natural language. This approach contrasts with reliance solely on unstructured text and embedding vectors, which are deemed insufficient for achieving the required level of justification and completeness for scholarly historical research. The infrastructure is designed to maintain scholarly expertise and ensure the long-term preservation and accessibility of valuable historical data.

9.2 LLM Evolution and Current State

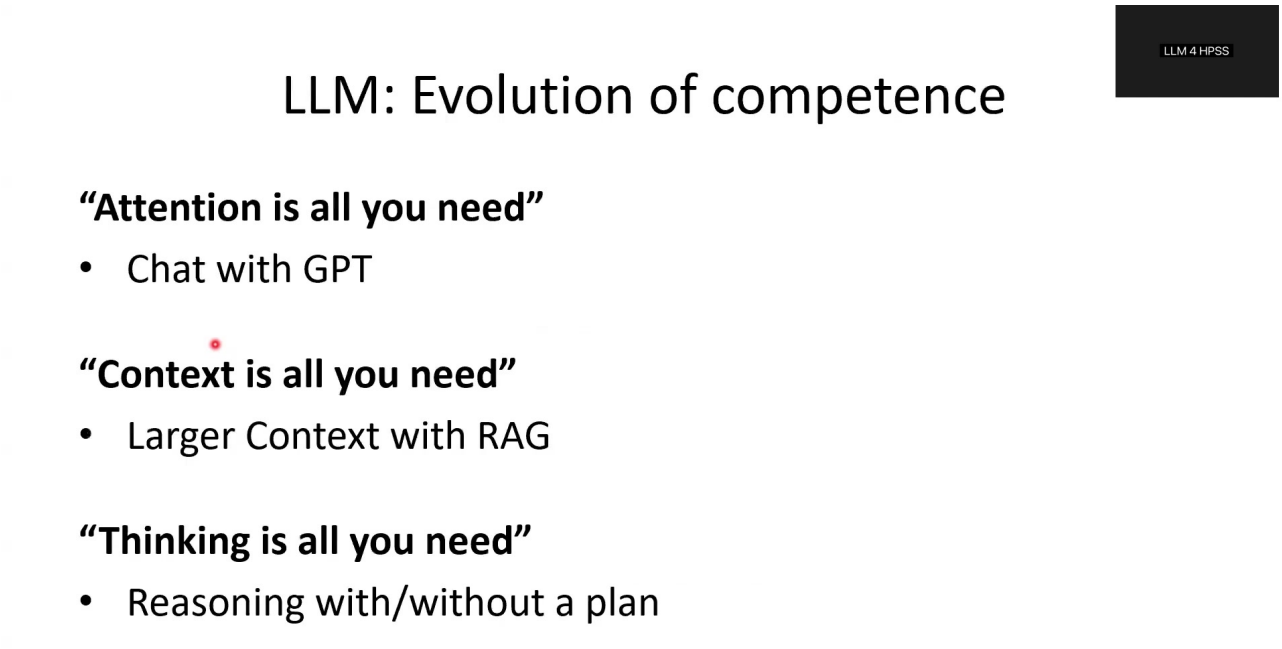

Figure 9.1: Slide 01

Large Language Models (LLMs) have undergone rapid evolution. The initial focus was captured by the phrase “Attention is all you need.” This capability was subsequently supplemented by the requirement for “Context is all you need,” necessitating larger context windows and methods such as Retrieval Augmented Generation (RAG).

The latest models now propose that “Thinking is all you need” is also required, focusing on reasoning capabilities, potentially with or without an explicit plan. This represents the current state of LLMs, integrating attention, context handling, and nascent reasoning abilities.

9.3 Missing Capabilities in Current LLMs

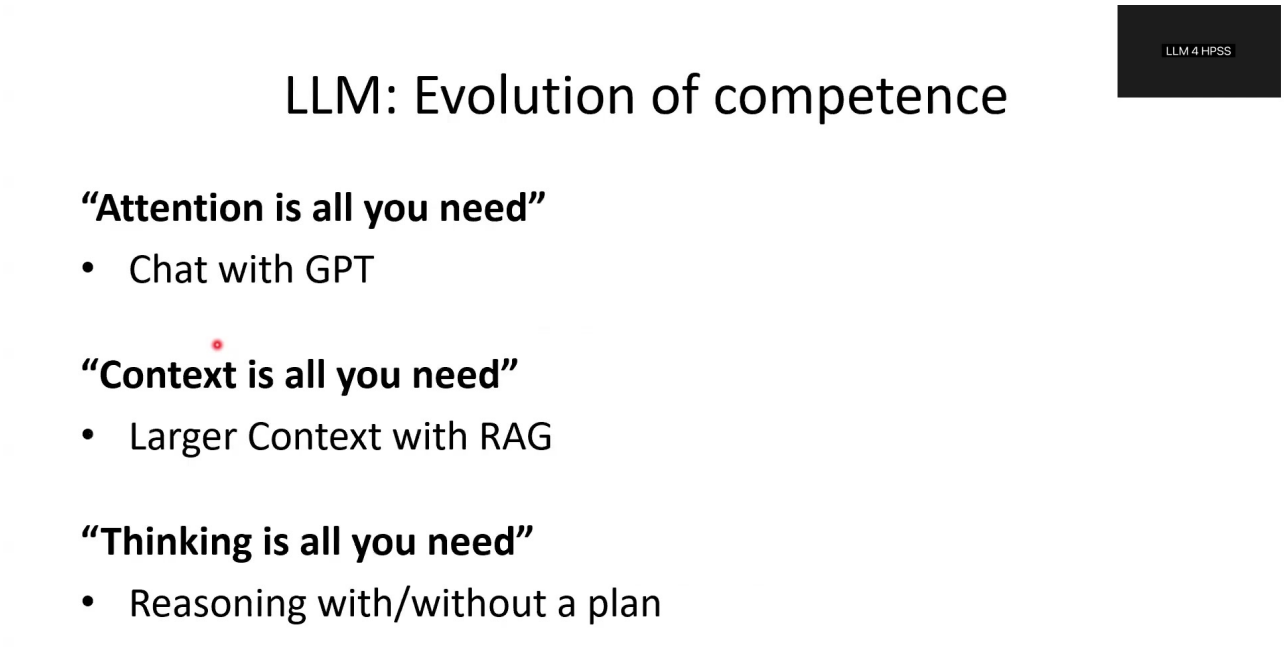

Figure 9.2: Slide 01

Current LLMs exhibit several critical missing capabilities essential for reliable scholarly work. A significant problem is the lack of an inherent mechanism or “opponent” to effectively counter hallucination, which is the generation of false or nonsensical information.

Furthermore, embedding vectors, a common representation method, are fundamentally not equivalent to the meanings of expressions. LLMs tend to formulate responses that may sound plausible or good but are factually false. They also frequently repeat content sourced from internet media without sufficient validation, treating it as knowledge.

Current models lack the ability to actively seek out and prioritize what is best justified based on evidence or reasoning. Finally, they are unable to formulate effective plans for conducting scientific inquiry. These identified capabilities are absent in existing LLM models, and it is asserted that current technologies offer no realistic prospect of achieving these missing goals.

9.4 Validation and Computational Epistemology

What is missing?

- Opponent to counter halluzination
- Embedding vectors are NOT meanings of expressions
- Do not formulate what sounds good but is false
- Do not repeat what others put out as internet media
- Seek what is best justified
- Make best plan for inquiry

Figure 9.3: Slide 02

A critical requirement identified for advancing LLM capabilities in scholarly domains is validation, summarized as "Validation is all you need." Validation is defined as the process of providing explicit reasons, arguments, and evidence that support or oppose the truth of a given proposition. It also involves providing reasons for or against the pursuit of specific actions.

To address this fundamental gap in current AI, a new discipline is proposed: Computational Epistemology. This field focuses on developing the methods and methodologies necessary to implement computational validation processes.

Achieving validation requires epistemic agency, which encompasses several key abilities. These include the capacity to identify propositions, which goes beyond merely processing sentences; the ability to analyze and understand argumentation structures within texts and sources; and the capability to identify the intentions, plans, and actions of historical persons as documented and traced in historical documents.

9.5 Working Environment and Inquiry Example

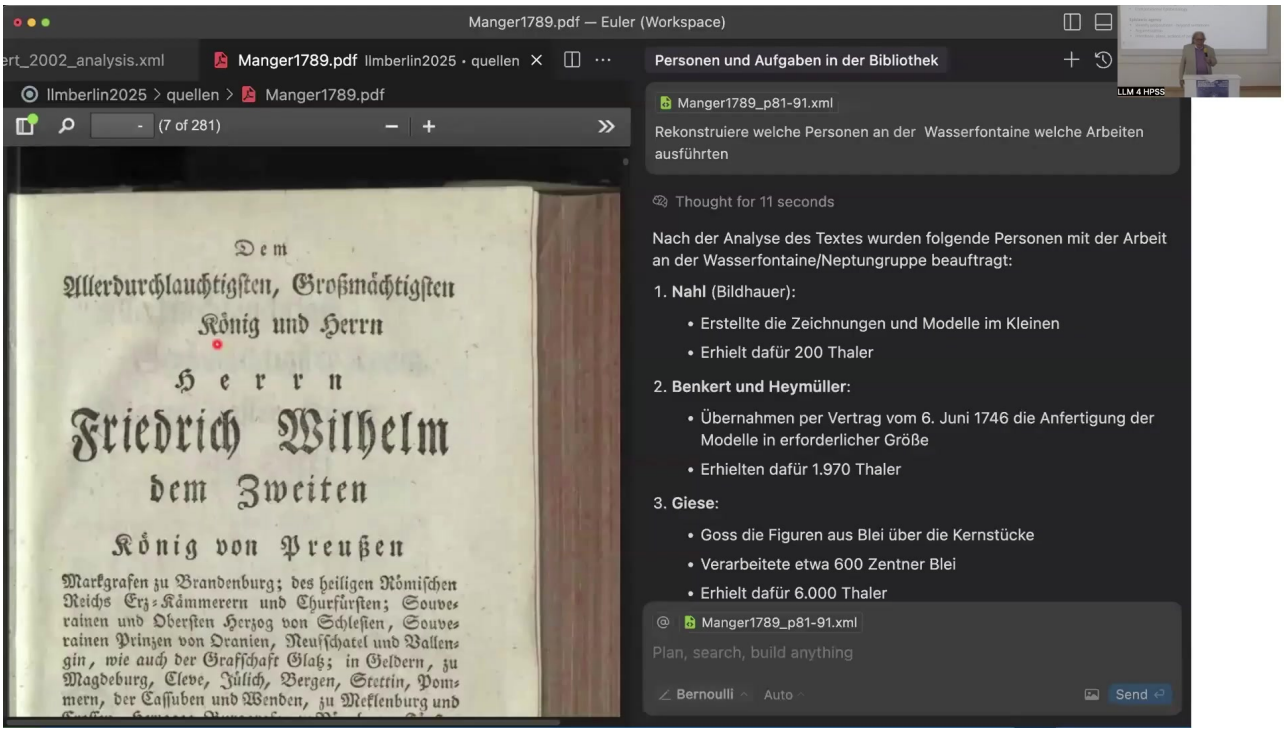

Figure 9.4: Slide 04

A working environment designed for validated historical inquiry is demonstrated through a screenshot of a web browser interface. The interface features a PDF viewer on the left, displaying a historical source document (*Manger1789.pdf*), specifically page 7 of 281, containing old German text related to construction under King Friedrich Wilhelm II.

On the right side, a dark-themed panel titled "Personen und Aufgaben in der Bibliothek" houses the search query input and results display areas. The example historical problem addressed is the long-standing dispute among historians of science regarding the involvement and potential responsibility of the mathematician Leonard Euler in the significant failure of the Sanssouci castle water fountain construction in the 18th century. The goal of the inquiry within this environment is to obtain a validated, reliable, and factually correct answer grounded in proven evidence, moving beyond mere hearsay.

An example query formulated in German is "*Rekonstruiere welche Personen an der Wasserfontaine welche Arbeiten ausführten*," which translates to "Reconstruct which persons performed which work on the water fountain." The system provides validated results in a numbered list format. For this specific query, the results identify the key individuals involved: Nahl, a sculptor, who created the drawings and small models and received 200 Thaler; Benkert and Heymüller, who took on the task of producing the models in the required size under a contract dated June 6, 1746, and were paid 1,970 Thaler; and Giese, who was responsible for casting the figures from lead over the core pieces, processing approximately 600 Zentner of lead, and received 6,000 Thaler.

The output references a specific XML file, *Manger1789_p81-91.xml*. The underlying platform for this working environment is the *Cursor* environment, which allows the use of AI agents. The specific agent employed for these historical inquiries is named "*Bernoulli*." A key challenge highlighted is that obtaining such validated answers requires more than simply reading a single PDF source; it necessitates the ability to search across *all* available sources, a task for which standard indexing and token-based approaches are deemed insufficient.

9.6 Scholarium: Curated Scholarly Evidence

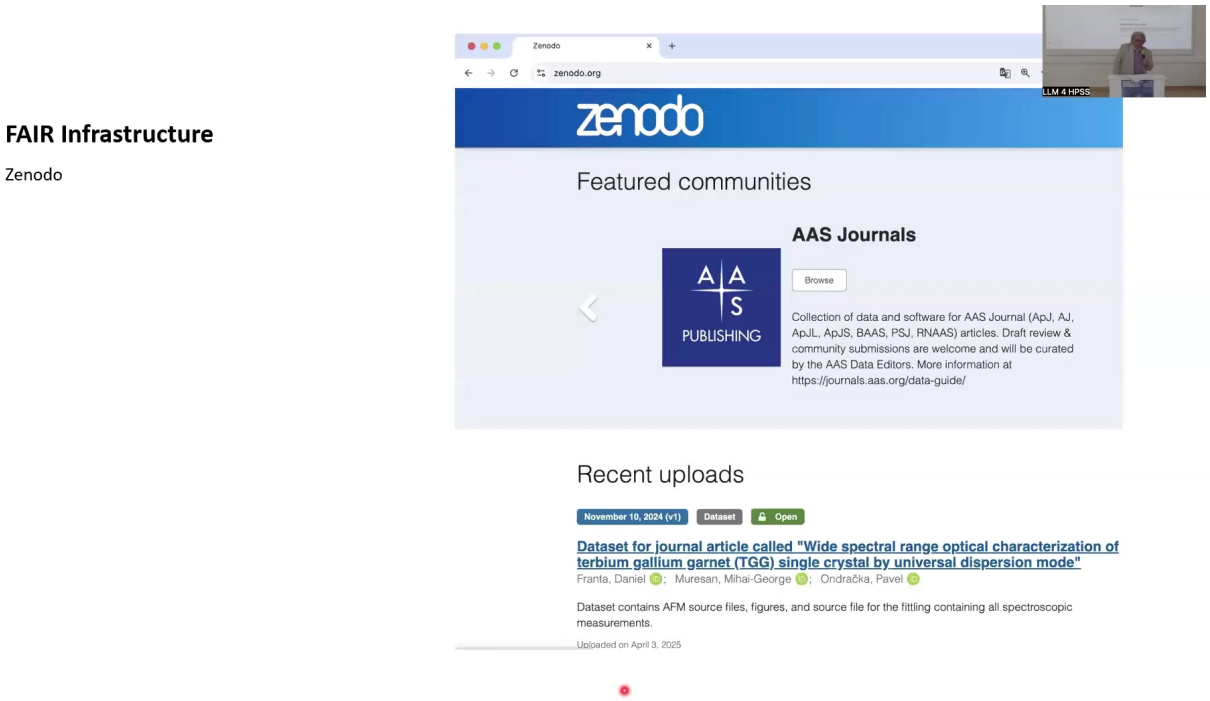

Figure 9.5: Slide 08

The first key component of the infrastructure is the *Scholarium* for Evidence. This component relies on the oversight of a Curated Scholarly Editorial Board, ensuring the scholarly reliability of the sources.

Examples of these foundational curated sources include the *Opera Omnia Euler*, a monumental collection comprising 86 volumes of Euler's work, which underwent scholarly editing for approximately 120 years by various scholars and was completed two years prior. This comprehensive edition includes all of Euler's letters and his 866 publications. Complementary scholarly reliable sources integrated into this component include the *Kepler Gesammelte Werke* and the *Brahe Opera Omnia*.

9.7 Scholarium: Registry instead of Embeddings

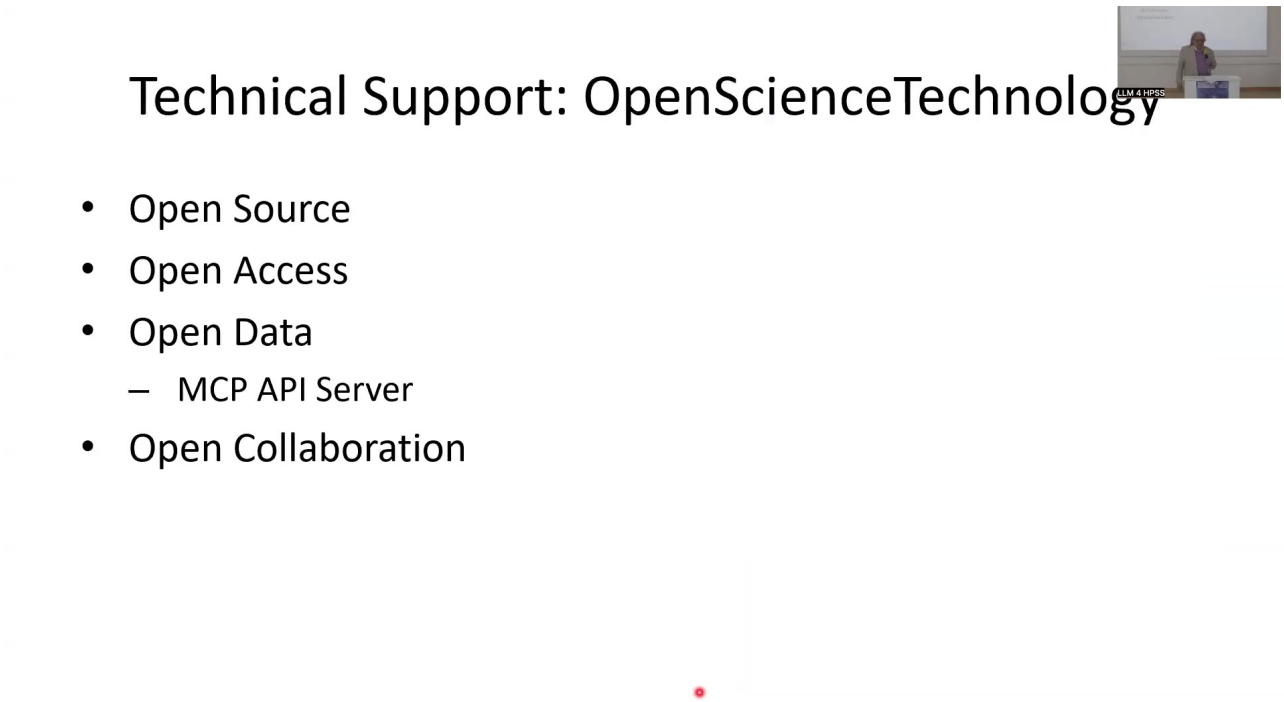

Figure 9.6: Slide 09

The second key component is the *Scholarium* as a Registry, which serves as a novel substitution for conventional embedding-based approaches. This component is a curated database containing a very detailed inventory of content items representing historically proven activities, all rigorously validated by sources.

The types of content recorded include personal actions, various communication acts such as letters, publications, and reports, as well as statements, implications, arguments, and inquiries. The registry also meticulously documents the use of language, terminology, and concepts by historical figures, along with their use of concepts, relations, models, methods, tools, devices, data, information, evidence, and sources. Access to this detailed and validated historical record is provided through both a dedicated AI API and the *Model Context Protocol* (*MCP*) API.

9.8 User Interface: AI Cockpit

The user interface component, referred to as the *AI Cockpit*, operates on the *LettreAI* platform within the *Cursor* environment. This interface integrates multiple accessible multimodal LLM models, including *Claude*, *Gemini*, and *Llama*.

The system leverages the capabilities of multimodal models, specifically mentioning *Gemini 2.5*, which can combine information from both text and images, a feature deemed beneficial for solving the requirements of the historical inquiry tasks.

9.9 FAIR Infrastructure

A crucial component is the *FAIR Infrastructure*, designed for the long-term storage and publication of the project's data while adhering to the FAIR principles (Findability, Accessibility, Interoperability, and Reusability).

The specific tool utilized for this purpose is *Zenodo*, which is hosted by CERN in Geneva. *Zenodo* provides the necessary capabilities to host the project's data reliably for many years.

9.10 Technical Support: OpenScienceTechnology

Technical support for the infrastructure is provided by the startup *OpenScienceTechnology*. This entity is responsible for running the system's infrastructure and specifically provides an *MCP API Server*.

The technical support operates under principles of Open Source, Open Access, Open Data, and Open Collaboration. The *MCP API Server* plays a key role in attempting to standardize AI access APIs to knowledge on a worldwide scale through a standardized interface, facilitating open collaboration.

Chapter 10

The Representation of SDG-Related Research in Bibliometric Databases: A Conceptual Inquiry via LLMs

The project investigates the representation of research related to the UN Sustainable Development Goals (SDGs) within major bibliometric databases, specifically Web of Science, Scopus, and OpenAlex. The core objective is to use Large Language Models (LLMs) as a tool to detect potential biases embedded in the SDG classification standards applied by these databases. The research considers the performative nature of bibliometric databases and their influence on research priorities, resource ...

10.1 Overview

The project investigates the representation of research related to the UN Sustainable Development Goals (SDGs) within major bibliometric databases, specifically Web of Science, Scopus, and OpenAlex. The core objective is to use Large Language Models (LLMs) as a tool to detect potential biases embedded in the SDG classification standards applied by these databases. The research considers the performative nature of bibliometric databases and their influence on research priorities, resource allocation, and policy decisions.

A case study focuses on five SDGs related to socioeconomic inequalities (SDG4, SDG5, SDG10, SDG8, SDG9). The methodology involves collecting a jointly indexed subset of publications from the three databases (15,471,336 publications from January 2015 to July 2023), classified according to each database's standard for the selected SDGs.

A key technical decision involves selecting a "light" pre-trained LLM, *DistilGPT2* (82M parameters), to avoid embedding existing knowledge biases present in larger models trained on extensive web data. This model is fine-tuned separately on subsets of publication abstracts corresponding to each database's classification for each of the five SDGs, resulting in 15 fine-tuned LLMs (*DistilGPT2* {bibDB, SDG}).

The fine-tuned LLMs are then used to analyze the content, specifically by extracting noun phrases from responses to prompts related to SDG targets. The analysis reveals a systematic overlook in the data (classified publications) of disadvantaged categories of individuals, poorest countries, and underrepresented topics explicitly mentioned in SDG targets. Conversely, significant attention is paid to economic superpowers and highly developing countries.

The findings highlight how bibliometric classification, despite appearing objective, decisively shapes the representation of SDG-related research. Limitations include the high sensitivity of results to model architecture, training data, hyperparameters, and decoding strategy, as well as the general framework employed. The study demonstrates the potential of LLMs as detectors of biases in research data infrastructures and serves as a proof-of-concept for their introduction in automating information extraction for research policy decision-making.

10.2 SDG Classification in Bibliometric Databases: Background and Implications

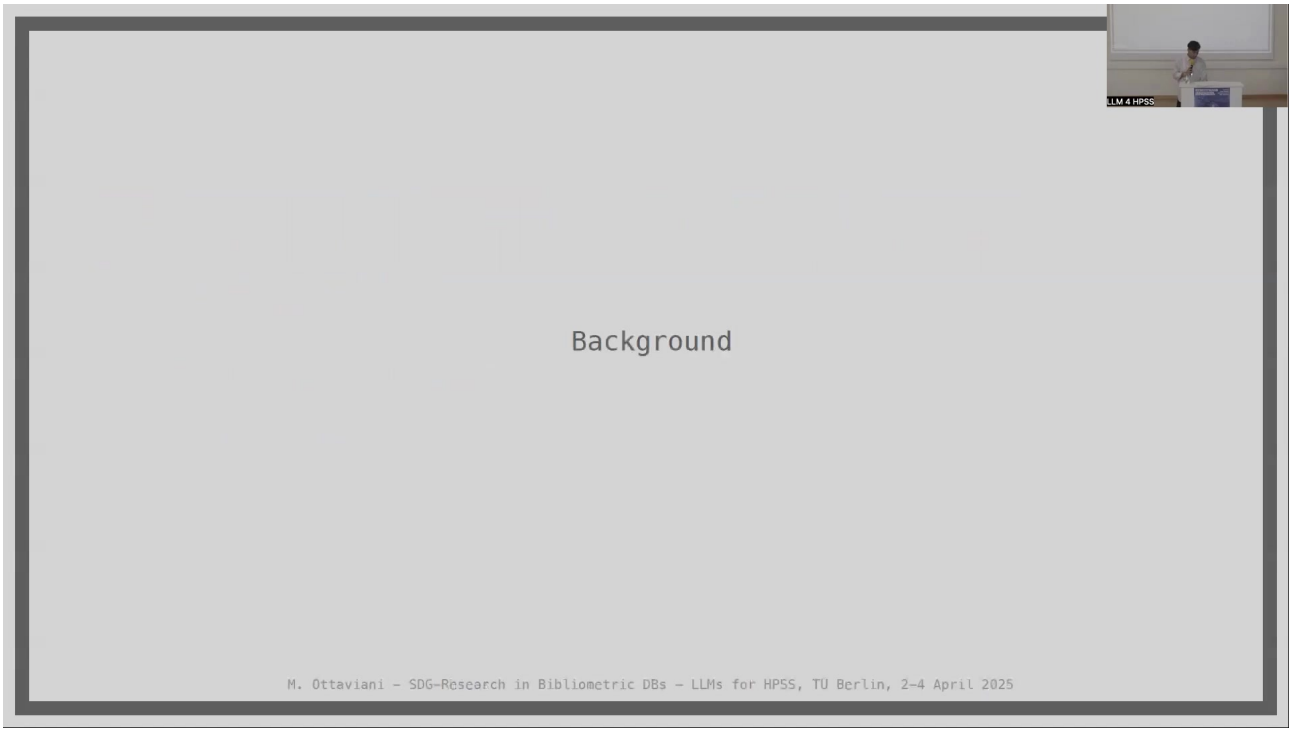

Figure 10.1: Slide 01

Bibliometric databases operate as critical digital infrastructures within the sociology of science, facilitating bibliometric analyses and impact assessments for the scientific community. These databases are not neutral but possess a performative nature, based on specific understandings of the science system and inherent value attributions, as discussed by Whitley (2000) and Winkler (1988).

Major platforms such as Web of Science, Scopus, and OpenAlex have introduced bibliometric classifications designed to align published research with the United Nations Sustainable Development Goals (SDGs). Prior investigations, including research by Armitage et al. (2020), have demonstrated that SDG labeling performed by different providers (such as Elsevier, Bergen, and Aurora) yields disparate results with notably little overlap in the sets of classified publications.

These discrepancies in classification standards result in varying perceptions regarding research priorities, which in turn can potentially influence decisions related to resource allocation and policy. The design and operation of these databases are also influenced by political and commercial interests.

10.3 Case Study Motivation and LLM Application

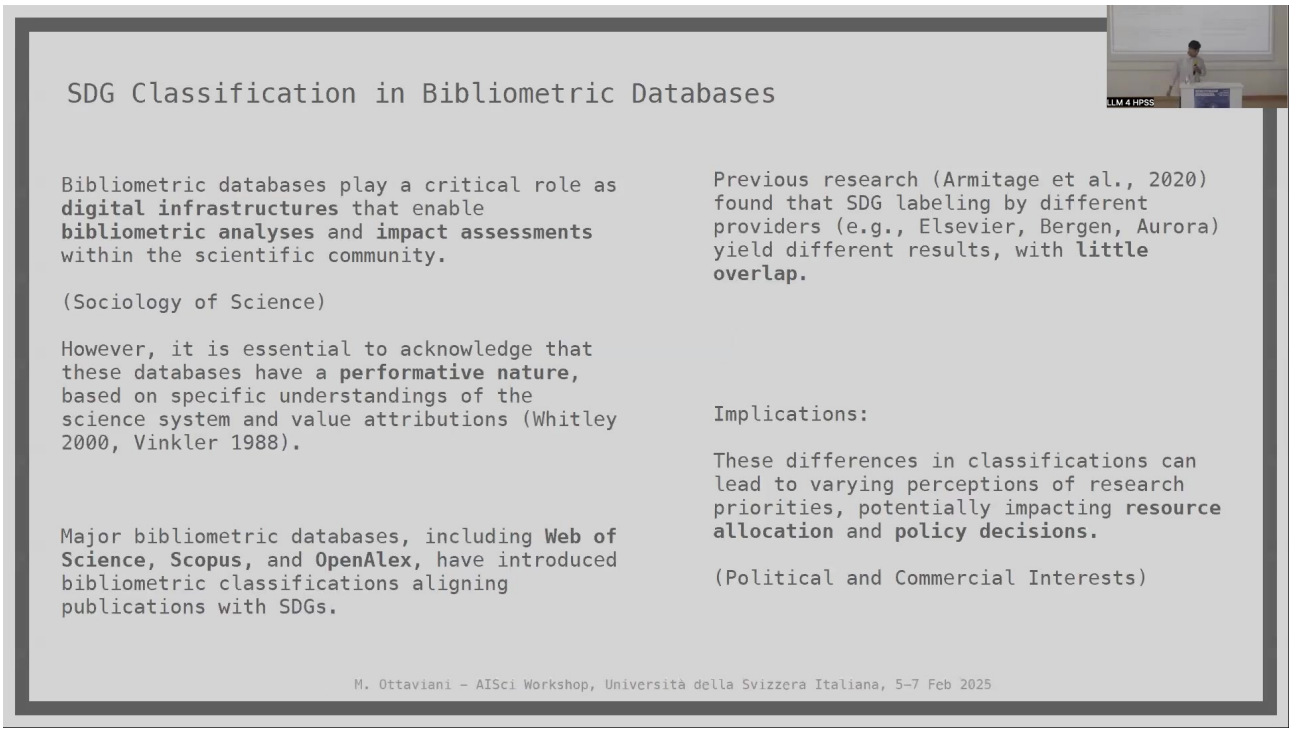

Figure 10.2: Slide 02

The project undertakes a case study examining the representation of UN Sustainable Development Goals within bibliometric data, documented in Ottaviani & Stahlschmidt (2024). The primary motivation for this study is to assess the aggregated effects on the representation of SDG-related research within bibliometric databases that could arise from the introduction of LLM-based tools.

The method employed involves the use of (Little) Pre-trained Large Language Models, specifically *DistilGPT2*. These LLMs are separately trained, or fine-tuned, on distinct subsets of publication abstracts. These subsets are derived from publications classified according to the SDG standards of diverse bibliometric databases.

The LLM technology is utilized in two primary capacities: firstly, as a detector of biases present within the data; and secondly, as a proof-of-concept exercise to demonstrate the potential for introducing such models to automate information extraction processes intended to inform decision-making in research policy.

10.4 Partial Chain of Dependencies and LLM Impact

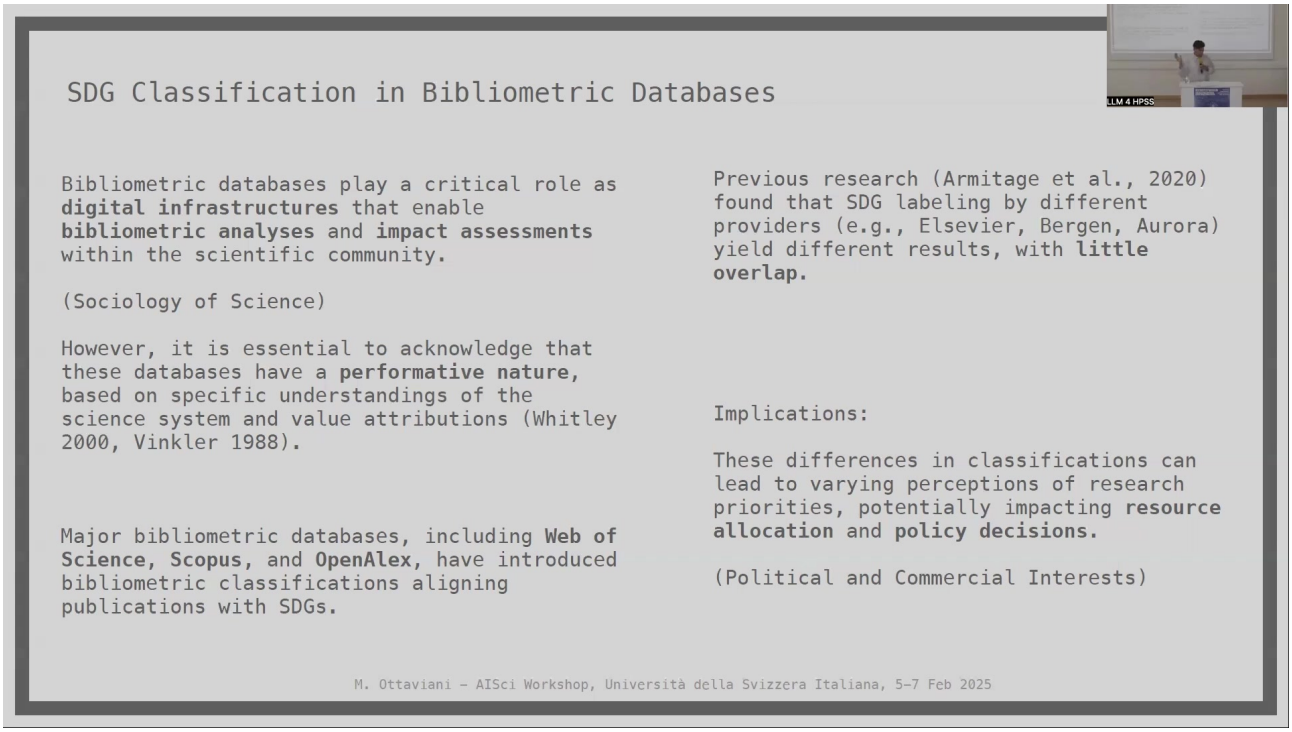

Figure 10.3: Slide 03

A partial chain of dependencies is considered within the study's framework. The SDG classification standards applied by databases are understood to define what constitutes "Research on SDGs". Various actors, including researchers, Small and Medium-sized Enterprises (SMEs), governments, and intermediate figures, process this defined "Research on SDGs". This research then serves to inform "Decision-making to align with SDGs". Subsequently, "Decision-making to align with SDGs" is depicted as impacting "Socioeconomic inequalities".

The introduction of LLMs into Research Policy is positioned as a mechanism for detecting "biases" present within the body of "Research on SDGs". This "Introduction of LLM in Research Policy" is also shown to impact "Socioeconomic inequalities". Within this chain, LLMs are considered to potentially alter the metadata associated with "Research on SDGs", and these changes in metadata can influence the advices, choices, indicators, and measures derived from the research.

10.5 Actors and Selected SDGs

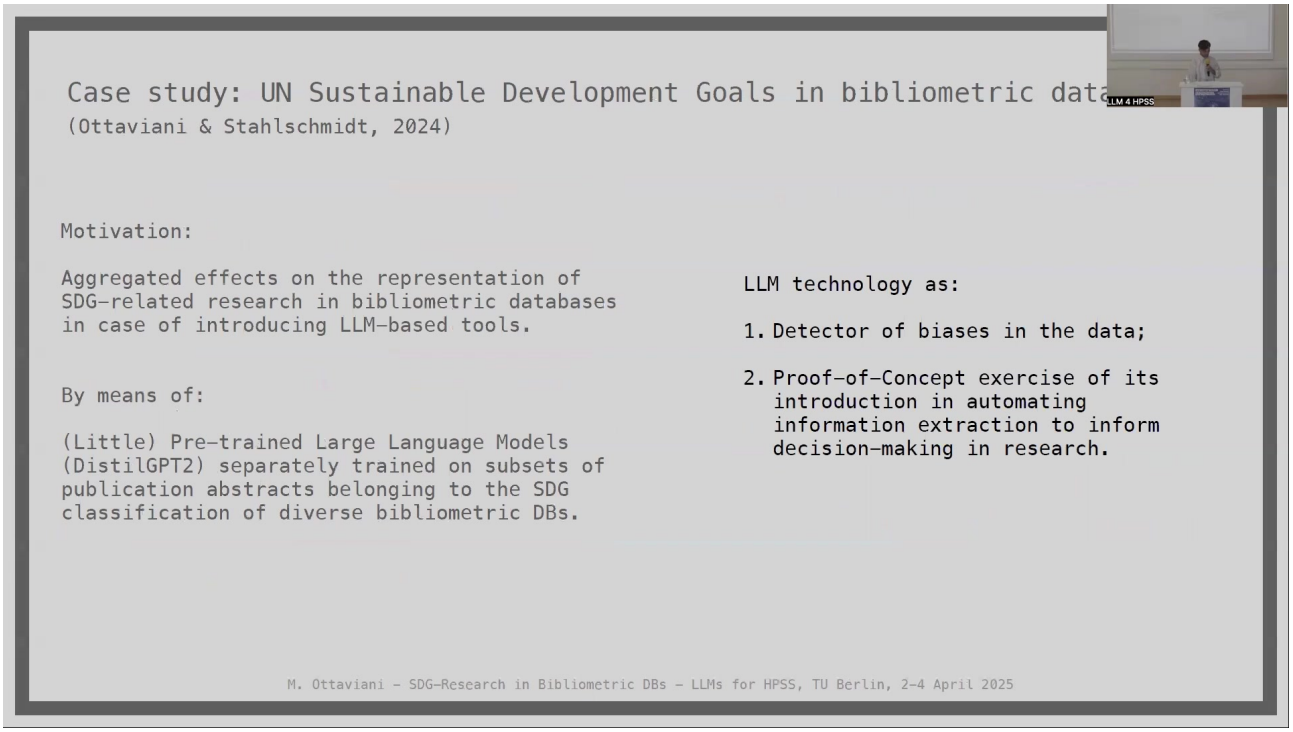

Figure 10.4: Slide 04

The study considers three primary bibliometric databases: Web of Science, managed by Clarivate (US); Scopus, managed by Elsevier (UK); and OpenAlex, which was formerly associated with Microsoft (US) but is now open source.

To investigate socioeconomic inequalities, five specific SDGs are selected: SDG4 (Quality Education), SDG5 (Gender Equality), SDG10 (Reduce inequalities), SDG8 (Decent Work and Economic Growth), and SDG9 (Industry, Innovation, and Infrastructure). These five SDGs are further categorized into two dimensions for analysis: the Equity or socio dimension, encompassing SDG4, SDG5, and SDG10; and the Economic and technological development or economic dimension, comprising SDG8 and SDG9.

10.6 Processed Data and Benchmark

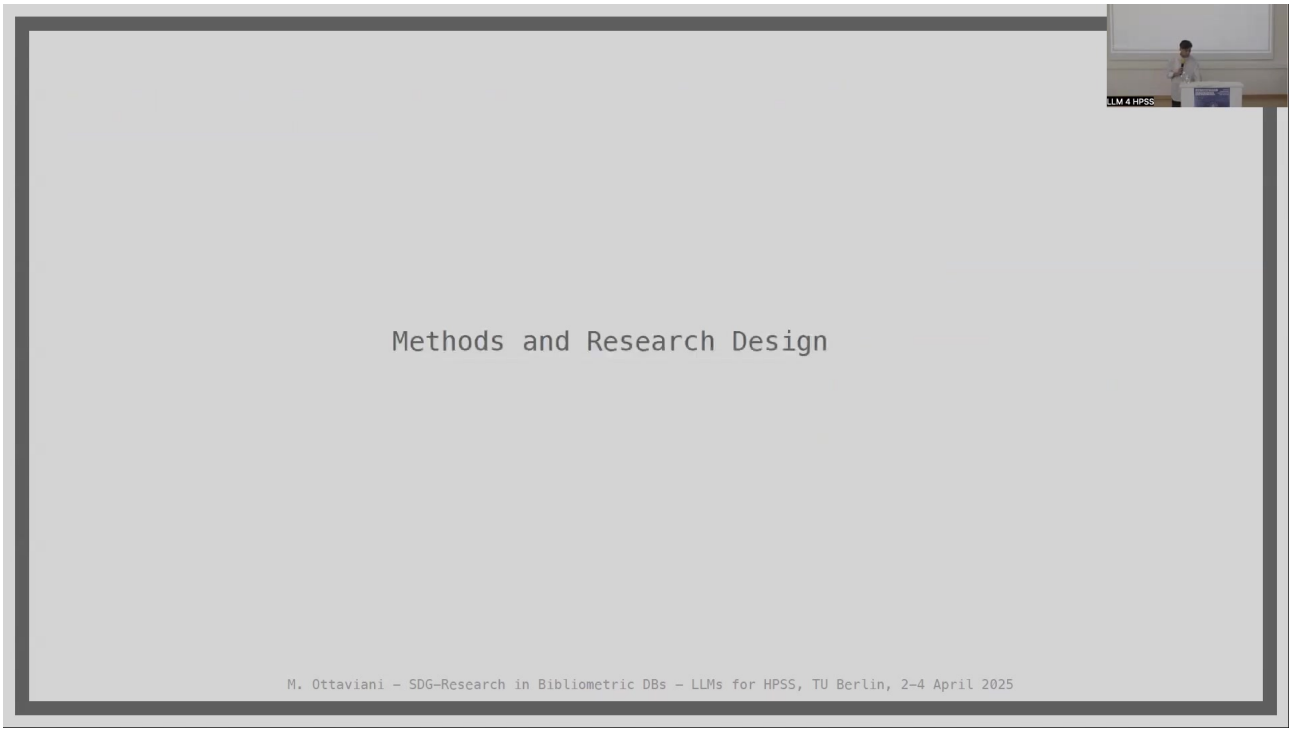

Figure 10.5: Slide 05

The study processes a jointly indexed subset of publications, totaling 15,471,336 publications. This subset consists of publications shared across all three bibliometric databases—Web of Science, Scopus, and OpenAlex—identified through exact DOI matching. The data collection spans the period from January 2015 to July 2023.

The analysis focuses on the application and performance of the three distinct classification standards used by these databases for the five selected SDGs. Crucially, this analysis is performed on the shared corpora, the jointly indexed subset, rather than the entirety of each database. This approach establishes a common benchmark for comparing the different classification outcomes. Consequently, for each specific SDG, three distinct subsets of publications are generated, each representing the set of publications classified under that SDG by one of the three bibliometric databases.

10.7 Comparing SDG Classifications and Identifying Dimensions of Bias

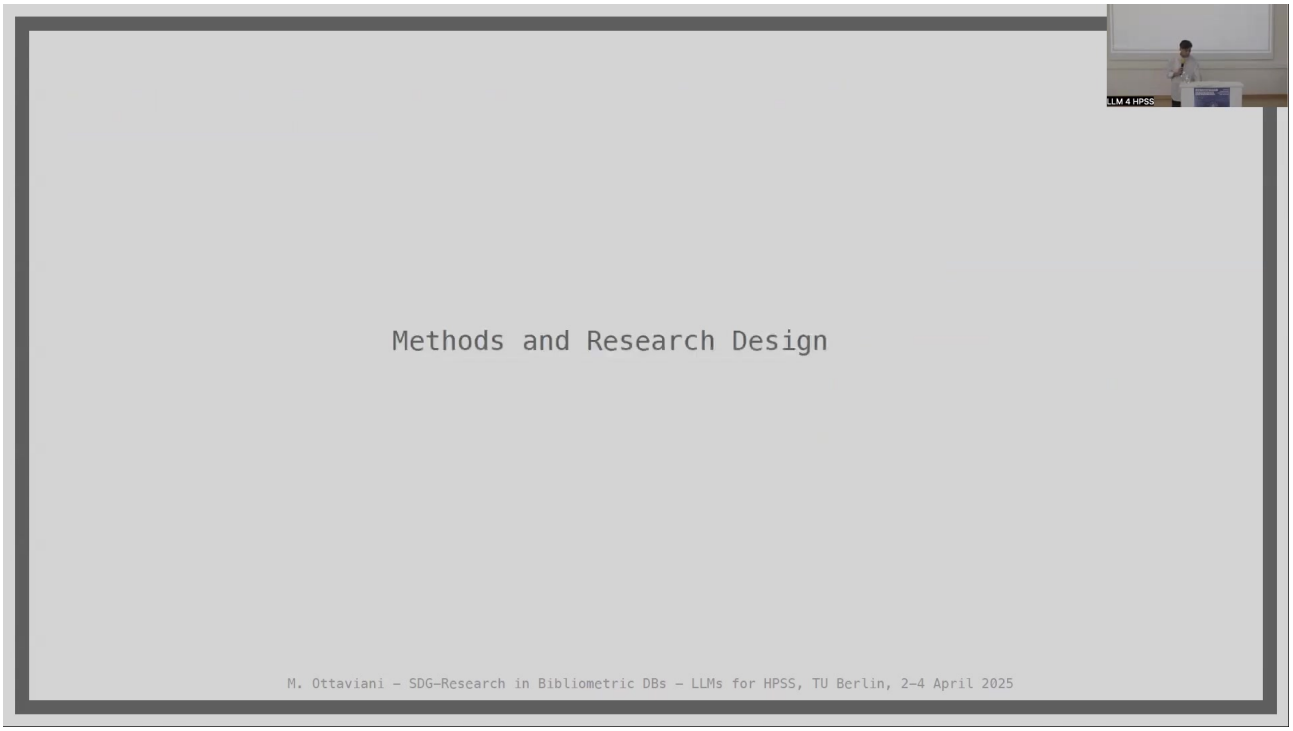

Figure 10.6: Slide 05

The study compares the sets of SDG-classified papers among the three bibliometric databases. This comparison is visually represented using Venn diagrams, which show the overlap of publications classified under specific SDGs, such as SDG4, SDG5, and SDG10 for the socio dimension, and SDG8 and SDG9 for the economic dimension.

The analysis involves determining which specific SDG targets are addressed within the classified publications and which are not. This process helps identify potential biases, including those that might be indirectly considered through the targets. Four main dimensions where biases are observed are identified:

- Locations, which are mentioned in targets across all SDGs
- Actors
- Data and metrics, which primarily emerge as part of the LLM responses
- Focuses, which are specific to each SDG

10.8 LLM Selection and Fine-tuning Strategy

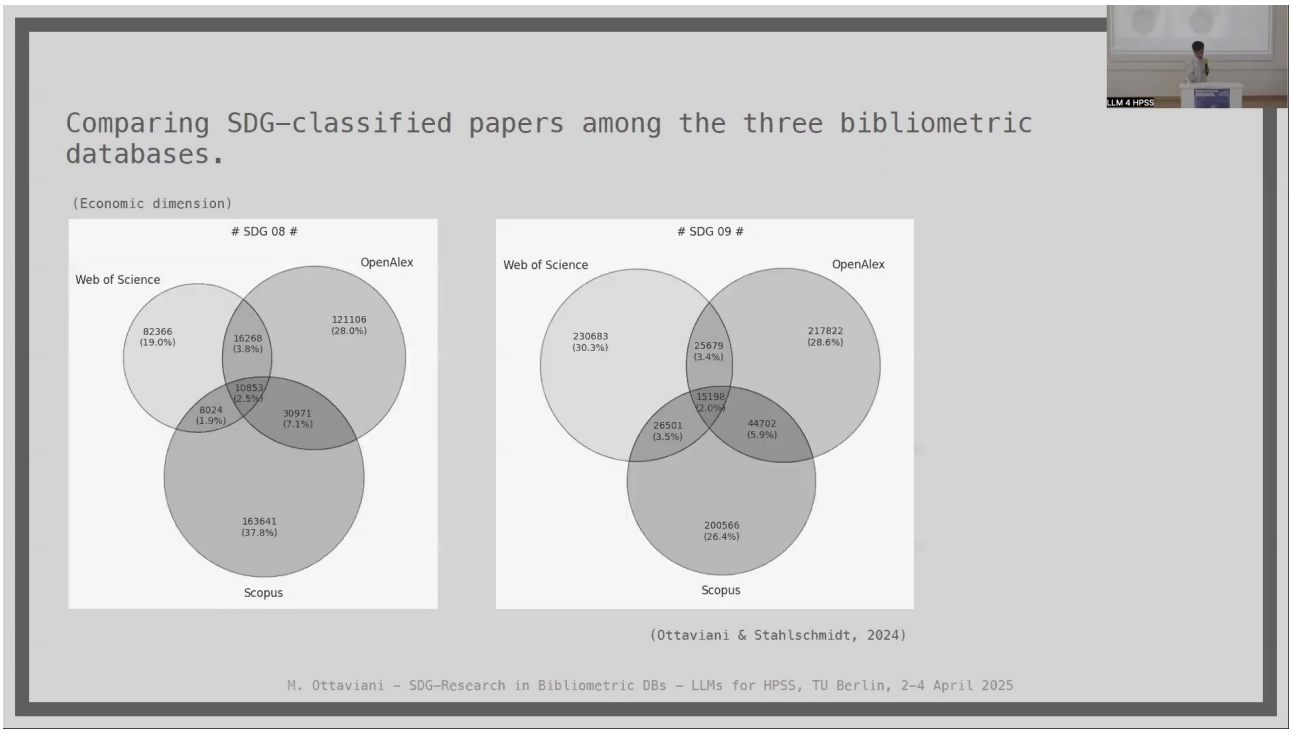

Figure 10.7: Slide 08

The study involves a specific choice of LLM technology and the subsequent fine-tuning of 15 separate LLMs. Leading commercial and open-source pre-trained LLMs are considered ineligible for this work. This is because their extensive pre-training datasets, which include sources like Wikipedia and Reddit conversations, embed existing knowledge about SDGs and strong semantic associations that could introduce bias.

A “fair compromise” is selected: *DistilGPT2*. This model is described as a “very light” pre-trained English-speaking variant of the open-source *GPT2*, utilizing a technique called “distillation” as described by Sanh (2019). *DistilGPT2* has 82 Million parameters, significantly fewer than models like *GPT4*, which has 1.76 Trillion parameters (MxMxdM).

The choice of *DistilGPT2* offers feasibility when working with proprietary data and its “little instructed” nature means its behavior is more closely aligned with the fine-tuning data. Fifteen distinct LLMs are fine-tuned, denoted as *DistilGPT2* {bibDB, SDG}, corresponding to each combination of the three bibliometric databases and the five selected SDGs.

10.9 Systematic Overlook by the LLM

Figure 10.8: Slide 14

Despite being stimulated by specific prompts related to SDG targets, the fine-tuned LLM consistently demonstrates a systematic failure to address certain categories of locations, actors, and focuses. Locations that are systematically overlooked include African countries (with the exception of South Africa), developing countries (with a question mark noted regarding China), least developed countries, and Small Island Developing States.

Similarly, specific categories of actors are systematically overlooked, such as vulnerable people, persons with disabilities, indigenous peoples, and children in vulnerable situations. For SDG4, as an illustrative example, specific focuses that are systematically missed include vocational training, scholarships, the concept of safe, non-violent, inclusive, and effective environments, sustainable lifestyles, human rights, the promotion of a culture of peace and non-violence, global citizenship, the appreciation of cultural diversity, free primary and secondary education, and tertiary education. This pattern of systematic overlook is identified as a recurrent result observed across all five of the SDGs examined in the study.

10.10 Considerations Across the 5 SDGs

Across the five SDGs studied, several consistent patterns and considerations emerge.

Regarding Locations, least developed countries are barely addressed, with South-Saharan Africa noted specifically in relation to SDG8. Beyond the undoubted monopoly of the United States in mentions, South Africa and China are the most frequently quoted locations, followed by the UK and Australia.

Concerning Actors, discriminated and vulnerable categories are systematically overlooked across the different SDGs, with no macro response observed for these groups.

In terms of Metrics, the analysis reveals that many different surveys are recalled as datasets, such as DHS (Demographic and Health Surveys) and WVS (World Values Survey). This indicates the presence of recurrent data from surveys within the semantic multi-layer networks formed after fine-tuning the LLMs. Various Research methodologies are also recalled, including theoretical, empirical, thematic analysis, market dynamics, and macroeconomics.

The Focuses identified are SDG-specific, but the most sensitive ones, such as Human Trafficking, human exploitation, and migration, are often missing. Furthermore, notable methodological differences are observed between the databases: Web of Science tends to show a very theoretical approach for certain SDGs, while Scopus and OpenAlex exhibit a significantly more empirical approach for the same goals.

10.11 Case Study Round-up: Findings and Limitations

The case study provides a round-up of key findings and acknowledged limitations. A primary finding is that introducing LLMs as an analytical AI tool positioned between the SDG classification process and the policymaker reveals a systematic overlook within the data, specifically the scientific publications classified by SDGs.

This systematic overlook pertains to the most disadvantaged categories of individuals, the poorest countries, and underrepresented topics that are explicitly focused on by SDG targets. Conversely, the data demonstrates full attention directed towards economic superpowers and highly developing countries. The results clearly underscore the decisive influence of the bibliometric classification of SDGs, highlighting that this practice, while appearing objective and science-informed, significantly shapes the representation of research.

The study acknowledges several limitations. High sensitivity is observed with respect to the model architecture used, the training data, the hyperparameters, and the decoding strategy. While the use of three different databases partially accounts for the sensitivity to training data, and the application of three different decoding strategies based on literature partially accounts for this sensitivity, these factors remain limitations. The framework employed is general, and the potential for exploring more developed model architectures is noted.

Chapter 11

Extracting Citation Data from Law and Humanities Scholarship

The project addresses the problem of extracting citation data from Law and Humanities scholarship, which heavily relies on complex footnotes. Existing bibliometric databases (*Web of Science*, *Scopus*, *OpenAlex*) have extremely poor coverage for historical and non-English SSH publications, are expensive, and have restrictive licenses. Traditional machine learning tools like *ExCite* perform poorly on complex footnote structures, exhibiting low extraction and segmentation accuracy (e.g., *ExCite* ...

11.1 Overview

The project addresses the problem of extracting citation data from Law and Humanities scholarship, which heavily relies on complex footnotes. Existing bibliometric databases (*Web of Science*, *Scopus*, *OpenAlex*) have extremely poor coverage for historical and non-English SSH publications, are expensive, and have restrictive licenses.

Traditional machine learning tools like *ExCite* perform poorly on complex footnote structures, exhibiting low extraction and segmentation accuracy (e.g., *ExCite* accuracy around 0.22-0.26 for extraction). Footnotes in SSH are often complex, containing commentary, abbreviations, and multiple references ("footnotes from hell"). Creating training data for traditional methods is laborious using annotation tools.

Large Language Models (*LLMs*) and *Vision Language Models* (*VLMs*) show promise for handling messy textual data and PDFs, but the primary challenge is trusting the results due to potential hallucinations (e.g., inventing non-existent citations, as seen in legal cases).

A robust testing and evaluation solution is required, necessitating a high-quality gold standard dataset, a flexible evaluation framework adaptable to fast-moving technology, and solid testing algorithms for comparable metrics.

The project develops a specialized gold standard dataset encoded in *TEI XML*, chosen for its well-established standard, detailed specification covering phenomena beyond mere reference management (including context for citation intention), and compatibility with existing digital humanities corpora and tools like *Grobid*. The dataset creation involves stages: screenshot of the PDF, segmentation of reference strings from non-reference text within footnotes, and parsed structured data.

The dataset currently includes over 1,500 references from open access journals to enable full publication from PDF to parsed data structures.

A *Python* package named *Llamore* (*Large Language Models for Reference Extraction*) is developed. *Llamore* is lightweight, acting as an interface to various *LLMs*/*VLMs* (compatible with *OpenAI API*, covering *Ollama*, *VLLM*, etc.). It takes text or PDFs as input and outputs references in *TEI XML* format.

It also provides an evaluation function using the *F1 score* metric to compare extracted references against gold standard references. The *F1 score* calculation involves counting exact matches of bibliographic elements (analytic title, monographic title, surname, publication date, etc.) and dividing by the number of predicted and gold elements.

The problem of aligning extracted references to gold references is solved using an unbalanced assignment problem solver (from *SciPy*), maximizing the total *F1 score* with unique assignments and penalizing missing or hallucinated references with an *F1 score* of zero.

Evaluation results show that *Llamore* performs comparably to *Grobid* on biomedical datasets (*PLOS 1000*) but significantly outperforms *Grobid* on the specialized humanities dataset, where *Grobid* struggles due to being out of distribution. While *LLMs* require significantly more compute than *Grobid*, their performance on complex SSH footnotes is superior.

Error analysis suggests issues include difficulty distinguishing volume/page numbers, misclassifying names in titles as authors, and misinterpreting terminology like “idem” as authors. The required *F1 score* for “good enough” data depends on the analysis goal (tendencies vs. accurate data). The current focus is on achieving reliable results before scaling up to avoid interpreting hallucinated data.

Future work could involve *retrieval augmented generation* (*RAG*) using disciplinary databases for validation and potentially human-in-the-loop workflows for error correction and iterative model improvement. The project aims for both specific research results (e.g., citation trends in a specific journal) and generic applicability to other projects.

11.2 Project Scope and Problem

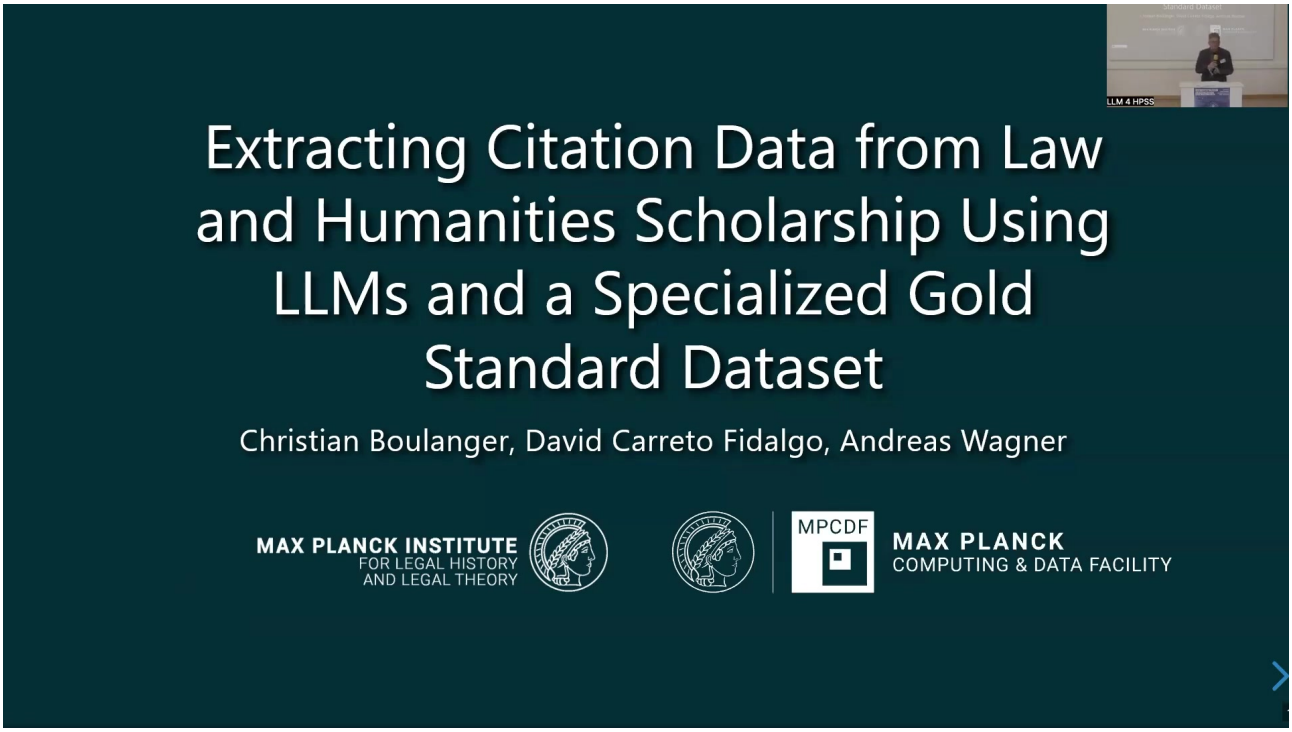

Figure 11.1: Slide 01

The project focuses on extracting citation data specifically from Law and Humanities scholarship, a domain characterized by extensive and complex footnotes. The primary challenge involves parsing these footnotes using *Large Language Models* (*LLMs*) or other algorithmic approaches.

The extracted data is intended for generating citation graphs, which are valuable tools in intellectual history and the history of science. Citation graphs enable the discovery of patterns and relationships within knowledge production, facilitate the reconstruction of intellectual influences, and allow for the measurement of the reception of specific ideas. An example application involves tracking the change in most-cited authors over time, such as an analysis conducted for the *Journal of Law and Society* between 1994 and 2003.

11.3 Problem: Bibliometric Database Coverage

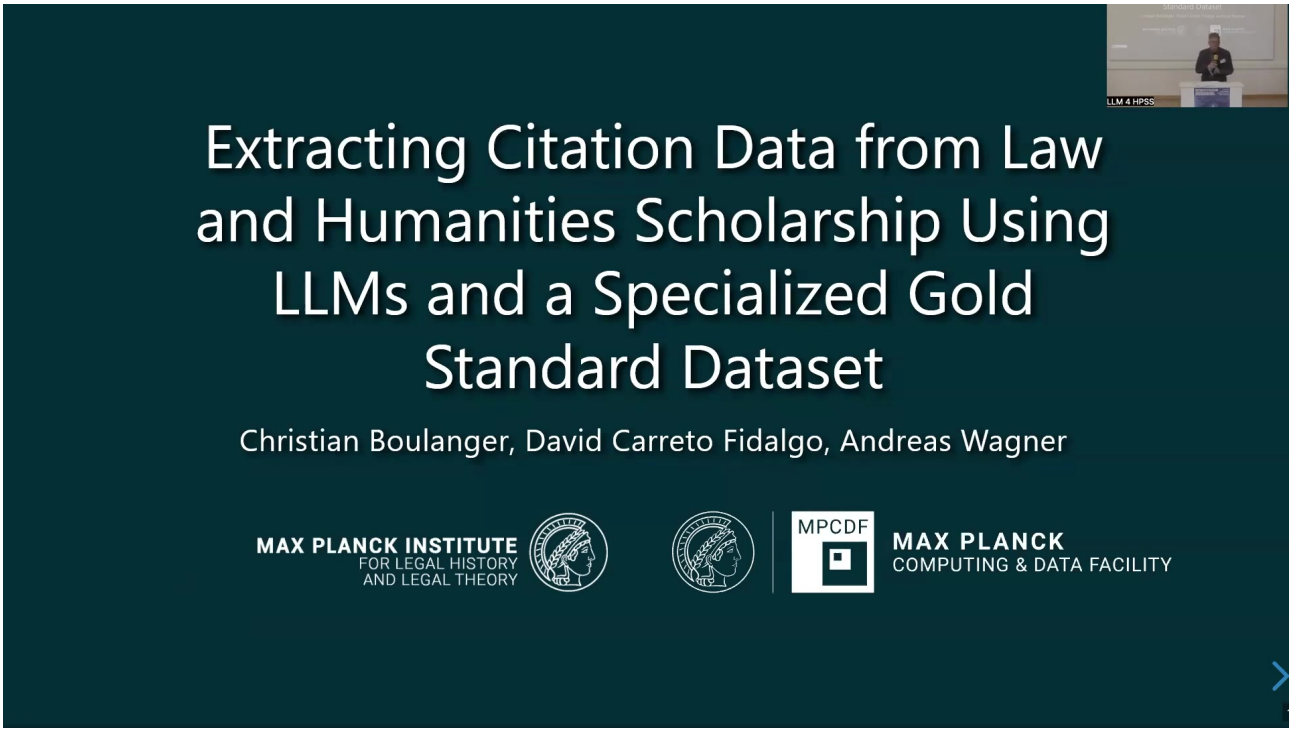

Figure 11.2: Slide 01

A significant problem is the extremely poor coverage of historical *Social Sciences and Humanities* (*SSH*) scholarship in existing bibliometric databases. The primary databases in this space include *Web of Science*, *Scopus*, and *OpenAlex*. *Web of Science* and *Scopus* are characterized by high costs and very restrictive licenses, making dependence on them undesirable.

While *OpenAlex* is preferable due to its open access nature, it also lacks sufficient coverage for the specific content required for this research.

Common coverage gaps across these databases for *SSH* literature include a lack of inclusion for journals not classified as "A-journals," insufficient data for publications from the pre-digital age, and a general lack of coverage for non-English language content.

An illustrative example is the *Zeitschrift für Rechtssoziologie*, a German journal established in 1980. Analysis of available citation data by decade shows very low coverage across *Dimensions*, *OpenAlex*, and *Web of Science* for the decades before the 2000s. Although coverage improves somewhat after 2000, it remains incomplete, particularly for *Web of Science*.

Several factors contribute to this poor coverage in *SSH*. Firstly, there is a perceived lack of financial incentive compared to *STEM*, medicine, and economics, which are typically well-represented.

Secondly, these databases often focus on metrics like the "impact factor," which aligns with science evaluation goals but not necessarily with the objectives of intellectual history research. Finally, the literature itself presents a technical challenge due to the extensive use of complex footnotes, which are difficult for automated systems to process accurately.

11.4 Complex Footnotes and Training Data

Figure 11.3: Slide 04

Problem 2 concerns the nature of the footnotes themselves, often referred to as "footnotes from hell." These are typical of humanities scholarship and contain complex structures, including commentary, messy formatting, abbreviations, parenthetical information, and multiple distinct references embedded within surrounding non-citation text. An example image displays a scanned page with a footnote exhibiting these characteristics, featuring content in both German and English, various abbreviations, and multiple citations within a single numbered entry.

Problem 3 highlights the difficulty in creating training data for citation extraction. The traditional approach involves a laborious manual annotation process. A web-based annotation tool is utilized for this purpose, allowing users to highlight segments of text within footnotes and assign specific labels corresponding to bibliographic elements such as:

- Author
- Title
- Date
- Journal
- Volume
- Pages
- DOI
- ISBN
- URL

- BackRef
- Signal
- Ignore

and indicators for whether a segment is a reference or not. This manual annotation requires a significant investment of time and effort.

11.5 Limitations of Existing Tools

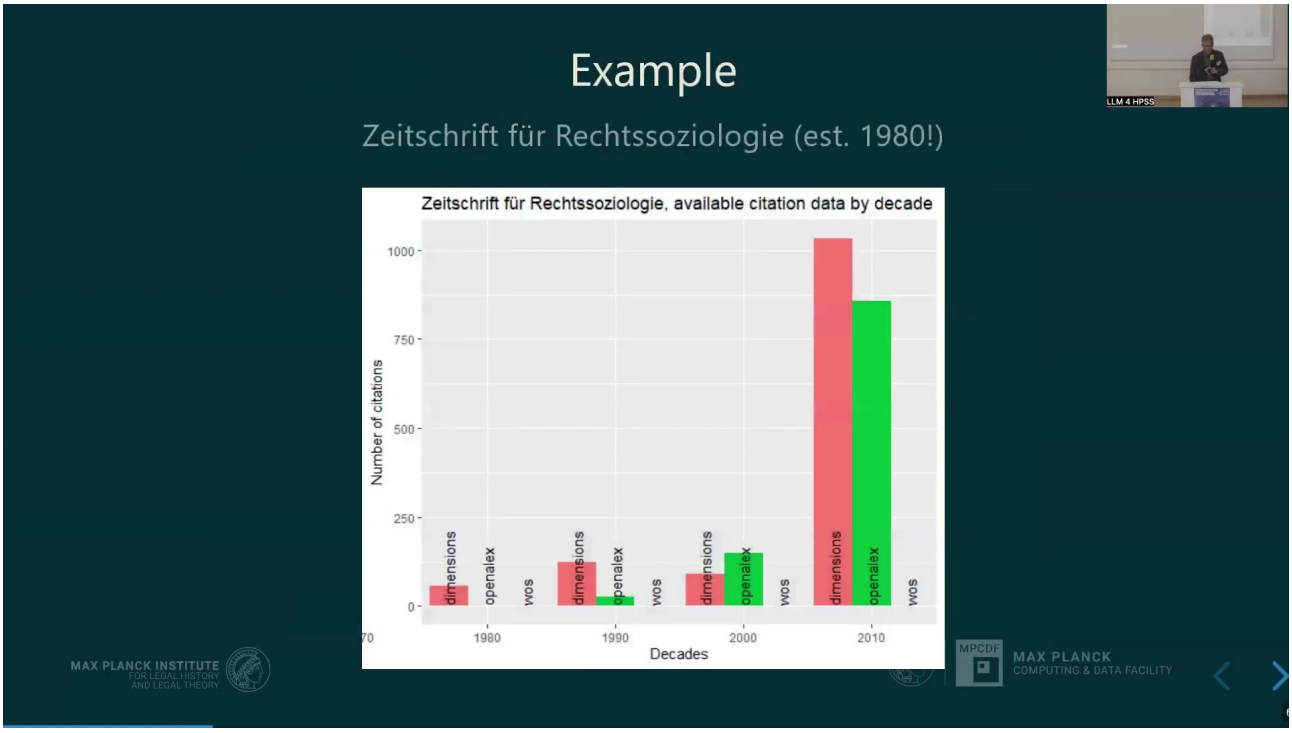

Figure 11.4: Slide 05

Problem 4 identifies a critical limitation: existing tools designed for citation extraction are unable to handle the complexity of humanities footnotes effectively. These traditional tools typically rely on machine learning methods such as *Conditional Random Forests*. However, their performance on the specific type of complex footnote data encountered in this domain is poor.

An example illustrating this limitation is the performance of the tool *ExCite*. Its performance is evaluated using metrics including Extraction Accuracy and Segmentation Accuracy across different training data configurations: Default, Footnoted, and Combined. The results show consistently low accuracy scores.

With Default training data, Extraction Accuracy is 0.24 and Segmentation Accuracy is 0.37. Using Footnoted training data yields Extraction Accuracy of 0.26 and Segmentation Accuracy of 0.37. The Combined training data results in Extraction Accuracy of 0.22 and Segmentation Accuracy of 0.47. These figures, sourced from Boulanger/Iurshina (2022), Table 1, demonstrate that *ExCite*, and by extension other similar traditional tools, struggle significantly with this task, regardless of the training data used.

11.6 LLMs as Solution: The Trust Problem

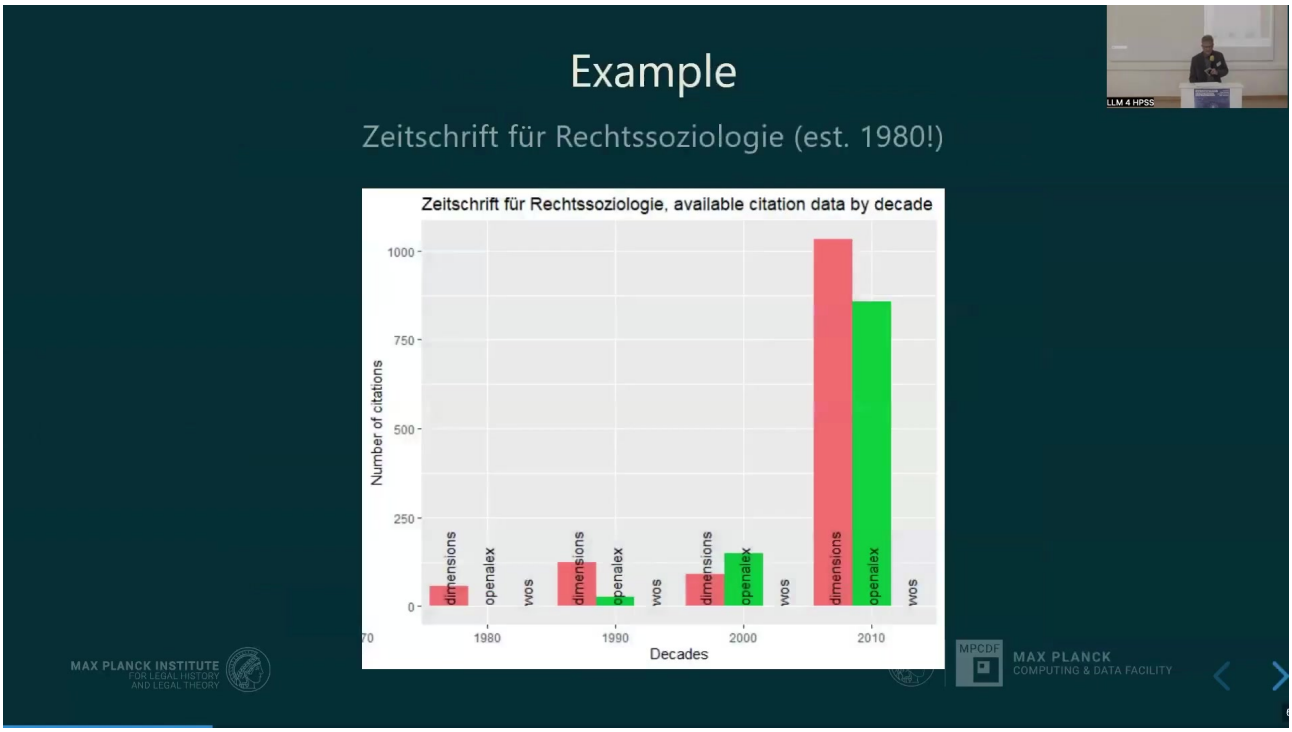

Figure 11.5: Slide 05

The question arises whether *Large Language Models* (*LLMs*) can offer a solution to the challenges of extracting citation data from complex footnotes. Early experiments conducted in 2022 using models such as *text-davinci-003* demonstrated the potential power of *LLMs* to extract references even from messy textual data. Newer models promise even better performance.

Furthermore, *Vision Language Models* (*VLMs*) possess the capability to process PDFs directly, which is advantageous given that source materials are frequently available in this format.

Potential methods for leveraging *LLMs* include *prompt engineering*, *Retrieval Augmented Generation* (*RAG*), and *finetuning*. However, a primary concern is the trustworthiness of the results produced by these models. A significant problem is the potential for hallucination, where *LLMs* invent information that does not exist, including fabricating citations.

A notable example illustrating this issue involved a lawyer who used *ChatGPT* for a federal court filing and cited non-existent cases invented by the model, resulting in severe consequences. This highlights the critical principle that analysis should not be undertaken unless the results are known to be correct and sufficient validation data is available to verify accuracy.

11.7 Solution Requirements

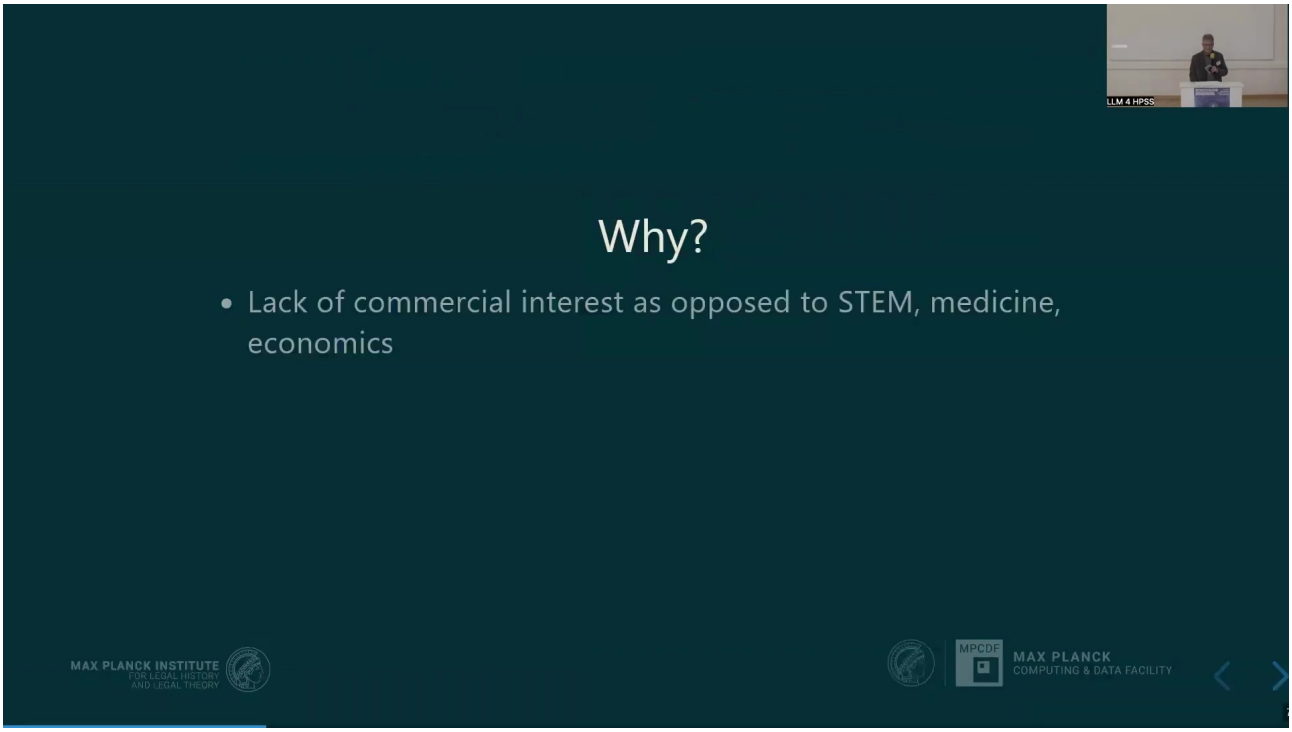

Figure 11.6: Slide 06

To address the trust issue and enable reliable citation extraction, a robust testing and evaluation solution is required. This solution must meet several key requirements.

- Firstly, it necessitates a high-quality Gold Standard dataset against which extracted results can be compared.
- Secondly, it requires a flexible framework capable of adapting easily to the rapidly evolving technology landscape of *LLMs* and *VLMs*.
- Finally, the solution must incorporate solid testing and evaluation algorithms designed to produce comparable metrics, allowing for objective assessment of different approaches and models.

11.8 Gold Standard Dataset

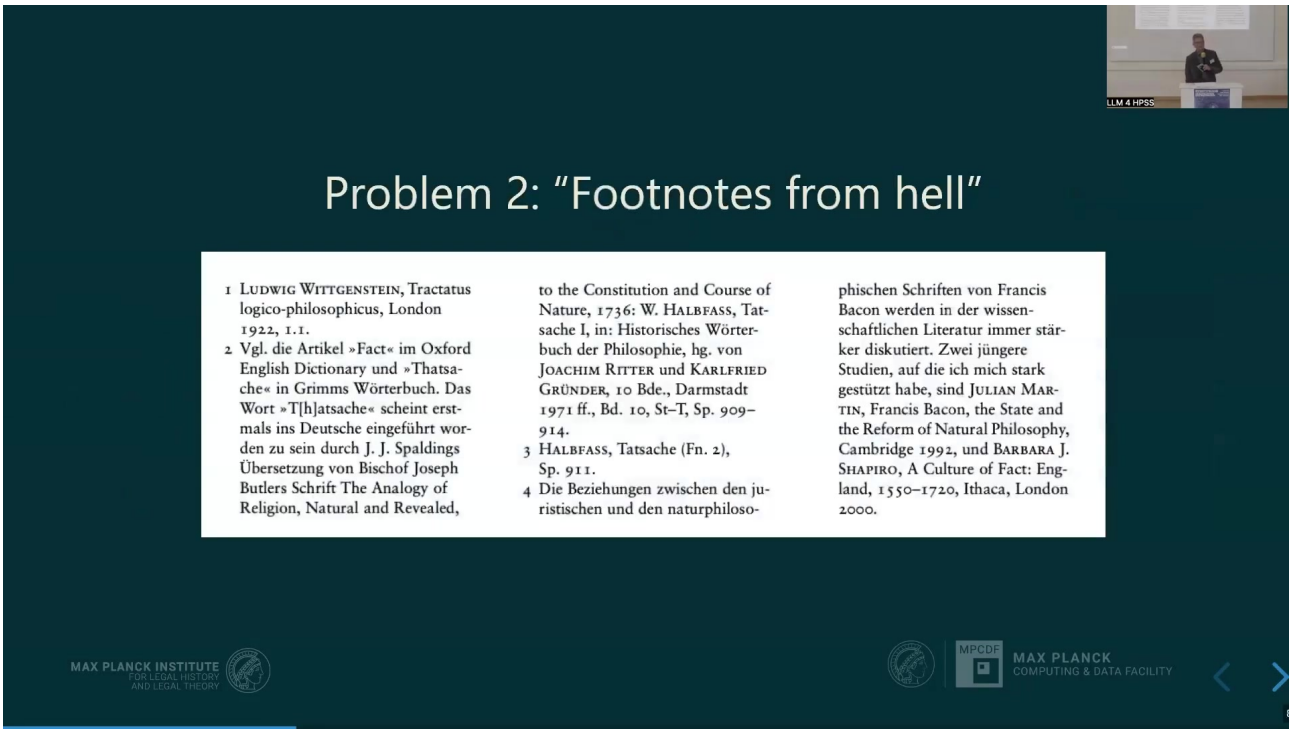

Figure 11.7: Slide 07

The project involves compiling a high-quality gold standard dataset intended for both training and evaluation purposes. The chosen encoding standard for this dataset is *TEI XML*. This standard was selected because it is well-established and precisely specified within the humanities and digital editorics fields.

TEI XML is more comprehensive than simpler bibliographical standards like *CSL* or *BibTeX*, covering a wider range of phenomena and extending beyond basic reference management. It allows for encoding contextual information, which can be valuable for tasks such as classifying citation intention.

Furthermore, adopting *TEI XML* enables the project to potentially utilize existing text collections and corpora from other digital editorics projects that publish their source data, including detailed reference encodings, in this format. Such existing corpora can also serve as resources for testing the generalization and robustness of the developed mechanisms.

The dataset establishment process is currently underway. The encoding involves several stages, illustrated by an example of footnote number four. These stages include capturing a screenshot of the PDF source, segmenting the text to distinguish the actual reference string from surrounding non-reference text within the footnote (such as introductory phrases), and finally creating a parsed structured data representation of the reference.

Midway through the project, the strategy for building the dataset was adjusted. Initially, the focus was on data directly relevant to the primary research question. More recently, the decision was made to include the source PDFs and structure the dataset to enable the use of *Vision Language Models* (*VLMs*). The goal is to be able to publish the entire dataset pipeline, from the source PDFs through to the parsed data structures.

To facilitate this, the selection of source journals was changed to focus on open access publications. The dataset currently includes the encoding of over 1,500 references. It is noted that this count refers to occurrences, meaning the same work referenced multiple times is encoded separately to capture the specific context of each mention.

A significant benefit of using the interoperable *TEI XML* standard is the availability of numerous tools. A particularly relevant tool for this project is *Grobid*, a popular system for reference and information extraction. *Grobid* utilizes *TEI XML* for its own training and evaluation processes. By using the same data format, the project can directly compare its performance against *Grobid*, potentially use *Grobid*'s existing training data, and contribute the newly created dataset to the *Grobid* development team.

11.9 Llamore: Extraction and Evaluation Tool

Figure 11.8: Slide 14

The project developed a tool named *Llamore*, which stands for *Large Language Models for Reference Extraction. Llamore* is implemented as a small *Python* package. Its core capabilities include taking text or PDF documents as input, extracting references from them, and exporting these extracted references as *TEI* formatted *XML* files.

Additionally, if gold standard references are provided, *Llamore* can evaluate the performance of the extraction process.

The design of *Llamore* prioritized two main objectives: being lightweight and ensuring broad compatibility. It is lightweight because it does not contain any language models internally; instead, it functions as an interface to a model selected by the user. This design also ensures compatibility with a wide range of both open and closed *LLMs* and *VLMs*.

Implementation details include its availability on *PyPI*, allowing simple installation via the `pip` package manager. The extraction workflow involves defining an extractor object specific to the chosen model, such as using the *OpenAIExtractor*. This particular extractor provides compatibility with many open models (like those served by *Ollama* or *VLLM*) by interacting with their *API* endpoints that are designed to be compatible with the *OpenAI API* specification.

The user provides a PDF or text input to the extractor, receives the extracted references, and can then export these references to an *XML* file. For evaluation, the user imports the *F1* class from the package and provides both the gold standard references and the extracted references to this class to compute performance metrics.

11.10 Evaluation Methodology

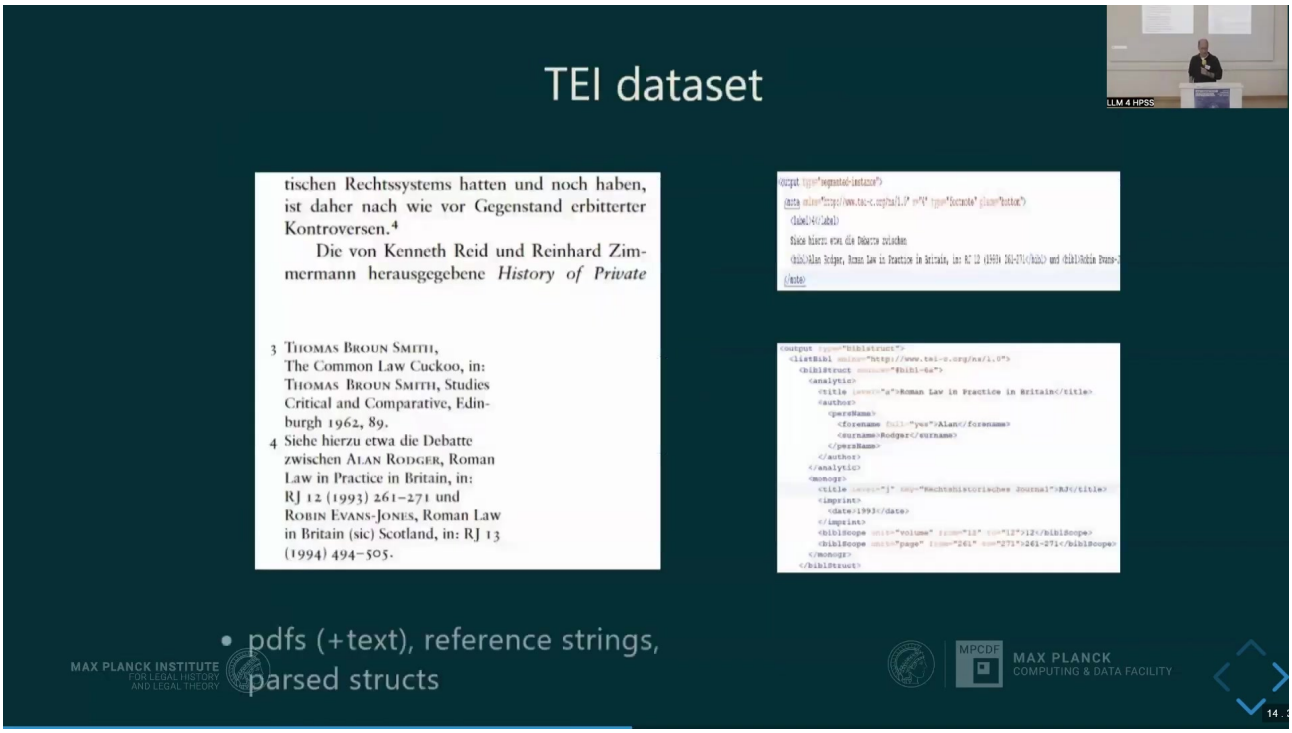

Figure 11.9: Slide 16

The evaluation methodology employed utilizes the *F1 score*, a well-established metric for comparing structured data. The *F1 score* is calculated based on Precision and Recall, which in turn depend on the number of matches between elements in an extracted reference and a corresponding gold standard reference. A match is defined based on the exact correspondence of specific bibliographic elements, such as analytic title, monographic title, surname, and publication date.

Partial matches, like a forename with an extra dot, might not count as an exact match depending on configuration. Precision is calculated as the number of matches divided by the total number of elements predicted in the extracted reference, while Recall is the number of matches divided by the total number of elements in the gold reference. The *F1 score* is the harmonic mean of Precision and Recall. An *F1 score* of 1 indicates perfect extraction, where the reference is perfectly captured, while an *F1 score* of 0 signifies no matches were found.

A key challenge in evaluating performance across a set of references is aligning the extracted references with their corresponding gold standard references. This is tackled by formulating the problem as an unbalanced assignment problem. *Llamore* uses a solver from the *SciPy* library internally to address this. The process involves computing the *F1 score* for every possible pairing between each extracted reference and each gold reference, constructing a matrix of scores.

The solver then finds the assignment of extracted references to gold references that maximizes the total sum of *F1 scores*, ensuring that each reference is assigned uniquely. The final overall score is the macro average of the *F1 scores* for the assigned pairs. The method penalizes missing gold references (those not matched by an extracted reference) and hallucinated extracted references (those not matched to a gold reference) by assigning them an *F1 score* of zero in the averaging process. This approach is noted as being similar to a method recently published by *Packet et al.*

11.11 Evaluation Results

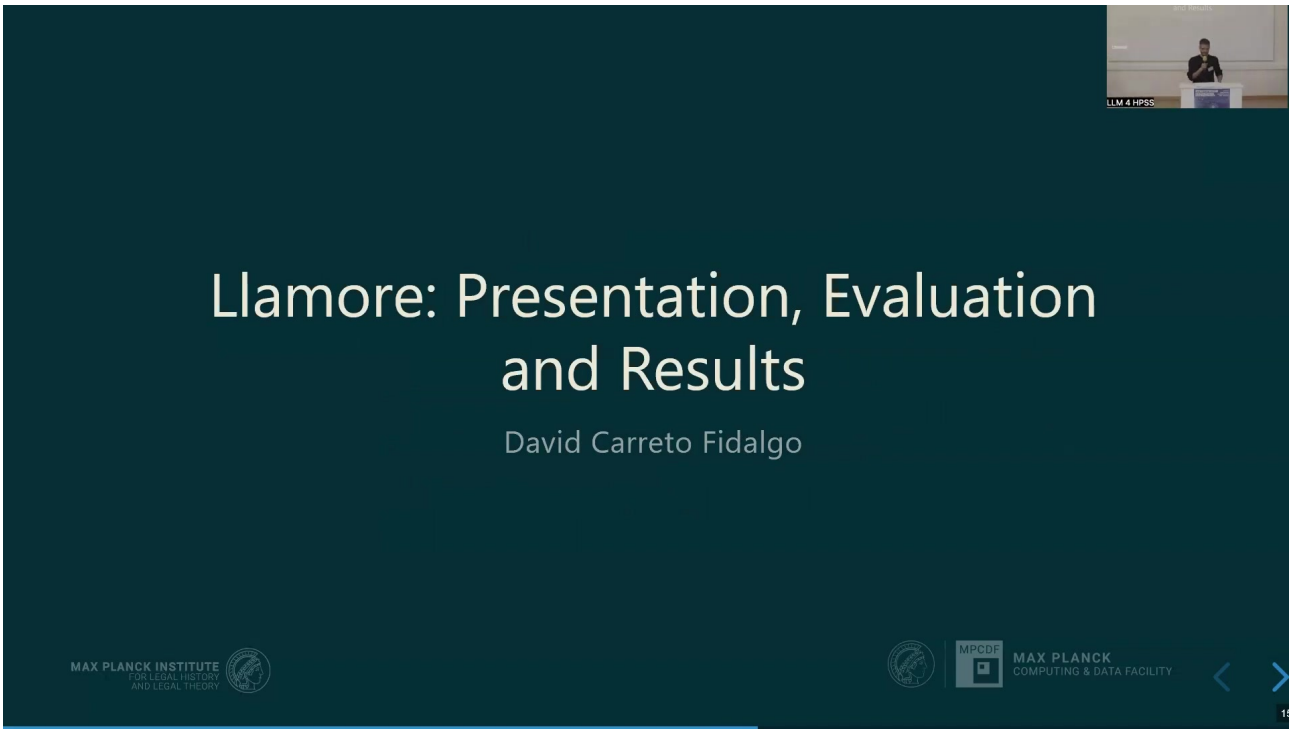

Figure 11.10: Slide 20

Evaluation was conducted to assess the effectiveness of the developed approach. Using the *PLOS 1000* dataset, which comprises 1,000 PDFs from the biomedical field, *Llamore*'s performance was found to be comparable to that of *Grobid*. It is noted that *Grobid* was specifically trained on data from similar journal articles.

However, in terms of efficiency and compute resources, *Grobid* is significantly superior, requiring orders of magnitude less computational power than *LLMs* like *Gemini*.

When evaluated on the specialized humanities dataset created by the project, *Grobid* struggled significantly to extract references, indicating that this data is out of its training distribution. In contrast, the *LLM*-based approach implemented in *Llamore* performed significantly better on this complex dataset.

The conclusion drawn is that while less computationally efficient on standard datasets, the *LLM*-based approach implemented in *Llamore* is more effective for extracting references from the challenging, out-of-distribution data characteristic of humanities footnotes compared to *Grobid*.

It is acknowledged that the *F1 scores* achieved are not considered high overall, suggesting substantial room for improvement. The current implementation utilizes a very basic approach, essentially sending the raw PDF text to a pre-trained model. Future work could explore potential improvements through methods such as *fine-tuning* models or incorporating more contextual information to enhance performance.

Chapter 12

Chatting with Papers

The project develops an AI solution for interacting with specific document collections, referred to as “chatting with papers.” The primary objective is to address the problem of information overload in science dynamics by improving information retrieval and knowledge production processes using AI. The system utilizes a mixed approach combining Large Language Models (LLMs) and semantic artifacts, specifically structured data represented as knowledge graphs and vector spaces derived from do…

12.1 Overview

The project develops an *AI* solution for interacting with specific document collections, referred to as “chatting with papers.” Its primary objective is to address the problem of information overload in science dynamics by improving information retrieval and knowledge production processes using *AI*.

The system utilizes a mixed approach, combining *Large Language Models* (*LLMs*) with semantic artifacts. These artifacts include structured data represented as knowledge graphs and vector spaces derived from document content. The core components are codenamed *Ghostwriter* (the interface) and *EverythingData* (the backend processing pipeline).

The approach is based on *Retrieval-Augmented Generation* (*RAG*), integrating vector embeddings of document content with a metadata layer. This metadata layer is represented as a knowledge graph, which incorporates ontologies and controlled vocabularies, including aspects of responsible *AI*. The graph is expressed using the *Croissant ML* standard.

The system aims for a “local” or “tailored” *AI* solution, functioning as a distributed *AI*. In this architecture, the *LLM* acts as an interface and reasoning engine, connected to the *RAG* library (graph) and consuming embeddings (vectors) as context. A key feature is the use of entity extraction pipelines that link terms to knowledge graphs, specifically *Wikidata*. This linkage provides ground truth, supports multilinguality, and enables validation of *LLM* responses against structured identifiers.

The system splits papers into small blocks, each with a unique identifier. It employs *LLM* techniques to connect and retrieve these blocks, applying weights and knowledge graph information to predict relevant text pieces. It provides summaries and references to original sources, avoids hallucination by relying solely on the ingested data, and indicates when information is not found.

The interface allows users to ask natural language questions, receive summaries and document lists, and add missing information. The system supports multilingual queries by linking terms to *Wikidata* identifiers, which have multilingual

translations. This approach is seen as a way to support the user's thought process and help find relevant research questions rather than providing definitive answers.

The system is being considered for open-source release under the *Linux Foundation*. It is also being explored for integration with various data sources, including *GitHub* content, manuals, and guidelines, with potential applications in building research infrastructure portals. Validation against other systems like *Neo4j Graph Builder* or *Microsoft Graph* is being considered. The approach of using knowledge organization systems linked to identifiers is proposed as a method for benchmarking future generations of *AI* models and ensuring sustainability. The system uses downscaled *LLMs* (e.g., 1 billion parameters) capable of running locally. Recency bias in results is acknowledged, with a proposed solution involving storing facts with timestamps in the knowledge graph to allow for processing linked to specific dates. The approach is considered similar to *Google Notebook ML* due to reliance on similar ideas and collaboration with the same teams.

12.2 Project Overview and Affiliations

Chatting with Papers

the mixed use of LLM's and semantic artifacts to support the understanding of science dynamics - and beyond

Slava Tykhonov [b], Philipp Mayr [a], Jetze Touber [b] Andrea Scharnhorst [b]

[a] GESIS, Cologne, Germany [b] DANS-KNAW, The Hague, The Netherlands

GESIS Leibniz-Institut für Sozialwissenschaften

Data Archiving and Networked Services

DANS

Figure 12.1: Slide 01

The project is titled "*Chatting with Papers*," with the subtitle "the mixed use of *LLM's* and semantic artifacts to support the understanding of science dynamics - and beyond." The authors involved are Slava Tykhonov, Philipp Mayr, Jetze Touber, and Andrea Scharnhorst.

Their affiliations are *GESIS*, Cologne, Germany for Philipp Mayr, and *DANS-KNAW*, The Hague, The Netherlands for Slava Tykhonov, Jetze Touber, and Andrea Scharnhorst. The presentation includes the logos for *GESIS* and *DANS*. The text "*LLM 4 HPSS*" is also present, indicating the context within which this work is situated.

12.3 Science Dynamics and Information Overload

Chatting with Papers

the mixed use of LLM's and semantic artifacts to support the understanding of science dynamics - and beyond

Slava Tykhonov [b], Philipp Mayr [a], Jetze Touber [b] Andrea Scharnhorst [b]

[a] GESIS, Cologne, Germany [b] DANS-KNAW, The Hague, The Netherlands

GESIS Leibniz-Institut für Sozialwissenschaften

Data Archiving and Networked Services

DANS

Figure 12.2: Slide 01

The evolution of sciences exhibits growth and increasing differentiation, presenting the challenge of reviewing, evaluating, and selecting relevant information. A fundamental precondition for creating new knowledge, whether in individual researchers or across academia, is the ability to find and understand existing information. Machines, particularly recent advancements in *AI*, have contributed to this growth in information volume.

The project investigates whether *AI* can also support the knowledge production process itself, framing this as a problem within the domain of *Information Retrieval*. The motivation stems from the need to manage the overwhelming volume of information researchers face.

The work is based on extensive experimentation by senior research engineer Slava Tykhonov at *DANS* across various projects, involving the construction of complex technical pipelines, characterized as a "back of things you can hardly unravel." The project aims to apply and illustrate this technical structure using a specific use case, making it understandable to a broader audience.

12.4 Talk Structure and System Architecture

Science dynamics and AI

In the evolution of the sciences we find growth and increasing differentiation.

Leaving us with the problem to review, evaluate and select.

Precondition to any creating of (new) knowledge (in individual brains or for larger parts of academia) is to find and understand.

Machines (latest AI) have fostered growth, can they also support the knowledge production process? This is a question of Information Retrieval.

Figure 12.3: Slide 02

The talk addresses the research question: Can an *AI* solution be constructed to facilitate interaction, or "chatting," with papers from a specific, selected collection? The introduction covers foundational concepts including information retrieval, the dynamics of human-machine interaction, and *Retrieval-augmented generation* (*RAG*) within the context of *generative AI.*

A specific use case involving papers from the *method-data-analysis* (*mda*) journal is presented. The workflow introduces a "local" or "tailored *AI* solution" architecture, comprising two main components known by the pet names *Ghostwriter* and *EverythingData*. *Ghostwriter* serves as the user interface, while *EverythingData* encompasses the entire backend processing pipeline. The presentation includes illustrations of both front end and back end operations, concluding with a summary and outlook on future directions.

12.5 *Ghostwriter*: New IR Interface and Query Models

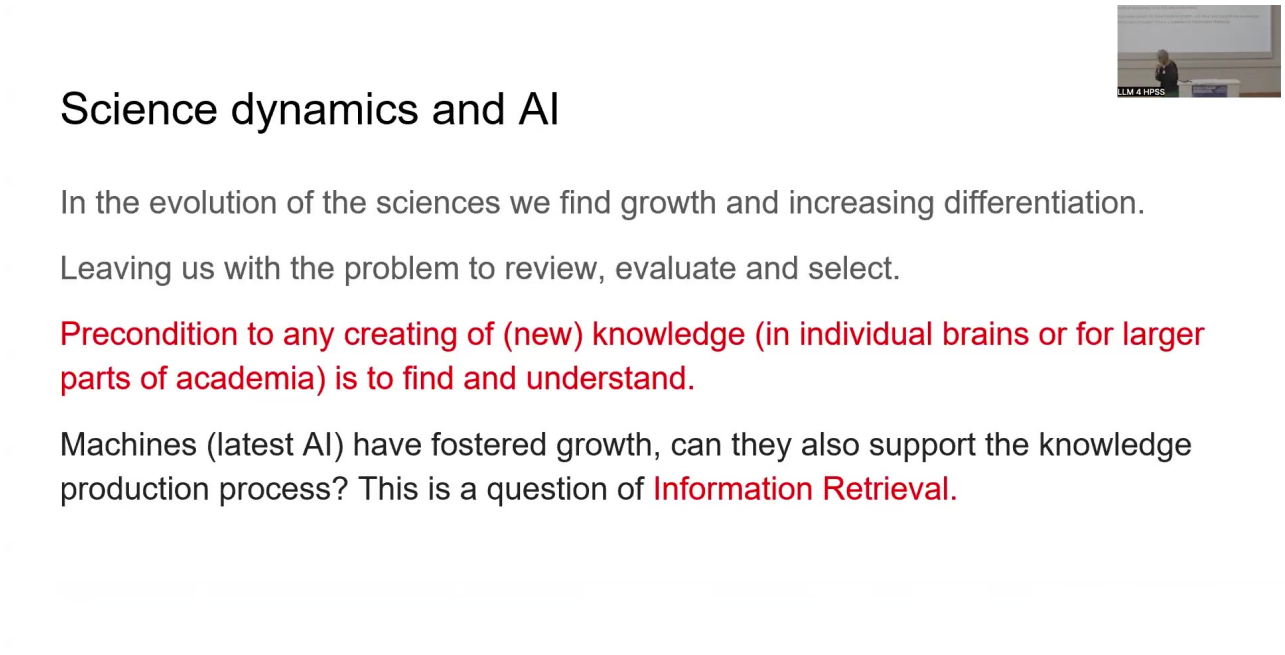

Figure 12.4: Slide 02

The *Ghostwriter* approach introduces a new interface for information retrieval. A primary challenge in this domain involves formulating the correct question, identifying the appropriate person or information source, and accurately interpreting the results. This is fundamentally linked to the classic *information retrieval* (*IR*) problem of finding the right query. The approach explores different models of query interaction, illustrated through comparisons.

Interacting with a database requires explicit knowledge of its schema and typical values to obtain results, representing the classic *IR* problem. A model involving querying connected structured data, such as databases or graphs, is likened to interacting with a *librarian*. The system suggests similar or improved queries based on schema connections and provides lists of potential results for different query variations. This is exemplified by features like *Google's schema.org* integration, which works well on the web but is less suited for local interactions.

Querying a *Large Language Model* is compared to interacting with a library or a round of *experts*. The *LLM* interprets the query as natural language input and provides suggestions for results, also expressed in natural language. The *Ghostwriter* approach combines a local *LLM* with a target data collection or space, embedding it within a network of additional data interpretation sources accessible via *APIs*. This is metaphorically described as chatting simultaneously with *experts* and *librarians*.

This combined approach creates a family of terms related to the query, identifies relevant structured information, and returns a list of results. When applied iteratively, this process assists users in reformulating their questions by enhancing their understanding of their actual query intent and the capabilities of the available data space. The metaphors of a "*librarian*" representing structured data, knowledge organization systems, and existing classifications, and an "*expert*" representing natural language, are central to describing these interaction models.

12.6 *Ghostwriter* and *EverythingData*: *RAG* Architecture

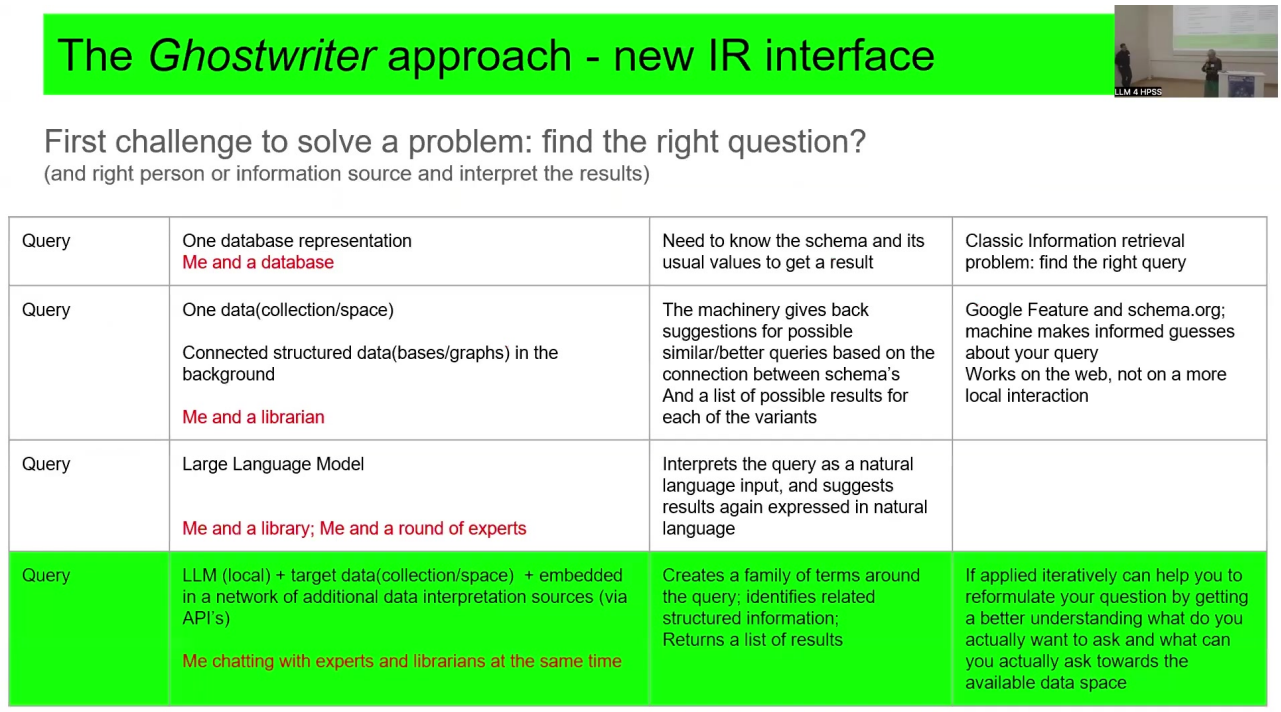

Figure 12.5: Slide 04

The *Ghostwriter* and *EverythingData* architecture is situated within the wider discourse of *Retrieval Augmented Generation* (*RAG*). The main ingredients of this system include a vector space and a graph. The vector space is constructed from the content of data files, with content encoded into embeddings that possess properties and attributes. These embeddings are computed using various machine learning algorithms and different *Large Language Models*.

The graph component represents a metadata layer that is integrated with various ontologies and controlled vocabularies, encompassing considerations for responsible *AI*. This graph is expressed using the *Croissant ML* standard. The vision behind this approach is to combine both the graph and vector components into a single model, a concept referred to as *GraphRAG*.

This is implemented locally as a form of *Distributed AI*, where the *LLM* serves as the interface between the human user and the *AI* system, simultaneously functioning as a reasoning engine. In implementation, the *LLM* is connected to a "*RAG library*," which is the graph component. It navigates through datasets and consumes the embeddings (vectors) as context to inform its responses.

Related concepts and resources mentioned include the *GenAI Knowledge Graph* and "*The GraphRAG Manifesto: Adding Knowledge to GenAI*," authored by *Philip Rathle*, *CTO* of *Neo4j*, with a link provided to the *Neo4j* blog post. The *Wikipedia* page for *Retrieval-augmented generation* is also referenced, along with a reference to *Arno Simons'* presentation on tool boxes.

12.7 *EverythingData* Backend and Vector Space

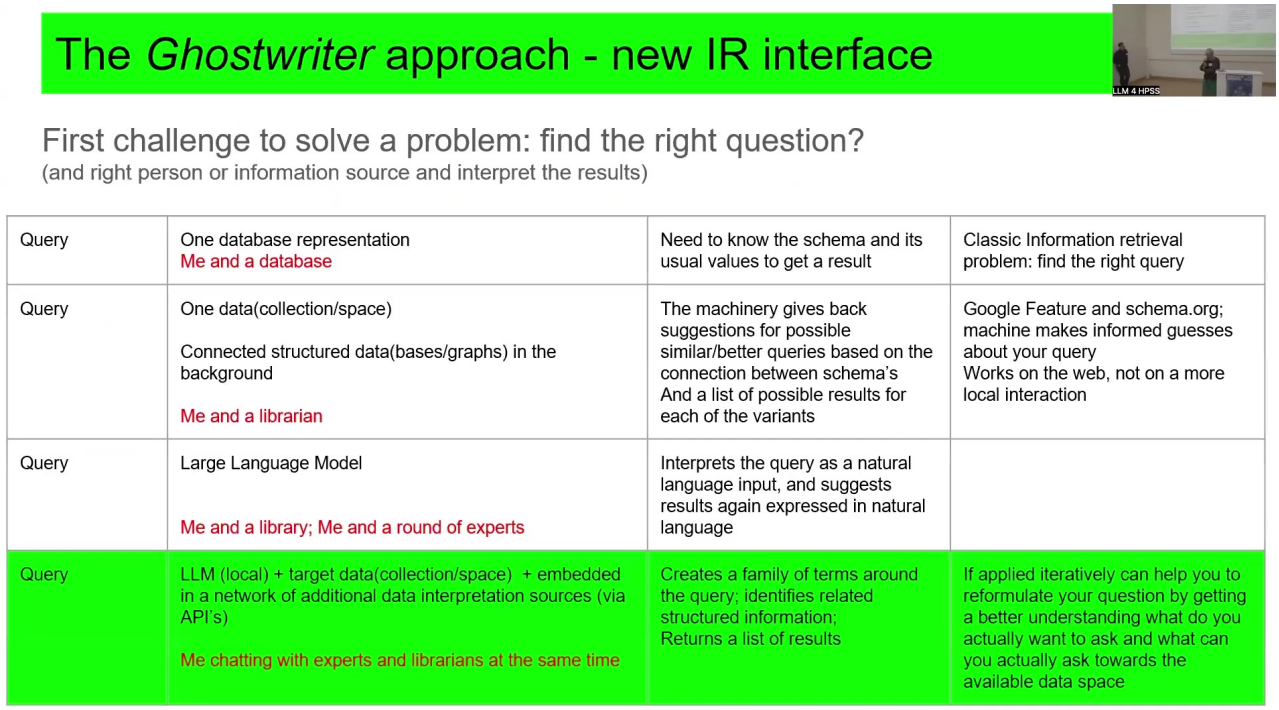

Figure 12.6: Slide 04

The system's input data consists of a collection of articles, specifically scraped from the *MDA* journal, although the system is designed to work with any collection of documents. This input is processed by the backend component, referred to as "*tamed EverythingData*." The backend executes various operations, including storing the information in a vector store utilizing *Quadrant*. Additional processing steps involve term extractions, constructing embeddings, and other related operations.

A crucial aspect of the backend is the integration of knowledge graphs, coupling the processed information to these graphs. This integration enhances the value of words, phrases, and embeddings by providing additional context and adding another layer of value to the existing context. The processed information is structured and fed into a vector space. The user interface interacts with this combined vector space and graph structure. Users formulate queries as natural language questions. The system responds by providing a list of relevant documents, consistent with standard information retrieval outputs, and a summary generated by the machinery based on the user's question.

12.8 *Ghostwriter* Functionality and Mechanisms

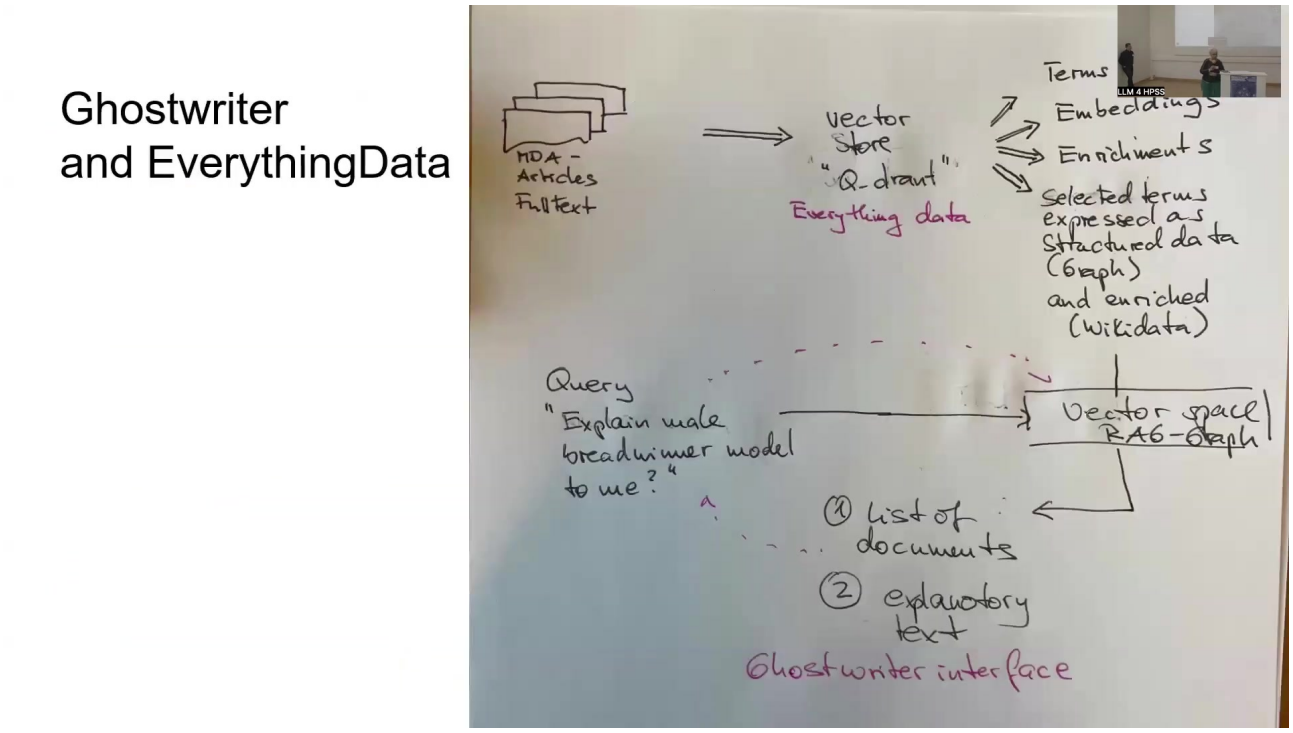

Figure 12.7: Slide 07

The *Ghostwriter* system is designed for chatting with papers, but its capabilities extend to interacting with any content from the web or even spreadsheets. When interacting with spreadsheets, the system can recognize specific values and provide responses without hallucinating, as it relies solely on the spreadsheet content as its source. The system utilizes a relatively simple *LLM* with 1 billion parameters, which is capable of answering complex questions by leveraging knowledge graphs.

By default, the system does not depend on knowledge pre-ingested into the *LLM*. Its primary goal is to provide answers based *only* on factual information present in the specific paper or papers that have been ingested. If the required information is not found within the ingested papers, the system explicitly states "I don't know." This strict reliance on the source material is the mechanism for avoiding hallucination, as the system has precise knowledge of where to locate information.

The underlying implementation involves splitting each paper into small blocks, with a separate identifier assigned to each block. *LLM* techniques are employed to intelligently connect and retrieve these blocks. The system applies weights and incorporates information from knowledge graphs to predict which specific pieces of text are most relevant to answer a given question. A user interaction feature includes an "Add paper" button, allowing users to contribute missing information; this added content will then be available for subsequent queries on the same topic.

The backend processing involves an entity extraction pipeline and annotations. Entities are linked to knowledge graphs, which is considered extremely important as it provides a ground truth mechanism for validating the accuracy of the *LLM*'s responses. Multilinguality support is a critical feature; the system can handle papers written in languages such as Chinese or German and respond to questions posed in English, aiming for reliable answers. The *LLM*'s final role is to synthesize the retrieved pieces of text to produce the results.

The fact extraction process involves splitting the user's question into smaller pieces and utilizing a knowledge organization system (*KOS*). This *KOS* is a repeatable process that can reveal new levels of related terms underneath the initial query term. A key step is linking everything to *Wikidata*. This process transforms free-text strings into identifiers that are linked to multilingual translations available in *Wikidata*. This linking provides access to all associated properties and enables the system to understand and respond to questions asked in various languages. The query construction process involves translating the user's question into potentially hundreds of languages, and all these translations are used as input to the *LLM*.

The knowledge graph linking provides a ground truth mechanism by decoupling knowledge from the questions and papers, storing this knowledge externally as a list of identifiers from *Wikidata*. This allows for a validation mechanism where different models, including those not yet trained, can be tested by asking the same questions and comparing the resulting lists of identifiers. Discrepancies in the identifier lists indicate that a particular model may not be suitable for the task. This approach is proposed as a method for creating benchmarks and supporting future generations of scientists. The project is collaborating with industry partners like *Google* and *Meta* to ensure the sustainability of this process, viewing the knowledge organization system as a potential future standard.

12.9 *Ghostwriter* Demonstration

A demonstration of the *Ghostwriter* interface is conducted within a browser environment. The first query example involves asking the system about "rational choice theory." The system processes this request by thinking and retrieving relevant pieces of information. The output consists of a summary compiled from different papers and includes references pointing directly to the original source papers, confirming that the results originate from the specified sources.

A second query example involves asking the system to "explain utility in Rational Choice Theory." The system responds by selecting different pieces of information from the ingested papers, presenting different results while still referencing the same source documents. The system provides an *API* that enables automatic mode operation, facilitating the construction of pipelines within an agentic architecture where the system can be prompted, results collected, and subsequent queries issued. This *API* can be used to analyze papers to identify new information or knowledge contributions.

The interface includes a feature allowing users to add a page or information if the system does not provide results for a query; this added information is then incorporated and will appear in responses to the same question in the future. A key demonstration of the system's capability is asking questions in English about a source paper that is written entirely in German, showcasing its multilinguality support.

12.10 Project Benefits and Philosophy

A significant benefit of the project's approach is the local availability of the system. This provides users with greater control compared to interacting with large, external systems, which can also be costly. The interaction with papers via the system is likened to chatting with an *invisible college*.

It is recommended that users approach this interaction with the same perspective as engaging with an *invisible college*, meaning the goal is not necessarily to find ultimate facts or definitive answers. Instead, the primary purpose of the system is to provoke and support the user's thinking process. The human user retains the role of understanding the question and identifying the appropriate research question. The system's function is to provide support for the user's own cognitive process. The recommended perspective is to view these technological possibilities as tools that enhance and support human thinking.

12.11 System Performance and Local Deployment

System performance has been improved by downscaling the *Large Language Models* used. The implementation transitioned from a complex *Llama* model with 70 billion parameters to a smaller model with only 1 billion parameters.

This current model is capable of running on a local computer. The ability to deploy and run *LLMs* locally on private or sensitive material is seen as a potential challenge to companies like *Nvidia* if this capability becomes widely known and adopted.

12.12 From Development to Production

The current status of the interface is described as a "playing ground," used primarily to gain a better understanding of the system's behavior and capabilities. However, similar underlying machinery is being applied in other, more serious projects intended for production environments. An example of such a production project is the *Odyssey* project in the Netherlands, which involves building a portal designed to bring together various data sources.

Projects like *Odyssey* necessitate considerations for long-term sustainability and the handling of diverse data sources, while still applying the same core principles developed in the *Ghostwriter*/*EverythingData* work. These aspects are actively discussed at a high level within research infrastructure discussions in the Netherlands.

12.13 Validation and Community Engagement

Future validation of the system is envisioned through its development as a community project under the *Linux Foundation*. The *Linux Foundation* has approached the project team with interest in publishing the work. The project is expected to be released as an open source project potentially within the current month.

The community is anticipated to play a crucial role in helping to validate and improve the system, reflecting the belief that significant progress is impossible without community involvement. Currently, the team is in an experimental phase regarding validation. The next steps involve engaging in scientific discourse and publishing scientific papers about the work, marking the beginning of the serious academic validation process.

12.14 Data Ingestion and Collections

Setting up a collection, such as a *Nodo* collection, is considered not hard. This assessment is based on observations of the system's capability to perform similar setup processes for information extracted from various other *APIs*. The system is designed to ingest data from any kind of source, including content from *GitHub*, manuals, guidelines, and papers.

An example collaboration involves building this system for *Harvard University*. The system deployed for *Harvard* currently contains approximately 300,000 documents, and *Harvard University* has commenced using it. The project team is receiving substantial feedback from users like *Harvard*. Based on this feedback, there is a strong belief that utilizing local models deployed on personal computers represents a preferable approach compared to being fully dependent on industry-provided solutions such as *ChatGPT*.

12.15 Project Goals and Collaboration

The project's primary goal is not centered on developing or selling software commercially. The preferred model is based on collaborations, typically triggered by individuals or groups who have concrete research questions that the system might help address.

The collaboration process involves seeking resources to conduct a try-out of the system for the specific use case, followed by handing over the system to the collaborating partners. These partners are then expected to tinker with, validate, and polish the system further. The team expresses anticipation for future collaborations.

12.16 Recency Bias Mitigation

A potential problem identified is the possibility of recency bias in the system's results. An example cited is querying a concept like "rational choice," which originated in the 1930s or 1940s, but potentially receiving results predominantly from the 2000s. This recency bias is acknowledged as true.

The proposed solution involves collecting facts and storing them within the knowledge graph. A key detail of the knowledge graph structure is the ability to store a fact along with a timestamp if one is available. This allows for separate processing based on these timestamps. For queries related to temporal aspects, the system can provide a list of all facts linked to specific dates, rather than a single, potentially biased answer, offering a way to mitigate recency bias.

12.17 Comparison to *Google Notebook ML*

When compared to *Google Notebook ML*, the system is assessed as being quite similar. This similarity is attributed to the reliance on the same underlying ideas and collaboration with the same development teams.

Chapter 13

RAG Systems in Philosophy and HPSS

The presentation details the application of Retrieval Augmented Generation (RAG) systems to philosophical research and teaching, specifically within the Humanities, Politics, and Social Sciences (HPSS) domain. The core objective is to address limitations of standard Large Language Models (LLMs) when applied to disciplines requiring high linguistic and semantic accuracy and deep engagement with specific textual corpora. Standard LLMs face problems with access to full texts (despite potenti…

13.1 Overview

The presentation details the application of Retrieval Augmented Generation (RAG) systems to philosophical research and teaching, specifically within the Humanities, Politics, and Social Sciences (HPSS) domain. The core objective is to address limitations of standard Large Language Models (LLMs) when applied to disciplines requiring high linguistic and semantic accuracy and deep engagement with specific textual corpora.

Standard LLMs face problems with access to full texts (despite potential inclusion in training data), limited context windows, and attribution of information. RAG systems are proposed as a solution by providing explicit access to domain-specific data sources, augmenting prompts with retrieved relevant text chunks, and enabling source citation.

The presentation outlines typical philosophical research questions that require detailed textual analysis. It describes the standard LLM query process and contrasts it with the RAG workflow.

The RAG workflow involves a retrieval query to data sources (such as vector databases or APIs), retrieval of relevant text chunks (typically via semantic search, potentially hybrid or classic search), and augmentation of the LLM prompt with these chunks before generation. This process directly addresses the problems of access, limited context window, and attribution.

Potential applications in philosophy include didactic use (chatting with philosophical corpora for instructive questioning) and research use (looking up facts in handbooks, exploring unexamined corpora, finding passages for close reading, and directly finding detailed answers to research questions).

An example RAG system is presented using the *Stanford Encyclopedia of Philosophy* (SEP) as the data source. The system was developed through a qualitative study involving theoretically grounded trial and error to optimize performance for philosophical queries.

Key aspects of this study included model choices (generative LLM, embedding model), tuning hyperparameters (number of documents to retrieve (top-k), max input/output token length, chunk size and overlap), and addressing methodological challenges like retrieval semantic mismatch through reranking. Evaluation of results, particularly for complex, unstructured philosophical answers, is identified as crucial and requiring domain expertise.

The implemented SEP RAG system features a frontend with configuration options for generative model (e.g., *gpt-4o-mini*), prompt token limits, number of texts to retrieve, and persona. It includes a comparative setup displaying answers from the LLM alone versus the RAG system for qualitative evaluation and benchmarking.

The output also lists retrieved texts, their source files, section headings, distance metrics, token lengths, and inclusion status based on prompt limits.

A specific hyperparameter tuning example, chunk size, is discussed. Options explored included fixed word counts, paragraphs, and sections.

The study found that chunking into main sections yielded the best results for the SEP corpus, despite section lengths often exceeding the embedding model's cutoff. This outcome is attributed to the highly systematic structure of the SEP, where section beginnings effectively summarize content. This highlights that effective chunking is corpus- and question-dependent.

Results indicate that RAG systems offer advantages in integrating verbatim corpora and domain knowledge, leading to more detailed answers and a dramatic reduction in hallucinations compared to standard LLMs. They also enable the citation of relevant documents supporting the generated answer. Overall, the RAG setup is identified as being very well suited for assisting in a wide range of scientific tasks.

However, several cautionary points are raised. RAG systems fundamentally require tweaking; appropriate settings for hyperparameters and methods are highly dependent on the specific corpus and the nature of the questions being asked. Evaluation is crucial and necessitates domain experts to define representative questions and expected answers. A key challenge is the decrease in answer quality when no relevant documents are found, requiring prompt adjustment.

Counterintuitively, RAG systems often provide worse results for widely discussed overview questions, such as inquiries about the central arguments against scientific realism, compared to more specific factual queries. A hypothesis for this phenomenon is that RAGs tend to focus on the local information present in the retrieved chunks. The prompt directs the model to answer based on this local information, which can inadvertently distract from a broader perspective. This suggests a need for prompt adjustments for different question types.

Ultimately, there is a need for more flexible systems, potentially agentic RAG systems, that can discern between different kinds of questions and adapt their strategy accordingly.

The discussion further explores challenges related to philosophical contentiousness and how RAG systems might represent diverse viewpoints, the potential for using LLMs as judges for evaluation, and the specific ways domain expertise influences RAG design, particularly in chunking and defining relevant arguments.

13.2 Introduction to RAG Systems

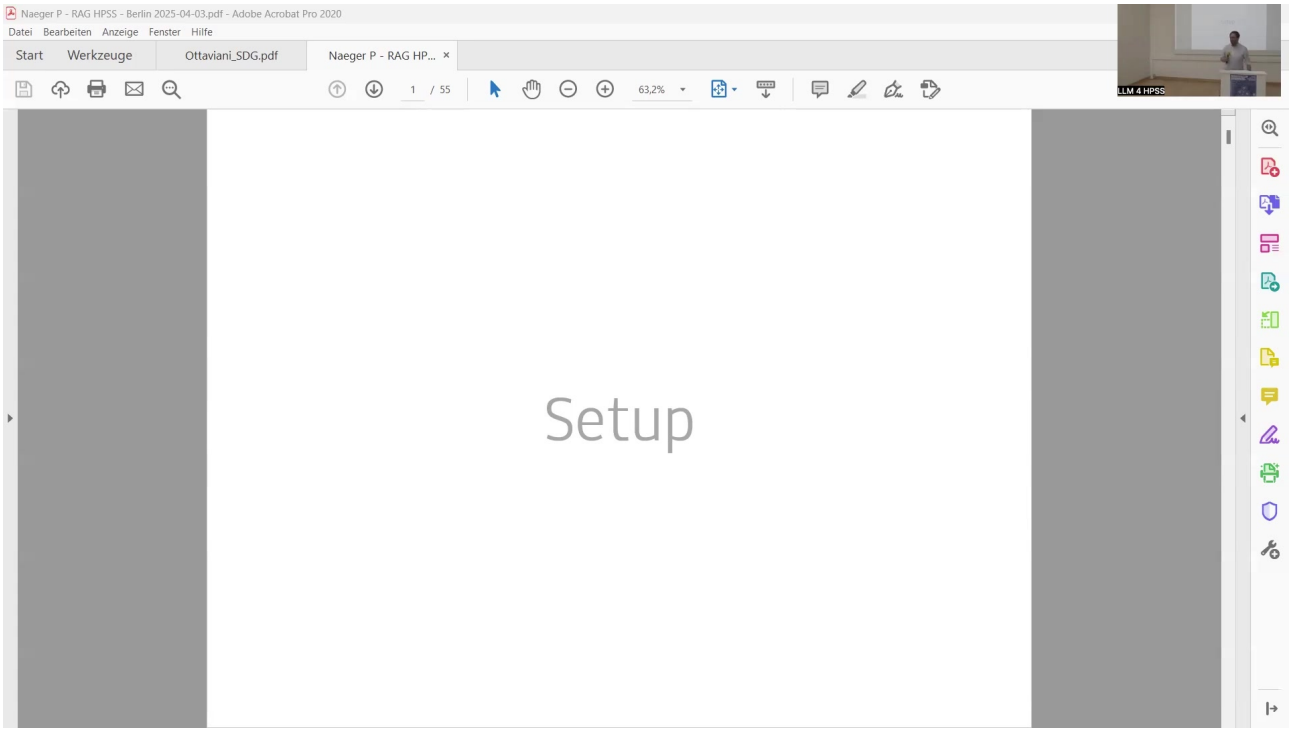

Figure 13.1: Slide 01

The presentation is delivered from the perspective of a philosopher of science who has engaged with the technical details of Large Language Model (LLM) systems. The focus is on applying these systems, specifically Retrieval Augmented Generation (RAG), within the domain of philosophy and the broader Humanities, Politics, and Social Sciences (HPSS).

A core requirement in philosophical research is a high degree of linguistic and semantic accuracy, demanding deep engagement with specific textual corpora.

Typical research questions in philosophy include inquiries such as "What is Aristotle's theory of matter in the *Physics*?" or "Does Einstein's idea of locality develop from his earlier to his later works?", referencing specific periods and texts like his relativity works and the 1948 paper on *Quantenmechanik und Wirklichkeit*. While standard LLMs like *ChatGPT* can provide decent, differentiated answers to such questions at a general level, they present several problems for rigorous philosophical research.

Standard LLMs lack direct access to the full text of source corpora. Although texts may have been included in their training data, the models cannot explicitly retrieve or quote them accurately. This often leads to hallucination when specific quotes are requested.

While online search features can sometimes provide access, they are subject to limitations such as copyright restrictions, as encountered with papers like the EPR paper. The training mechanism of LLMs is designed to prevent verbatim learning, focusing instead on generalizable statistical rules of text production, which is counter to the philosophical need for direct engagement with original text sources and their fine-grained formulations.

Furthermore, standard LLMs have a limited context window, such as the 128,000 tokens available in *ChatGPT-4*, which is insufficient for processing large philosophical corpora. Finally, there is a significant attribution problem, as standard

LLMs do not provide sources or citations for the claims made in their answers. RAG systems are presented as a suitable setup to address these specific problems.

13.3 RAG System Architecture and Problem Solving

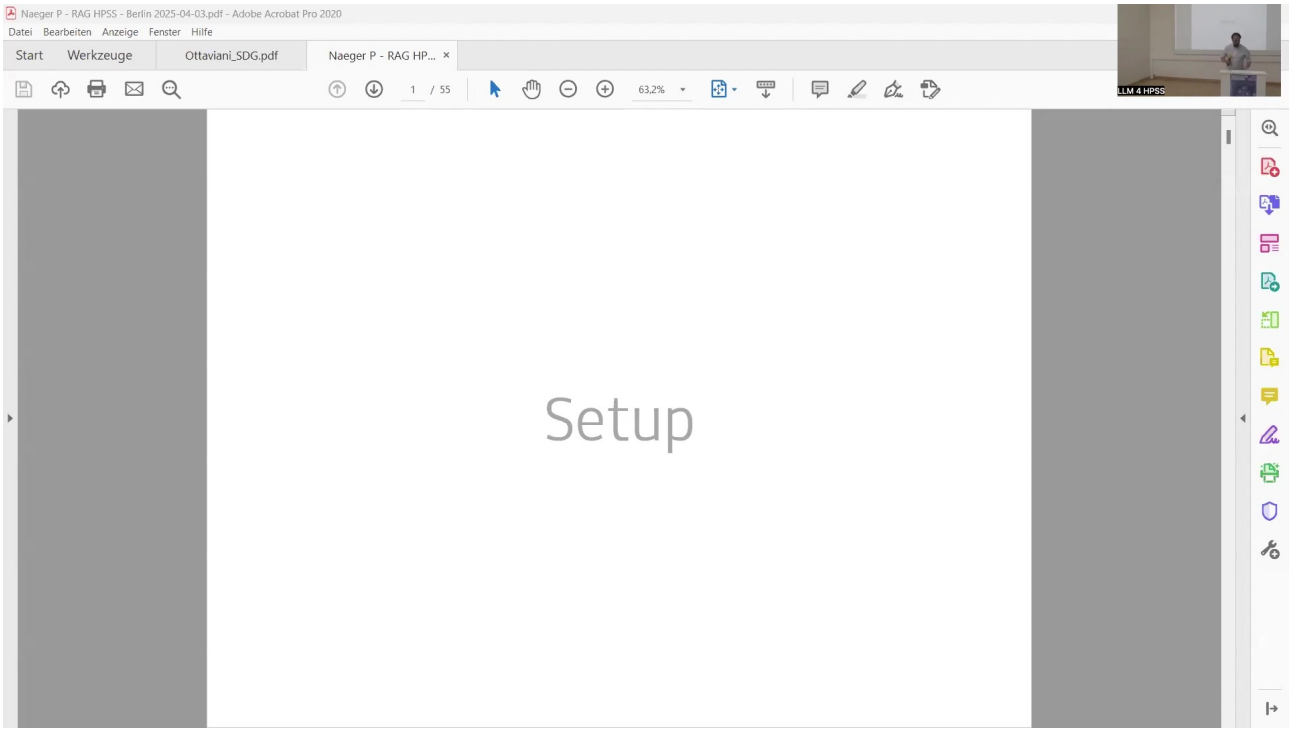

Figure 13.2: Slide 02

The RAG system architecture is described as a setup capable of solving the identified problems with standard LLMs in philosophical contexts. The process begins with a user initiating a query through an application (APP).

The APP then sends a retrieval query to designated data sources. These data sources contain the appropriate corpus, such as Aristotle's corpus or Einstein's corpus, and can be implemented using technologies like vector databases or APIs. The retrieval mechanism typically employs semantic search to find relevant text chunks, although hybrid or classic search methods are also viable options.

The data sources return the retrieved chunks of text to the APP. The APP then augments the original LLM query by incorporating these retrieved chunks into the prompt.

This augmented query is sent to the LLM, which performs text generation based on the provided information. The LLM returns the generated answer to the APP, which finally delivers the answer to the user.

This RAG setup directly addresses the problems faced by standard LLMs. It solves the problem of access by providing explicit access to specific texts within the defined corpus, ensuring that the LLM works with the actual source material.

It mitigates the problem of the limited context window by providing only the most relevant text chunks to the LLM, effectively managing the input size within the model's capacity. Furthermore, the RAG system solves the attribution problem by enabling the citation of sources for the provided text chunks, allowing users to verify the basis of the generated claims.

13.4 Applications in Philosophy

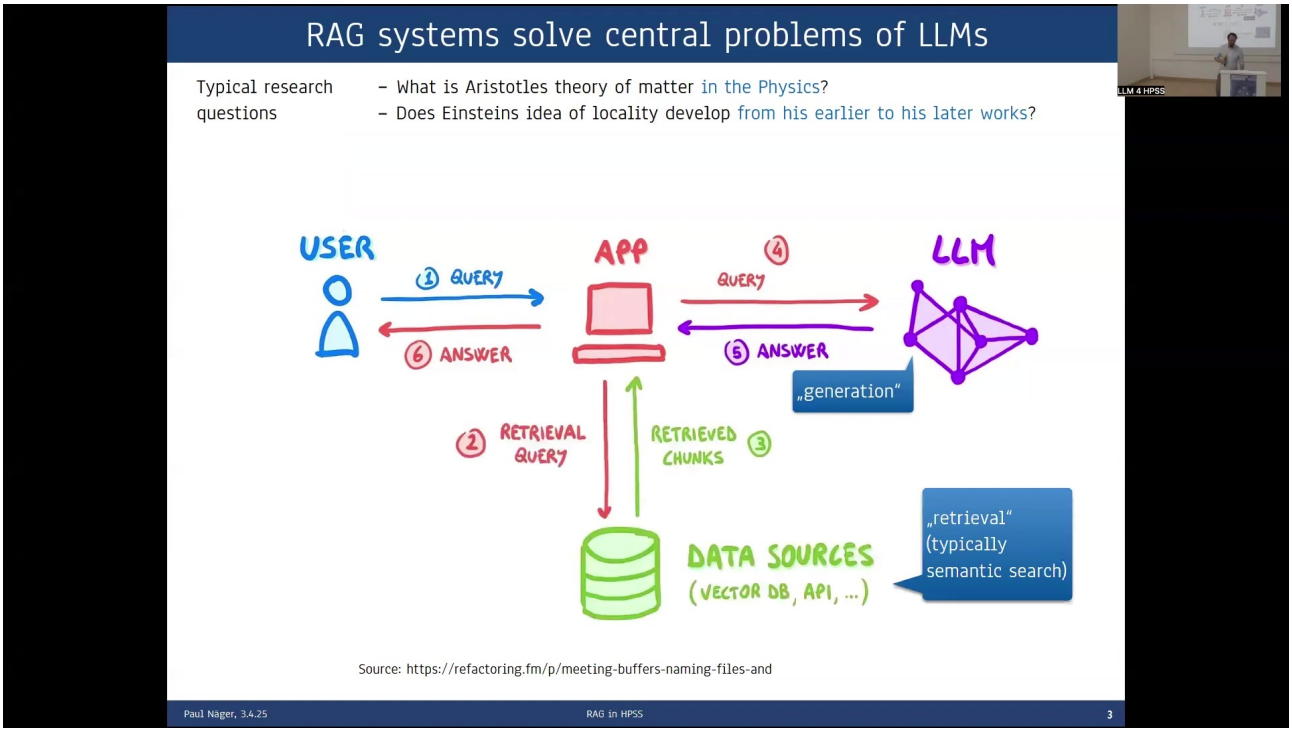

Figure 13.3: Slide 07

The general idea behind applying RAG systems in philosophy is to enable users to chat with philosophical corpora. This interaction style is similar to using *ChatGPT* but provides significantly more detailed domain knowledge and is grounded in a verbatim text basis from the specified corpus.

Didactic applications are a key area. RAG systems are useful for students approaching complex philosophical texts, such as Locke's *Essay Concerning Human Understanding*.

They allow for repeated questioning, enabling students to start with general ideas and progressively delve deeper into specific details, like Locke's epistemology or his theory of matter. This interactive process provides an instructive method for students to gain a deeper understanding of the texts.

Research applications are also emphasized. RAG systems are expected to be important for looking up facts in handbooks, serving functions previously performed by manually consulting books for orientation, remarks, or footnote information.

This addresses the unreliability of factual information obtained solely from standard LLMs and necessitates the development of high-quality RAG systems for reliable factual retrieval. Other research uses include exploring corpora that have not been extensively studied, efficiently finding specific passages for close reading, and potentially, in the future, directly finding detailed answers to complex research questions.

13.5 Example: Stanford Encyclopedia of Philosophy RAG System

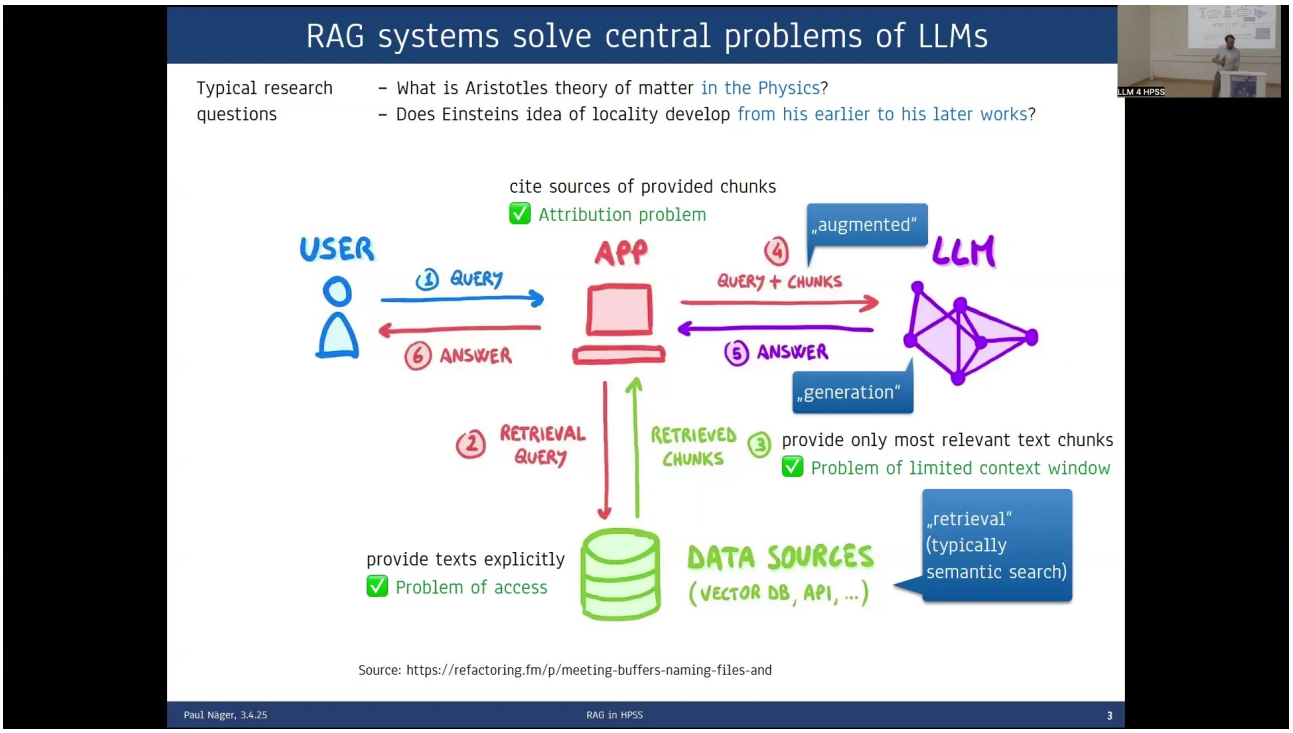

Figure 13.4: Slide 10

An example RAG system was developed using the *Stanford Encyclopedia of Philosophy* (SEP) as the data source. The content of the SEP was prepared by scraping it into markdown format. The initial aim of this project was to create a useful tool for the philosophical community.

The development process evolved into a qualitative study employing theoretically grounded trial and error. An observation during the initial setup was that a standard RAG configuration, based on typical textbook descriptions involving only retrieval and generation components, produced poor answers.

These answers were found to be worse than those obtained by querying a standard LLM like *ChatGPT* directly without retrieval augmentation.

This led to an iterative improvement process involving significant tweaking and optimization. This included tweaking the choice of models, specifically selecting the appropriate generative LLM and embedding model.

Hyperparameters were also tuned, including the number of documents to retrieve (top-k), the maximum input and output token lengths, and the chunk size and overlap used for text processing. Furthermore, methodological complexities were added, such as implementing reranking mechanisms to address issues of retrieval semantic mismatch, where initially retrieved chunks might not be the most relevant.

The method for evaluating these improvements was theoretically grounded trial and error, assessing by which measures the answers improved. Sound evaluation standards were identified as crucial.

A key challenge in evaluating philosophical RAG systems is that the desired answers are typically free, unstructured text rather than simple atomic facts (like asking for Wittgenstein's last place of living, which yields a city name). Evaluating complex propositions for their factual accuracy is not straightforward and requires significant domain expertise.

13.6 SEP RAG System Implementation

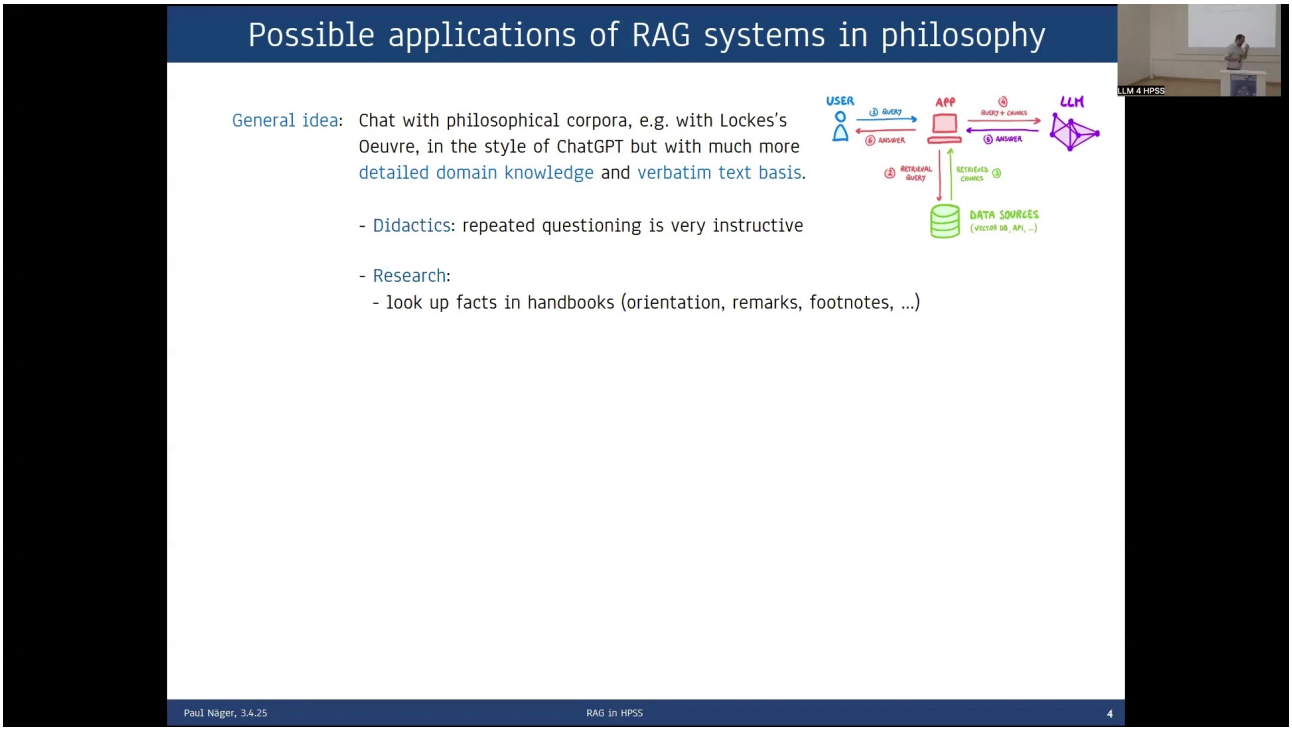

Figure 13.5: Slide 13

The implemented SEP RAG system consists of a frontend and a backend, which is written in Python code. The frontend provides a web interface with several configurable options.

Users can select the Generative Model to be used, such as *gpt-4o-mini*. There are settings for the Prompt Token Limit, including the maximum limit supported by the chosen model (e.g., 128000) and a configurable limit for the current session (e.g., 15000). The number of texts to retrieve, corresponding to the top-k value, is also configurable (e.g., 15). A Persona text area allows defining the desired behavior for the LLM, for instance, instructing it to act as "an expert philosopher" who answers "meticulously and precisely."

The frontend includes an input field for the Philosophical Question, where users enter their query (e.g., "What is priority monism?"). A "Generate answer" button triggers the RAG process.

The output section is designed for qualitative evaluation, featuring a comparative setup. It displays the "Answer with LLM alone" as a benchmark on one side and the "Answer with RAG" on the other. The RAG answer includes source citation indicators, such as "[Text 0]", linking parts of the answer to the retrieved texts. A "Benchmark" button is available for comparative evaluation.

Below the answers, a "Retrieved Texts Overview" table lists the texts found during the retrieval phase. This table includes columns for file names, section headings, distance metrics, token lengths (length_token for the chunk, total_token for the full section/file), and a flag indicating whether the text was included in the final prompt based on token limits. This table shows the article names and specific section headings that were retrieved and indicates which ones were utilized in the prompt and which were excluded due to prompt length constraints. The backend functionality is implemented in Python code, described as comprising a few thousand lines.

13.7 Hyperparameter Tuning: Chunk Size

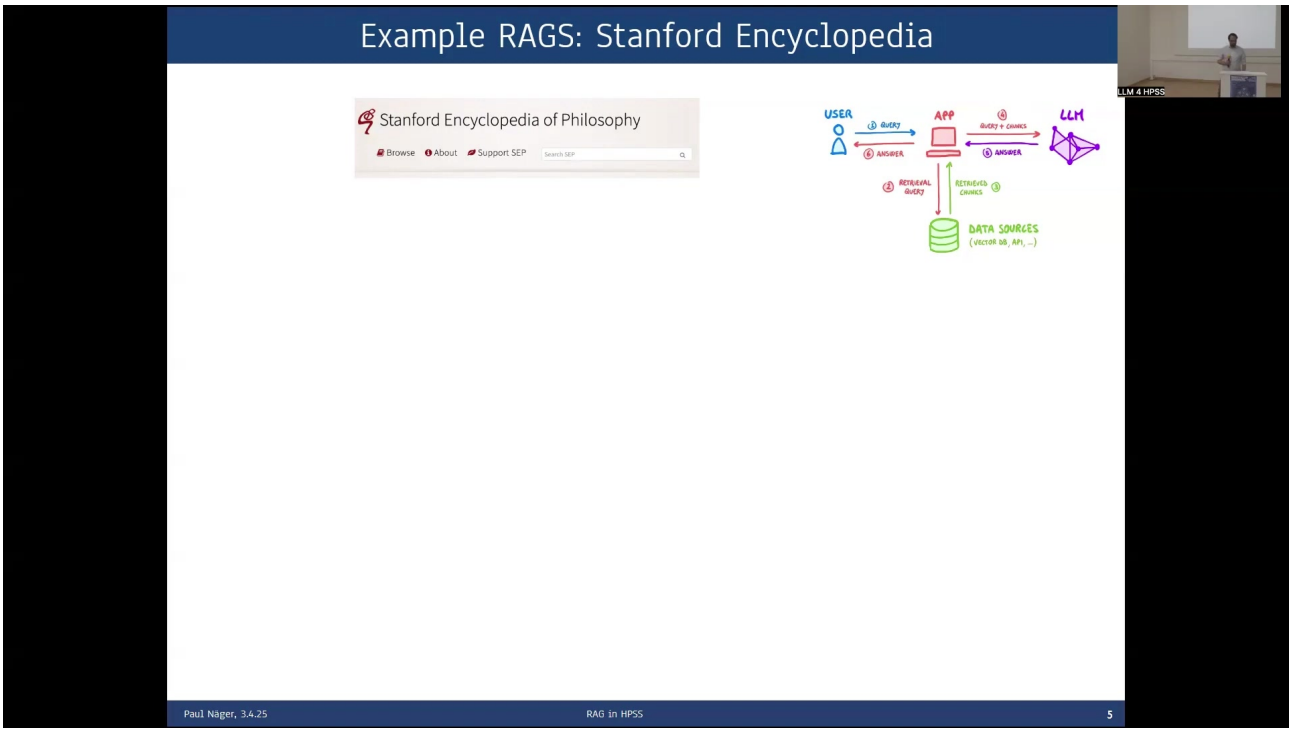

Figure 13.6: Slide 15

Chunk size is identified as a key hyperparameter requiring optimization in RAG system development. Several options exist for defining text chunks.

One approach is using a fixed number of words, such as 500 tokens or words. This provides a clean criterion but disregards the inherent structure of the document, such as headings or sections. Alternative methods involve chunking by semantic units like paragraphs or sections, potentially considering different hierarchical levels of sections.

For the specific case of the SEP corpus, the optimization process revealed a surprising result: the best performance was achieved by chunking the text into its main sections, including their headings.

This finding was unexpected because the average length of SEP sections (approximately 3000 words) significantly exceeded the typical cutoff length of the embedding model used (512 words). Despite this discrepancy, chunking by main sections yielded superior results compared to smaller, fixed-size chunks or paragraphs.

A hypothesis for this surprising outcome is that the SEP documents are highly systematically ordered. The initial parts of each section often effectively summarize the main theme and key ideas. These crucial introductory segments likely fall within the effective context window of the embedding model, even if the entire section is much longer.

This suggests that the structure and organization of the corpus play a significant role in determining the optimal chunking strategy, and this approach might not be as effective for less structured or heterogeneous texts.

Future work is planned to explore the use of embedding models with longer context windows, such as *Cohere Embed 3*, to see if they further improve performance with larger chunks.

The key lesson derived from this optimization process is that effective chunking is not a one-size-fits-all solution; it highly depends on the specific characteristics of the corpus being used and the nature of the questions being posed.

13.8 Results, Discussion, and Challenges

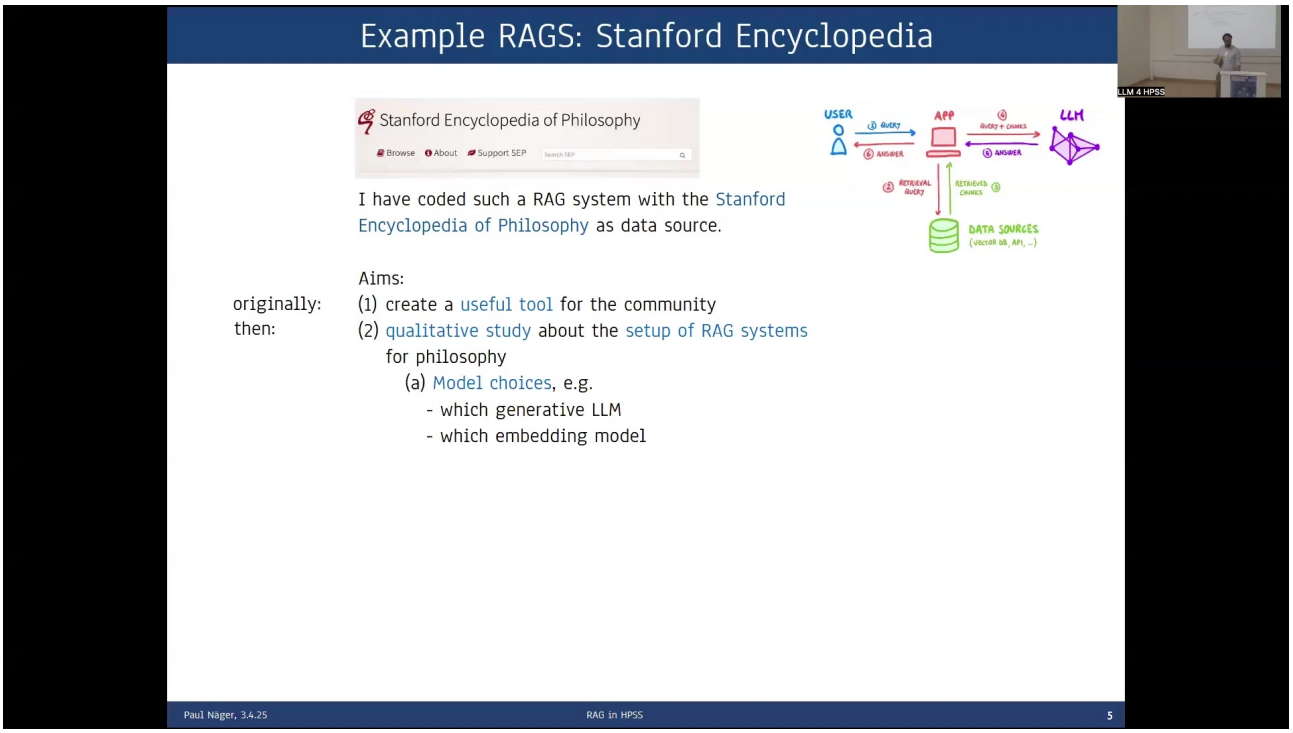

Figure 13.7: Slide 18

The results and discussion highlight several advantages of RAG systems. They effectively integrate verbatim corpora and domain-specific or special knowledge, leading to more detailed answers and a dramatic reduction in hallucinations compared to standard LLMs.

RAG systems also enable the citation of relevant documents supporting the generated answer. Overall, the RAG setup is identified as being very well suited for assisting in a wide range of scientific tasks.

However, several cautionary points are raised. RAG systems fundamentally require tweaking; appropriate settings for hyperparameters and methods are highly dependent on the specific corpus and the nature of the questions being asked.

The evaluation of RAG systems is crucial and necessitates a representative set of questions along with expected answers. This process essentially requires domain experts, such as philosophers in this context, for both evaluation and initial setup, as the optimal configuration is specific to the domain, the type of corpus, and the kind of questions. A challenge remains regarding how to effectively evaluate RAG performance when dealing with unexplored corpora.

Several challenges are also identified. A decrease in answer quality occurs if no relevant documents are found during retrieval, indicating a need to adjust the prompt in such cases.

Counterintuitively, RAG systems often provide worse results for widely discussed overview questions, such as inquiries about the central arguments against scientific realism, compared to more specific factual queries. A hypothesis for this phenomenon is that RAGs tend to focus on the local information present in the retrieved chunks. The prompt directs the model to answer based on this local information, which can inadvertently distract from a broader perspective. This suggests a need for prompt adjustments for different question types.

Ultimately, there is a need for more flexible systems, potentially agentic RAG systems, that can discern between different

kinds of questions and adapt their strategy accordingly.

Chapter 14

Plural pursuit across scales

The presentation investigates the structure of quantum gravity research using computational methods to address questions in the philosophy of science, specifically the concept of “plural pursuit”. The core problem is formulating a quantum theory of gravity, which has led to multiple attempted solutions and a situation characterized as plural pursuit. Plural pursuit is defined as distinct yet concurrent instances of normal science dedicated to a common problem-solving goal, where each inst…

14.1 Overview

The presentation investigates the structure of quantum gravity research using computational methods to address questions in the philosophy of science, specifically the concept of “plural pursuit”. The core problem is formulating a quantum theory of gravity, which has led to multiple attempted solutions and a situation characterized as plural pursuit. Plural pursuit is defined as distinct yet concurrent instances of normal science dedicated to a common problem-solving goal, where each instance is articulated by a community tied to an intellectual disciplinary matrix. The research empirically tests whether quantum gravity research is an instance of plural pursuit, meaning independent communities pursuing different paradigms in parallel.

The methodology involves a bottom-up reconstruction of the research landscape and a top-down comparison with physicists’ intuitions. The bottom-up approach utilizes a dataset of 228,748 abstracts and titles from theoretical physics literature listed on Inspire HEP.

This involves a two-step clustering pipeline: linguistic analysis and social network analysis. Linguistic analysis spatializes documents into an embedding space, performs unsupervised clustering to identify 611 fine-grained topics, and assigns specialties to authors based on their most common topic. Social network analysis constructs a co-authorship graph of approximately 30,000 physicists and applies community detection to recover 819 communities.

A key challenge is that computational notions of topics and communities are scale-dependent. To address this, hierarchical clustering is applied: Ward agglomerative clustering for topics and hierarchical stochastic block modelling for communities, yielding multi-level partitions. An adaptive topic coarse-graining strategy is introduced, based on the Minimum Description Length (MDL) criterion, to select an appropriate scale by merging topics as long as it retains useful information for understanding the social structure. This process reduces the initial 611 topics to 50 coarse-grained topics.

The bottom-up results show that the relationship between communities and topics is complex, exhibiting nested structures and lacking a clear one-to-one mapping at fine scales. The coarse-grained topics, derived using the MDL criterion, reveal

that some small-scale linguistic topics are preserved due to their relevance to social structure, while others are merged. The analysis of the correlation matrix between these 50 topics and communities across different hierarchical levels indicates that some topics (e.g., string theory) correspond to communities at higher hierarchical levels, while others (e.g., loop quantum gravity) correspond to communities at lower, more fine-grained levels. Observations suggest instances where communities are tied to multiple topics or nested within larger structures, indicating a departure from a simple plural pursuit configuration.

The top-down approach surveys founding members of the International Society for Quantum Gravity to elicit their intuitive list of structuring approaches (e.g., string theory, supergravity, causal sets, loop quantum gravity, holography). An *SVM* classifier is trained on text embeddings (*all-MiniLM-L6-v2*) using hand-coded labels based on this list to predict which papers belong to which approach.

Comparing the top-down (supervised) approaches to the bottom-up (coarse-grained) topics shows agreement for approaches considered well-defined and conceptually autonomous. However, disagreement exists for phenomenological or less conceptually integrated frameworks. A notable finding is the convergence of the bottom-up analysis regarding string theory, supergravity, and holography. While historically and conceptually distinct, the bottom-up analysis lumps them together in the coarse-grained topic structure, aligning with the intuition of some physicists that the communities working on these areas have significant overlap and are not meaningfully separable at a certain scale.

The conclusions state that socio-epistemic systems operate at multiple scales, and notions of communities and disciplinary matrices are scale-dependent. Identifying plural pursuit configurations requires matching these structures across scales. The bottom-up reconstruction in quantum gravity research can confirm or re-assess physicists' intuitions. The work demonstrates that computational methods can revisit and challenge philosophical insights, acting as a continuation of philosophy by other means.

14.2 Introduction

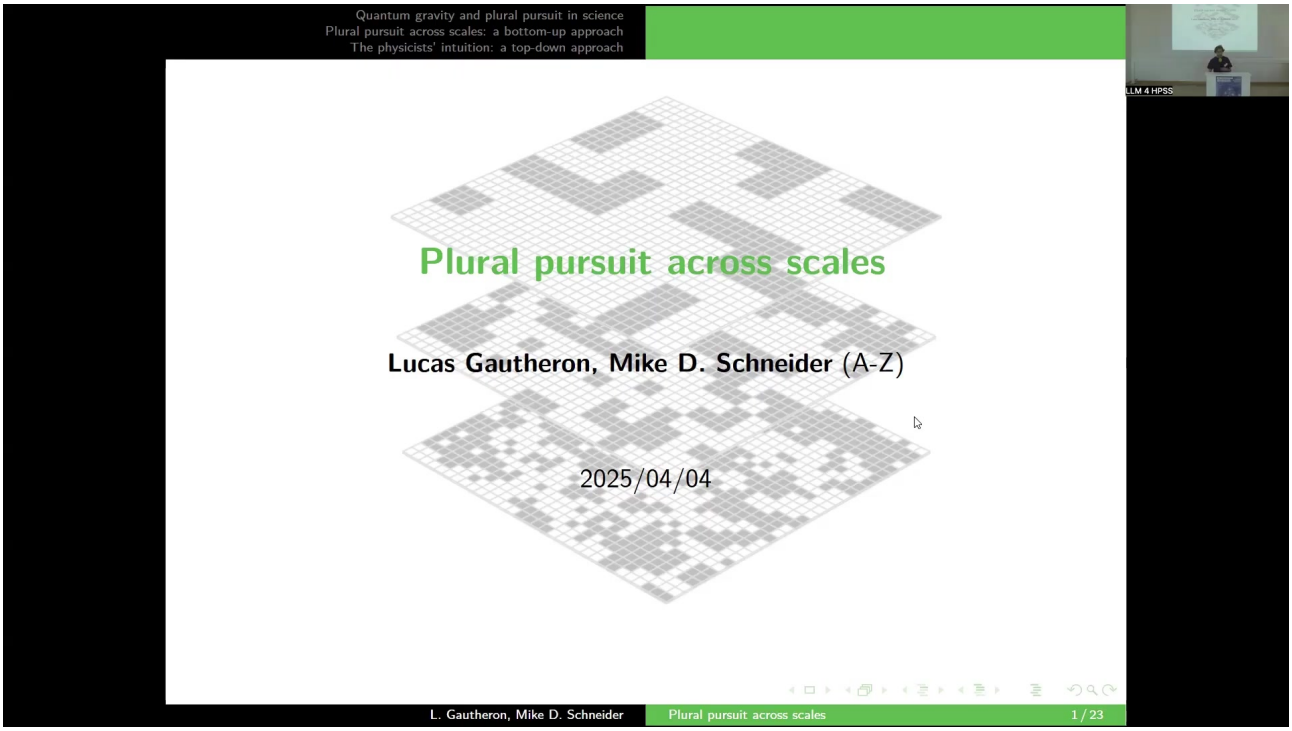

Figure 14.1: Slide 01

The research represents a joint effort with Mike Schneider from the University of Missouri. The primary objective is to address general questions within the philosophy of science by employing a combination of computational methods and social network analysis techniques. The specific case study chosen for this investigation is quantum gravity.

A long-standing and significant problem in fundamental physics is the formulation of a quantum theory of gravity. This problem involves reconciling the established knowledge of physical phenomena at small scales, described by quantum mechanics, with the understanding of phenomena at very large scales, described by general relativity.

Numerous attempted solutions exist to address this challenge. These attempted solutions include prominent approaches such as String theory, Supergravity, Loop quantum gravity, spin foams, Causal set theory, and Asymptotic safety, among others. The existence of multiple concurrent approaches attempting to solve the same fundamental problem leads to the introduction of a conceptual framework termed “plural pursuit”.

14.3 Plural Pursuit: Definition and Empirical Question

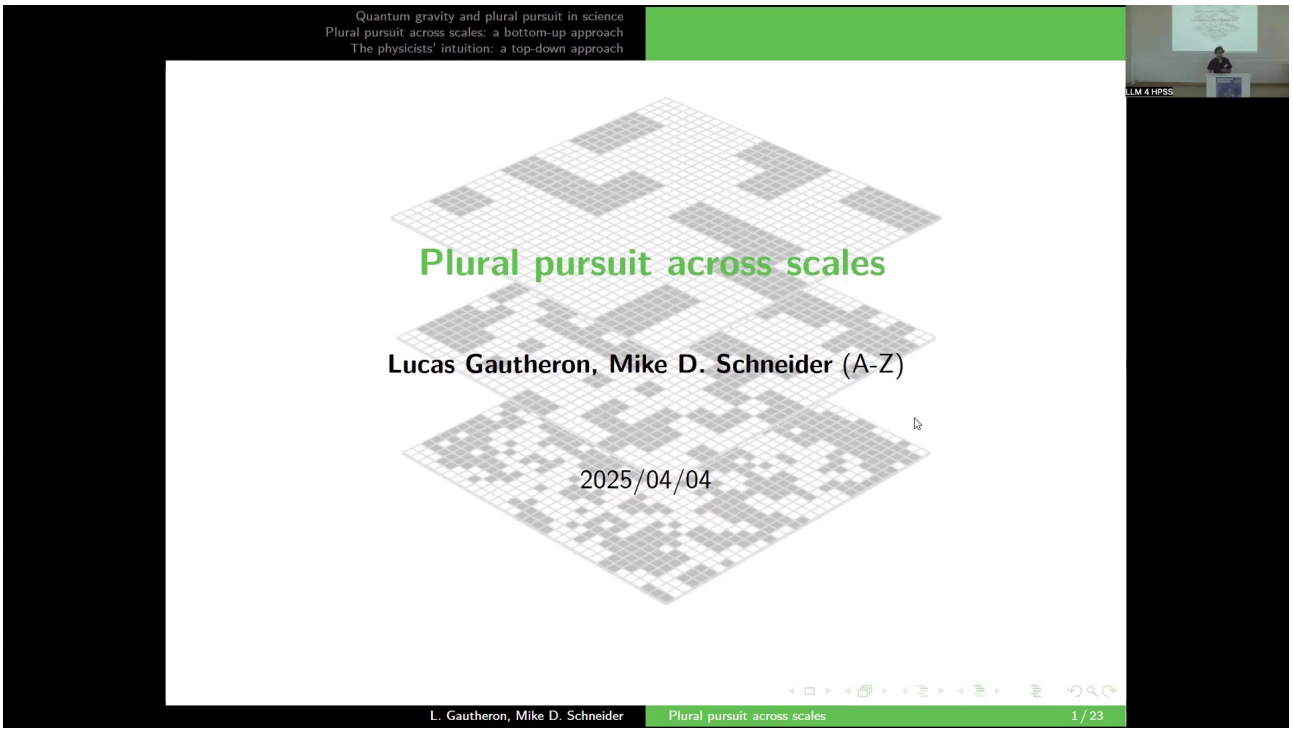

Figure 14.2: Slide 01

Plural pursuit is defined as a situation characterized by distinct yet concurrent instances of normal science. These instances are dedicated to achieving a common problem-solving goal. In the context of the quantum gravity case study, this common goal is the reconciliation of quantum mechanics and gravitation.

Each instance of normal science within this framework is articulated by a specific community that is tied to an intellectual disciplinary matrix. This concept aligns with established ideas in the philosophy of science, such as Kuhn’s paradigms, Laudan’s research traditions, and Lakatos’ research programmes, which describe the structured intellectual and social frameworks guiding scientific research.

Based on this definition, the research poses an empirical question: Is quantum gravity research an instance of plural

pursuit? This question is interpreted as investigating whether the field is composed of independent communities that are pursuing different paradigms in parallel.

14.4 Bottom-Up Methodology: Data and Pipeline

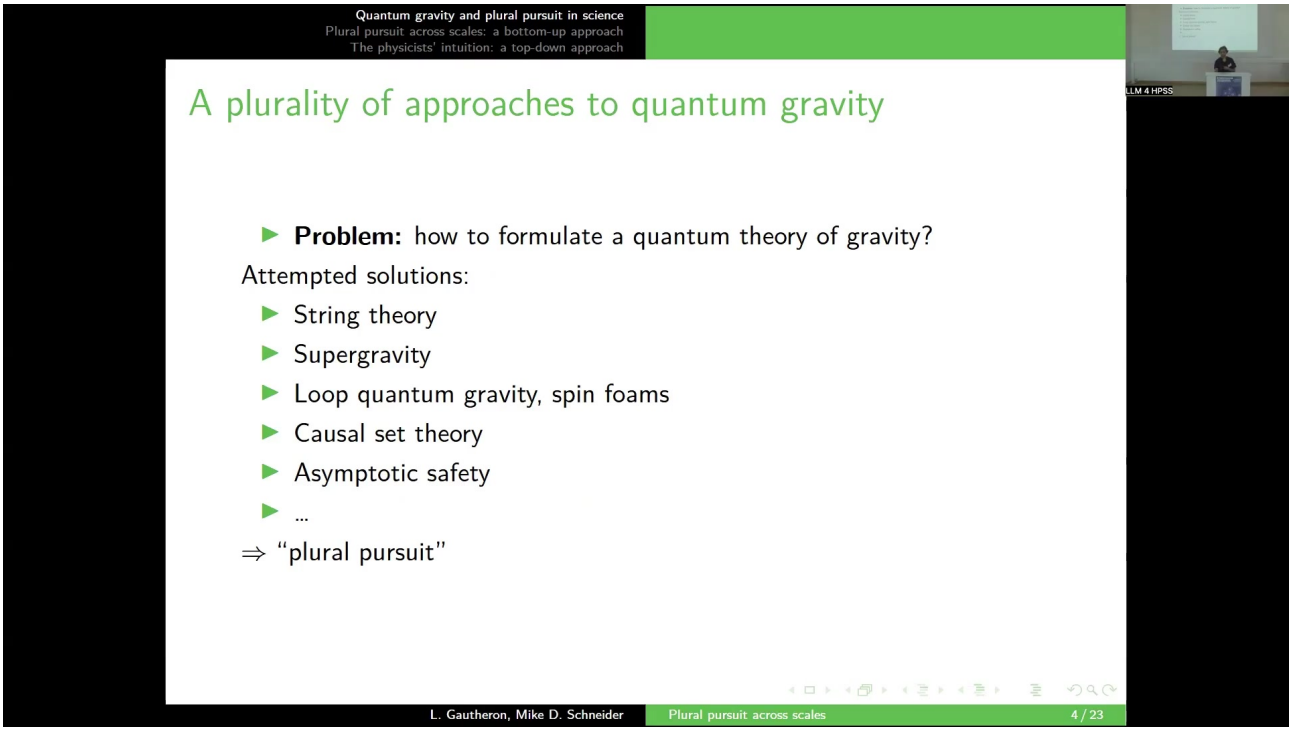

Figure 14.3: Slide 03

To address the empirical question, the research proposes performing a bottom-up reconstruction of the research landscape within quantum gravity. This reconstruction aims to capture both the linguistic and intellectual structure of the field, as well as its social structure.

The data source for this reconstruction is the Inspire HEP database, from which a dataset comprising 228,748 abstracts and titles of theoretical physics literature was gathered. The analysis proceeds through a two-step clustering pipeline: Linguistic analysis and Social network analysis.

The Linguistic analysis component (L) involves several steps. Step L.1 spatializes the documents into an embedding space. Step L.2 performs unsupervised clustering on this embedding space, resulting in an initial partition of the literature into K = 611 topics.

This clustering is performed at a very fine-grained level, which is considered necessary to identify niche approaches within quantum gravity that may involve a relatively small number of papers, potentially as few as 100. Step L.3 involves specialty assignment, where each scientist or physicist is assigned a specialty corresponding to the topic that is most common across their publications. This process yields a partition of authors based on the linguistic and intellectual structure derived from the literature.

The Social network analysis component (S) begins with Step S.1, constructing a co-authorship graph. In this graph, nodes represent individual physicists, and edges represent co-authorship relationships between them. The network includes

approximately 30,000 physicists.

Step S.2 applies a community detection method to this co-authorship graph, recovering an initial partition of the network into C = 819 communities. This provides a partition of authors that reflects the social structure of the field as captured by collaborative relationships.

14.5 Plural Pursuit: Mapping and Challenges

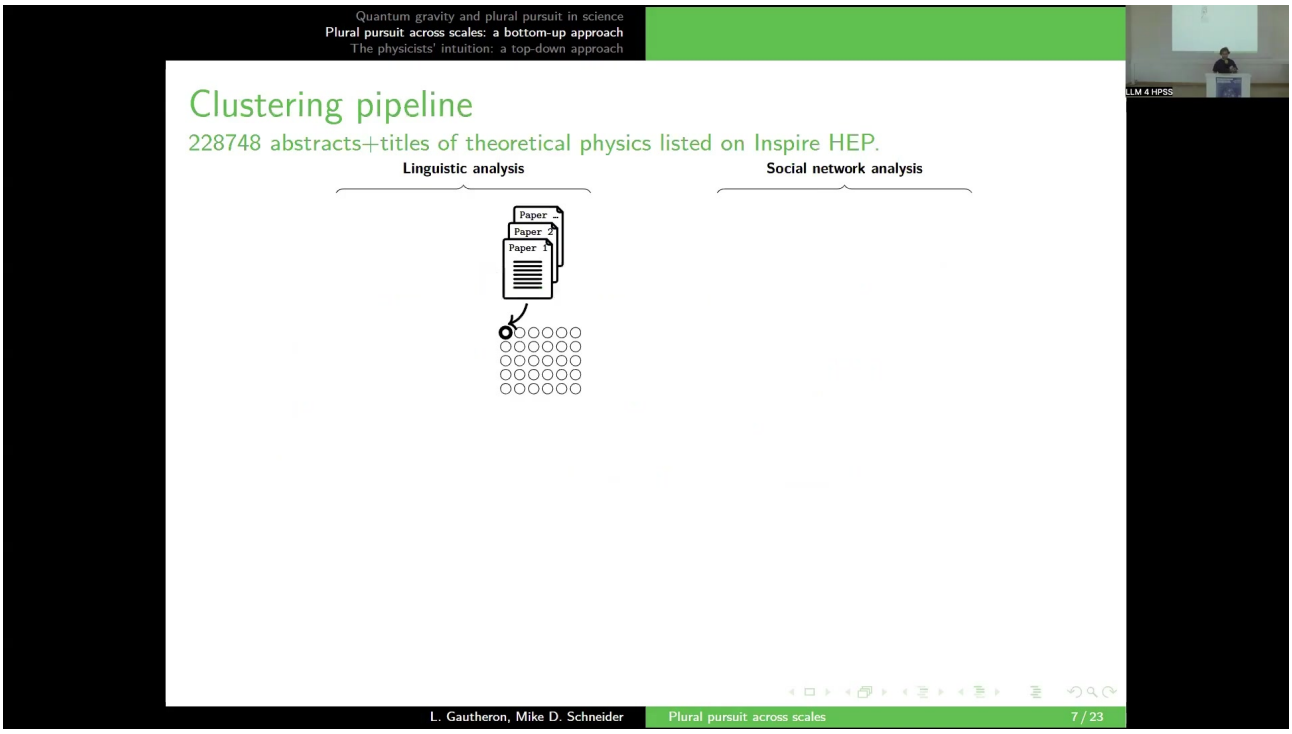

Figure 14.4: Slide 05

Within the framework of the computational constructs derived from the bottom-up analysis (communities and topics), plural pursuit is conceptualized as a one-to-one mapping between communities and topics. Ideally, this configuration would be represented by a block-diagonal correlation matrix where communities are listed on one axis and topics on the other. Such a matrix structure would imply that each community is entirely dedicated to a single topic, signifying a clear division of labor within the field.

However, applying this concept directly to the very fine-grained partitions (611 topics, 819 communities) results in a correlation matrix that is complex and difficult to interpret, exhibiting a high degree of structure without clear block-diagonal patterns. This complexity arises from inherent issues with the fine-grained partitions.

The first issue is the arbitrary nature of the level of fine-graining for topics. For example, a broad research program like string theory, which is intuitively understood as a coherent entity, might be fragmented and scattered across numerous fine-grained topics in the initial clustering. The second issue is that large research programs may be pursued by multiple communities in parallel. This occurs because social communities are shaped not only by intellectual similarity but also by various micro-social processes, such as geographical proximity or local collaborations.

These issues fundamentally stem from the fact that the computational notions of both topic and community are scale-

dependent. Furthermore, this technical challenge reflects a deeper conceptual problem: research programs themselves are often hierarchically nested. For instance, string theory can be conceptually divided into families and subfamilies like Superstring Theory, which further branches into Type II, Heterotic, Bosonic String Theory, Type I, Type IIA, Type IIB, Heterotic SO(32), and Heterotic $E_8 \times E_8$. Therefore, to accurately identify instances of plural pursuit, it is necessary to address this issue of scale dependence and the inherent ambiguity it introduces.

14.6 Hierarchical Reconstruction

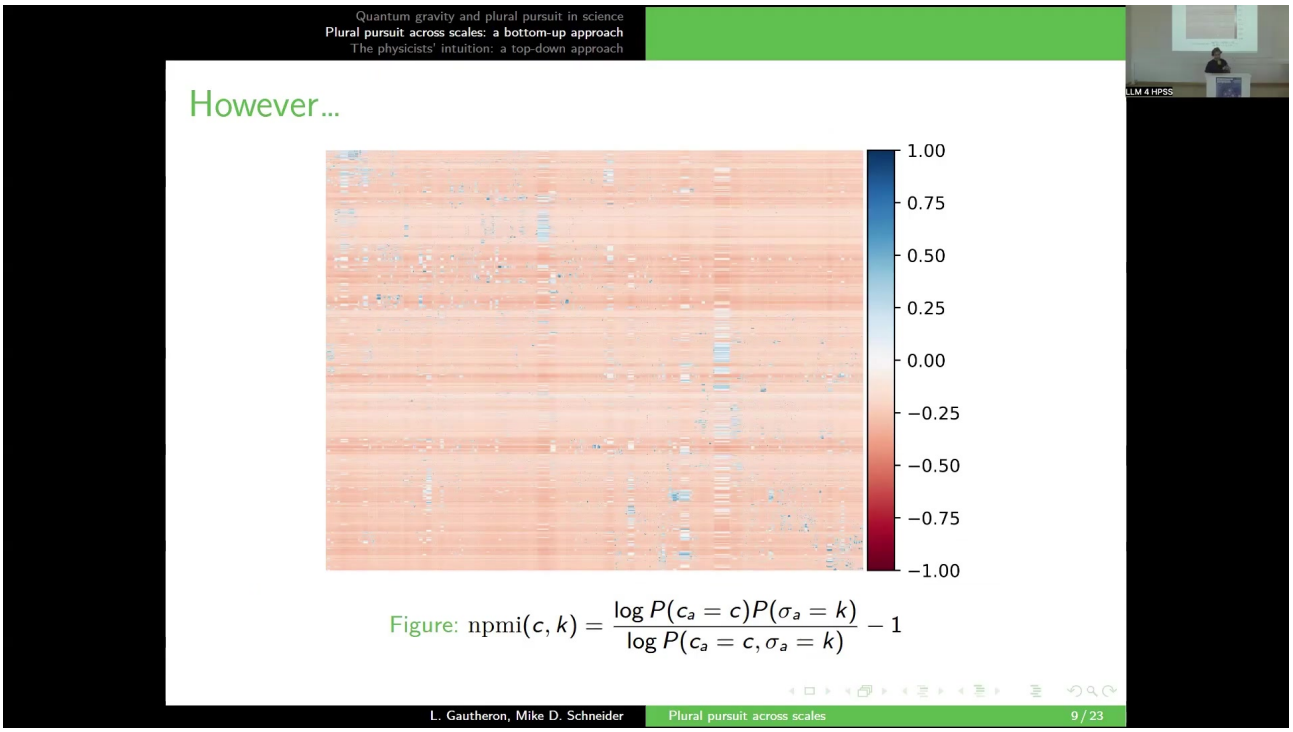

Figure 14.5: Slide 08

To address the scale dependence inherent in the initial fine-grained partitions, the research proposes building a hierarchical reconstruction of the quantum gravity research landscape. This involves applying hierarchical clustering techniques to both the topic and community structures.

For the topics, a hierarchical clustering approach is used, specifically the Ward agglomerative clustering algorithm. This method starts with the initial 611 fine-grained topics and iteratively merges them one by one based on an objective function. This process builds a dendrogram representing the hierarchical relationships between topics at different levels of granularity.

For the community structure, a hierarchical clustering is also constructed. This is achieved using a hierarchical stochastic block model, which is capable of learning a multi-level partition of the network directly. This model identifies coarser and coarser communities at increasing levels of the hierarchy. The model used is referenced as Peixoto 2014.

The application of these hierarchical methods results in hierarchical structures for both topics and communities. These structures induce a notion of scale, which allows researchers to observe and analyze the socio-epistemic system of quantum gravity research at various levels of coarse-graining, moving from very fine distinctions to broader groupings.

14.7 Adaptive Topic Coarse-Graining: MDL Criterion

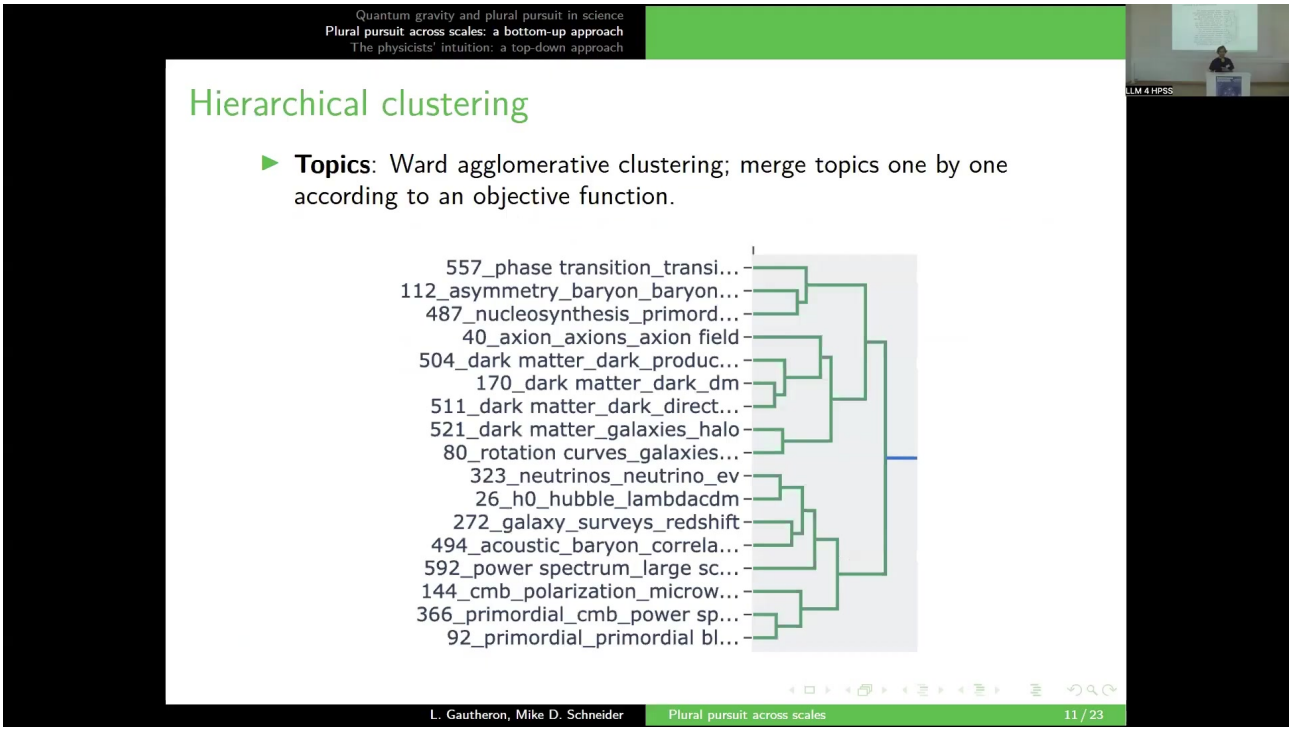

Figure 14.6: Slide 09

A challenge arises because the selection of a specific scale at which to observe the community and topic structures remains arbitrary at this stage. The choice of scale significantly impacts the resulting correlation matrix between communities and topics, leading to potentially different interpretations of the research landscape.

To address this, the research proposes an adaptive topic coarse-graining strategy. The core idea is to merge the initial K=611 fine-grained topics. This merging process continues only as long as it does not remove information deemed useful for understanding the social structure of the field.

This strategy is based on an information theoretic criterion: the Minimum Description Length (MDL) criterion. The MDL criterion seeks to find the partition $\square$ that minimizes the quantity given by the formula $\texttt{arg min}_{\square}\ [-\log P(G|\square) - \log P(\square)]$. This formula balances two factors: the first term, $-\log P(G|\square)$, represents how well the linguistic partition $\square$ explains the social structure, as captured by the graph G; the second term, $-\log P(\square)$, represents the complexity of the partition $\square$ itself. The objective is to find a partition that is complex enough to explain the social structure effectively but not overly complex or fine-grained.

In practice, this involves navigating the hierarchical topic dendrogram. The process "zooms in" or maintains finer distinctions as long as doing so improves the MDL criterion. It stops coarse-graining when further complexity (finer partitions) no longer yields sufficient information gain about the social structure. Applying this strategy reduces the initial 611 topics to a set of 50 coarse-grained topics.

An observation regarding these resulting topics is that certain small-scale linguistic topics from the initial fine-grained partition are preserved. This indicates that these specific nuances, captured at a fine level, are important for understanding the social structure. Conversely, many other topics are lumped together into much larger groupings. This outcome justifies the initial step of starting with a very fine-grained classification, as some of these small topics are indeed relevant

for explaining the social structure.

14.8 Bottom-Up Results: Topics and Communities

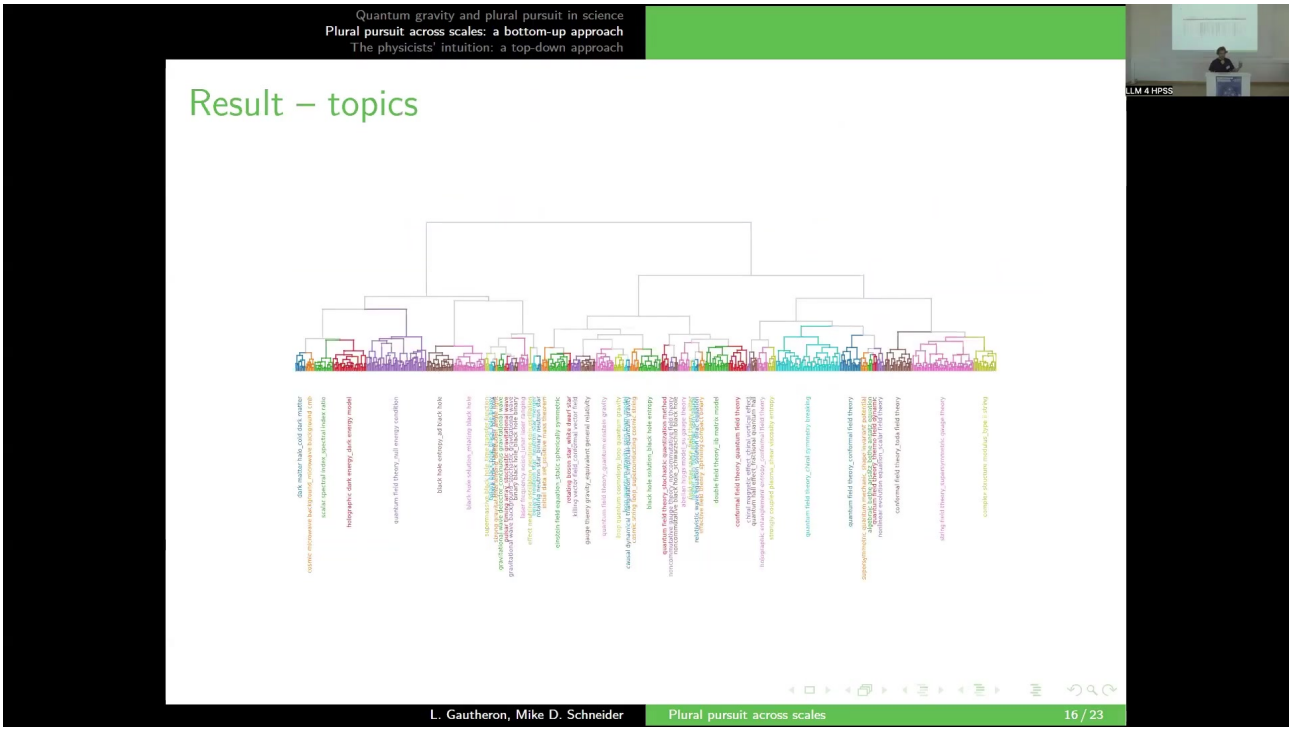

Figure 14.7: Slide 13

The adaptive coarse-graining strategy results in a set of 50 topics. These topics are assigned labels by retrieving representative engrams, providing some interpretability. The subsequent analysis focuses specifically on those topics identified as relevant to quantum gravity.

The core analysis involves confronting these 50 coarse-grained topics with the community structures derived from the hierarchical stochastic block model. This is visualized and analyzed using a correlation matrix, where the columns represent the 50 resulting topics and the rows represent communities identified at different levels of the community hierarchy. For each topic, the analysis attempts to identify the community that best corresponds to or explains that topic across the various hierarchical levels of the community structure.

Observations from this analysis reveal several patterns. Some topics, such as a large topic identified in purple, do not appear to be tied to a specific community, suggesting they represent concepts or areas broadly relevant across the entire field rather than being the exclusive domain of a particular group. However, other topics show a strong correspondence with specific community structures at certain hierarchical levels.

For instance, the topic identified as string theory corresponds well to a community structure found at the third level of the community hierarchy. In contrast, other research programs in quantum gravity, such as loop quantum gravity, appear to correspond to communities found at much lower, more fine-grained levels of the hierarchy.

Furthermore, the analysis provides evidence that challenges a simple model of plural pursuit characterized by a clear one-to-one mapping and division of labor. Nested structures are observed, such as a smaller community tied to the

topic of holography that is itself part of a larger community associated with string theory. This indicates that different scales are entangled and that a clear separation of communities pursuing distinct intellectual domains is not consistently present across the landscape.

14.9 Top-Down Approach: Survey and Classification

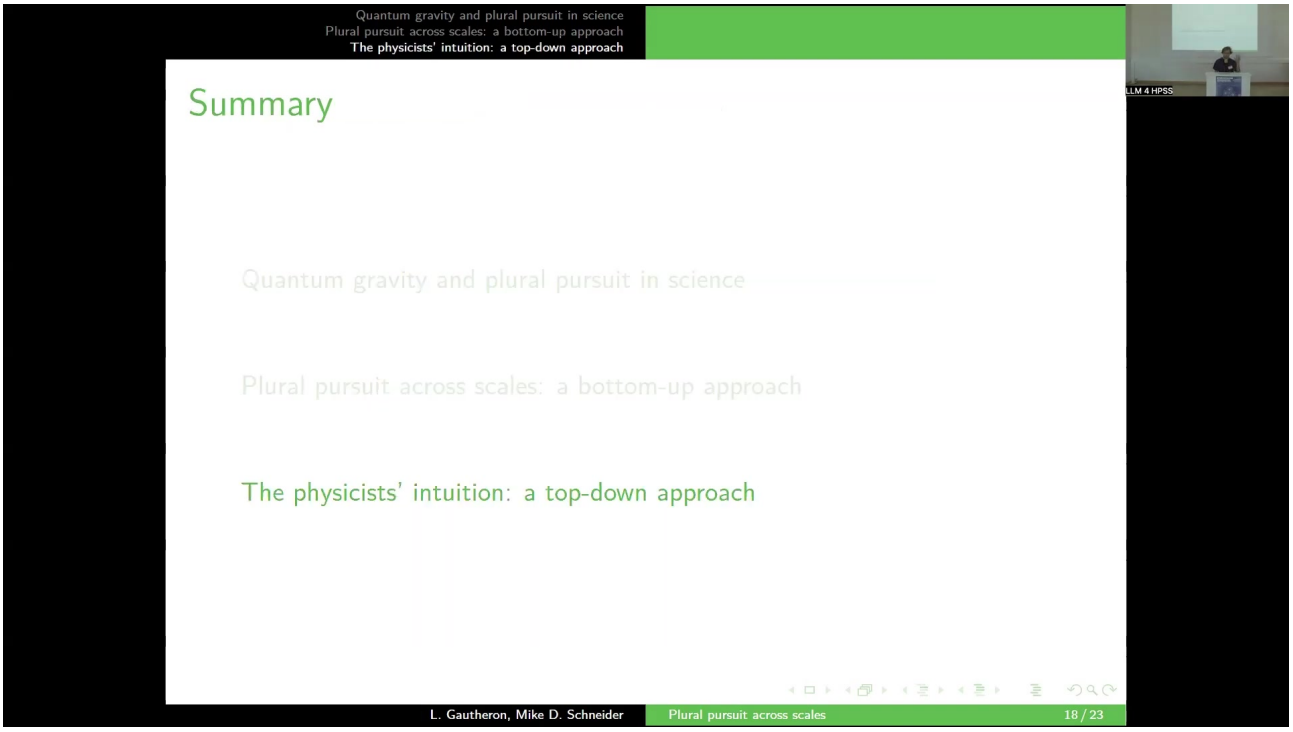

Figure 14.8: Slide 15

The second part of the research involves confronting the bottom-up reconstruction of the quantum gravity research landscape with the intuitions held by physicists themselves regarding how their field is structured.

This top-down perspective was gathered through a survey administered to the founding members of the International Society for Quantum Gravity. Participants were asked to provide a list of approaches to quantum gravity that they perceived as structuring the overall research landscape. The feedback from the survey, while not entirely unanimous, resulted in a comprehensive list of intuitive approaches, including asymptotic safety, causal sets, dynamical triangulations, group field theory, lqg (loop quantum gravity), spin foams, noncommutative geometry, swampland, modified dispersion relation, dsr, quantum modified bh, shape dynamics, tensor models, string theory, supergravity, and holography.

For a more detailed comparison, the research specifically focused on the last three approaches: string theory, supergravity, and holography. This particular focus was motivated by the observation that some physicists surveyed expressed disagreement about whether these should be considered separate approaches, despite their distinct historical origins and conceptual differences.

To facilitate the comparison between this intuitive, top-down view and the bottom-up analysis, a classifier was trained to predict which papers belong to which of these intuitive approaches. The classifier used was a *Support Vector Machine* (*SVM*). It was trained using text embeddings derived from the titles and abstracts of papers, specifically utilizing the

all-MiniLM-L6-v2 model for generating these embeddings. The training data consisted of a set of papers with hand-coded labels indicating their corresponding intuitive approach.

14.10 Top-Down vs. Bottom-Up Comparison

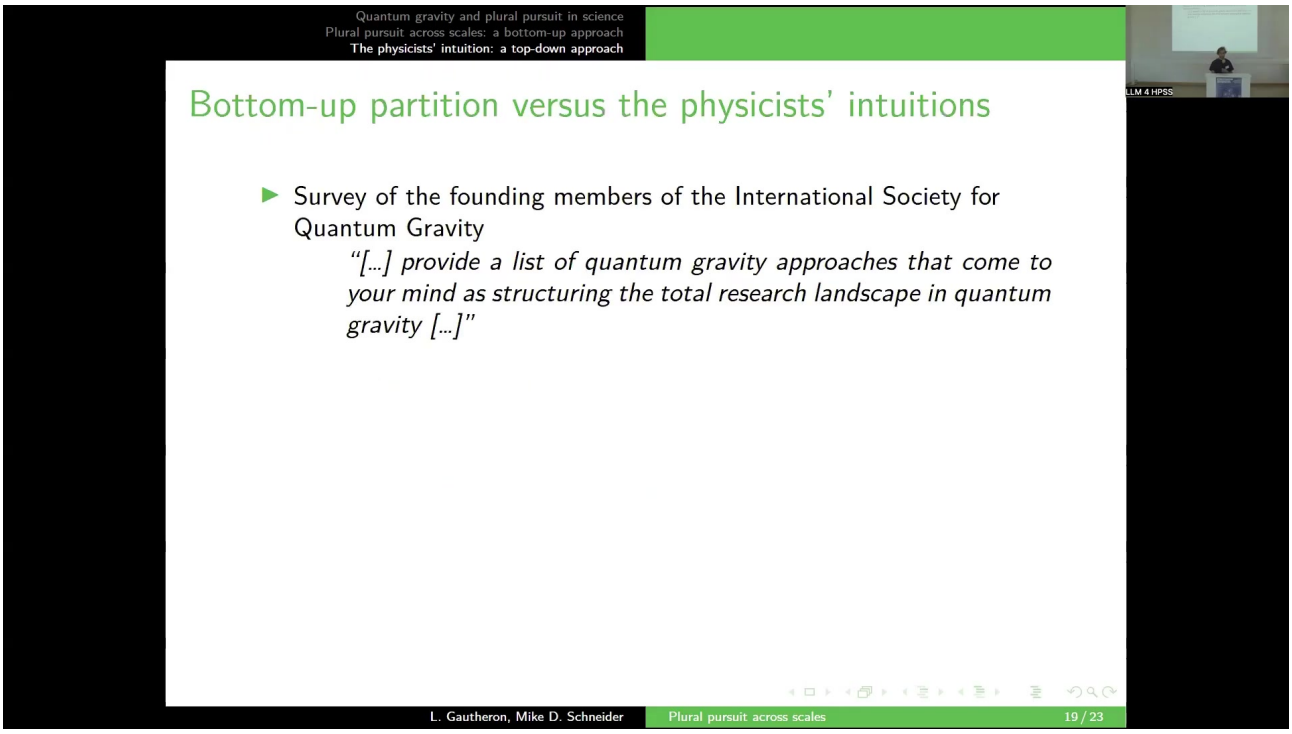

Figure 14.9: Slide 16

The research proceeds to compare the top-down perspective, represented by the supervised classification into intuitive approaches, with the bottom-up reconstruction, specifically the 50 coarse-grained topics. This comparison is visualized using a heatmap where rows represent the top-down (supervised) approaches and columns represent the coarse-grained bottom-up topics.

The findings from this comparison reveal varying degrees of agreement. For certain top-down approaches, there is a strong correspondence with specific topics that emerged from the bottom-up analysis. However, for other approaches, the correspondence is weak or non-existent. An explanation for this disagreement is that the bottom-up analysis tends to align well with approaches that are considered well-defined and conceptually autonomous, while it shows less correspondence with approaches that are primarily phenomenological or not yet developed into full-fledged conceptual frameworks.

A particularly notable finding concerns the relationship between string theory, supergravity, and holography. The bottom-up analysis reveals a large cluster identified as string theory that appears to encompass both supergravity and string theory. This observation converges with the intuition expressed by some physicists in the survey, who struggled with whether supergravity and string theory should be considered separate entities, noting the significant overlap between the communities working on them and questioning whether they could be meaningfully separated. The interpretation is that the coarse-graining process, by stripping away linguistic nuances that do not have consequences for the social structure, groups together areas like supergravity and string theory, even though they are conceptually

distinct, thereby reflecting the social reality of highly overlapping communities.

14.11 Conclusions

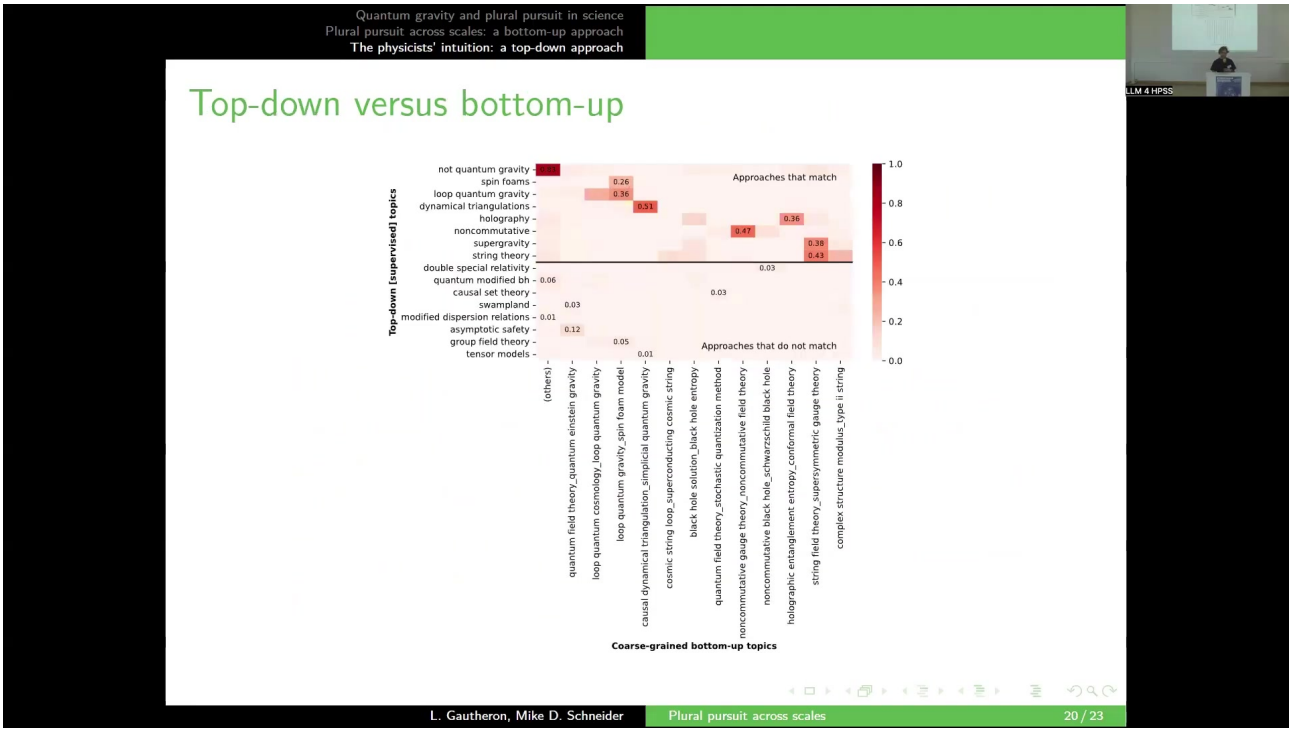

Figure 14.10: Slide 18

The research draws several conclusions regarding the structure of scientific fields and the application of computational methods in the philosophy of science.

Firstly, it concludes that socio-epistemic systems, which encompass both the social and intellectual aspects of scientific research, can be observed and analyzed at multiple scales. This implies that the fundamental notions of communities and disciplinary matrices, central to understanding the structure of science, are inherently scale-dependent.

Secondly, identifying configurations of plural pursuit, which ideally involve a one-to-one mapping between communities and their intellectual substrate, necessitates matching these structures across different scales. A simple analysis at a single, arbitrary scale is insufficient to capture the complex relationships present.

Thirdly, specifically in the case of quantum gravity, the bottom-up reconstruction of the research landscape, utilizing data-driven methods, can serve to either confirm or re-assess certain intuitions held by physicists about how their field is structured. The analysis provides empirical evidence that can validate or challenge subjective perceptions.

More broadly, the research highlights the increasing power of computational methods as tools that can help revisit or challenge long-standing philosophical insights that were previously based primarily on intuition. This includes intuitions about concepts such as what constitutes a paradigm or a community in a given scientific context. Drawing an analogy, the research suggests that computation can be seen as the continuation of philosophy by other means, paraphrasing Clausewitz.

Chapter 15

Text Granularity and Topic Model Performance

This study investigates the impact of text granularity (titles, abstracts, full-texts) on the performance of two distinct topic modeling approaches, Latent Dirichlet Allocation (LDA) and BERTopic. The research addresses the practical challenge of significant resource requirements for obtaining, preprocessing, and analyzing full-text corpora by comparing topic models derived from different text levels. A corpus of scientific articles in Astrobiology serves as the material. Six topic models...

15.1 Overview

This study investigates the impact of text granularity (titles, abstracts, full-texts) on the performance of two distinct topic modeling approaches: *Latent Dirichlet Allocation* (*LDA*) and *BERTopic*. The research addresses the practical challenge of significant resource requirements for obtaining, preprocessing, and analyzing full-text corpora by comparing topic models derived from different text levels.

A corpus of scientific articles in Astrobiology serves as the material for this study. Six topic models are generated: *LDA* on titles, abstracts, and full-texts, and *BERTopic* on titles, abstracts, and full-texts. These models are then analyzed and compared qualitatively and quantitatively using metrics such as Adjusted Rand Index, Topic Diversity, Joint Recall, and Coherence CV.

The qualitative analysis involves comparing topic coherence and the stability of topics across models, referencing a previously established *LDA* full-text model with 25 topics and 4 thematic clusters. Quantitative results indicate that title-based models generally perform poorly, while abstract models show better coherence and diversity. Full-text models demonstrate superior joint recall.

Specifically, *BERTopic* Abstract emerges as a strong performer in coherence, and *BERTopic* Title in diversity, while *LDA* Fulltext and *BERTopic* Fulltext excel in joint recall. The study concludes that the optimal choice of text level and topic model depends on specific research objectives. Abstract-based models offer a good balance and consistency with full-text models, while title-based models, despite limitations, can identify robust core topics. The potential for leveraging structural information (titles, abstracts, full-texts) in future models is also discussed.

15.2 Introduction

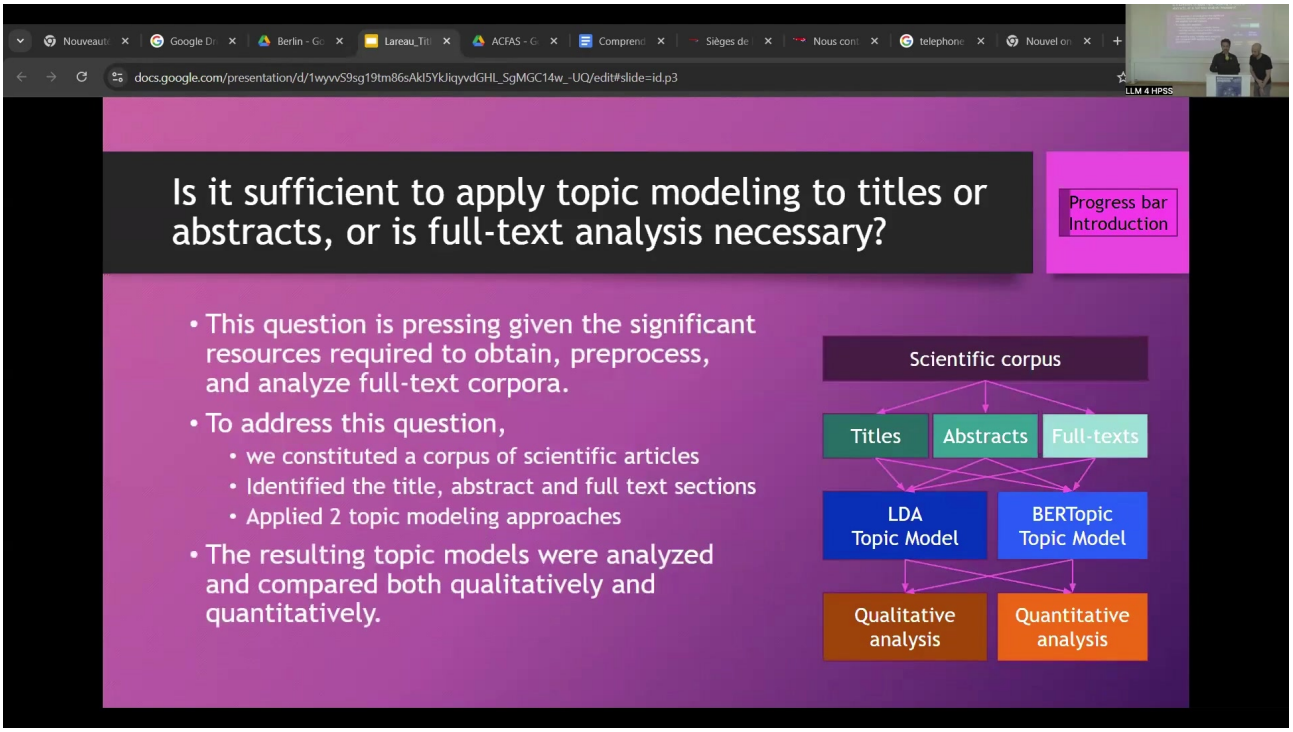

Figure 15.1: Slide 01

This presentation is delivered by Francis Lareau, a Postdoctoral Fellow affiliated with the University of Sherbrooke and the University of Quebec in Montreal (UQAM). This work is a comparative study conducted with Christophe Malaterre from the University of Quebec in Montreal.

The study focuses on topic modeling, a technique for extracting themes from a corpus. Topic modeling is recognized as an important tool for analyzing large volumes of scientific literature, especially within the history, philosophy, and sociology of science (HPSS).

A problem arises because existing studies utilize different textual structures for topic modeling, namely titles, abstracts, and full text. Obtaining, preprocessing, and analyzing full-text corpora demand significant resources. This prompts the central research question: Is applying topic modeling to titles or abstracts sufficient, or is full-text analysis necessary?

15.3 Study Design

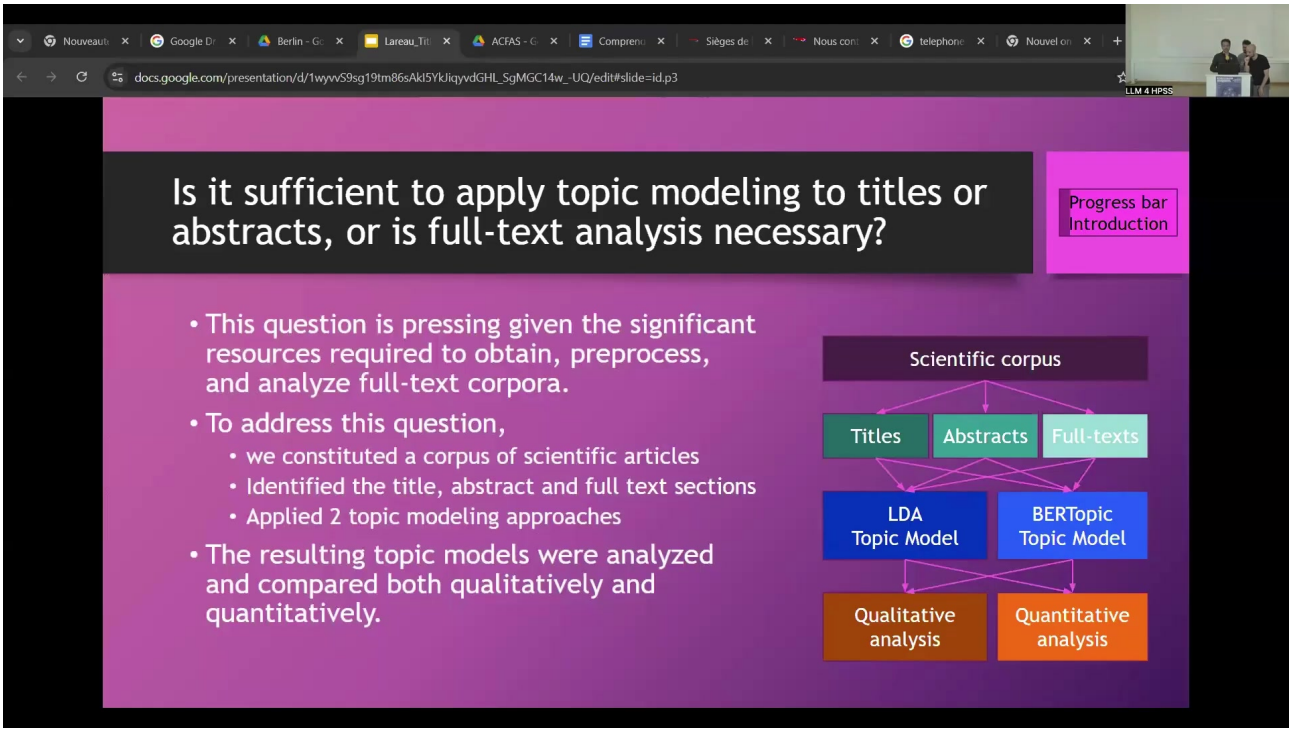

Figure 15.2: Slide 02

This study addresses the pressing question of whether analyzing titles or abstracts is sufficient for topic modeling, given the substantial resources needed for full-text corpora acquisition, preprocessing, and analysis. The methodology involves a structured workflow.

First, a corpus of scientific articles is constituted. Second, the distinct title, abstract, and full text sections are identified within this corpus. Third, two different topic modeling approaches, *Latent Dirichlet Allocation* (*LDA*) and *BERTopic*, are applied separately to each of the three identified text levels: titles, abstracts, and full texts.

This process generates a total of six distinct topic models. Finally, these six resulting topic models undergo both qualitative and quantitative analysis and comparison to evaluate their performance across the different text levels.

15.4 Topic Modeling Approaches

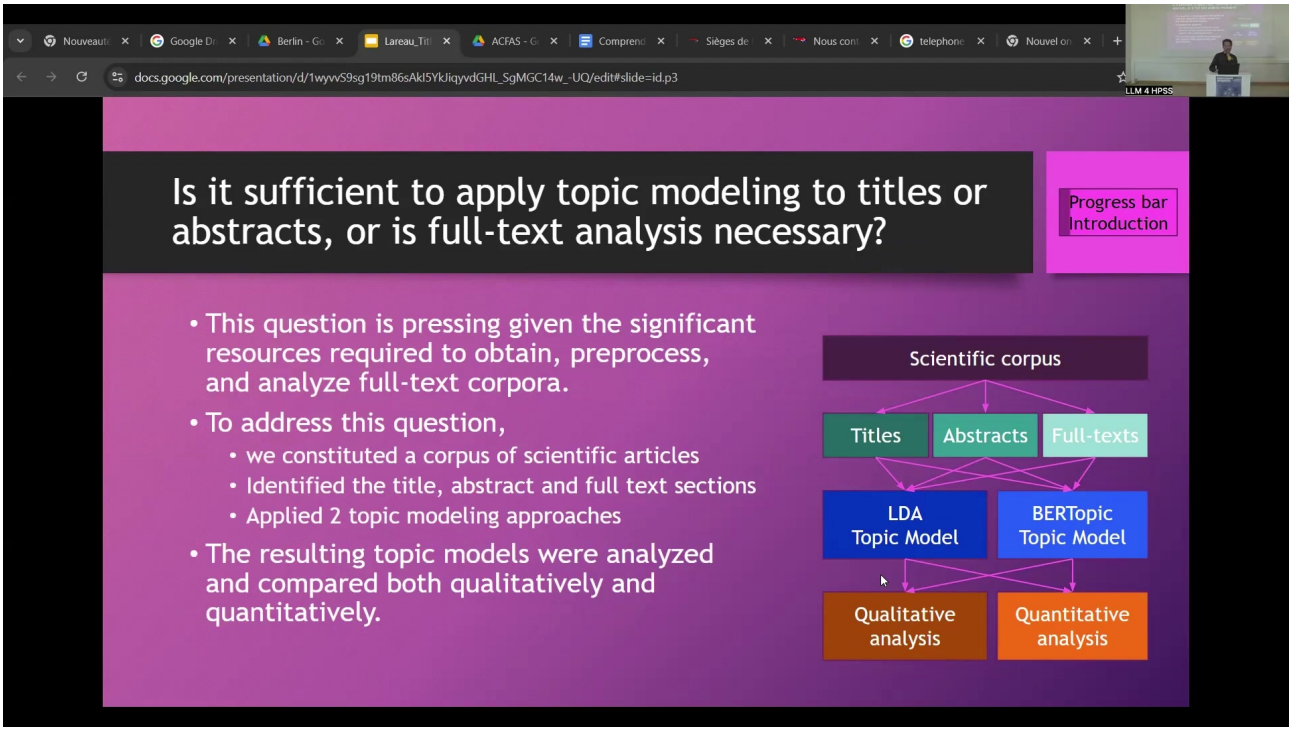

Figure 15.3: Slide 03

The study compares two distinct topic modeling approaches: *Latent Dirichlet Allocation* (*LDA*) and *BERTopic*. Both approaches share fundamental postulates: documents can be represented by numerical vectors, topics are identifiable through linguistic regularities manifested as repetitions, and machine learning facilitates the automatic detection of these regularities.

Latent Dirichlet Allocation (*LDA*) is characterized as a classical statistical method. It employs a classical vector representation technique based on counting words within documents. In the *LDA* framework, topics are conceptualized as latent variables that adhere to Dirichlet's law. A key advantage of *LDA* is its ability to handle long texts, making it suitable for analysis across titles, abstracts, and full texts.

In contrast, *BERTopic* is described as a modern, modular approach, developed by Martin Grootendorst. It utilizes an LLM-based vector representation method, originally based on *BERT*, which gives the approach its name. Topics in *BERTopic* correspond to topological densities of documents, typically identified using clustering algorithms like *HDBSCAN*.

Historically, *BERTopic* did not handle long texts efficiently, but recent advancements have addressed this limitation. For this study, a specific embedding model, *Stella EN 1.5B V5*, was selected for the *BERTopic* implementation. This model was chosen based on its high ranking on the Massive Text Embedding Benchmark on Hugging Face and its capacity to handle approximately 131,000 tokens, addressing the long text limitation.

15.5 Material and Qualitative Analysis

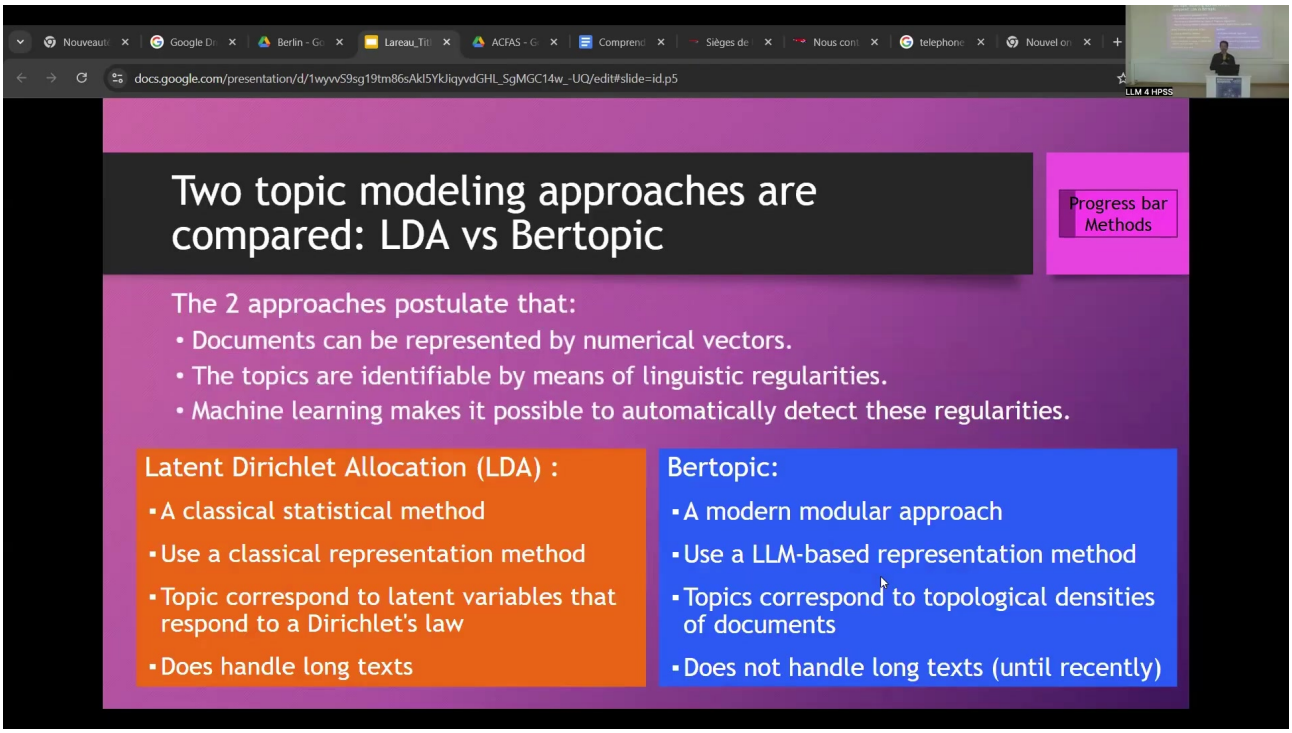

Figure 15.4: Slide 06

The material utilized in this study is an Astrobiology corpus, which was previously subjected to an in-depth topic analysis. This prior analysis, documented in *Malaterre & Lareau (2023)*, resulted in the selection of a full-text *LDA* model comprising 25 topics. This model serves as a reference for the current comparative study.

The 25 topics of the reference model were analyzed meticulously. This analysis involved examining the most representative words and documents associated with each topic. Based on this examination, each topic was assigned a descriptive label derived from its key terms.

The relationships between topics were then assessed by calculating their mutual correlation, determined by the co-occurrence of topics within documents. Subsequently, a community detection algorithm was applied to the correlation data, identifying four distinct thematic clusters. These clusters were designated with letters (A, B, C, D) and assigned corresponding colors (red, green, yellow, blue) for visualization.

The results of this reference analysis are represented visually as a graph illustrating the correlations between the 25 topics. The graph displays the topic labels and the color variations indicating their thematic clusters. The thickness of the lines connecting topics represents the strength of their correlation, while the size of the circles representing topics indicates their overall presence across all documents in the corpus. This established and analyzed reference model provides a basis for qualitatively comparing the six topic models generated and investigated in the current comparative study.

15.6 Quantitative Metrics

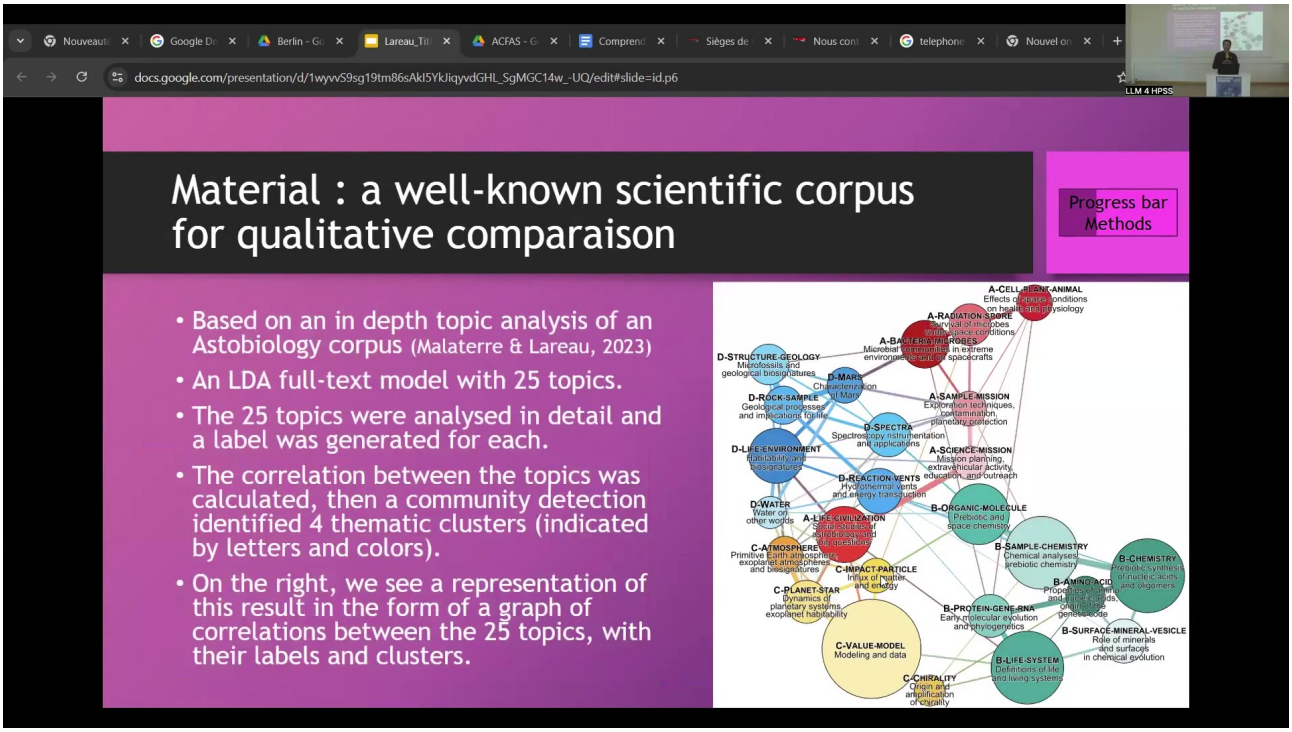

Figure 15.5: Slide 07

The quantitative analysis compares the six topic models using four specific metrics:

- *Adjusted Rand Index*: This metric assesses the similarity between two different document clusterings. It is corrected for chance, meaning a value of zero corresponds to a random clustering.

- *Topic Diversity*: This measures the proportion of distinct top words utilized to describe the topics within a single topic model. A higher diversity indicates that topics are characterized by different sets of words.

- *Joint Recall*: This evaluates the extent to which the top words collectively represent the documents assigned to each topic. It provides an average measure of document-topic recall, indicating how well the topic's representative words can retrieve the documents belonging to that topic.

- *Coherence CV*: This evaluates the semantic meaningfulness of the topic's top words. It is computed as the average of the cosine relative distance between the top words within each topic, where a higher value suggests greater semantic relatedness among the words and thus more coherent topics.

15.7 Adjusted Rand Index Results

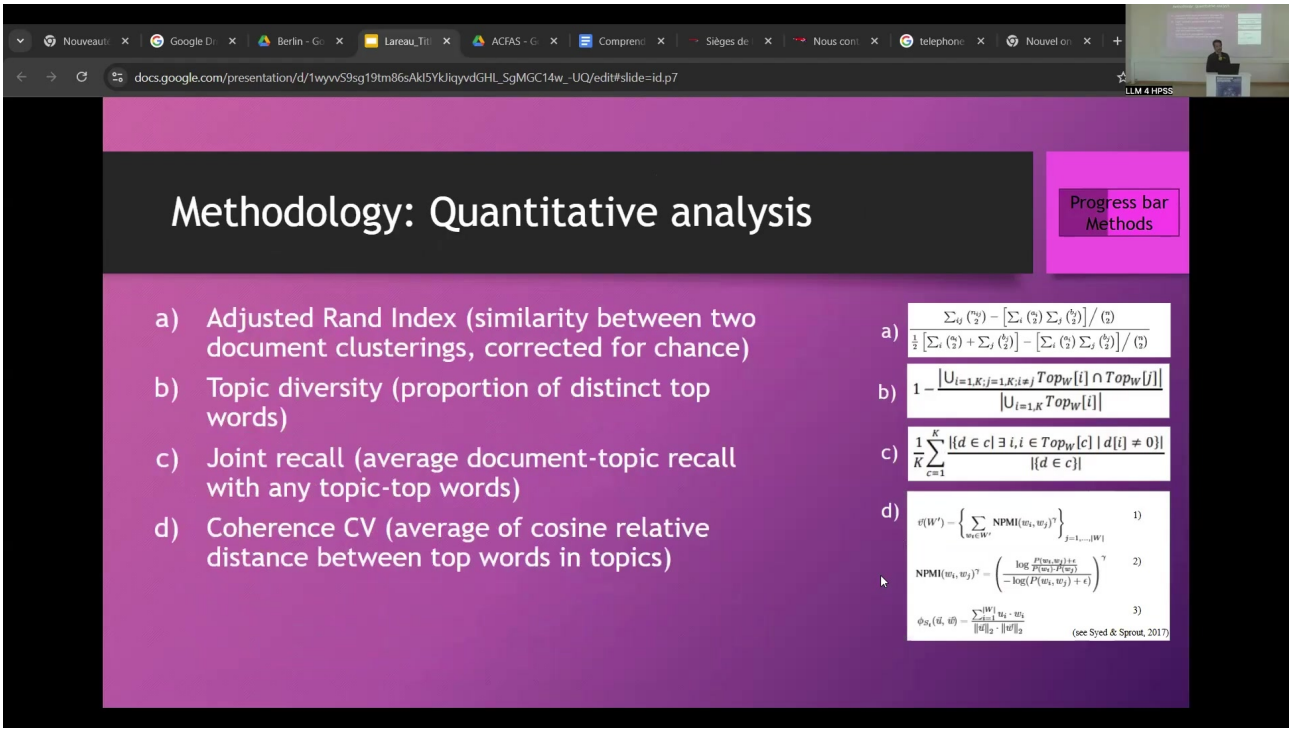

Figure 15.6: Slide 08

The *Adjusted Rand Index* is employed to quantitatively assess the similarities among the six generated topic models by comparing their document clusterings. A value of zero for this metric signifies a random clustering.

The results indicate that the *LDA* model trained on titles is the most distinct among all models, as evidenced by low *Adjusted Rand Index* values, specifically under 0.2, shown in the heatmap comparison.

Conversely, all other models exhibit a better overall match with each other, demonstrating values exceeding 0.2. A notable observation is that the *BERTopic* models tend to correspond more strongly with each other, with *Adjusted Rand Index* values generally above 0.35. Furthermore, the *BERTopic* Abstract model appears to be more central in its similarity profile, showing good correspondence to every other model, with values over 0.30, except for the *LDA* Title model.

15.8 LDA Model Comparison

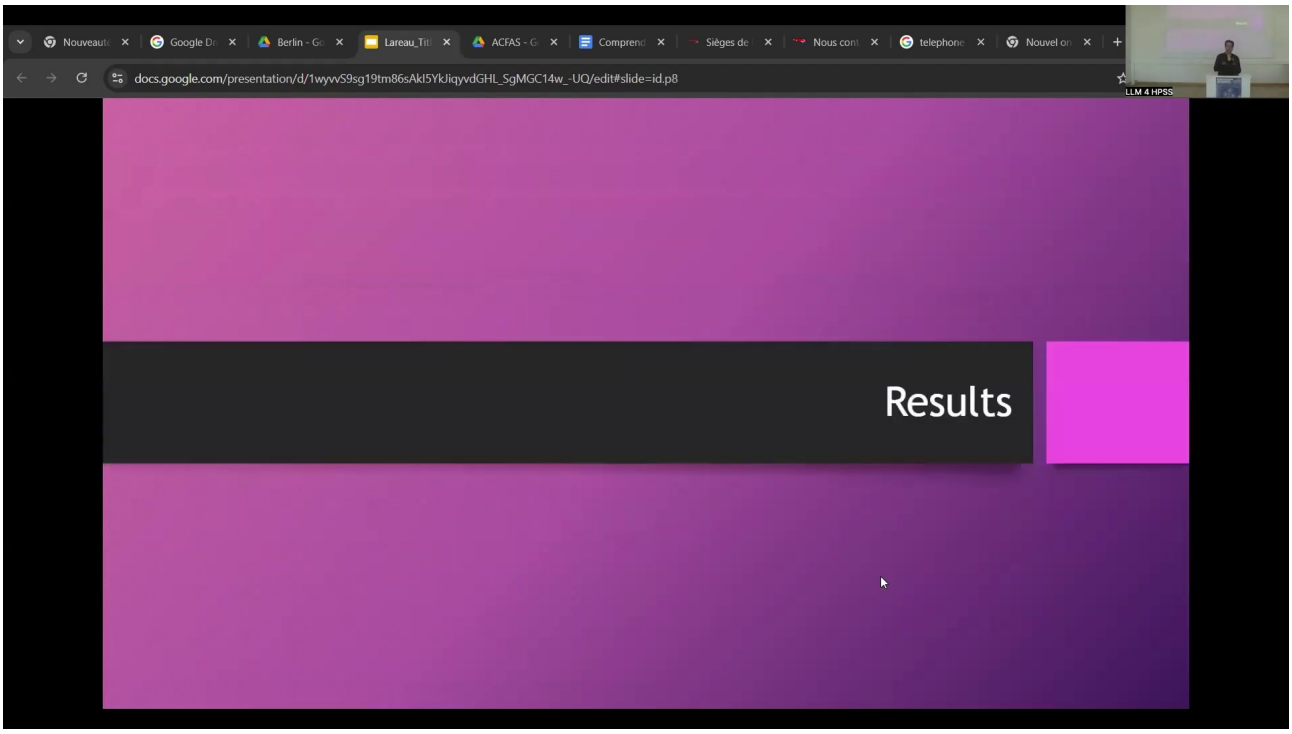

Figure 15.7: Slide 09

A more detailed qualitative analysis focuses on comparing the *LDA* Full-text model with the *LDA* Abstract and *LDA* Title models, using heatmaps that visualize the number of shared documents between topics. A reddish diagonal pattern in these heatmaps signifies a good correspondence between topics across models.

Comparing the *LDA* Full-text model with the *LDA* Abstract model (Table A) reveals a good overall fit. This is evident from the prominent reddish diagonal, indicating that topics in one model largely correspond to single topics in the other with a high proportion of shared documents.

However, the analysis also identifies specific topic transformations: three full-text topics disappear entirely in the abstract model (represented by long horizontal dark gray lines), three full-text topics split into multiple topics in the abstract model (short horizontal dark gray lines), and three abstract topics are formed by the merger of multiple full-text topics (short horizontal dark gray lines). Additionally, the *LDA* Abstract model exhibits one small class containing fewer than 50 documents.

In contrast, the comparison between the *LDA* Full-text model and the *LDA* Title model (Table B) indicates a poor overall fit, characterized by substantial reorganization of topics. This is visually represented by numerous vertical and horizontal dark lines in the heatmap, signifying that many full-text topics disappear and many new topics emerge in the title model, with little direct correspondence.

15.9 BERTopic Model Comparison

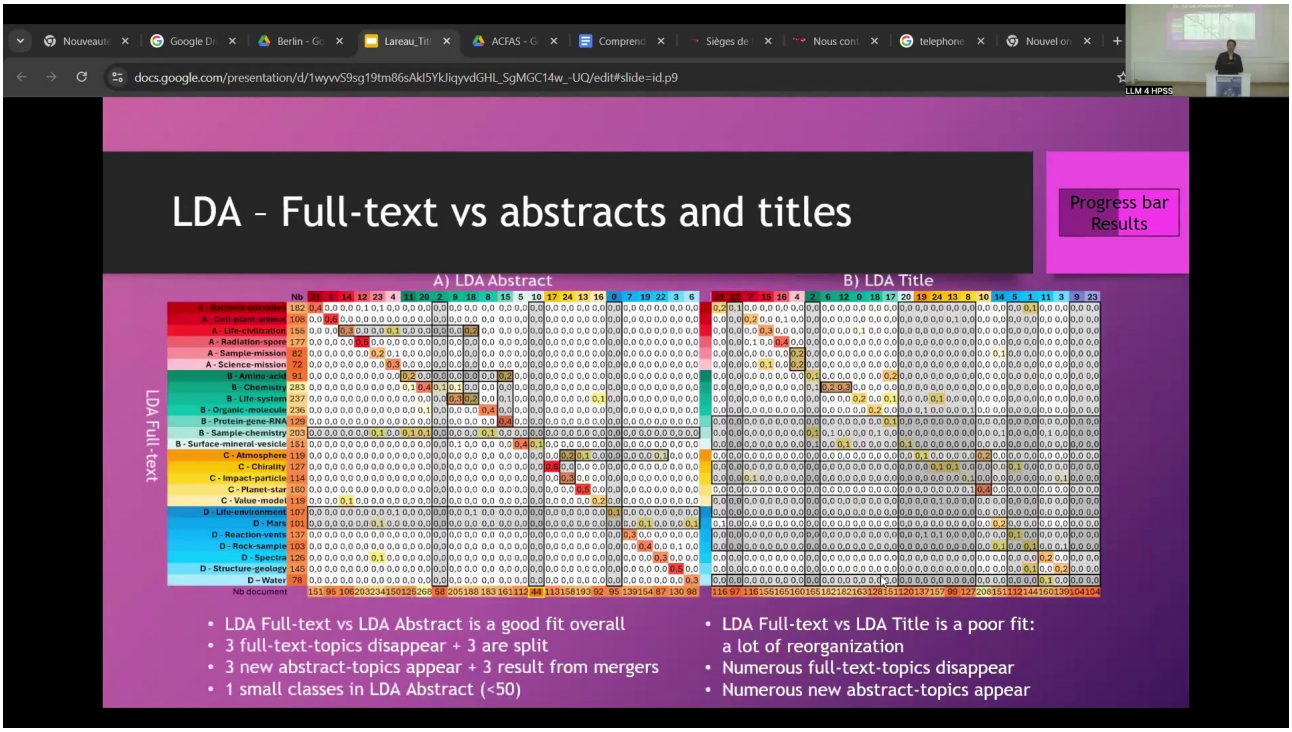

Figure 15.8: Slide 11

The qualitative comparison extends to the *BERTopic* models, assessing their correspondence with the *LDA* Full-text reference model using heatmaps showing shared documents.

Comparing the *LDA* Full-text model with the *BERTopic* Full-text model (Table C) shows an average overall fit. The analysis reveals that 8 topics from the *LDA* Full-text model disappear, and 6 topics split into multiple topics in the *BERTopic* Full-text model (indicated on the horizontal axis). On the vertical axis, 5 new topics appear in the *BERTopic* model, and 1 topic results from the merger of *LDA* topics. The *BERTopic* Full-text model also exhibits issues with class size, including 4 small classes and 1 very large class.

The comparison between the *LDA* Full-text model and the *BERTopic* Abstract model (Table D) indicates a relatively good overall fit. Four *LDA* topics disappear, and 6 topics split in the *BERTopic* Abstract model (horizontal axis). Two new topics appear, and 4 topics result from mergers in the *BERTopic* Abstract model (vertical axis).

Finally, comparing the *LDA* Full-text model with the *BERTopic* Title model (Table E) shows an average overall fit. Seven *LDA* topics disappear, and 1 topic splits in the *BERTopic* Title model (horizontal axis). Seven new topics appear, and 1 topic results from a merger in the *BERTopic* Title model (vertical axis). The *BERTopic* Title model also presents class size issues, with 3 small classes and 1 large class.

15.10 Comparing Top Words

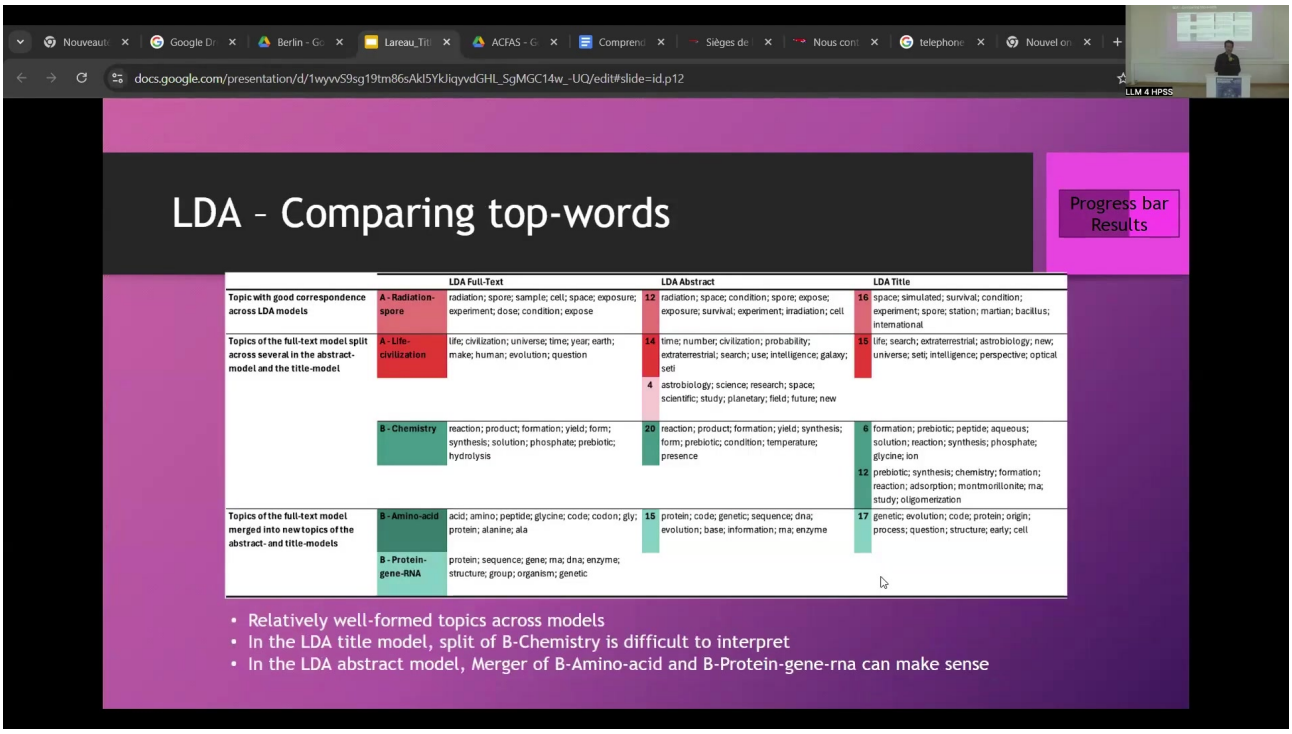

Figure 15.9: Slide 13

A qualitative assessment involves comparing the top words associated with selected topics across the different models to understand the nature of the topics generated.

Within the *LDA* models (Full-text, Abstract, and Title), topics are observed to be relatively well-formed. A robust topic, exemplified by "A radiation spore," demonstrates good correspondence in its top words across all three *LDA* models.

Splitting of topics is also observed: the "A life civilization" topic from the full-text model splits across the abstract and title models, which is considered sensible as it relates to a general theme of research in astrobiology. The "B chemistry" topic from the full-text model also splits across the abstract and title models, though this particular split is noted as being more challenging to interpret without deeper analysis. Merging of topics occurs as well, such as the "B amino acid" and "B protein gene RNA" topics from the full-text model merging into a single topic in other models, which is deemed sensible as it forms a more general thematic area.

Comparing the *BERTopic* models (Full-text, Abstract, and Title) with the *LDA* Full-text model also reveals relatively well-formed topics across all *BERTopic* models. The robustness of the "A radiation spore" topic is again observed, appearing consistently across all *BERTopic* models and the *LDA* Full-text reference. The "A life civilization" topic is relatively stable across the *BERTopic* models, although some splitting occurs, leading to narrower topics specifically focused on extraterrestrial life. The "B chemistry" topic also splits across the *BERTopic* models, resulting in more narrow thematic topics.

15.11 Coherence, Diversity, and Joint Recall Results

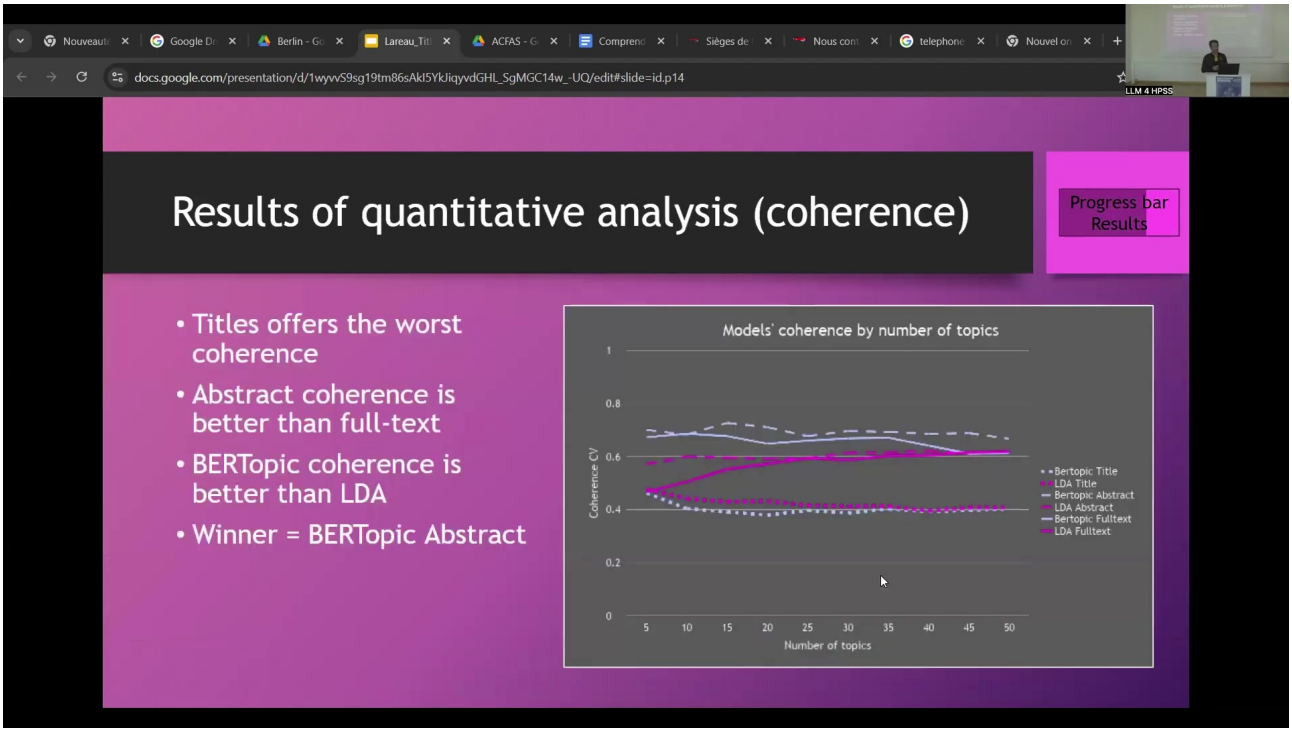

Figure 15.10: Slide 15

Quantitative performance metrics are evaluated for all six topic models across a range of topic numbers, specifically from 5 to 50.

The *Coherence CV* metric, which assesses the meaningfulness of the topic top words, yields several findings:

- Models trained on titles exhibit the worst coherence.
- Abstract models demonstrate better coherence compared to full-text models.
- Overall, *BERTopic* models show better coherence than *LDA* models when applied to abstracts and titles, although this difference becomes less pronounced as the number of topics increases.
- Based on this metric, *BERTopic* Abstract is identified as the clear winner.

Regarding *Topic Diversity*, which measures the proportion of distinct top words, the results show that diversity generally decreases as the number of topics increases:

- Models trained on titles offer the best diversity.
- *BERTopic* models exhibit better diversity than *LDA* models.
- The winner for diversity is *BERTopic* Title, closely followed by *BERTopic* Full-text.

The *Joint Recall* metric evaluates how effectively the top words collectively represent the documents classified within each topic:

- Titles yield the worst joint recall.

- Full-text models perform better than their abstract and title counterparts.
- *LDA* models generally show better joint recall than *BERTopic* models.
- The winners for *Joint Recall* are *LDA* Fulltext and *BERTopic* Fulltext, with *BERTopic* Abstract performing very closely behind.

15.12 Model Performance Summary

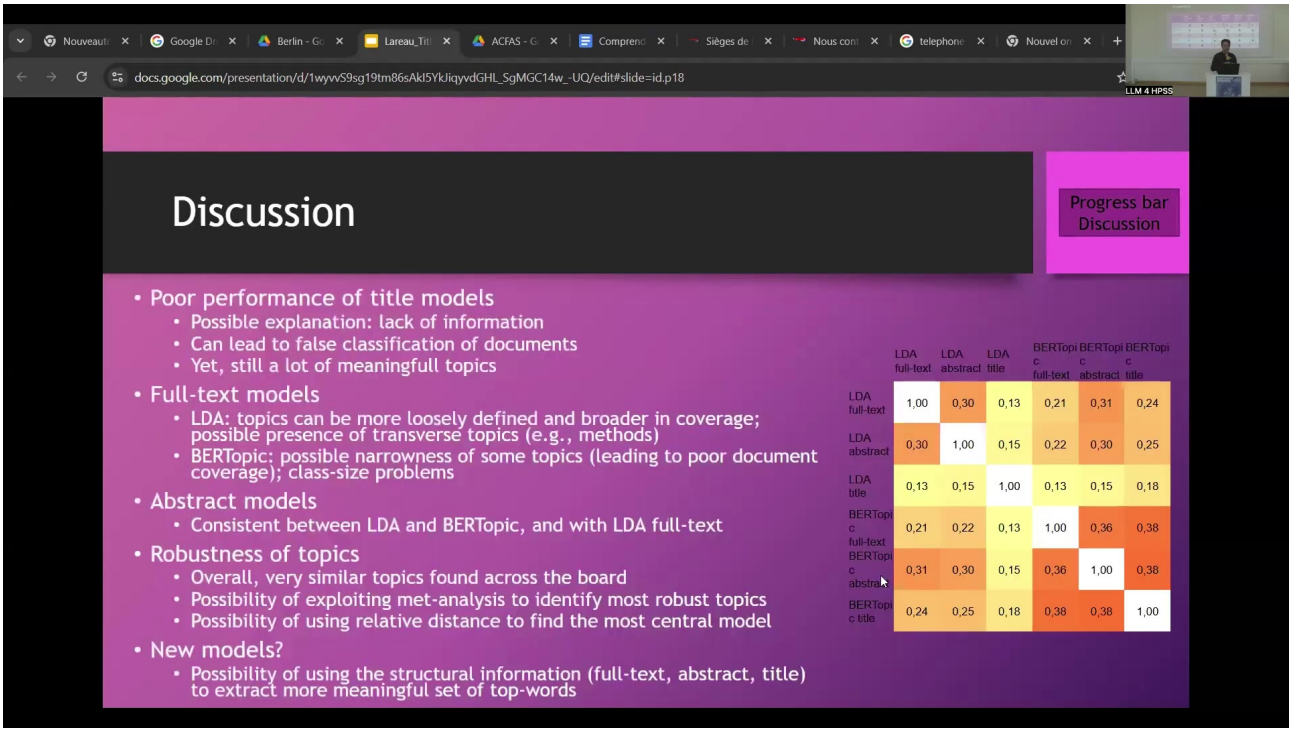

Figure 15.11: Slide 17

A summary table consolidates the performance results, offering an overall view of each model's strengths and weaknesses across various assessment criteria. The models evaluated are:

- *LDA* Full-text (rated 4*)
- *LDA* Abstract (4.5*)
- *LDA* Title (2.5*)
- *BERTopic* Full-text (4.5*)
- *BERTopic* Abstract (4.5*)
- *BERTopic* Title (3*)

The assessment criteria include Overall fit, Top-words quality, Coherence, Diversity, and Joint recall. Performance scores are visually represented using circle icons, where a full black circle signifies the highest score and an empty circle indicates a low score. Red crosses highlight specific problems, such as class imbalance.

The analysis indicates that there is no single absolute best model; the optimal choice is contingent upon the specific research objectives. Different objectives necessitate different model characteristics. For instance, if the primary goal is the discovery of main topics without requiring precise classification of every document, issues like poor recall or large classes might be acceptable. In such a scenario, the *BERTopic* Full-text model performs well, although it exhibits some class imbalance. The *BERTopic* Title model, while generally less optimal, is capable of producing some robust topics that are also identified by the other models.

15.13 Discussion and Future Directions

The discussion highlights several key observations and potential future directions. The poor performance observed in title-based models is primarily attributed to the inherent lack of information in titles compared to abstracts or full texts. This limitation can lead to inaccurate classification of documents, although title models are still capable of identifying meaningful core topics.

Full-text models exhibit distinct characteristics depending on the approach. *LDA* models applied to full text tend to produce topics that are more loosely defined and broader in coverage, potentially capturing transverse themes such as research methods. *BERTopic* full-text models, on the other hand, may result in some topics being too narrow, leading to poor document coverage, and can suffer from class-size imbalance problems.

Abstract models demonstrate notable consistency, both between the *LDA* and *BERTopic* implementations and in their correspondence with the *LDA* full-text model. A significant finding is the overall robustness of topics, with very similar thematic areas being identified across the board, regardless of the specific model or text level used.

Future research possibilities include exploiting meta-analysis techniques to systematically identify the most robust topics that consistently appear across multiple models and text levels. Another direction involves using relative distance metrics to determine which model is the most central or representative among the set. Furthermore, the study suggests the potential for developing new topic modeling approaches that explicitly leverage the structural information present in documents (i.e., the distinction between full text, abstract, and titles) to extract more meaningful sets of topics or top words.

Chapter 16

Time-Aware Language Models

The presentation describes the development and evaluation of a novel architecture for creating time-aware language models (TALMs), specifically targeting applications in historical analysis. The core problem addressed is the implicit nature of temporal understanding in current Large Language Models (LLMs), which is derived statistically from training data. The proposed solution involves explicitly adding a temporal dimension to the latent semantic token features within a *Transformer* architecture.

16.1 Overview

The presentation describes the development and evaluation of a novel architecture for creating time-aware language models (*TALMs*), specifically targeting applications in historical analysis. The core problem addressed is the implicit nature of temporal understanding in current *Large Language Models* (*LLMs*), which is derived statistically from training data. The proposed solution involves explicitly adding a temporal dimension to the latent semantic token features within a *Transformer* architecture.

The technical approach modifies a standard *Transformer* decoder model by injecting time data, represented as a non-trainable, min-max normalized day of the year, into the token embeddings. This allows the model to learn how the probability distribution of tokens depends on time.

A proof-of-concept implementation utilizes a small generative *LLM* trained on a specific dataset: daily weather reports from the UK Met Office digital archive for the years 2018-2024. This dataset consists of approximately 2,500 reports, each 150-200 words long, characterized by a limited vocabulary and repetitive language. The text processing involves text vectorization with standardization (lower and strip punctuation) and no sub-word tokenization, resulting in a vocabulary of 3,395 words.

The model architecture is a modest-sized *Transformer* decoder with 4 multihead attention blocks, totaling 39 million parameters (150 MB). This is significantly smaller than models like *GPT-4* (1.8 trillion parameters across 120 layers). Training is performed on 2 x A100 GPUs, taking 11 seconds per epoch. The code for the vanilla and time-aware *Transformer* models is available on GitHub at https://github.com/j-buettner/time_transformer.

Two experiments demonstrate the model's ability to learn temporal drift:

- *Synonymic Succession:* Synthetic drift is injected by time-dependent replacement of "rain" with "liquid sunshine" following a sigmoid probability curve over the year. The model successfully reproduces this time dependence in predicted token sequences.

- *Changing Co-occurrence/Collocation Fixation:* Synthetic time-dependent change is injected where "rain" followed by any word except "and" is replaced by "rain and snow" with increasing probability over the year. The model learns this changing co-occurrence pattern, demonstrating the fixation of the "rain and snow" collocation over time. Attention analysis shows increased attention from "snow" to "rain" in the predicted sequences.

The proof of concept indicates that *Transformer*-based *LLMs* can be made time-aware efficiently by adding a temporal dimension to the token embedding. Potential applications include providing a foundation for downstream tasks on historical data, enabling instruction-tuned models to "talk to a specific time," and modeling dependence on other metadata dimensions (country, genre). Challenges include uncertainty regarding the efficiency of fine-tuning due to architectural changes, the need for data curation (including timestamping token sequences), and the loss of metadata-free self-supervised learning benefits. An alternative approach involving a targeted encoder model or changing the training task (e.g., predicting document date) is also considered.

Discussion points include the potential for modeling semantic shift over time using this approach, the importance of persistent identifiers for source tracking, existing literature on time-aware *LLMs* and semantic change detection (specifically mentioning encoder-based models and "temporal heads" in foundational models), and a theoretical discussion on whether explicit time injection is necessary given that temporal information is implicitly present in other factors.

16.2 Motivation for Time-Aware Language Models

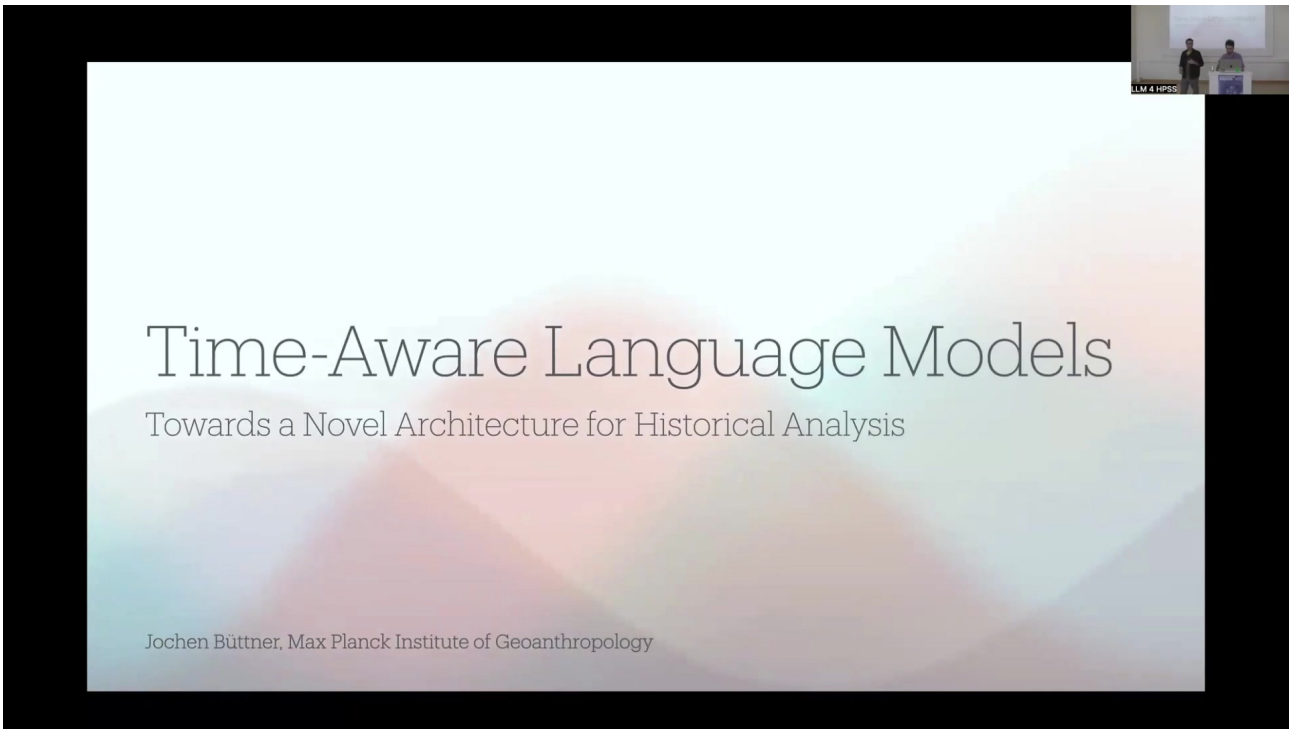

Figure 16.1: Slide 01

Current *Large Language Models* (*LLMs*) exhibit only an implicit understanding of time. This temporal understanding is derived statistically from the patterns observed within their training data. Introducing explicit time awareness into these models is identified as a beneficial enhancement, particularly for their application in historical analysis and potentially other fields where temporal context is crucial.

16.3 Text Processing Architectures

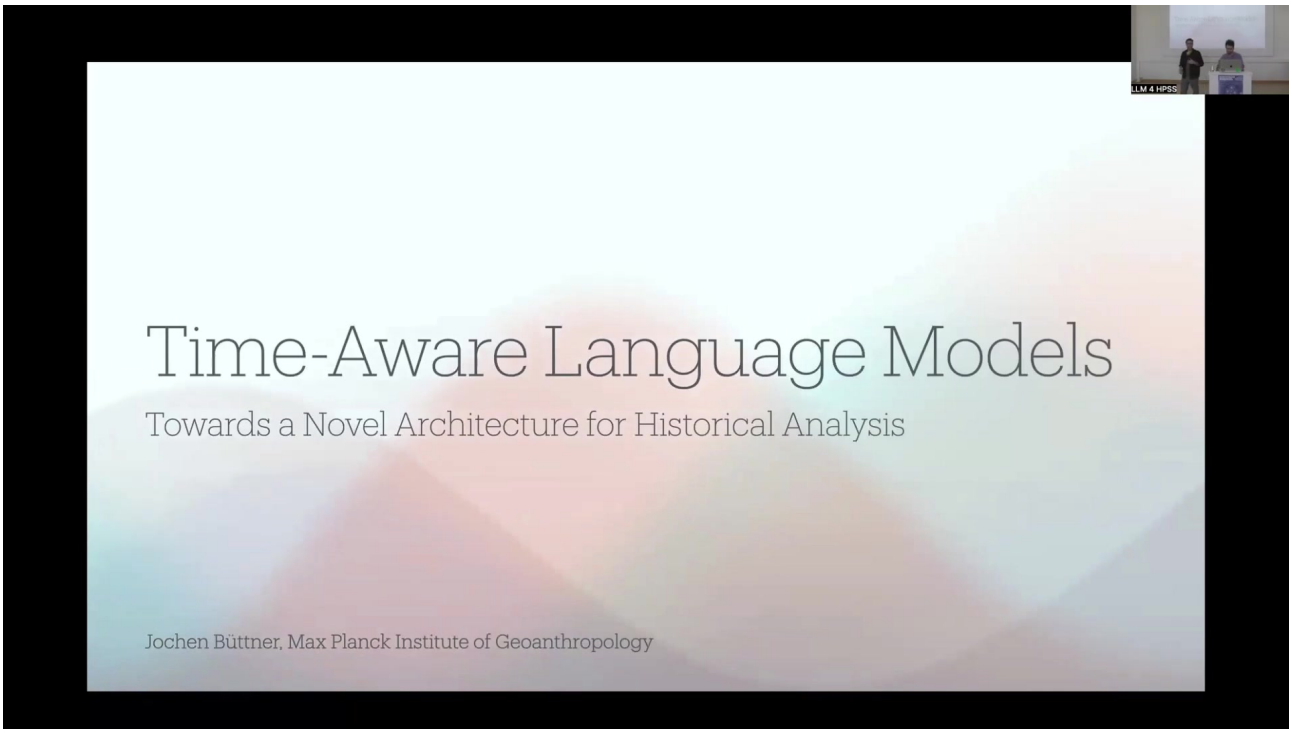

Figure 16.2: Slide 01

The primary neural network architectures employed for processing text have evolved. Historically, around 2017, *Long Short-Term Memory* (*LSTM*) networks were the dominant architecture for tasks such as next-token prediction. As of approximately 2025, the landscape has shifted, with *Transformer* networks becoming the primary architecture utilized for next-token prediction and other text processing tasks.

16.4 Explicit Time Awareness

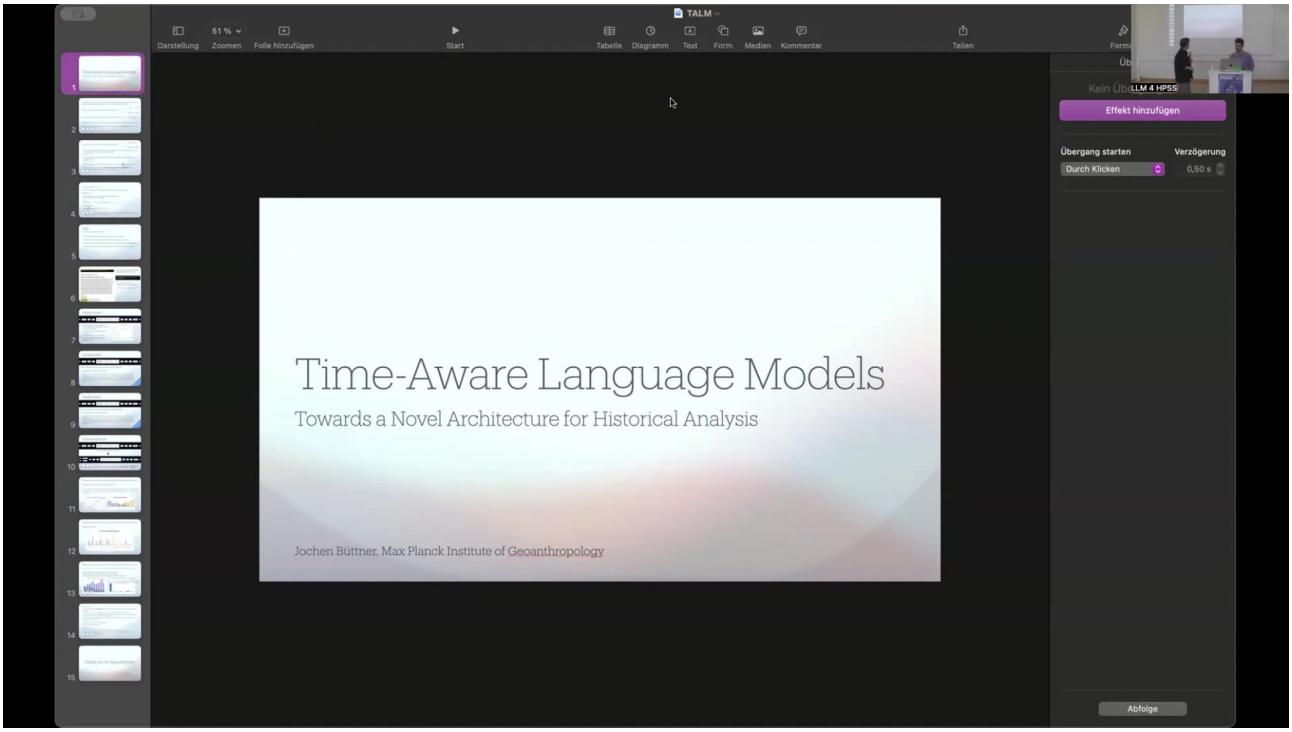

Figure 16.3: Slide 04

The concept is that *Large Language Models* can be endowed with explicit time awareness. This involves enabling the models to learn and subsequently reproduce patterns within their training data that change as a function of time. A proof of concept for this approach is based on the implementation using a small generative *LLM*.

16.5 Temporal Dependence of Token Probabilities

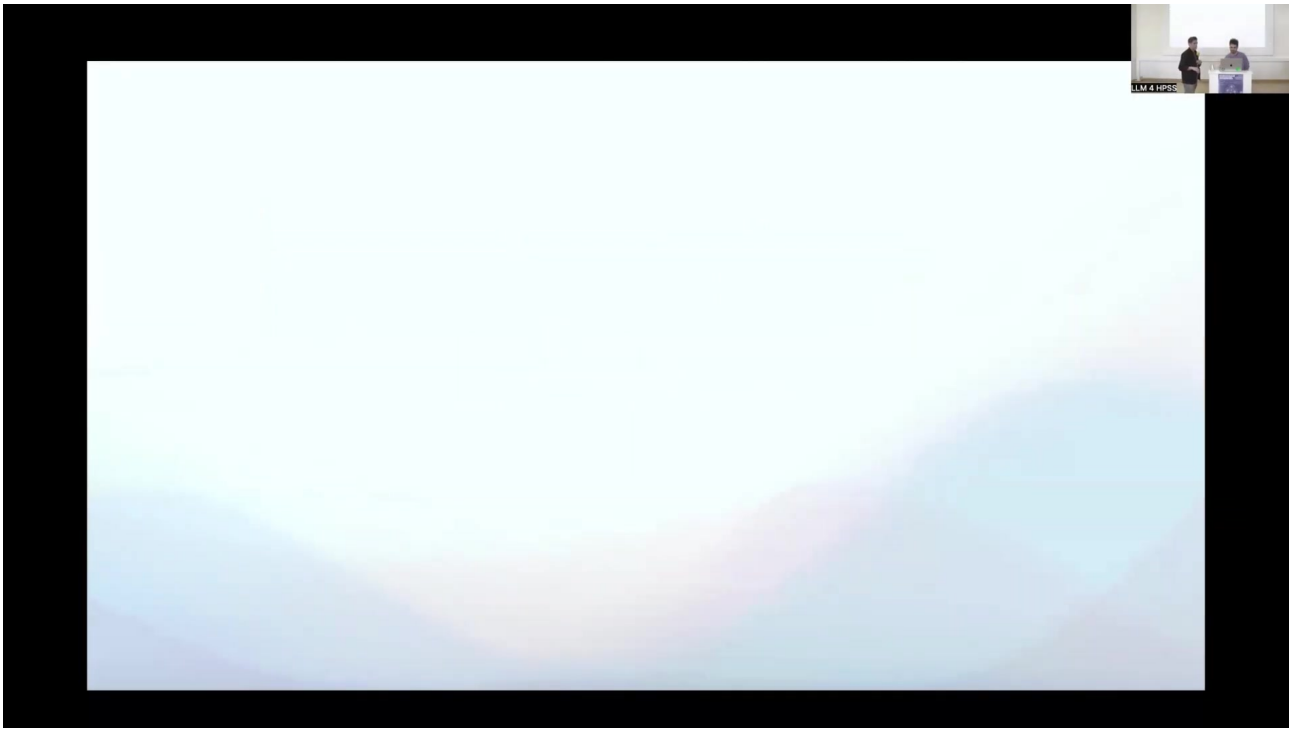

Figure 16.4: Slide 05

Standard *Large Language Models* estimate the probability distribution over their vocabulary for the next token, x_n, conditioned on a sequence of preceding tokens, $x_1, ..., x_{n-1}$. This is formally expressed as $p(x_n|x_1, ..., x_{n-1})$.

However, in real-world scenarios, the probability of a token given its context is not static; it is dependent on time, t. This temporal dependence is represented as $p(x_n|x_1, ..., x_{n-1}, t)$. Consequently, the probability of an entire sequence of tokens, $x_1, x_2, ..., x_n$, generated at a specific time t, is the product of these time-dependent conditional probabilities for each token in the sequence: $p(x_1, x_2, ..., x_n|t) = \prod_{k=1}^{n} p(x_k|x_1, x_2, ..., x_{k-1}, t)$. During inference, current *LLMs* can only reflect the temporal drift observed in the underlying distribution of token sequences through in-context learning, which is an implicit mechanism.

16.6 Modeling Time-Dependent Probabilities

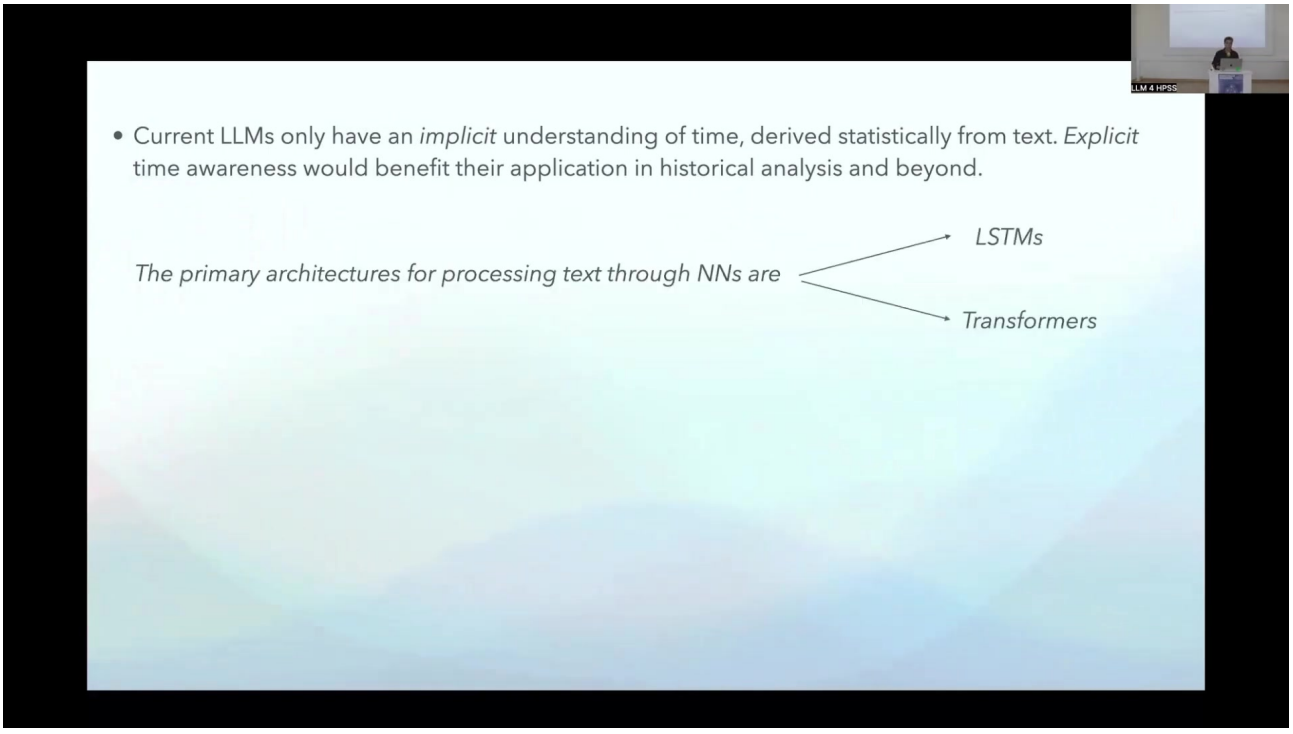

Figure 16.5: Slide 07

A key challenge is explicitly modeling the time-dependent probability of the next token, $p(x_n|x_1, ..., x_{n-1}, t)$. An approach involving time slicing, where separate models are trained for distinct time periods, is considered extremely data inefficient.

The proposed solution is a *Time Transformer* architecture. This method involves adding a temporal dimension to the latent semantic features of each token. The combined embedding for a token x at time t is represented as a vector $E(x, t) = \{e_1(x), e_2(x), ..., e_{d-1}(x), \phi(t)\}$, where $e_i(x)$ are the standard semantic features and $\phi(t)$ is a feature representing time. This sequence of time-aware embeddings, $[E(x_1, t), E(x_2, t), ..., E(x_{n-1}, t)]$, is then fed into a *Transformer* model to predict the time-dependent probability of the next token, $p_\theta(x_n|x_1, ..., x_{n-1}, t)$.

The training objective for this model is to minimize the negative log-likelihood across the entire dataset, given by $\min_\theta - \sum_{i=1}^{N} \sum_{k=1}^{n^{(i)}} \log p_\theta(x_k^{(i)}|x_1^{(i)}, ..., x_{k-1}^{(i)}, t^{(i)})$, where the summation is over all sequences i in the dataset, each with length $n^{(i)}$ and associated time $t^{(i)}$. This approach injects time directly into the representation of every token, enabling the model to learn precisely how strongly or weakly the temporal dimension influences each token's probability.

16.7 Data Source and Preparation

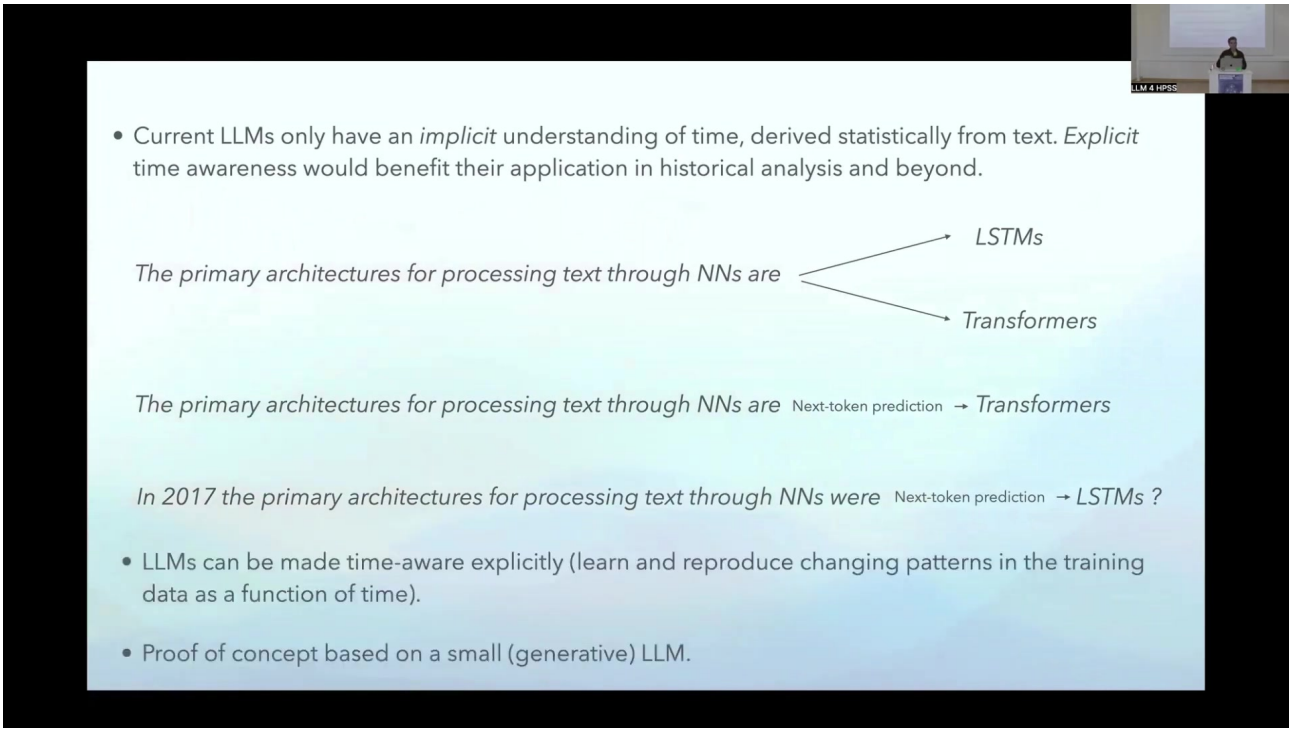

Figure 16.6: Slide 10

The data source utilized for the proof-of-concept implementation consists of Met Office Weather reports. This dataset is characterized by a limited vocabulary and the use of simple, repetitive language, making it suitable for initial experimentation with a small model. The reports are provided by the Met Office, the UK's national meteorological service, and past reports are accessible through their digital archive at https://digital.nmla.metoffice.gov.uk/.

The specific dataset used comprises daily reports covering the years 2018 through 2024, totaling approximately 2,500 reports. Each report is between 150 and 200 words in length. Text processing is performed using `tf.keras.layers.TextVectorization` with the standardization setting `standardize="lower_and_strip_punctuation"`. This process involves neglecting case and interpunctuation, and notably, no sub-word tokenization is applied. This results in a vocabulary size of 3,395 unique words. The choice of a small model and dataset relates conceptually to research exploring the capabilities of small language models, such as the work described in the paper "*TinyStories: How Small Can Language Models Be and Still Speak Coherent English?*".

16.8 *Transformer* Model Architecture and Training

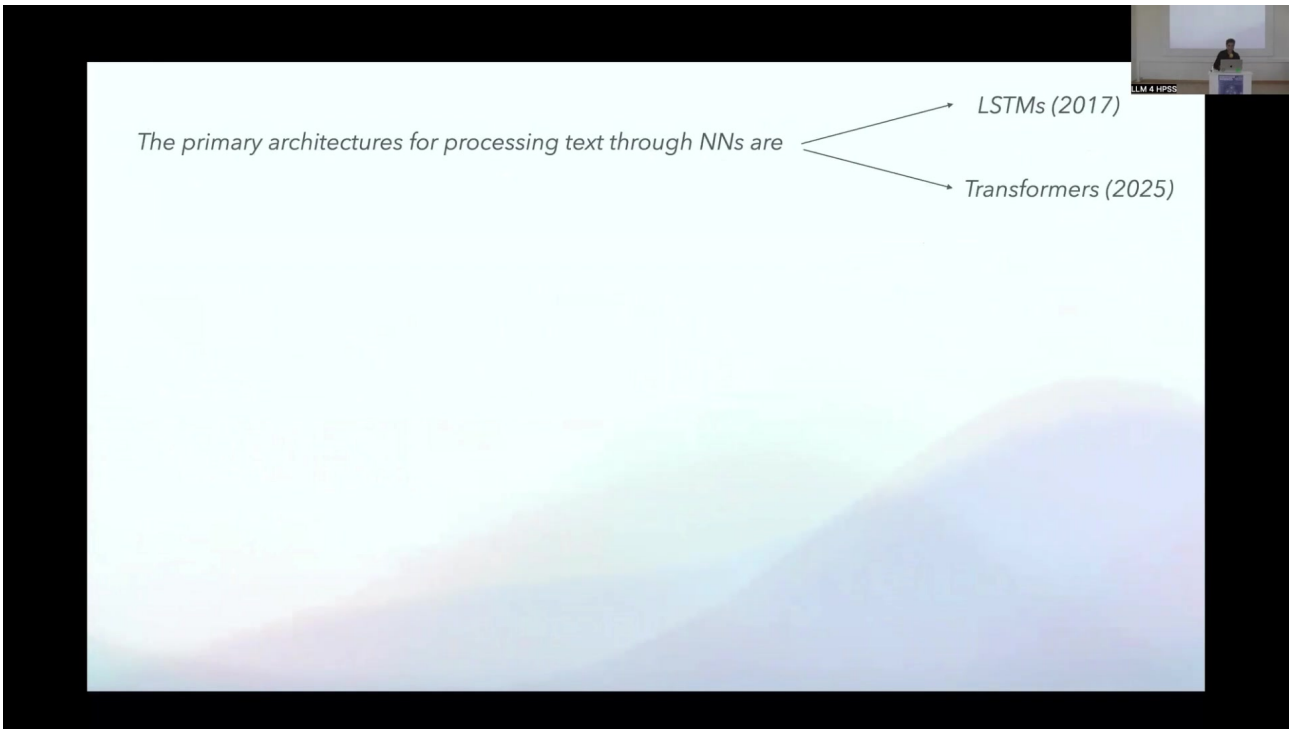

Figure 16.7: Slide 12

The baseline model architecture is a modest-sized *Transformer* decoder. It consists of an Embedding Layer, Positional Encoding, Dropout, four Decoder Layers, and a Final Dense Layer. Each Decoder Layer incorporates Multi-Head Attention, followed by Add & Norm, a Feed-Forward Network (FFN), and another Add & Norm step.

The model has a total of 39 million parameters, occupying approximately 150 MB, which is considerably smaller than large models like *GPT-4*, which has 1.8 trillion parameters distributed across 120 layers. Training is conducted using 2 x A100 GPUs, achieving a speed of 11 seconds per epoch. The code implementation for this model is available on GitHub at https://github.com/j-buettner/time_transformer.

The training process demonstrates that the model learns to reproduce the language style and patterns of the weather report dataset effectively. For instance, given the seed sequence "During the night, a band ...", the model generates text such as "... of rain moved into scotland northern ireland and northern england outbreaks of rain continued to move across northern england and wales it stayed largely dry with clear spells and a few scattered showers in the north and west elsewhere there were plenty of clear spells and a few fog patches and overall it was a mild night across the south of the uk ....". Model performance is tracked using accuracy, visualized in a line graph showing training and validation accuracy over 50 epochs. The training accuracy exhibits a steady increase, while the validation accuracy increases initially before plateauing.

16.9 *Time Transformer* Architecture

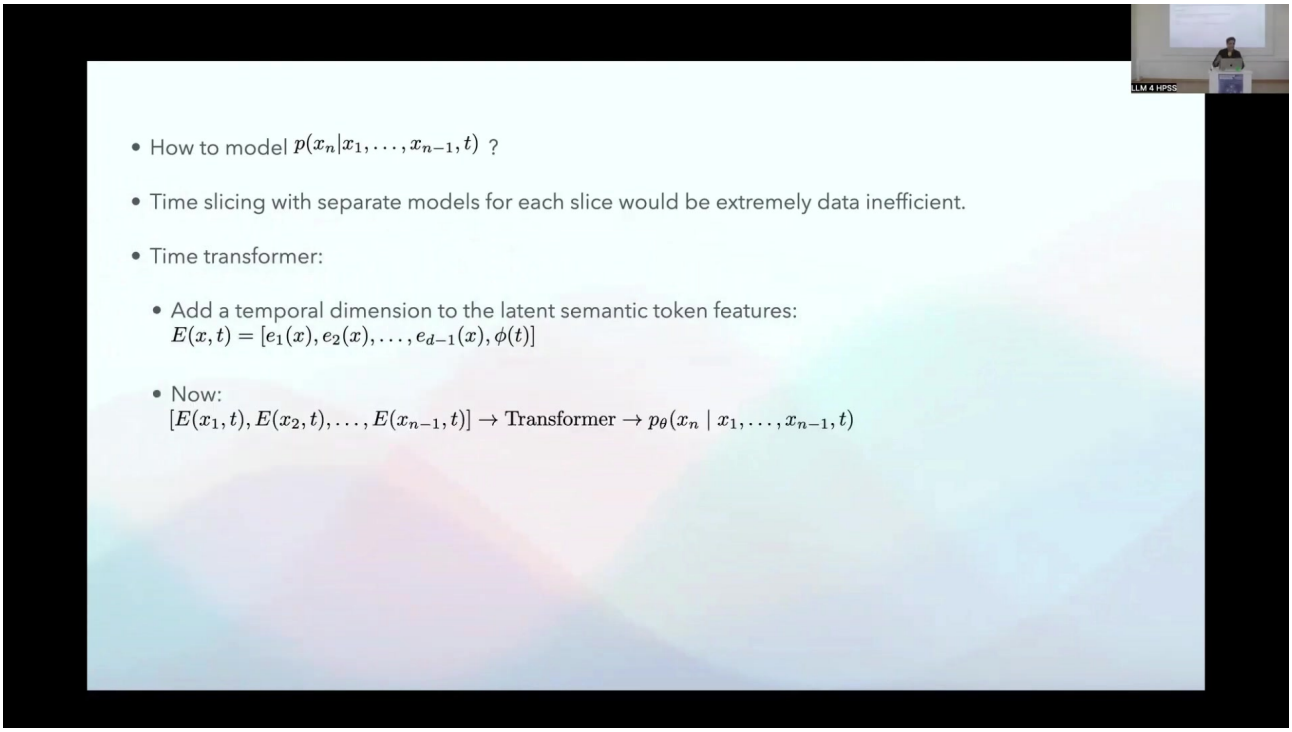

Figure 16.8: Slide 15

The *Time Transformer* architecture is created with a minimal adjustment to the vanilla *Transformer* model. The input now includes both Text Input and Time Data. The Text Input is processed by an Embedding Layer, while the Time Data is processed by a dedicated Time Embedding layer.

The outputs from the standard Embedding Layer and the Time Embedding layer are combined. This combined embedding is then fed into the Casual Masking and Positional Encoding layers. The subsequent layers, including the Decoder Layers and the Final Dense Layer, retain the same structure as the vanilla model. The time dimension is represented as a non-trainable, min-max normalized value corresponding to the day of the year. The time embedding is calculated using the formula `time embedding = (day of year - 1) / (365 - 1)`, normalizing the day of the year (1 to 365) to a range between 0 and 1.

16.10 Experiment 1: Learning Synonymic Succession

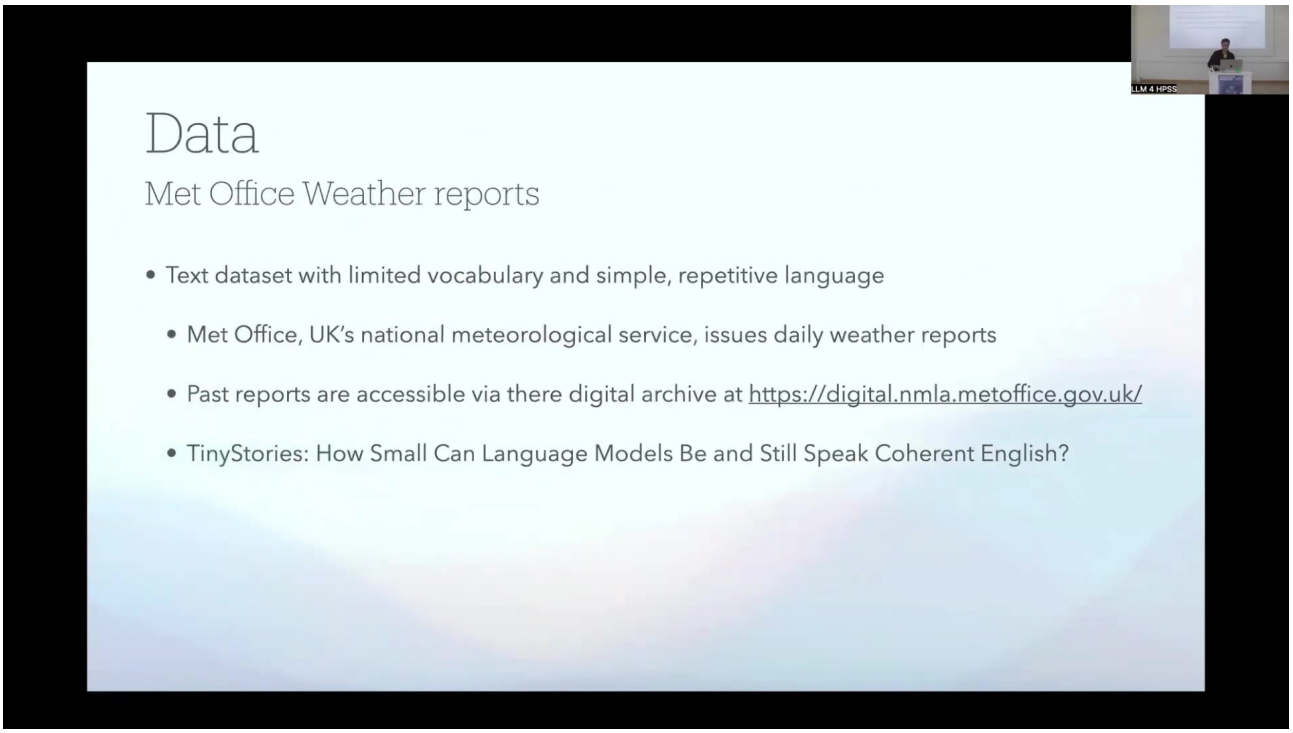

Figure 16.9: Slide 16

The first experiment aims to evaluate the model's ability to efficiently learn temporal drift within the underlying data distribution, specifically focusing on synonymic succession. This involves injecting synthetic drift into the training data by implementing a time-dependent replacement rule: the word "rain" is replaced by the phrase "liquid sunshine". The probability of this replacement occurring follows a sigmoid curve across the days of the year, ranging from Day 0 to Day 365. The probability starts near 0.00 at the beginning of the year and increases to near 1.00 by the end of the year, with the most significant increase occurring around Day 180.

The evaluation method involves checking whether this injected time dependence is accurately reproduced in the token sequences predicted by the model, generating a separate sequence for each day of the year. The presentation also notes observed seasonal patterns in the real weather data, illustrated by bar charts. One chart shows the monthly occurrences of "Rain and Snow" versus "Rain Only", indicating that "Rain and Snow" is more frequent in the later months (September to December), while "Rain Only" is more frequent in the earlier months (January to April). Another chart depicts the monthly occurrences of terms related to heat ("Hot", "Warm") versus terms related to cold and snow ("Snow", "Sleet", "Wintry"), clearly showing that "Snow", "Sleet", and "Wintry" terms are more prevalent in the winter months (January, February, March, November, December), while "Hot" and "Warm" terms are more frequent in the summer months (June, July, August).

16.11 Experiment 2: Learning Changing Co-occurrence

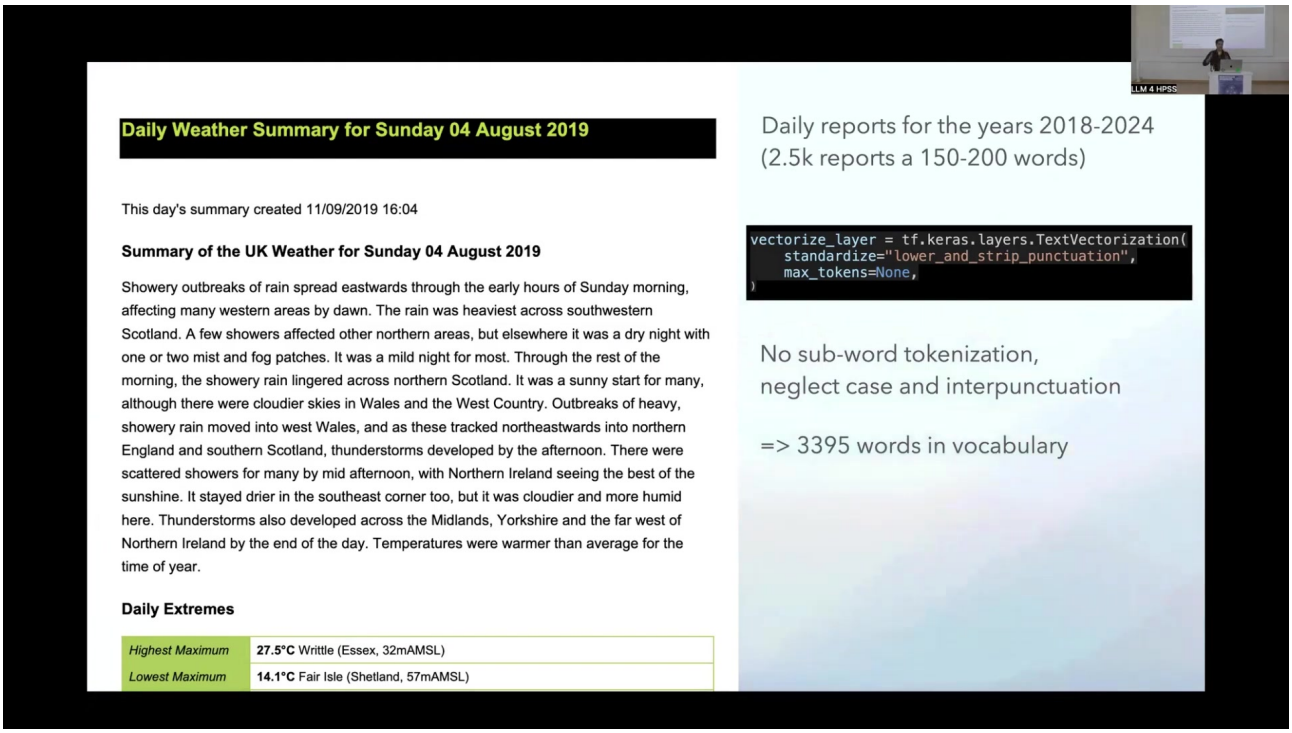

Figure 16.10: Slide 18

The second experiment also aims to evaluate the efficient learning of temporal drift, focusing on a changing co-occurrence pattern, specifically termed 'fixation of a collocation' from a variable to an obligatory relationship. This involves injecting a synthetic time-dependent change into the training data: any instance of the word "Rain" followed by any word other than "and" is synthetically altered to become "rain and snow". This process is described as analogous to the linguistic concept of the fixation of a collocation, citing "*bread and butter*" as a common example.

The evaluation is presented through a bar chart titled "Monthly Comparison of"Rain and Snow" vs. "Rain Only" Occurrences" based on the predicted sequences generated for each day of the year. The y-axis quantifies the occurrences of "rain" only versus "rain and snow" in these predictions. The chart visually demonstrates that the blue bars representing "Rain and Snow" occurrences are higher in the later months (September to December), while the green bars representing "Rain Only" occurrences are higher in the earlier months (January to April), effectively reproducing the injected synthetic drift pattern. Example predicted sequences for Day 1 and Day 363 illustrate this learned pattern, both showing "heavy rain and snow". Further analysis is presented via an attention chart titled "Attention from 'snow' to previous 10 tokens (Head 5)". This bar chart displays the attention weights from the token 'snow' to the preceding 10 tokens. The bar corresponding to "rain" exhibits the highest attention weight, approximately 0.45, indicating that the model has learned a strong associative link between "rain" and "snow" within this specific context.

16.12 Proof of Concept, Applications, and Challenges

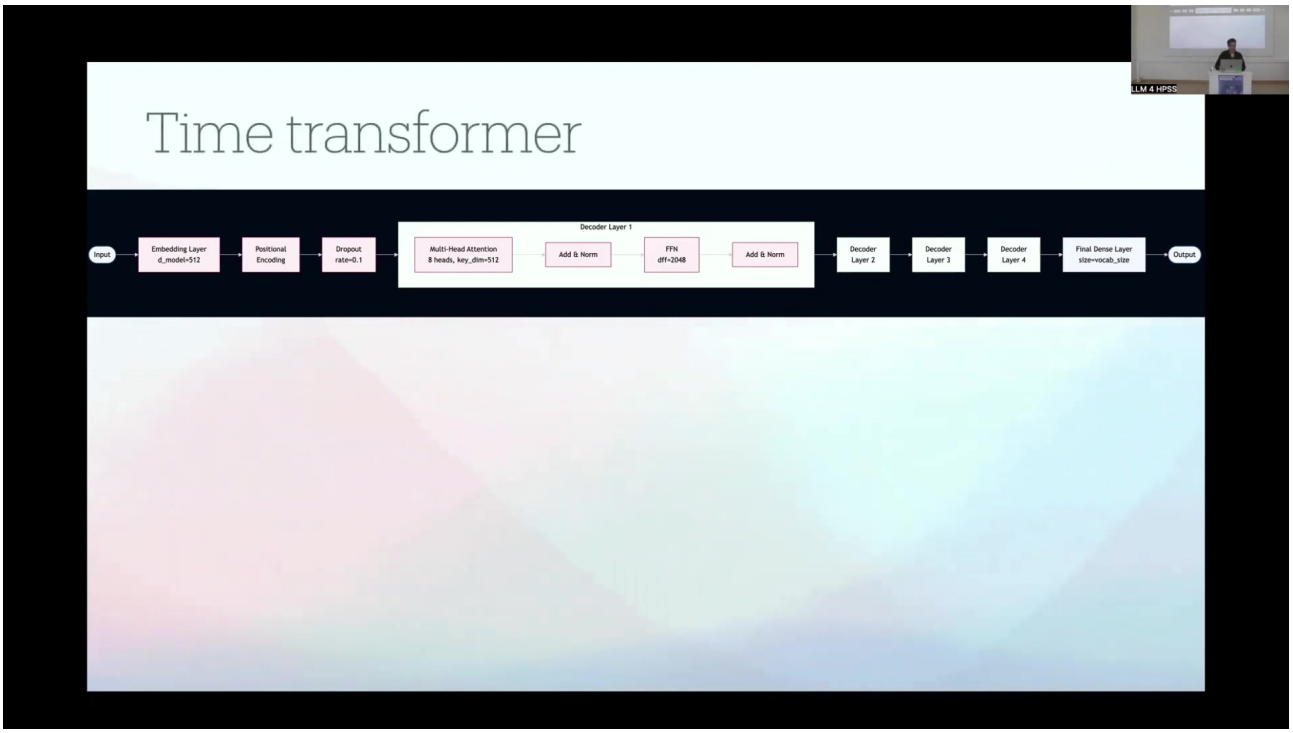

Figure 16.11: Slide 21

The proof of concept demonstrates that *Transformer*-based *Large Language Models* can be efficiently made time-aware by augmenting the token embedding with a temporal dimension. This approach offers several potential applications. A foundational *Time Transformer* could serve as an excellent base model for various downstream tasks involving historical data.

Furthermore, an instruction-tuned *Time Transformer* could enable users to interact with the model by specifying a particular time period, effectively allowing them to "talk to a specific time." This capability might also lead to improved results in standard usage scenarios where the model is expected to reflect the present state of knowledge or language ("talk to the present"). The methodology is also potentially generalizable, suggesting that the dependence of underlying token sequence distributions on other contextual or metadata dimensions, such as country or genre, could be modeled using a similar approach.

Potential next steps for this research include benchmarking the *Time Transformer* against alternative methods, such as an explicit time-token approach, and testing whether the proposed architecture leads to an increase in training efficiency. Further exploration of other aspects is also warranted.

However, several challenges must be addressed for practical application. It is currently unclear whether fine-tuning the modified architecture is possible and efficient. The approach also necessitates significant data curation, particularly the accurate determination of the generation time for each token sequence, which represents a loss of the metadata-free benefit typically associated with self-supervised learning. Accurately timestamping historical data is identified as a key challenge. An alternative direction considered is the development of a modest, targeted encoder model for time-aware tasks.

Chapter 17

LLMs for Chemical Knowledge Analysis

The work focuses on leveraging Large Language Models (LLMs) for metadata enrichment and diachronic analysis of chemical knowledge within a large corpus of historical scientific texts. The primary objective is divided into two parts. Part I involves using LLMs to improve metadata for historical texts, specifically focusing on categorizing articles by scientific discipline and semantic tags (topics), and generating abstractive summaries. Part II presents a case study analyzing the evolution…

17.1 Overview

The work focuses on leveraging Large Language Models (LLMs) for metadata enrichment and diachronic analysis of chemical knowledge within a large corpus of historical scientific texts. The primary objective is divided into two parts. Part I involves using LLMs to improve metadata for historical texts, specifically focusing on categorizing articles by scientific discipline and semantic tags (topics), and generating abstractive summaries. Part II presents a case study analyzing the evolution of the chemical space over time across different scientific disciplines, aiming to identify periods of interdisciplinarity and knowledge transfer.

The research utilizes the *Royal Society Corpus* (RSC) 6.0 Full, a diachronic corpus spanning over 300 years of scientific writing (1665-1996). This corpus contains almost 48,000 texts and nearly 300 million tokens. It was created with steps to improve OCR quality and correct spelling.

For metadata enrichment (Part I), the *Hermes-2-Pro-Llama-3-8B* model, a fine-tuned variant of *Llama 3* optimized for structured output (JSON, YAML), is employed. The LLM acts as a "librarian" following a detailed system prompt that defines its role, objective, input format (OCR text, existing metadata like title, author, date, journal, text snippet), and specific tasks.

The tasks include suggesting alternative titles, writing 3-4 sentence TL;DR summaries for a high school level, identifying exactly five main topics (like Wikipedia keywords), and classifying the primary scientific discipline from a predefined list of nine categories (Physics, Chemistry, Environmental & Earth Sciences, Astronomy, Biology & Life Sciences, Medicine & Health Sciences, Mathematics & Statistics, Social Sciences & Humanities) and a suitable second-level sub-discipline (not from the primary list). The expected output format is YAML.

Validation checks show 99.81% valid YAML output and 94% of predicted primary disciplines fall within the predefined set. Some instances of hallucination were observed, such as "Earth Sciences" instead of "Environmental & Earth Sciences" or inventing "Music" as a discipline. Initial analysis of the LLM-classified data shows discipline distribution over time,

highlighting the rise of chemistry around the late 18th century (chemical revolution) and the prominence of biology, physics, and chemistry from the 19th century onwards. A t-SNE projection of TL;DR summaries visualizes the distribution and overlap of disciplines in semantic space.

For the diachronic analysis of the chemical space (Part II), the focus is on chemistry, biology, and physics. Chemical terms are extracted using *ChemDataExtractor*, a Python module. A two-pass application of *ChemDataExtractor* is used: first on the whole text, then a second pass on the initial list of extracted substances to reduce noise, particularly in earlier periods.

Kullback-Leibler divergence (KLD) is the primary method for analyzing the chemical space evolution. KLD is applied in two ways: independently per discipline to trace evolution over time (comparing 20-year windows sliding by 5 years) and pairwise between disciplines (chemistry vs. physics, chemistry vs. biology) using 50-year periods. Results show similar KLD trends across disciplines with peaks and troughs, and decreasing KLD towards the timeline end, indicating less variation between past and future periods.

Analysis of substance contributions to KLD divergence reveals differences between periods. For example, there was a focus on elements in the late 18th century versus biochemistry in biology and noble/radioactive gases in chemistry/physics in the late 19th century. Pairwise KLD analysis using word clouds confirms thematic differences (biochemical substances in biology, organic chemistry substances in chemistry, metals/noble gases in physics). The pairwise comparison also helps detect "knowledge transfer" cases, where an element's distinctiveness shifts from one discipline to another over time (e.g., tin shifting from chemistry to physics in the 18th century).

Future work includes evaluating the LLM output quality, comparing results with other LLMs, adding more disciplines to the chemical space analysis, and conducting more fine-grained diachronic analysis by adjusting time windows and comparison periods. Challenges include handling historical terminology and OCR quality issues in older texts, and the potential for LLM hallucinations or artifacts in classification. The extracted metadata is intended to feed into knowledge graphs for further structured analysis.

17.2 Introduction and Research Objectives

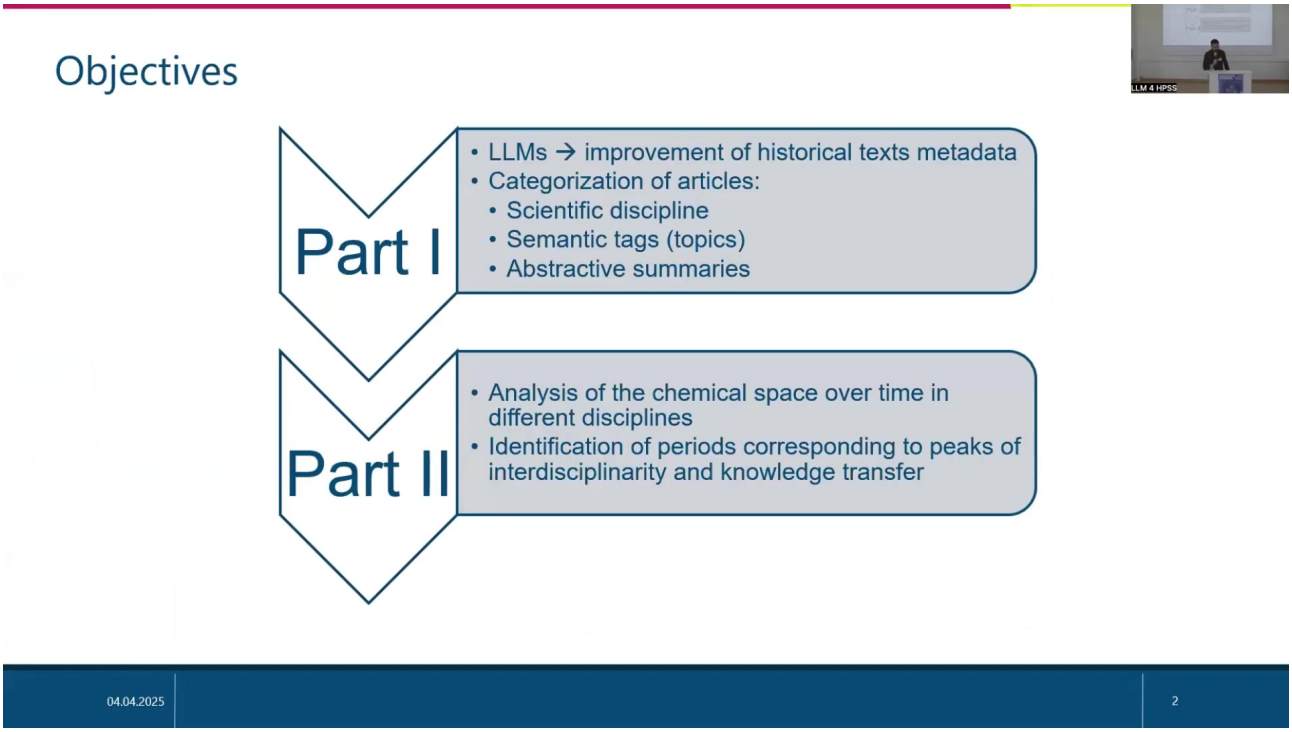

Figure 17.1: Slide 01

The presentation is titled “Leveraging Large Language Models for Metadata Enrichment and Diachronic Analysis of Chemical Knowledge in Historical Scientific Texts”. The authors are Diego Alves, Sergei Bagdasarov, and Badr M. Abdullah, affiliated with the Department of Language Science and Technology at Saarland University. The work is presented at the Large Language Models for the History, Philosophy, and Sociology of Science (LLM 4 HPSIS) workshop.

The research is structured into two main parts:

- Part I investigates the application of Large Language Models (LLMs) to enhance the metadata associated with historical scientific texts. This involves categorizing articles based on their scientific discipline and identifying relevant semantic tags or topics. Additionally, the process includes generating abstractive summaries for the articles.

- Part II constitutes a case study focused on analyzing the evolution of the chemical space across different scientific disciplines over time. The objective is to pinpoint specific periods characterized by heightened interdisciplinarity and significant knowledge transfer between fields.

17.3 Data Source: The Royal Society Corpus

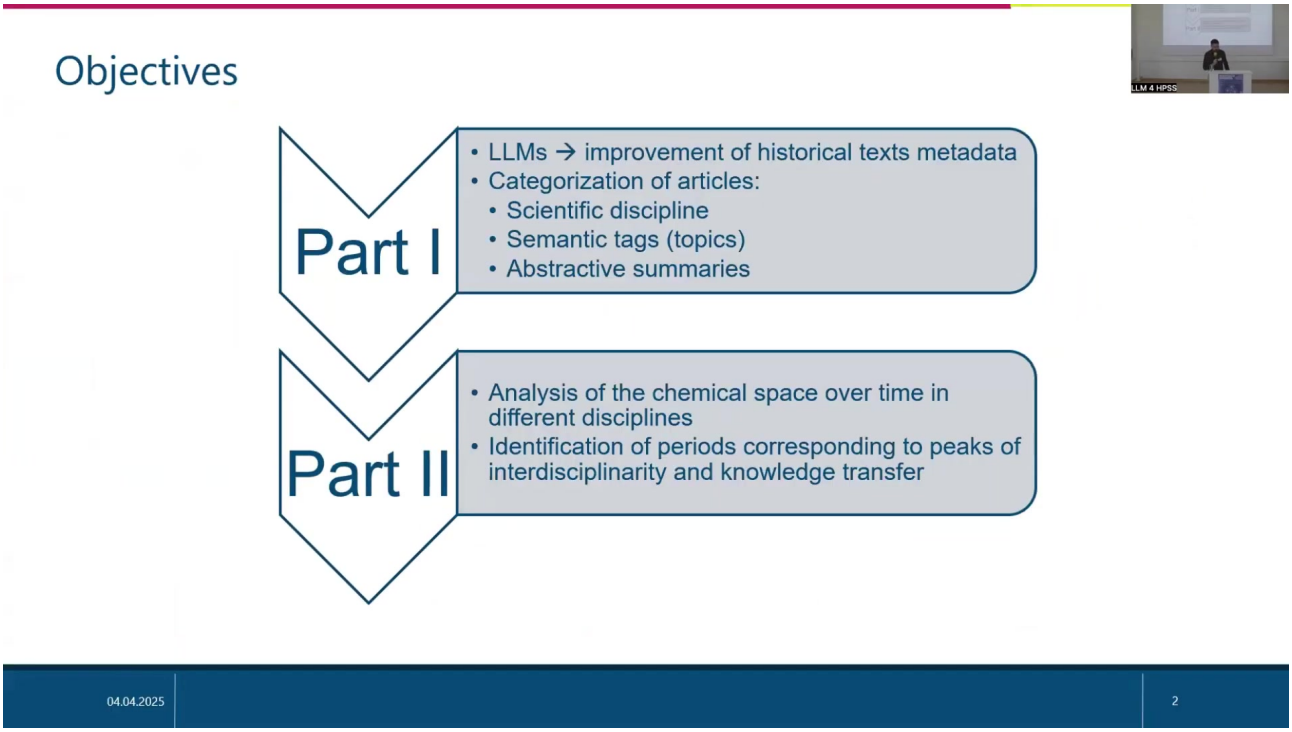

Figure 17.2: Slide 01

The research investigates how scientific English evolved over time, particularly its optimization for expert-to-expert communication. It also analyzes phenomena such as knowledge transfer and the identification of influential papers and authors.

The primary data source is the *Philosophical Transactions of the Royal Society of London*. This journal was first published in 1665 by the *Royal Society of London* and holds the distinction of being the oldest scientific journal in continuous publication, maintaining a high reputation today. It played a pivotal role in the development of scientific communication, notably by establishing the peer-reviewed paper publication model.

The corpus contains numerous influential contributions throughout history. Examples include Isaac Newton's "*New Theory about Light and Colours*" from the 17th century (1672), Benjamin Franklin's description of the "*Philadelphia Experiment*" (the Electrical Kite) in an 18th-century letter (1752), and James Clerk Maxwell's work "*On the Dynamical Theory of the Electromagnetic Field*" from the 19th century (1865). The corpus also includes less conventional papers, such as speculations about inhabitants of the Moon, though the research focuses on linguistic and thematic analysis rather than scientific validity.

The specific version of the corpus utilized is the *RSC 6.0 Full*. This version covers a period exceeding 300 years, from 1665 to 1996. It comprises almost 48,000 individual texts and contains nearly 300 million tokens. The corpus includes some pre-existing metadata attributes such as author, century, year, and volume. A previous study applied LDA topic modeling to infer research field categories, but this approach resulted in categories that sometimes mixed scientific disciplines, sub-disciplines, and types of text like "observations" and "reporting".

17.4 LLMs for Metadata Enrichment

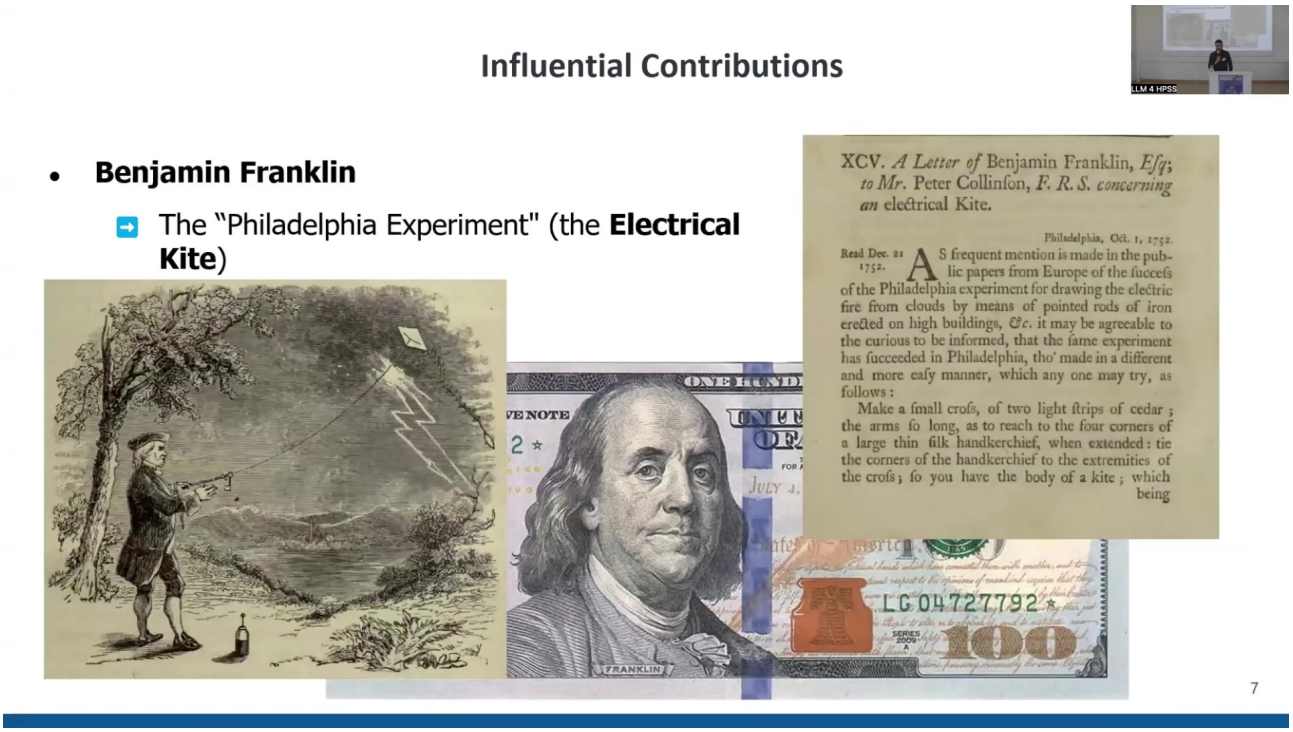

Figure 17.3: Slide 05

The project aims to enhance the existing metadata and generate new metadata for the corpus using Large Language Models (LLMs). LLMs are applied for various information management tasks including text clean-up, summarization, facilitating access and retrieval, information extraction, categorization, and ultimately, feeding knowledge graphs.

For each article in the corpus, the desired outputs from the LLM processing include a hierarchical categorization specifying the primary discipline and a sub-discipline, a list of index terms or semantic tags representing the main topics, and a concise TL;DR summary.

The specific LLM utilized for this task is *Hermes-2-Pro-Llama-3-8B*, which belongs to the *Llama 3* family developed by *Meta*. *Llama 3* is available in different parameter sizes, including 8 billion (8B) and 70 billion (70B), with a 400 billion (400B) version currently under training. The model is accessible via *Hugging Face* and is reported to perform significantly better than previous models like *Mistral* and *Llama 2* for instruction-following tasks. The *Hermes-2-Pro-Llama-3-8B* variant is specifically fine-tuned with instruction tuning, making it particularly adept at producing structured output formats such as JSON and YAML.

The LLM is instructed to act as a "librarian" responsible for organizing a collection of historical scientific articles from the *Royal Society of London* published between 1665 and 1996. Its objective is to read, analyze, and organize this large corpus to create a structured database that facilitates search, retrieval, and analysis by researchers, historians, and scientists. The input provided to the LLM consists of OCR-extracted text of the original articles along with existing metadata like the title, author(s), publication date, journal, and a short text snippet. The prompt acknowledges potential issues with OCR quality in the historical texts.

The LLM is assigned four specific tasks:

- A. Read and analyze the article to understand its content and context, then suggest an alternative title that better reflects the content.

- B. Write a short, 3-4 sentence TL;DR summary that captures the article's essence and main findings. The summary must be concise, informative, highlight key points, and be written in simple language suitable for a high school student.

- C. Identify exactly five main topics for the article. These topics are conceptualized as Wikipedia Keywords for categorizing the text into scientific sub-fields or thematic groups within a scientific journal.

- D. Based on the identified topics, determine the primary scientific discipline from a predefined list of nine categories: Physics, Chemistry, Environmental & Earth Sciences, Astronomy, Biology & Life Sciences, Medicine & Health Sciences, Mathematics & Statistics, and Social Sciences & Humanities. The LLM must also identify a suitable second-level sub-discipline that is a branch of the primary discipline and not one of the primary disciplines itself.

The required output format is a valid YAML file. An example input based on an article by Isaac Newton is provided, showing the existing metadata and a text snippet. A corresponding example output in YAML format demonstrates the expected structure, including the article ID, revised title, a list of five topics (e.g., Optics, Refraction, Spectroscopy), the TL;DR summary, the primary scientific discipline (e.g., Physics), and the scientific sub-discipline (e.g., Optics & Light). The prompt explicitly states that the output must be valid YAML and contain no additional text.

Validation checks were performed on the LLM's output. A high percentage, 99.81%, of the produced outputs were valid YAML files (17486 out of 17520). Furthermore, 94% of the predicted primary scientific disciplines fell within the predefined set of nine categories. However, some instances of hallucination or incorrect assignments were observed. These included predicting "Earth Sciences" instead of the full "Environmental & Earth Sciences" category, inventing novel categories such as "Music", and occasionally including the index number from the predefined list as part of the discipline string (e.g., "3. Environmental & Earth Sciences"). The LLM also sometimes assigned sub-disciplines like "Neurology" or "Zoology" as primary disciplines. Despite these issues, the majority of papers were correctly assigned to the predefined primary disciplines.

Initial analysis was conducted using the LLM-classified data. A stacked area chart visualizes the distribution of scientific disciplines over time across the corpus. This analysis reveals a more homogeneous distribution of disciplines up to the end of the 18th century. A notable peak in chemical articles is observed in the late 18th century, which is associated with the historical chemical revolution. From the 19th century onwards, biology, physics, and chemistry emerge as the three main pillars represented in the Royal Society's publications.

A first analysis of the generated TL;DR summaries was performed using t-SNE projection to visualize the semantic space. The t-SNE plot shows the distribution of articles based on their summary texts, colored by their assigned discipline. It indicates overlap between chemistry, physics, and biology, with chemistry appearing centrally located in the overlapping region. Disciplines such as humanities, astronomy, and mathematics tend to form more isolated clusters in this projection space. The analysis can be performed diachronically to observe shifts and changes in the semantic overlap between disciplines over time.

17.5 Diachronic Analysis of the Chemical Space

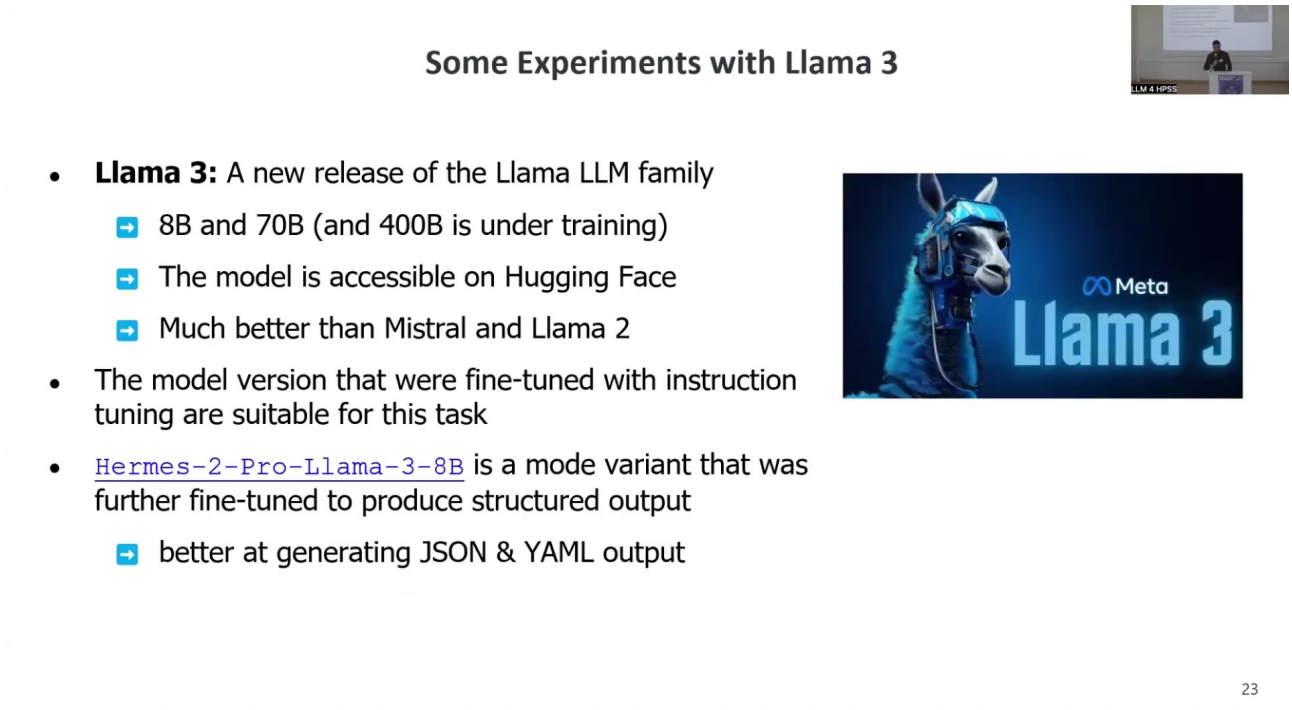

Figure 17.4: Slide 13

Part II of the research presents a diachronic analysis of the chemical space, focusing specifically on three disciplines: Chemistry, Biology, and Physics, as these are the most frequently represented in the corpus.

Chemical terms are extracted from the texts using *ChemDataExtractor*, a Python module designed for the automatic identification of chemical substances. The application of *ChemDataExtractor* involved a two-pass method to mitigate noise, particularly prevalent in earlier periods where the initial pass on the whole text tagged non-chemical entities (like animals) as chemical terms. The second pass applied *ChemDataExtractor* specifically to the list of substances identified in the first pass, which helped to reduce the amount of noisy output.

Kullback-Leibler divergence (KLD) is the method employed to analyze the evolution and comparison of the chemical space. KLD is applied in two distinct ways. The first method involves an independent analysis for each discipline to trace the evolution of its chemical space along the timeline. This is done by comparing the word distributions of chemical terms within two time windows: a period of 20 years before a specific date is compared with a period of 20 years after that date. This comparison window is then slid along the timeline in 5-year increments. This process yields KLD values for each discipline over time, illustrating internal changes in their chemical vocabulary.

The second method involves pairwise comparisons between disciplines, specifically comparing chemistry with physics and chemistry with biology. This analysis is based on 50-year periods. The output consists of KLD values that quantify the difference in chemical term distributions between the two disciplines within those periods.

The results from the independent KLD analysis per discipline show a similar trend across chemistry, biology, and physics, with peaks and troughs occurring in roughly the same historical periods. Towards the end of the timeline, the KLD plots become flatter, and the overall KLD decreases. This indicates less variation in the chemical space between past and future periods within each discipline during later years.

Analysis focusing on the peak observed in the late 18th century reveals that KLD allows for zooming in to identify the specific chemical substances contributing most significantly to the divergence. In biology and physics during this period, one or two elements exhibit extremely high KLD values, acting as primary drivers of change. Across chemistry, biology, and physics, the same elements are observed to be responsible for the changes seen in the late 18th century.

The analysis of a later period, the second half of the 19th century, shows significant changes. The graphs for biology and physics become much more populated, indicating a wider range of substances contributing to the divergence. The individual contributions of elements are also more uniform. Thematic differences in substances emerge: biology's chemical space evolves towards biochemistry, while chemistry and physics show a focus on noble gases and radioactive elements, which were discovered towards the end of the 19th century.

Pairwise comparisons, such as between chemistry and biology or chemistry and physics in the second half of the 20th century, further confirm these thematic differences. Word clouds generated from the distinctive substances in these comparisons show that the biology word cloud contains more substances related to biochemical processes in living organisms, while the chemistry word cloud features substances associated with organic chemistry, such as hydrocarbons and benzene. Comparing chemistry and physics reveals more metals, noble gases, and various types of metals, including rare earth, semi-metals, and radioactive metals, in the physics word cloud.

The pairwise comparison method also proves useful for detecting instances of "knowledge transfer". This is defined as a case where an element is ranked as distinctive of one discipline in an earlier period but becomes more distinctive of another discipline later in time. An illustration using tin (Sn) shows its distinctiveness shifting from chemistry to physics between the first and second halves of the 18th century. Similar shifts are observed for other elements in the early 20th century. Elements becoming distinctive of biology in the 20th century are typically related to biochemical processes.

17.6 Conclusion and Future Work

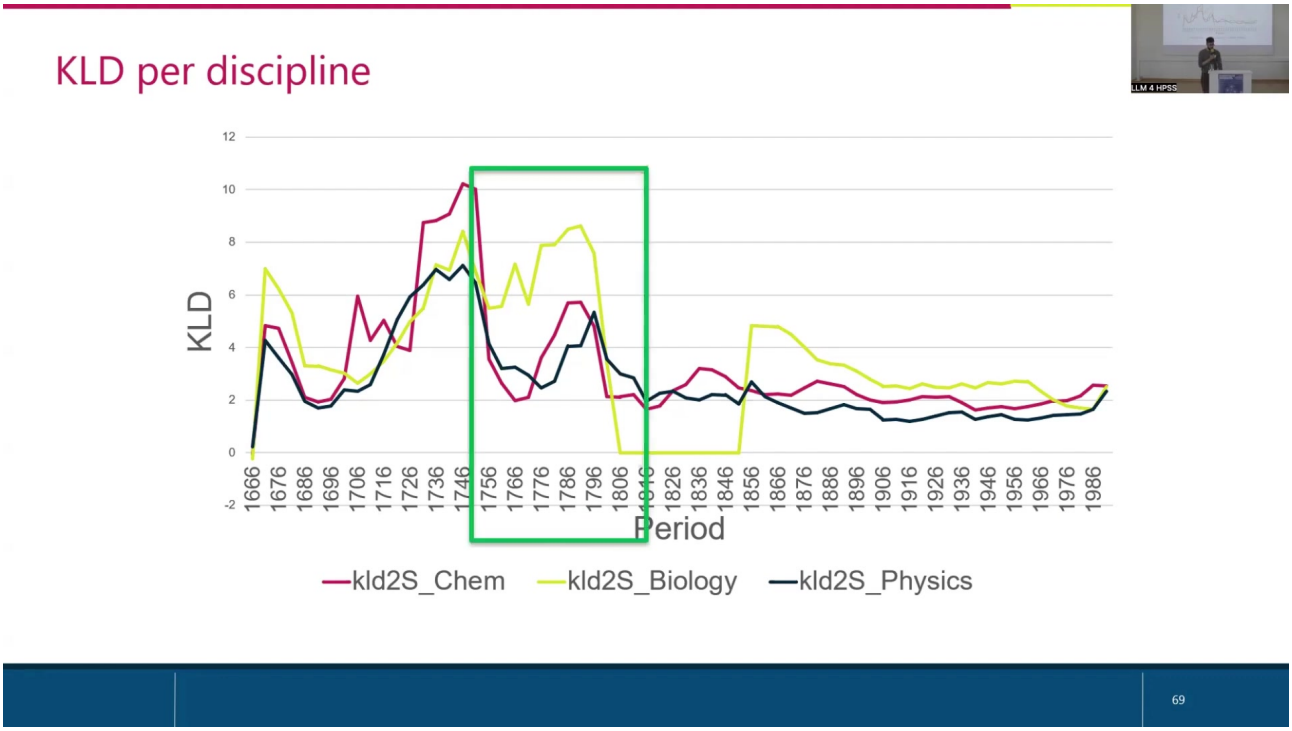

Figure 17.5: Slide 22

In conclusion, the research successfully utilized a Large Language Model (LLM) to improve the categorization and topic modeling of texts within the corpus. Building upon the metadata generated by the LLM, a diachronic analysis of the chemical space was conducted across three disciplines.

Future work includes several areas for improvement and further exploration:

- It is necessary to evaluate the quality of the output produced by the *Llama* model to assess its accuracy and reliability.
- Comparing the results obtained from *Llama* with those from other LLMs is also planned to understand model-specific variations.
- The chemical space analysis can be expanded by including more disciplines, such as conducting a direct comparison between chemistry and biology.
- Furthermore, a more fine-grained diachronic analysis is intended, which will involve experimenting with different time sliding windows and comparison periods for the Kullback-Leibler divergence calculations.

Chapter 18

Interpretable Models for Linguistic Change

The presentation details a research project focused on modeling context and the interplay between different types of context to trace linguistic change, specifically in English scientific writing. The project utilizes methods from both traditional linguistic analysis and deep learning. The core objective is to develop interpretable models that bridge these approaches to understand how language changes over time and across different contextual dimensions. The research investigates the chem...

18.1 Overview

The presentation details a research project focused on modeling context and the interplay between different types of context to trace linguistic change, specifically in English scientific writing. The project utilizes methods from both traditional linguistic analysis and deep learning. The core objective is to develop interpretable models that bridge these approaches to understand how language changes over time and across different contextual dimensions. The research investigates the chemical revolution period (1760s-1820s) in the Royal Society Corpus (RSC) as a case study, focusing on the shift from the phlogiston theory to the oxygen theory.

Previous work involved modeling context using separate approaches. The current work aims to combine these approaches and analyze their interactions. The theoretical framework draws upon language variation and register theory (Halliday 1985, Biber 1988), which posits that situational context determines language use and linguistic context exhibits variation. It also incorporates principles from rational communication and information theory (Jaeger and Levy 2007, Piantadosi et al. 2011), suggesting that linguistic variation modulates information content for efficient communication.

Methods for detecting periods of change include continuous comparison using Kullback-Leibler Divergence (KLD) on probability distributions of linguistic units (words, POS trigrams) over time (Degaetano-Ortlieb and Teich 2018, 2019). This method identifies periods of increased divergence, indicating significant linguistic shifts. Analysis of *what* changes involves examining the specific lexical items and grammatical patterns contributing to high KLD, revealing "waves of increased expressivity" potentially linked to new concepts. The effects of change are observed across linguistic levels, including lexical items (lemmas) and grammatical units (POS trigrams).

Paradigmatic context and change are analyzed using semantic space models (Fankhauser et al. 2017, Bizzoni et al. 2019), visualizing semantic similarity and frequency of terms like "phlogiston" and "oxygen" across different time periods. Identifying *who* leads or spreads change utilizes *Cascade models* (*Hawkes processes*) (Bizzoni et al. 2021), which model influence spread within a network, identifying innovators (e.g., Priestley) and spreaders (e.g., Pearson).

Investigating *how* change is realized linguistically and *why* it occurs from a communicative perspective involves analyzing Surprisal (Shannon 1949), which correlates with cognitive effort (Hale 2001, Levy 2008, Crocker et al. 2016). Linguistic structures that reduce surprisal and encoding effort, such as shifts from prepositional phrases ("consumption of oxygen") to compounds ("oxygen consumption"), are analyzed over time in relation to community adoption (number of authors).

The proposed framework for modeling context for language variation and change addresses limitations of current methods (e.g., static network approaches) by treating context as a central signal. It proposes using *Graph Convolutional Networks* (*GCNs*) for modeling complex relational data. A pilot study on the chemical revolution outlines a multi-stage process:

- Data Sampling: Using the RSC, applying tf-idf and KLD to identify keywords in the target period (1760s-1820s).

- Network Construction: Building time-aware networks. This involves creating word- and time-aware feature vectors using *BERT* for word embeddings and one-hot encoding for categorical metadata (author, journal, period). Node feature matrices are created for 20-year periods. Change in node features is measured using KLD across periods, resulting in a diachronic series of graphs. Network size is managed using community detection algorithms (e.g., Riolo & Newman 2020).

- Link Prediction: Predicting how, when, and by whom words are used. Word profiles are augmented with semantic embeddings (from *BERT*), contextual metadata (author, journal, period), and grammatical information (POS, syntactic role). A *Transformer-GCN* model learns patterns in these profiles to predict new links, with *GCN* capturing structural relationships and *Transformer* attention highlighting influential contextual features.

- Entity Alignment: Inspecting and interpreting change. This involves identifying Network Motifs (small, overrepresented subgraphs) using the *Kavosh algorithm*, which groups isomorphic graphs to find motifs across networks. Entity alignment (e.g., using *GCNs* for tasks like aligning concepts across different datasets or languages) is a future perspective.

Limitations and perspectives include computationally tracing conceptual versus linguistic change, integrating metadata as a core signal, determining the optimal unit of language change (word, concept, grammar, discourse), identifying recurring linguistic pathways for concept emergence, and ensuring interpretability of complex models. Future work includes expanding to multilingual corpora (e.g., French, German journals) and other text types (letters, monographs) and investigating the expression of attitude (positive/negative) towards concepts within the network structure.

18.2 Context and Theoretical Framework

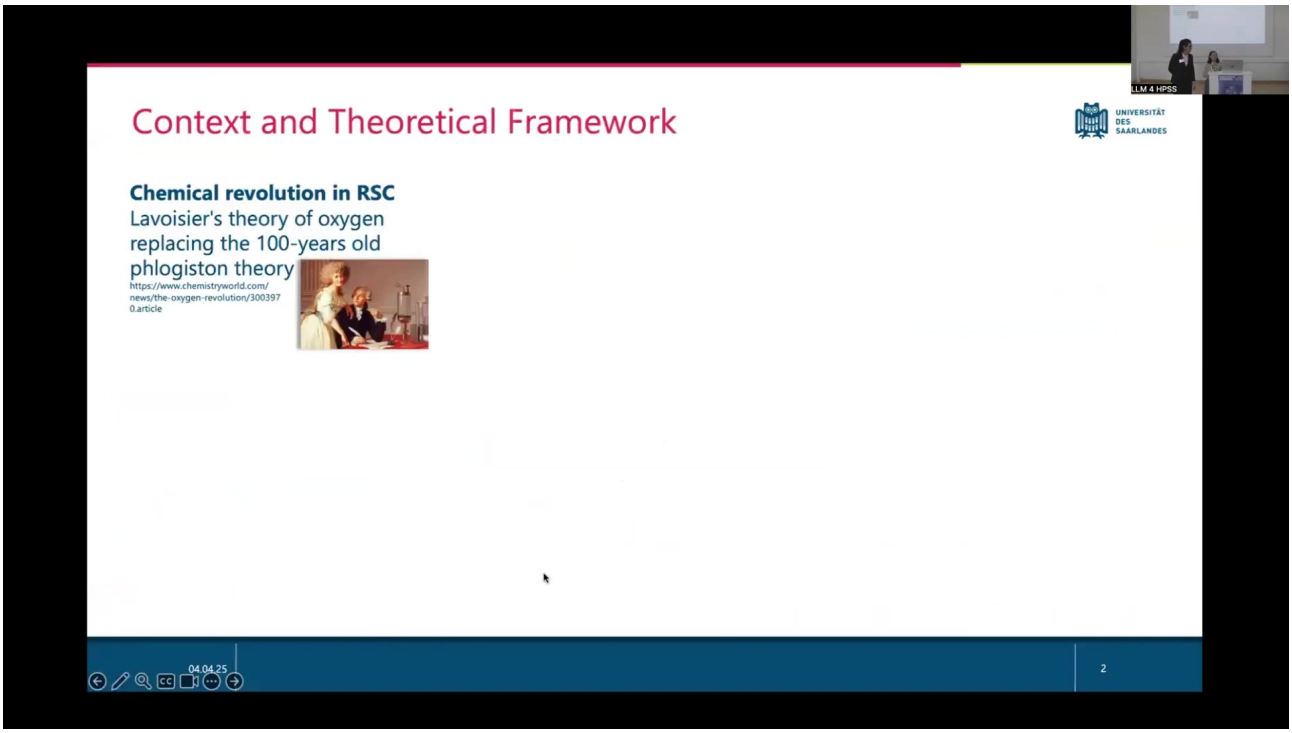

Figure 18.1: Slide 01

The research focuses on the computational analysis of semantic change across different environments, specifically modeling context and the interplay between various types of context. A key case study involves the chemical revolution as documented in the Royal Society Corpus (RSC). This historical event centers on the replacement of the 100-year-old phlogiston theory by Lavoisier's theory of oxygen. Previous research efforts modeled context using separate approaches, while the current work aims to combine these methods to analyze the interactions between different contextual dimensions.

The framework identifies six key types of context: Situational (Where), Temporal (When), Experiential (What), Interpersonal (Who), Textual (How), and Causal (Why). The theoretical foundation draws upon two main areas. Firstly, language variation and register theory, as described by Halliday (1985) and Biber (1988), posits that situational context dictates language use and that linguistic context inherently exhibits variation. Examples of such variation include phrases like "…air which was dephlogisticated…", "…dephlogisticated air…", and "…oxygen…". Secondly, rational communication and information theory, developed within the IDeaL SFB 1102 project and referenced in works by Jaeger and Levy (2007) and Piantadosi et al. (2011), suggests that linguistic variation serves to modulate information content, leading to optimization effects that facilitate efficient communication with reasonable effort.

18.3 Detecting Linguistic Change

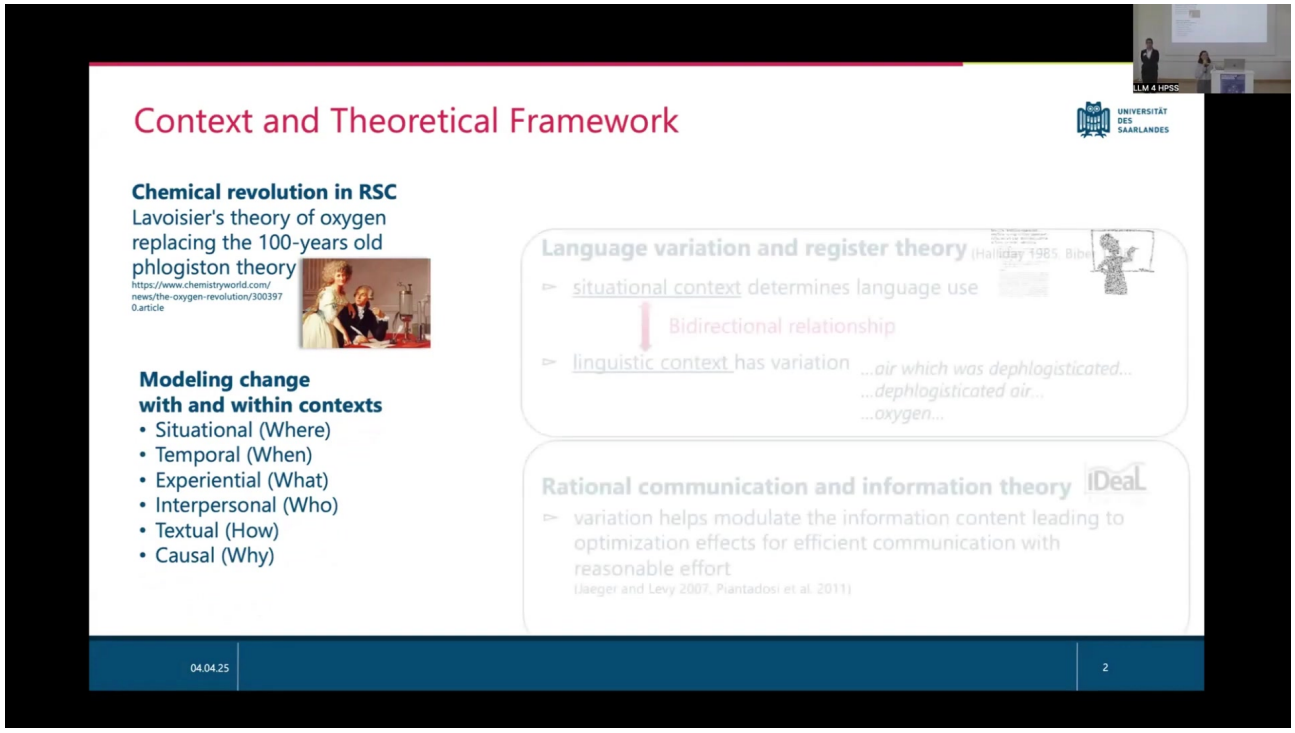

Figure 18.2: Slide 03

The research addresses the problem of detecting periods of change in language use by identifying these periods directly rather than relying on comparisons between predefined time segments. The primary method for detecting change utilizes Kullback-Leibler Divergence (KLD). This approach compares the probability distributions, p(unit|context), of linguistic units over time using a continuous comparison method (Degaetano-Ortlieb and Teich 2018, 2019). The interpretation of KLD values is direct: similar distributions result in low divergence, while differing distributions yield higher divergence. The continuous comparison employs a sliding time window, for instance, comparing a 20-year period designated as "PAST" with the subsequent 20-year period labeled "FUTURE".

To analyze *what* changes, the method plots KLD over time for various linguistic items, including both lexical items and POS trigrams. Peaks observed in these KLD plots are interpreted as "waves of increased expressivity," suggesting the emergence of new concepts or significant shifts in the linguistic treatment of existing ones. The analysis includes a wide range of lexical items such as "electricity", "electrify", "'s", "limb", "ditto", "air", "dephlogisticated experiment", "nitrous", "acid", "gas", "oxide", "be", "hydrogen", "current", "urine", "cell", "corpuscule", "glacier", "tide", "the", "of", "sin", and "cos". A specific period of interest, approximately from 1765 to 1805, is highlighted, encompassing terms like "dephlogisticated experiment", "nitrous", "acid", "air", "gas", "oxide", "be", "hydrogen", "current", "urine", and "cell". This period aligns with significant historical events like the discovery of hydrogen (inflammable air) by Henry Cavendish in 1766 and the discovery of oxygen (dephlogisticated air) by Joseph Priestley in 1774.

The analysis observes effects across different linguistic levels. KLD is applied to lexical items, using the lemma as the unit of analysis, and also to grammatical units, specifically POS trigrams. The findings indicate that peaks in KLD for lexical items, occurring around 1775-1805, correspond roughly to peaks observed in the KLD analysis of POS trigrams. Examples of POS trigrams analyzed include "NN NN IN (zenith distance of)", "VBZ JJR IN (is greater than)", "DT NN IN (the end of)", "NN NN NN (thunder and lightning)", "IN JJ NN (of dephlogisticated air)", "DT NNS IN (the effects of)", "NN

NN DT (oxide of iron)", "NN NN IN (the quantity/number of)", "VBZ JJR IN (is greater than)", "NN NN IN (unite edge of)", and "IN DT JJ (for the same)". This suggests that linguistic change during this period manifested across both vocabulary and grammatical structure.

18.4 Paradigmatic Context and Influence

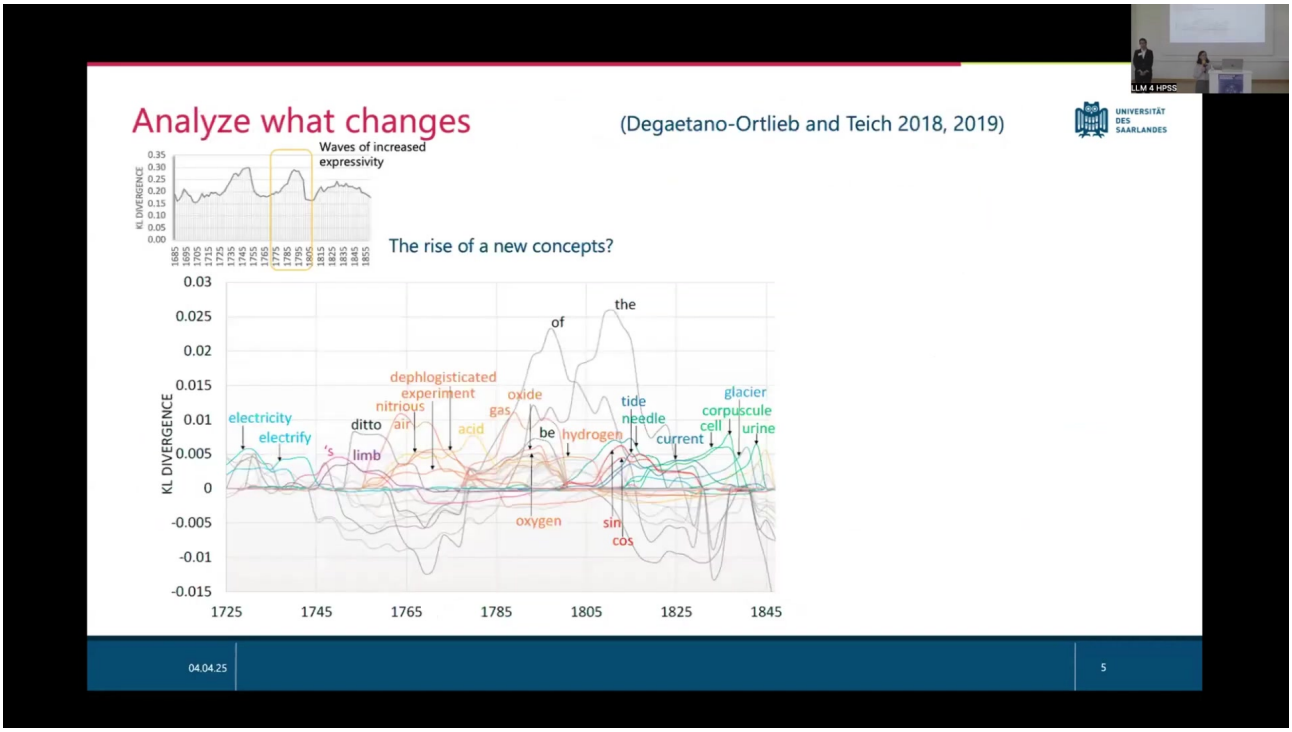

Figure 18.3: Slide 06

The analysis of paradigmatic context and change employs methods described by Fankhauser et al. (2017) and Bizzoni et al. (2019). This technique involves visualizing semantic space at different time periods, such as 1780, 1800, and 1840. Terms are represented as points within this space, where their position indicates semantic similarity. The visualizations provide additional details: the size of the circle representing a term indicates its relative frequency, and color is used to represent clusters of terms. Observing the shifts in term positions and clustering over time reveals semantic change, exemplified by the appearance of "oxygen" and the changing position and frequency of terms like "phlogiston" and "dephlogisticated". The corpora used for this analysis are available at corpora.ids-mannheim.de.

To identify *who* is leading or spreading change, the research utilizes *Cascade models*, specifically *Hawkes processes*, as detailed by Bizzoni et al. (2021). These models are applied to model the spread of influence or linguistic innovations within a network, such as a network of authors. The models enable the identification of individuals acting as "Innovators," such as Priestley, and those acting as "Early Adopters" or "Spreaders," including Pearson and Davy. The results are visualized using a heatmap that shows author influence over time. In this visualization, the color intensity represents the degree of influence, and dashed lines are used to indicate the spread of influence across different time points.

18.5 Linguistic Realization and Communicative Perspective

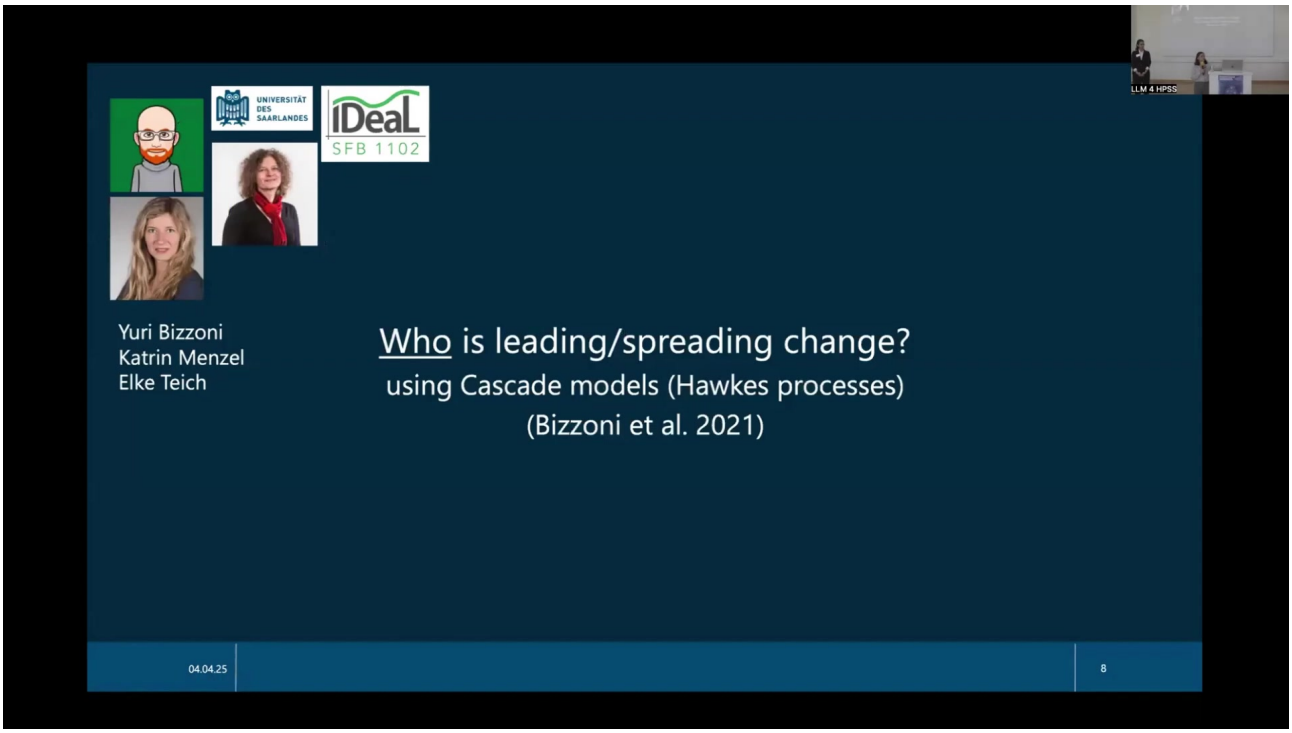

Figure 18.4: Slide 09

The research investigates how linguistic change is realized and why these changes occur from a communicative perspective. The approach involves analyzing change within the linguistic context using the concept of Surprisal, as introduced by Shannon (1949). The underlying principle is that the surprisal of a linguistic unit is proportional to the cognitive effort required to process it, a relationship supported by work from Hale (2001), Levy (2008), and Crocker et al. (2016). A core hypothesis is that linguistic changes occur to reduce cognitive effort and facilitate more efficient communication.

The analysis tracks the surprisal of different linguistic structures over time. Examples of structures examined include clausal forms like "…the oxygen (which was) consumed", prepositional phrases such as "…the consumption of oxygen…", and compound forms like "…the oxygen consumption…". The observation is that the surprisal of longer, more effortful structures, exemplified by the prepositional phrase "Prepositional consumption of oxygen", tends to decrease over time. This decrease coincides with the emergence and establishment of shorter, less effortful structures, such as the compound "Compound oxygen consumption", within the linguistic community. A correlation is observed between the decrease in surprisal for these shorter forms and an increasing number of authors adopting and using them. This process indicates that shorter encoding emerges and becomes established in the community, effectively reducing cognitive effort as reflected by lower surprisal values. This work is related to the MA thesis of Viktoria Lima-Vaz (2025) and a submission by Degaetano-Ortlieb et al. (July 2024).

18.6 Framework for Context and Language Dynamics

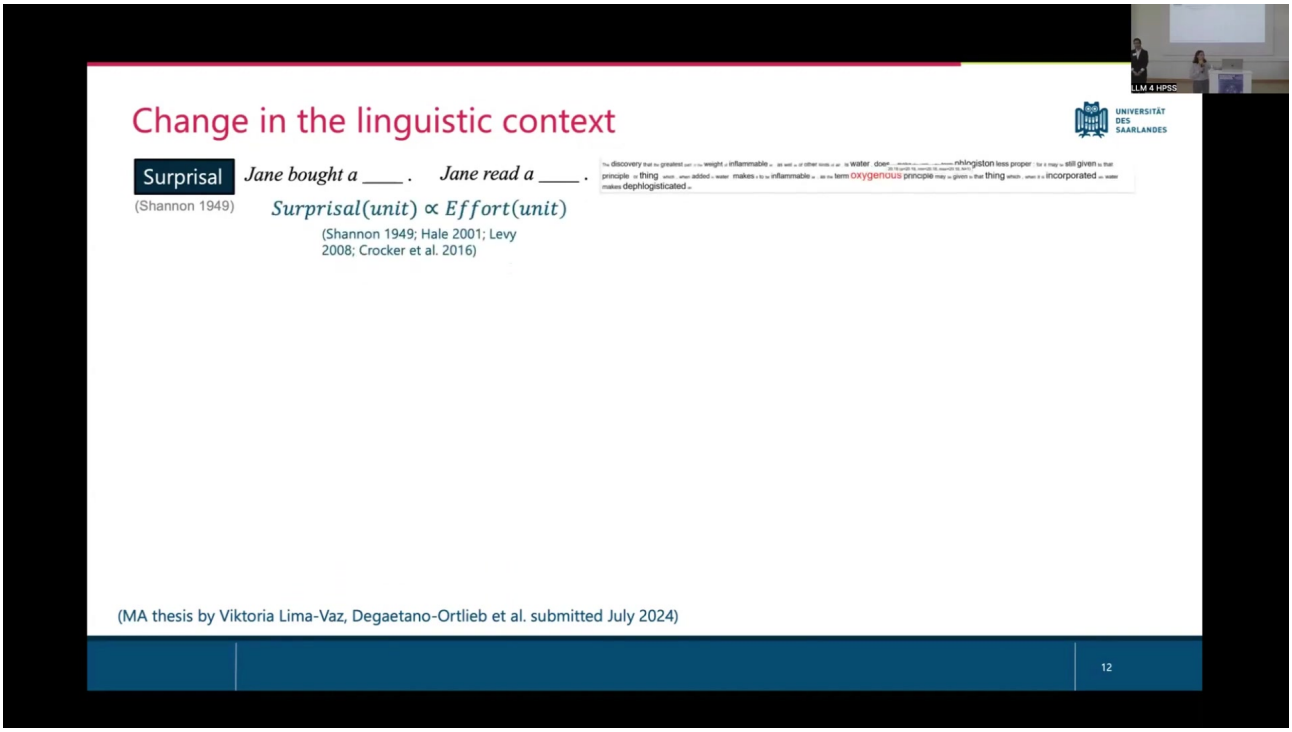

Figure 18.5: Slide 12

The proposed framework for modeling context for the analysis of language variation and change is motivated by the understanding that language change is driven by shifts in social context, including evolving goals, social structures, and domain conventions. Current limitations in the field include that existing semantic change studies and KLD applications often track shifts but do not adequately model the interaction between various contextual signals. Furthermore, static network approaches are limited in their ability to capture dynamic interactions over time.

The proposed framework posits that context serves as a central signal for modeling language dynamics. *Graph Convolutional Networks* (*GCNs*) are proposed as one possible technological direction due to their powerful capability for modeling complex relational data. A pilot study focusing on the chemical revolution is outlined, utilizing the Royal Society Corpus (RSC) and targeting the period between the 1760s and 1820s.

18.6.1 Stage I: Data Sampling

This stage employs methods such as tf-idf and KLD to identify relevant keywords within the target period. KLD is used to define words that are distinct for each period, with words contributing highly to KLD being considered relevant.

18.6.2 Stage II: Network Construction

This stage aims at building word- and time-aware feature vectors. This involves using *BERT* for generating word vectors and one-hot encoding for categorical metadata such as Author, Journal, and Period. A node feature matrix is created for each 20-year period. Change in these node feature vectors is measured using KLD to assess dissimilarity across periods, resulting in a diachronic series of graphs. To manage network complexity, community detection algorithms, such as those described by Riolo & Newman (2020), are used for network size definition.

18.6.3 Stage III: Link Prediction

This stage seeks to predict how, when, and by whom words are used. This involves using word profiles augmented with semantic embeddings (e.g., from *BERT*), contextual metadata (e.g., author, journal, period), and grammatical information (e.g., part of speech, syntactic role). A *Transformer-GCN* model is employed, which learns patterns in these augmented profiles and predicts new links. The *GCN* component captures structural relationships within the network, while the *Transformer* attention mechanism highlights the most influential contextual features.

18.6.4 Stage IV: Entity Alignment

This stage is designed to inspect and interpret the observed change. It utilizes Network Motifs, defined as small, overrepresented subgraphs that reflect meaningful interaction structures. The *Kavosh algorithm* is used, which groups isomorphic graphs to identify these motifs within networks. Entity alignment is also considered as a future application, potentially involving tasks like aligning similar graphs across different time periods (t1, t2).

18.7 Limitations and Future Work

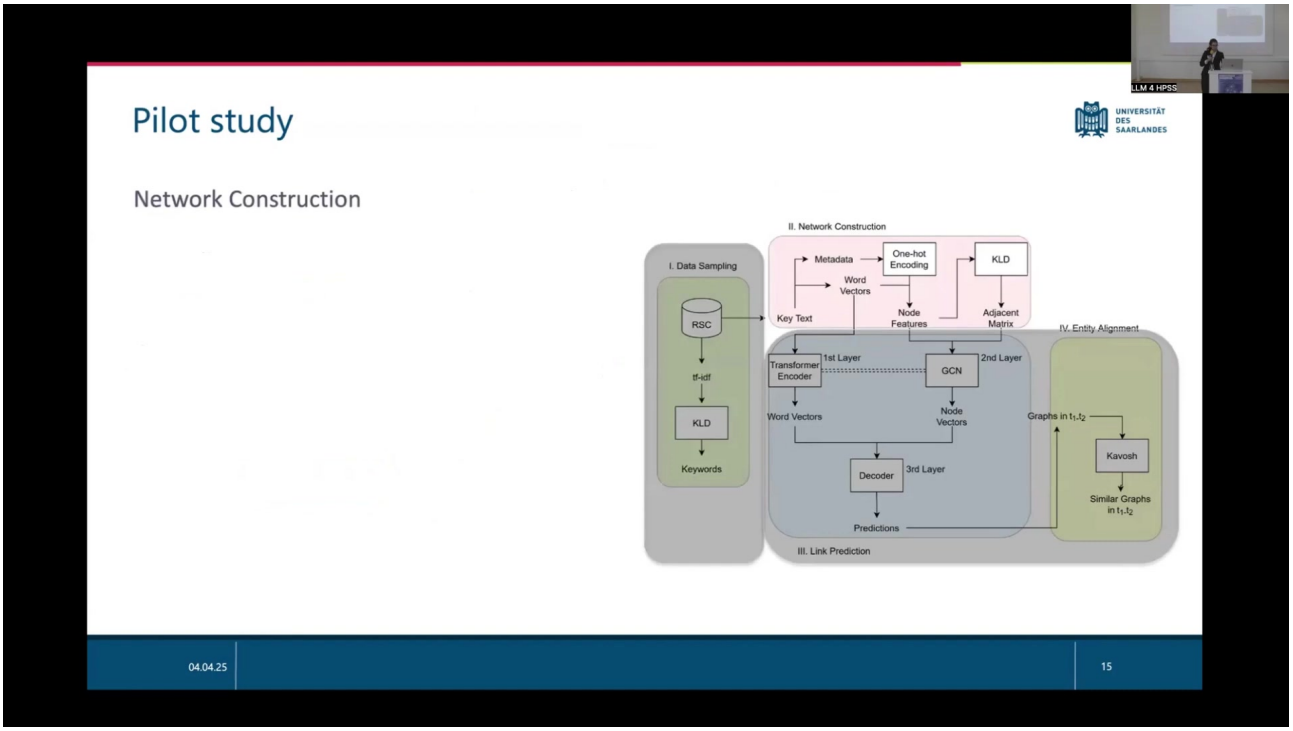

Figure 18.6: Slide 16

The research acknowledges several limitations and outlines future perspectives. Key questions include what it truly means to computationally trace conceptual change and whether models can capture deeper epistemic shifts beyond mere linguistic drift. Another challenge is understanding how context becomes integrated into the meaning represented by language models, and whether metadata should be treated as external noise or a core signal. The optimal 'unit' of language change remains a question: are shifts best observed at the level of words, concepts, grammar, or discourse patterns? The possibility of identifying recurring linguistic pathways for concept emergence and determining if new

ideas follow predictable linguistic trajectories is also explored. Finally, the limits of interpretability in complex models are considered, emphasizing the need to ensure that explanations are meaningful rather than merely plausible.

Future perspectives include expanding the data sources beyond the Royal Society Corpus to include multilingual corpora, such as French and German journals, and incorporating other text types like monographs and letters. A significant area for future work is addressing the expression of attitude or stance in language use, particularly the challenge of differentiating between positive and negative usage of terms like "oxygen" or "dephlogisticated air" within the context of heated historical debates. Potential approaches involve analyzing differences in network structure based on usage context, such as linking terms to critiques or "dispective" adjectives.

Matching linguistic patterns to authors known to advocate for or against specific theories, potentially leveraging external historical knowledge from philosophy of science texts, is another avenue. Furthermore, insights from work on propaganda analysis, such as in the context of the Russian-Ukraine war, could be mapped onto historical texts to identify propagandistic strategies. The method would involve comparing network structures over time, focusing on structural features and identifying which nodes promote more edges.

Applying the developed methods to current era corpora, such as a quantum gravity corpus, is also a future perspective. The goal is to observe community building and changing language in real-time or near real-time. This would require establishing a protocol for structuring the data, potentially in a relational database like SQL, to facilitate its translation into a graph format. Processing the data to leverage categorical values, such as author, journal, and topics, as features within the graph structure would be necessary. Depending on the structure of the dataset, data engineering may be required.

Finally, Entity Alignment is identified as a future perspective, particularly for enabling multilingual or multi-corpus comparisons. This involves tasks such as aligning concepts, for example, the "oxygen" subgraph, across different datasets, such as those from English versus French journals. The method would utilize a *graph convolutional network* specifically for an entity alignment task, distinct from a link prediction task. The objective is to determine if entities are identified as the same based on their neighboring nodes and overall network structure.

Chapter 19

LLM for HPS Studies: Analyzing the NHGRI Archive

The presentation addresses the limited understanding of science funding processes, which often relies solely on public outputs like publications and grants. It proposes analyzing born-physical archives of funding agencies to gain insights into the internal processes of science funding and innovation. The case study focuses on the archive of the National Human Genome Research Institute (NHGRI), a key funder of the Human Genome Project and subsequent genomics initiatives. The NHGRI archive ...

19.1 Overview

The presentation addresses the limited understanding of science funding processes, which often relies solely on public outputs like publications and grants. It proposes analyzing born-physical archives of funding agencies to gain insights into the internal processes of science funding and innovation.

The case study focuses on the archive of the National Human Genome Research Institute (NHGRI), a key funder of the Human Genome Project and subsequent genomics initiatives. The NHGRI archive contains over 2 million pages of diverse internal documents, including meeting notes, handwritten correspondence, presentations, spreadsheets, newspapers, forms, proposals, and emails. This archive presents challenges due to its scale, complexity, and the presence of sensitive information like PII and handwriting.

The research employs a suite of computational methods and tools to process this archive. These include training a synthetic-data informed handwriting model for removal and recognition, utilizing multimodal models (vision, text, layout) for tasks like entity extraction and synthetic data generation, and implementing entity disambiguation and PII masking techniques.

The extracted data is used for various analyses, including reconstructing correspondence networks and computationally modeling funding decisions.

Key findings include the identification of informal leadership structures within NHGRI, such as a “kitchen cabinet” during the International HapMap Project, discovered through unsupervised network community detection. Analysis of brokerage roles reveals differences in communication patterns between formal and informal groups.

A computational model predicting organism sequencing funding decisions demonstrates that biological, project, reputation, and linguistic features are all informative. This highlights effects like the Matthew effect, where higher H-index and community size correlate with funding success.

The work underscores the importance of preserving born-physical archives and developing computational tools to make their content accessible and analyzable for historical and sociological research on science. The project is part of a larger consortium, "Born Physical, Studied Digitally," which seeks collaborators to apply these methods to other archives, including federal court records and seismology data.

The research aims to inform policy, increase data accessibility, and answer fundamental questions about how science works, particularly regarding the influence of funding and the emergence of innovation.

19.2 Limitations of Current Understanding of Science Funding

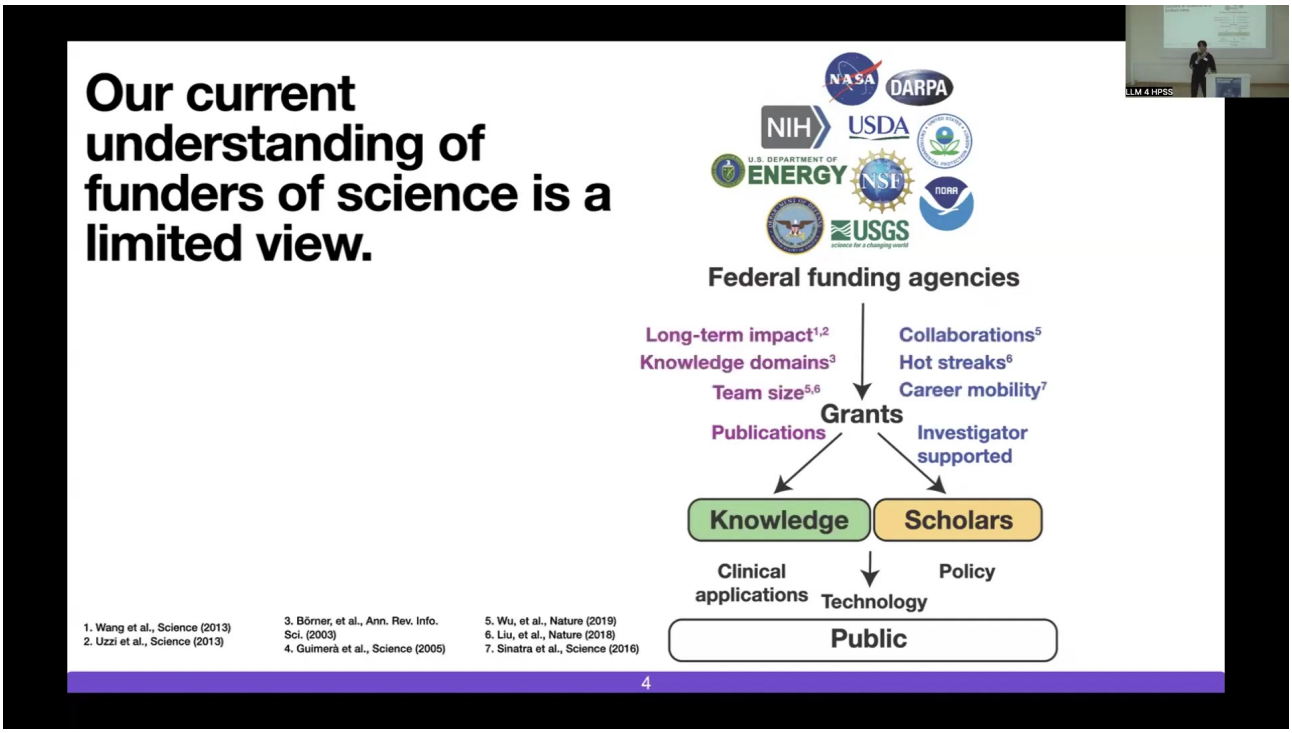

Figure 19.1: Slide 01

The current understanding of science funders is limited, primarily focusing on public outputs. Since World War II, state-sponsored research has been the dominant model, based on a social contract where funders, acting on behalf of the public, invest in research with the expectation that it will translate into informing policy, clinical applications, and new technology.

This framework, analyzing public outputs such as publications and the activities of scholars, has provided insights into aspects of science, including the long-term impact of research, the size of research teams, the origins of interdisciplinary domains, and the career mobility of scientists. Bibliometric analysis of scientific articles has also contributed to this understanding.

However, the scientific article offers a skewed and incomplete view of the scientific enterprise. Relying solely on

bibliometrics to define science oversimplifies its inherent complexity. A deeper understanding requires examining the processes behind scientific outputs, moving beyond the flawed picture provided by the scientific article alone.

The conventional model depicts a flow from federal funding agencies (e.g., NASA, DARPA, NIH, USDA, DOE, NSF, USGS, NOAA) providing grants, which support scholars and lead to publications. Both scholars and publications contribute to knowledge, which in turn informs clinical applications, technology, and policy, ultimately impacting the public. Research drawing on these public outputs is supported by studies such as Wang et al. (2013), Uzzi et al. (2013), Börner et al. (2003), Guimera et al. (2005), Wu et al. (2019), Liu et al. (2018), and Sinatra et al. (2016).

19.3 Research Questions and Expanded Model of Science Funding

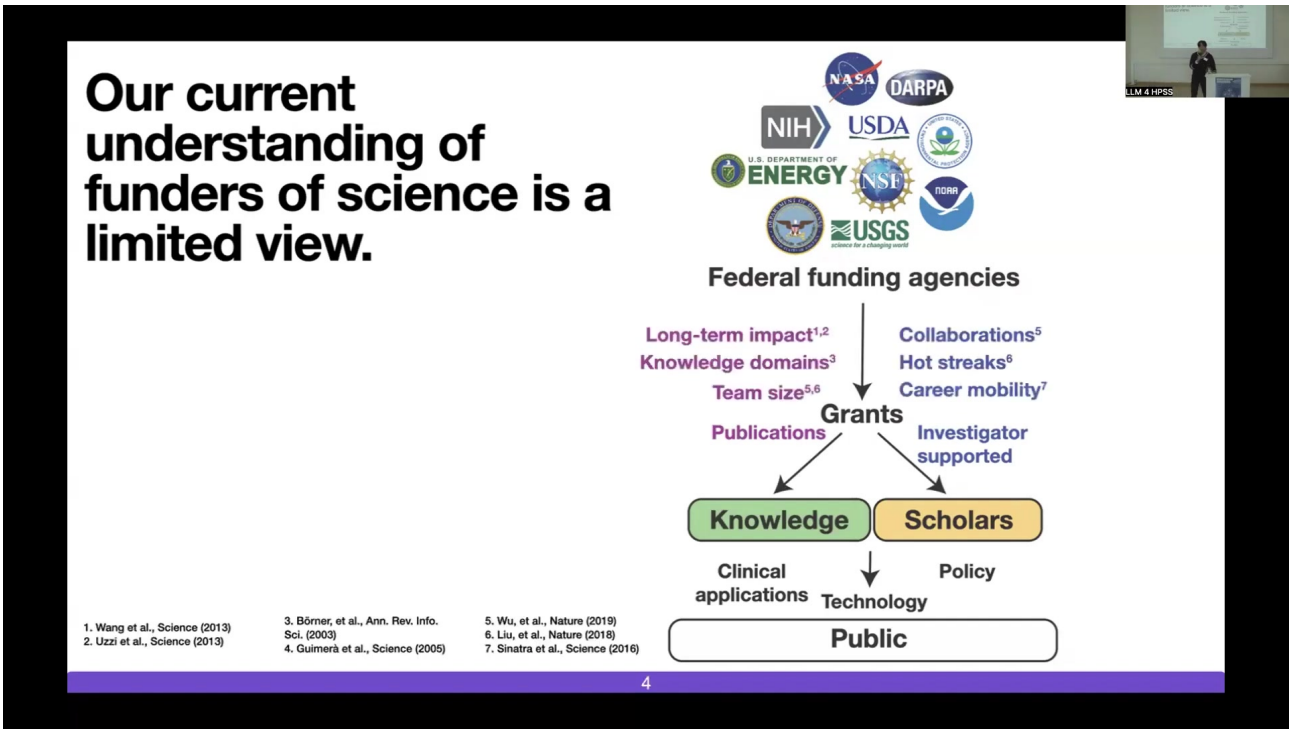

Figure 19.2: Slide 01

Understanding the role of funders is central to comprehending "how science works." This requires shifting the focus from the products of science, such as published articles, to the underlying processes.

Key research questions emerge from this perspective, including whether science drives funding decisions or if funding shapes the direction of science. The analysis also seeks to identify points within the innovation pipeline, from initial ideation to long-term impact, where innovation emerges, spills over into other areas, or fails and falls through the cracks. A significant challenge is that failed projects, which do not typically result in publications, remain largely invisible in analyses focused solely on articles. Further questions concern the nature of assistance provided by funders beyond financial support and the potential presence of biases in funding allocation.

An expanded model of science funding incorporates additional elements to capture these processes. Public data serves as an input informing grant decisions. A circular relationship exists between grants and technology development, indicating that grants lead to technology development, which can in turn influence future grants. Similarly, knowledge informs community engagement, which can feed back into the knowledge creation process. Cooperative agreements are

introduced as an output from grants that also serve as an input to community engagement, highlighting collaborative mechanisms.

19.4 Case Study: The Human Genome Project and NHGRI

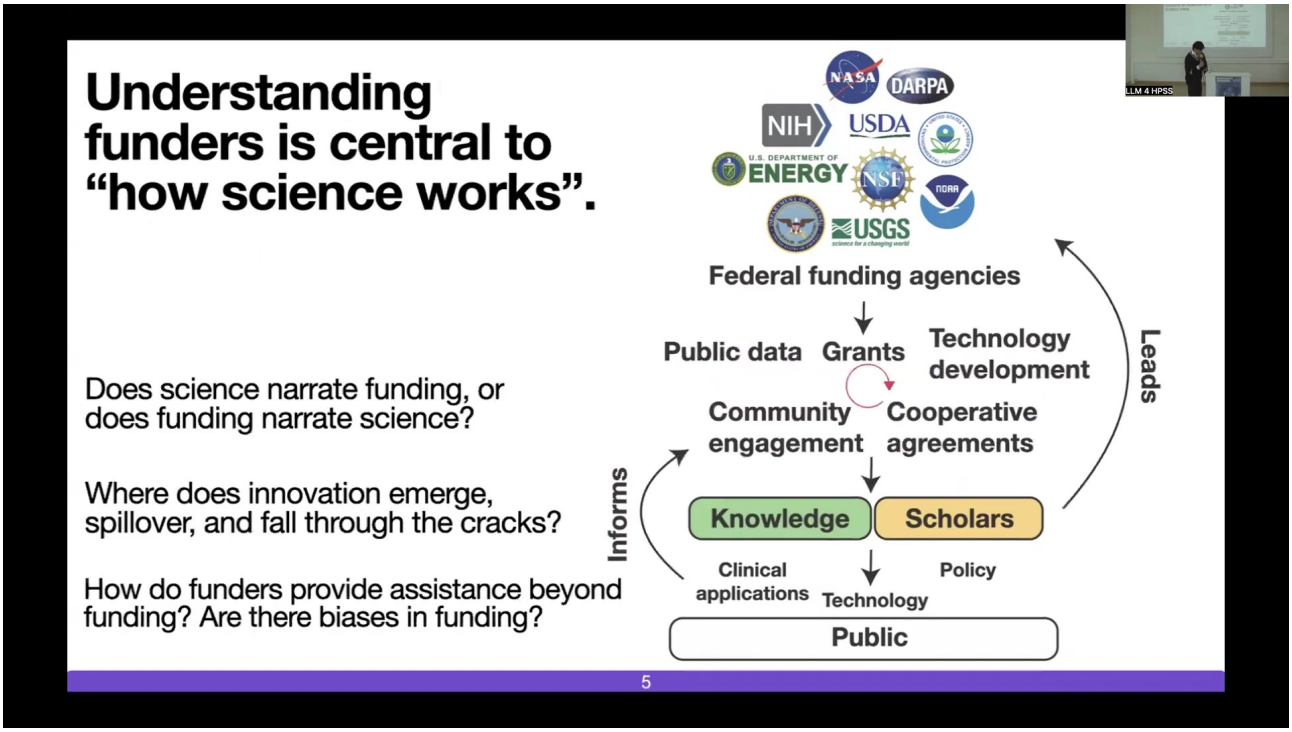

Figure 19.3: Slide 02

The Human Genome Project (HGP) serves as a key case study, recognized as the first "big science" initiative in the field of biology. This project is relevant in the context of Netpi, which examines big science in particle physics, by providing a parallel in biology. The HGP era involved the collaboration of tens of countries and thousands of researchers with the primary goal of sequencing the human genome.

The HGP is notable for several reasons. It garnered unprecedented public attention for a biology project, shifting focus from laboratory organisms like *Drosophila* and *C. elegans* to a project with direct human relevance. Its impact continues today, as most omics methods in modern biology rely on a reference genome, and the field of genomics itself largely arose from the sequencing of the human genome.

The project also pioneered new data sharing practices that are now widely adopted and marked a significant integration of computational methods with biology. The HGP was primarily led by two major organizations: the Welcome Trust in the UK and the National Human Genome Research Institute (NHGRI), which served as the US National Institutes of Health (NIH) arm specifically for the HGP. Francis Collins, who led the NIH and directed NHGRI, was a key figure in this endeavor.

19.5 NHGRI as an Innovative Funding Agency

Figure 19.4: Slide 04

Analysis indicates that NHGRI stands out as one of the most innovative funding agencies within the National Institutes of Health. This assessment is based on several quantitative metrics used for comparison across various NIH institutes. NHGRI consistently shows the highest performance in the share of its funded manuscripts that rank among the top 5% most cited. It also leads in the number of citations its funded research receives from patents and the total citations accumulated after ten years. Furthermore, NHGRI's funded research scores highest on a disruption metric, which measures the extent to which subsequent publications cease citing earlier work, suggesting a shift in research trajectories.

Based on these metrics, NHGRI is consistently identified as a leader in innovation within the biomedical community. However, while these metrics demonstrate that NHGRI is innovative, they do not explain the underlying reasons or processes that contribute to this innovation. Understanding *why* NHGRI is innovative requires a deeper examination of its internal operations and decision-making processes.

19.6 Interdisciplinary Team and Research Goals

Figure 19.5: Slide 05

The research is conducted by an interdisciplinary team comprising individuals with diverse expertise, including engineers, historians, physicists, ethicists, and computer scientists. Notable members of the team include former leaders of the NIH and NHGRI, such as Francis Collins. The project involves partnerships with organizations such as the NIH National Human Genome Research Institute, NVIDIA, and the NSF.

The research pursues several key goals. It aims to understand the specific factors and processes that contributed to the rise of the field of genomics. The team also seeks to identify failure modes within research and funding processes and analyze how innovation diffuses or spills over into different areas. A central objective is to study the dynamics of interaction between scientific funding agencies and academic scholars and scientists to understand how these relationships can foster better scientific outcomes.

19.7 The NHGRI Archive: Content and Challenges

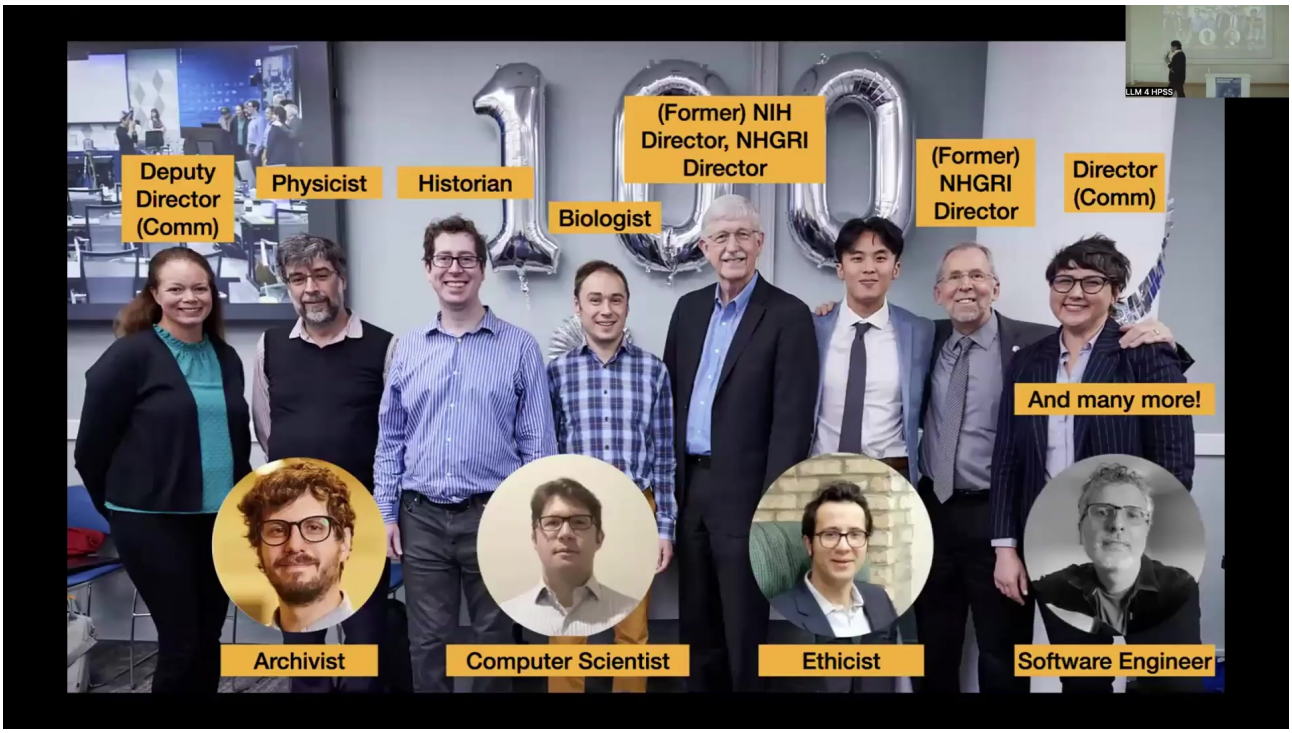

Figure 19.6: Slide 06

The NHGRI Archive constitutes a rich collection of content, though its structure is complex. Due to the notable historical nature of the Human Genome Project, many internal forms and documents from the 1980s, 1990s, and subsequent years were preserved.

This archive contains a variety of document types, including meeting notes detailing the daily coordination of the genome project, handwritten notes from correspondences, agendas, and conferences, as well as presentations, spreadsheets, newspaper clippings remarking on the period, forms, proposals, and printed copies of emails.

The archive is substantial in scale, exceeding 2 million pages, and continues to grow by 5% annually through ongoing digitization efforts. A significant challenge arises from the difficulty of studying this born-digital and born-physical artifact at scale using traditional methods, posing a barrier for individual researchers or even teams.

19.8 Distinction Between Internal Documents and Public Data

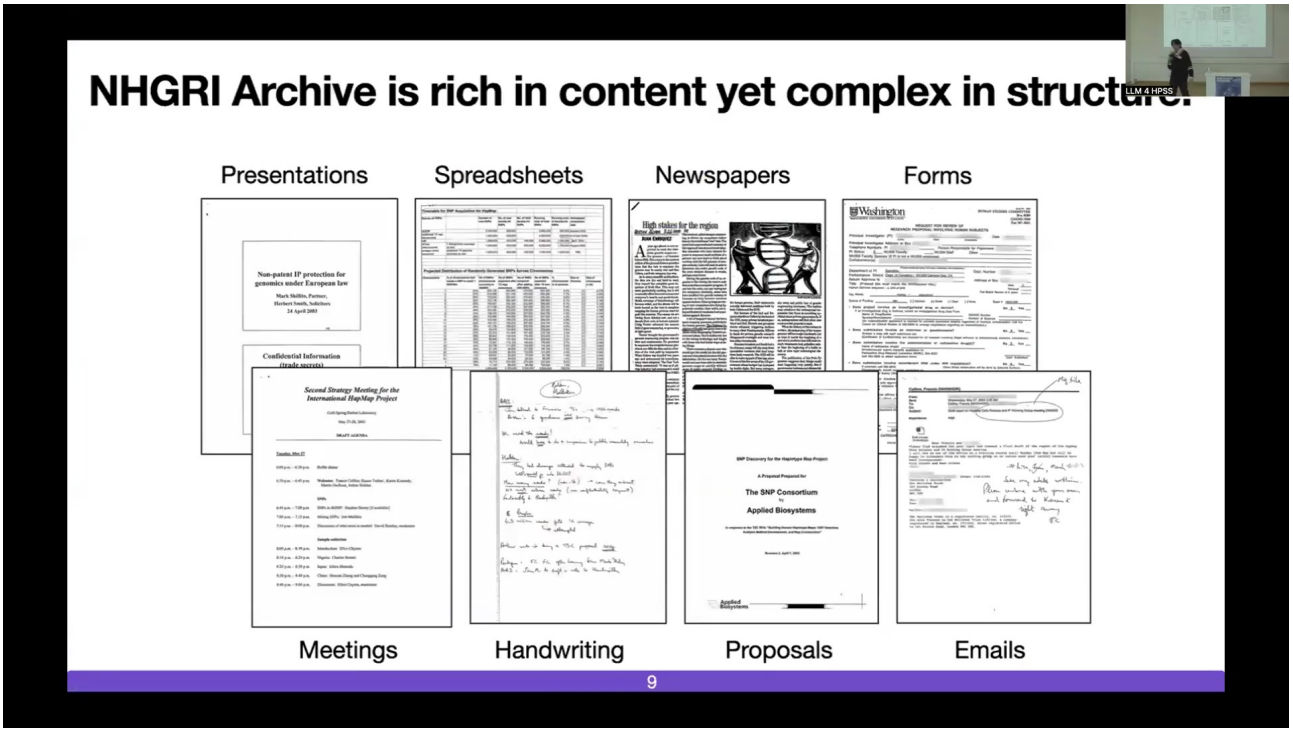

Figure 19.7: Slide 07

Internal documents within the NHGRI archive are fundamentally different from the data publicly available to scholars. Publicly accessible data primarily consists of Requests for Applications (RFAs) and publications, found in resources like PubMed and NIH RePORTER. The content of these public sources differs significantly from that found internally.

The internal documents provide detailed descriptions of the numerous large-scale genomic projects funded by NHGRI. These projects, often involving tens or hundreds of millions of dollars and thousands of researchers, were designed to create essential resources for the genomics community, thereby contributing significantly to the rise of the field.

A t-SNE plot visualizing the document space illustrates this distinction, showing distinct clusters corresponding to various genomic projects such as LSAC, modENCODE, eMERGE, ENCODE, Ethical, Legal, and Social Implications Research, NHGRI-EBI GWAS Catalog, H3Africa, International HapMap Project, Human Genome Project, and PAGE. The clusters representing RFAs and Publications are clearly separated from those representing the internal project documents, highlighting the unique nature of the archive's content.

19.9 Methodology: Handwriting Processing

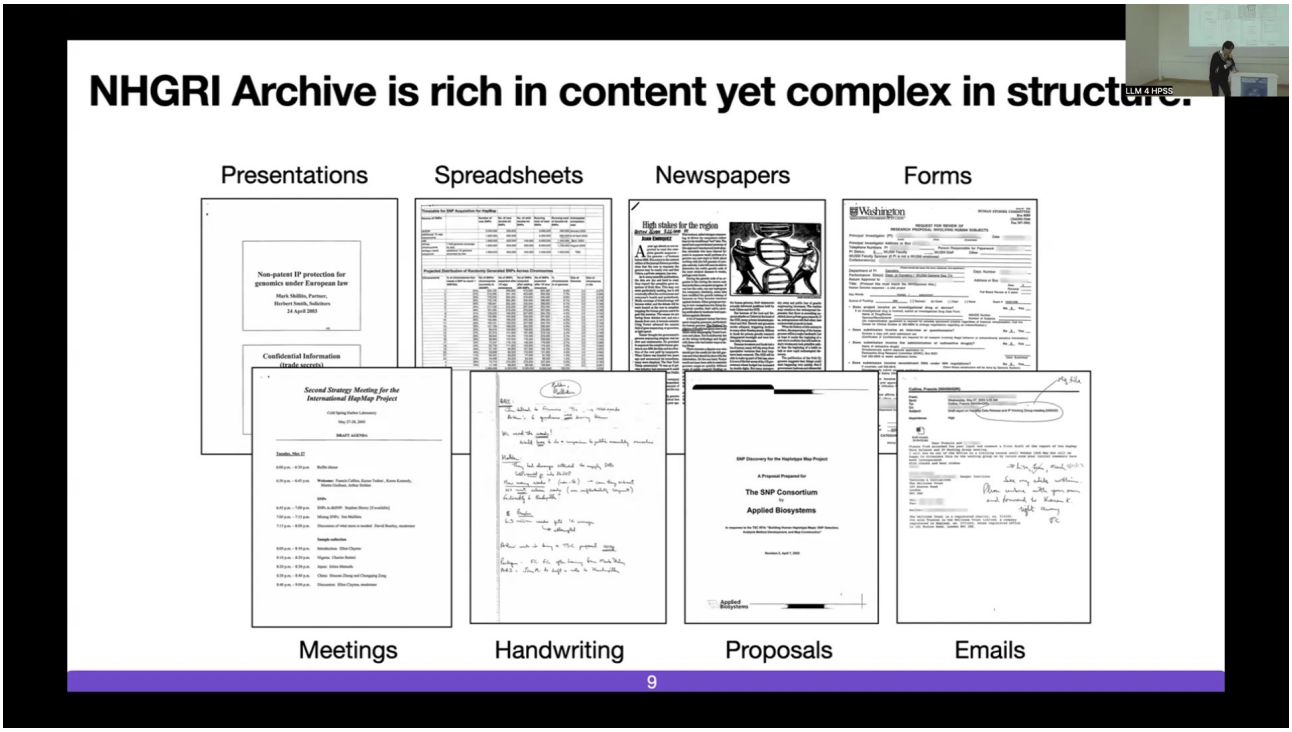

Figure 19.8: Slide 08

A significant methodological challenge arises from the presence of a large volume of handwriting within the NHGRI archive, a consequence of its born-physical origins. Processing handwriting with AI presents difficulties, not only in terms of technical proficiency but also ethically, due to the unknown nature of the content it may contain. To address this, a custom handwriting model is trained.

The purpose of this model is twofold: it can remove handwriting from documents, which aids in improving the accuracy of Optical Character Recognition (OCR) for the remaining printed text, and it enables the creation of a dedicated pipeline specifically for handwriting recognition. The model architecture utilized includes components resembling a *U-Net* architecture, as depicted in a diagram. The ethical considerations surrounding the use of AI with handwriting are further explored in a separate ethics case study.

19.10 Methodology: Multimodal Models and Synthetic Data Generation

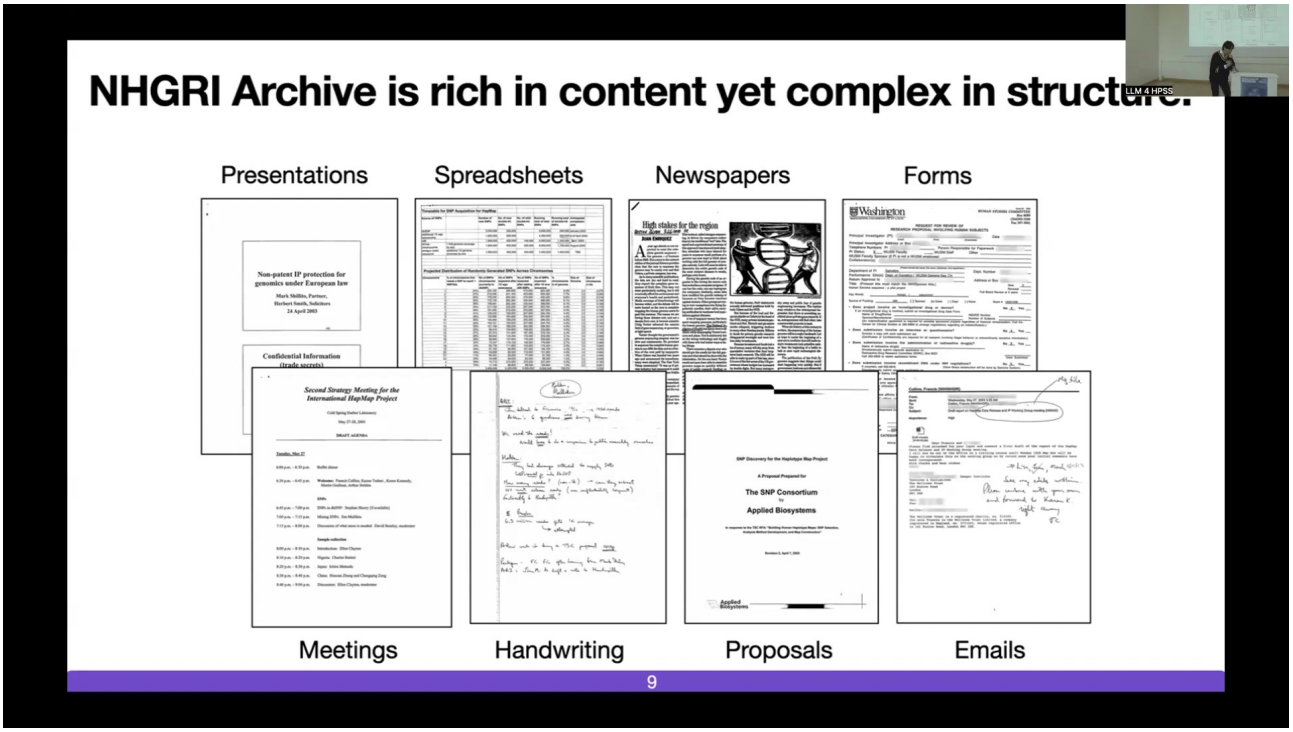

Figure 19.9: Slide 08

The methodology incorporates multimodal models, drawing upon research from the document intelligence community, including work by Huang et al. (2022) and Zhang et al. (2022). These models are designed to smartly combine multiple modalities: vision (image), text, and layout. The layout modality plays a crucial role by supervising the joint embedding process and discretizing the document structure, represented internally using angle brackets and numbers.

The models typically include components such as a joint embedding space, a Masked Autoencoder Training Objective, a Text Decoder, and a Vision Decoder. This multimodal approach enables various tasks, including entity extraction, where specific pieces of information are identified and highlighted within a document. It also facilitates the generation of synthetic data, producing artificial documents or images. The synthetic data generation capability is utilized to create synthetic training datasets, which are valuable for developing and training new classifiers.

19.11 Methodology: Entity and PII Recognition and Disambiguation

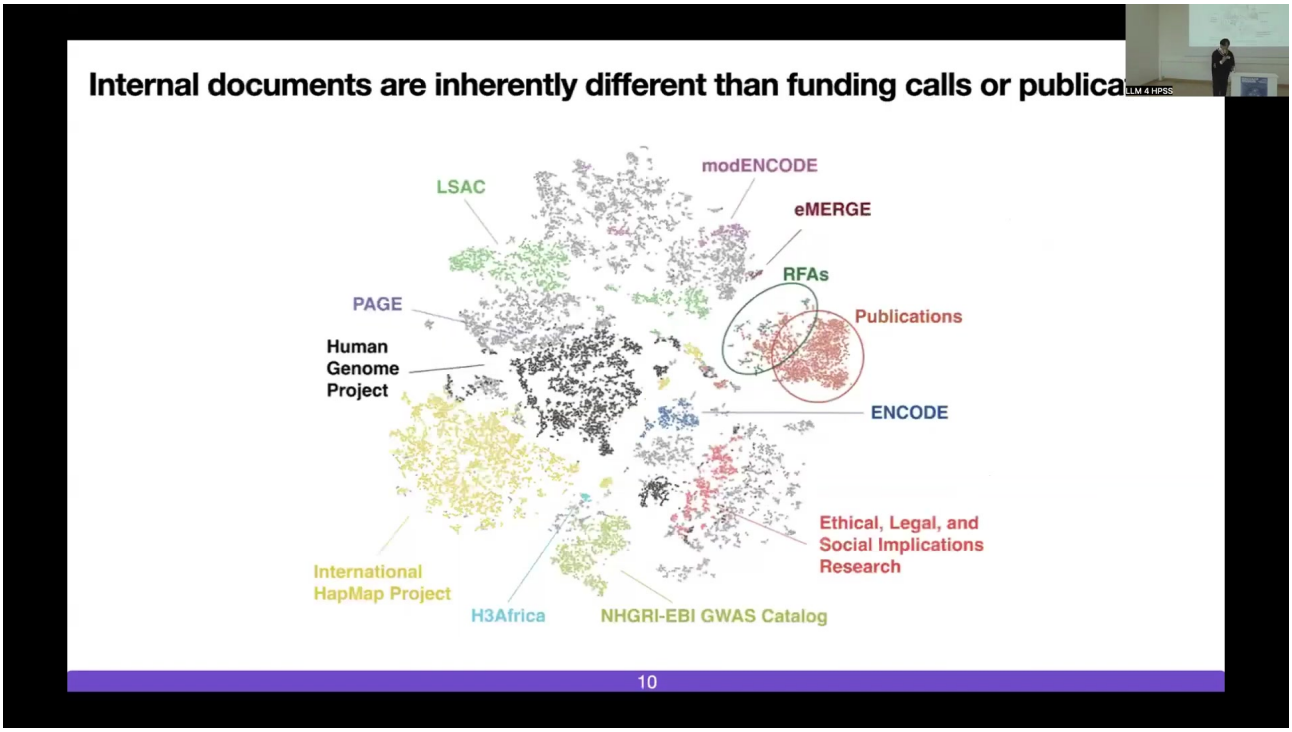

Figure 19.10: Slide 09

A critical task involves the recognition of entities and Personally Identifiable Information (PII). The NHGRI archive is considered a "living archive" because it contains information about real individuals, including sensitive data such as credit card numbers and social security numbers. A particular challenge is that some of these individuals remain active in government and academia today. Therefore, it is essential to implement rigorous processes for removing, masking, and disambiguating individuals within this large archive.

The methods employed for entity and PII recognition demonstrate good performance, as indicated by F1 scores. Performance metrics show that F1 scores for different entity types—PERSON, ORG, EMAIL, LOC, and IDNUM—increase significantly as more samples are used for finetuning. The F1 scores for EMAIL, IDNUM, and LOC approach 1.0 with sufficient finetuning, indicating high accuracy. Visual examples show highlighted entities on documents, such as an "identifier" on a boarding pass and various entities like "Organization," "Address," "Name," "Email," "IDNUM," and "LOC" on a letter.

19.12 Case Study: Reconstructing a Correspondence Network from Emails

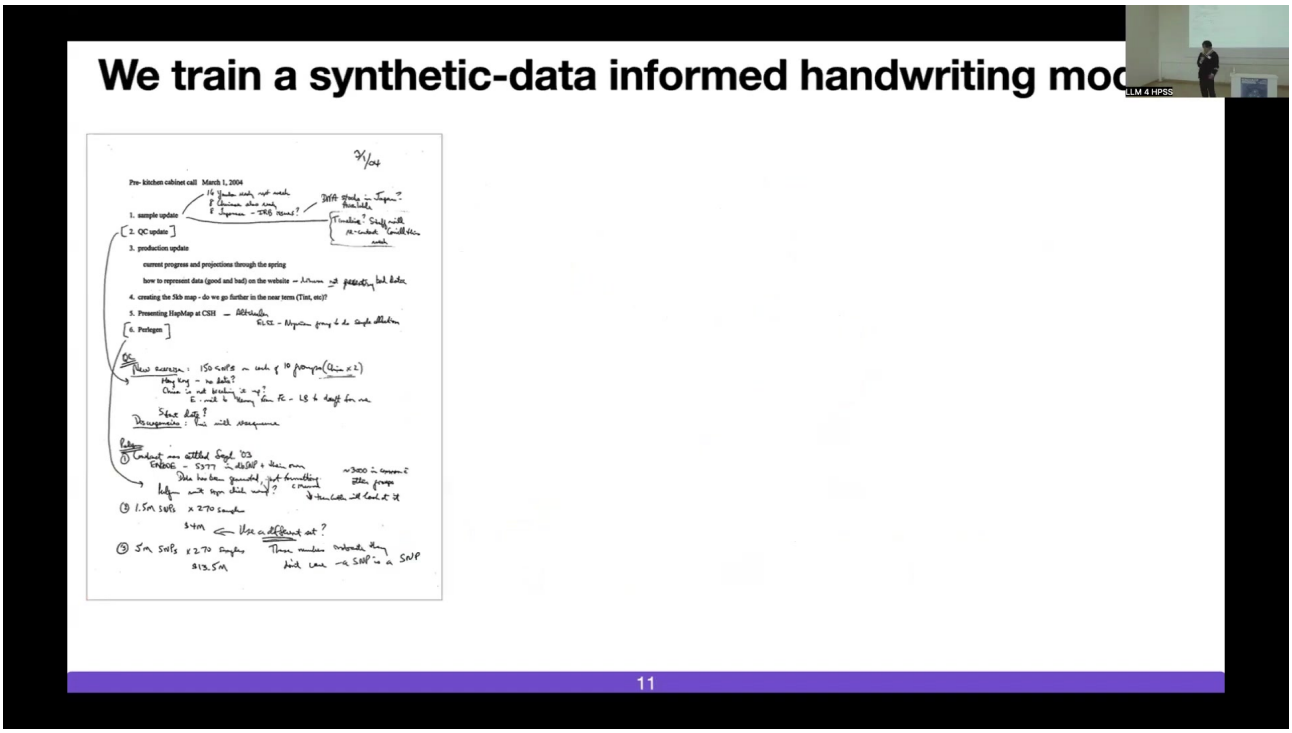

Figure 19.11: Slide 10

One case study involves reconstructing a correspondence network within NHGRI based on the archive's content. The data source for this analysis consists of printed scanned copies of emails. The method involves extracting entities from these scanned documents and linking them to build the network.

The analysis utilized 62,511 email conversations, which were derived from a collection of 5,414 scanned emails. In the resulting network visualization, each node represents a physical paper copy of an email from the archive.

19.13 Network Analysis: Affiliation Association

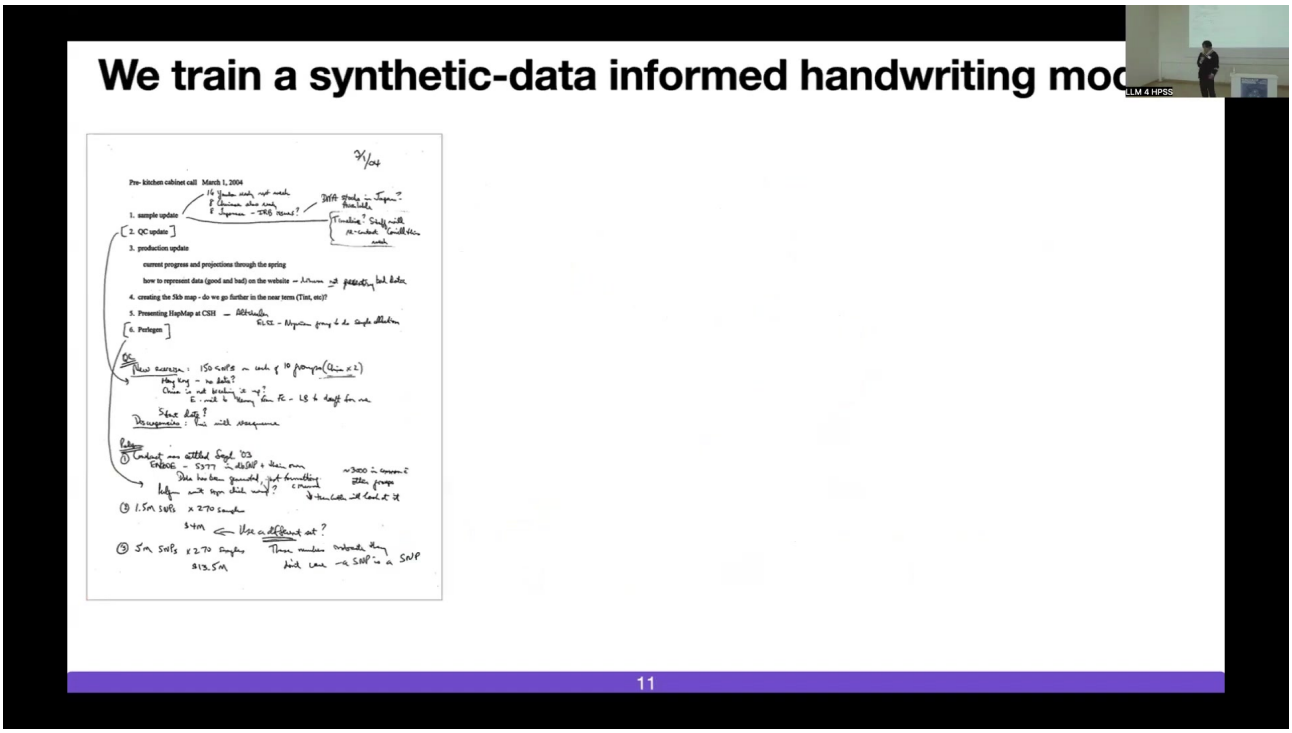

Figure 19.12: Slide 10

The network analysis includes associating the nodes, representing individuals or emails, with their respective affiliations. Two primary affiliation types are identified: Non-NIH Affiliation, visually represented by red nodes in the network graph, and NIH Affiliation, represented by blue nodes. These affiliations are linked to the specific organization, funding agency, or company with which individuals were associated during the Human Genome Project era.

19.14 Network Analysis: Community Detection and Informal Structures

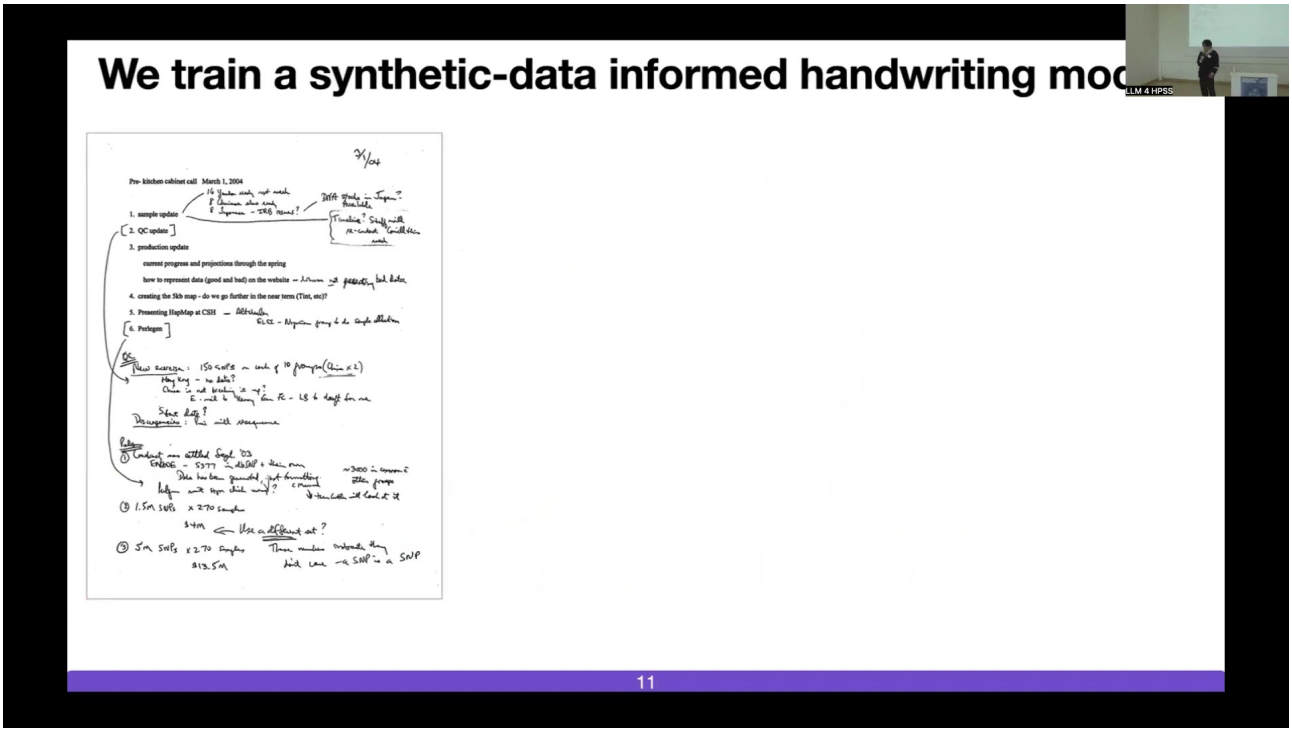

Figure 19.13: Slide 10

Network analysis techniques, including community detection methods such as the stochastic block model, are applied to the reconstructed correspondence network. This analysis focuses on understanding the interactions between academia and NIH personnel. A specific case study examines emails exchanged during the International HapMap Project.

The International HapMap Project is described as another "big science" genomics initiative that followed the HGP. Unlike the HGP's focus on sequencing, the HapMap Project concentrated on genetic variation and is significant as the basis for genome-wide association studies (GWAS). This was a large-scale project involving numerous universities and agencies, posing a challenge for the funding agency in terms of coordination and management.

The formal structure for managing such projects typically includes a steering committee with representatives from participating universities. However, the analysis computationally discovered an informal structure not previously discussed: the "Kitchen Cabinet." This informal leadership circle met prior to the official steering committee meetings, functioning to identify and address potential problems preemptively. This "Kitchen Cabinet" was identified from the archive data using unsupervised computational methods.

19.15 Network Analysis: Brokerage Roles and Leadership Comparison

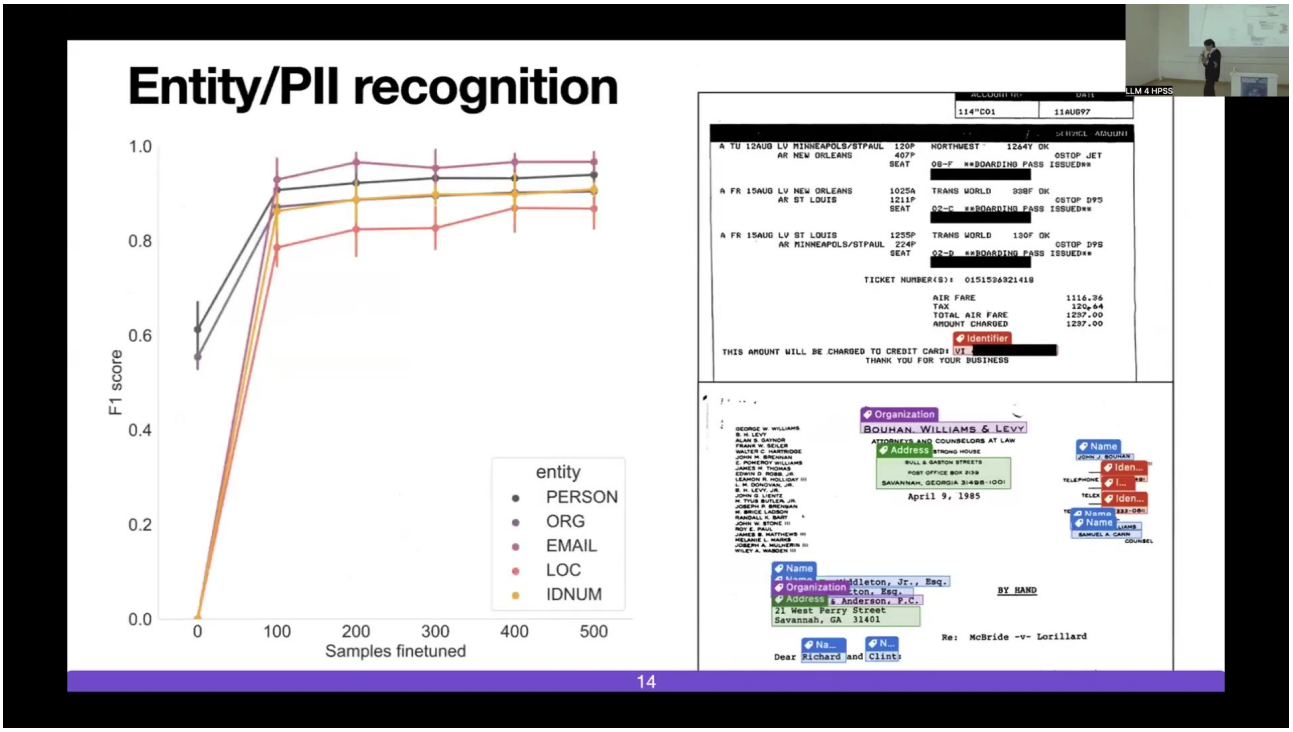

Figure 19.14: Slide 12

The analysis extends to comparing different leadership circles based on their brokerage roles within the network. Brokerage roles are defined by the interaction patterns of each node (individual) with others in the network. Examples of these roles include a consultant, who receives information and disseminates it back within the same group, and a gatekeeper, who receives information but does not pass it back to the originating group. Other roles include coordinator, liaison, and representative.

The analysis compares the distribution of brokerage roles among three groups: the “Kitchen Cabinet” (representing informal leadership), the Steering Committee (representing formal leadership), and the Rest of HapMap (representing other project participants). The findings indicate that the “Kitchen Cabinet” primarily functioned in a consultant role, a pattern distinct from that observed in the other leadership circles during the project. Francis Collins is specifically noted as acting in a consultant role within the “Kitchen Cabinet.” A box plot visualization illustrates the distribution of brokerage roles for these three comparison groups.

19.16 Portfolio Analysis: Modeling Funding Decisions

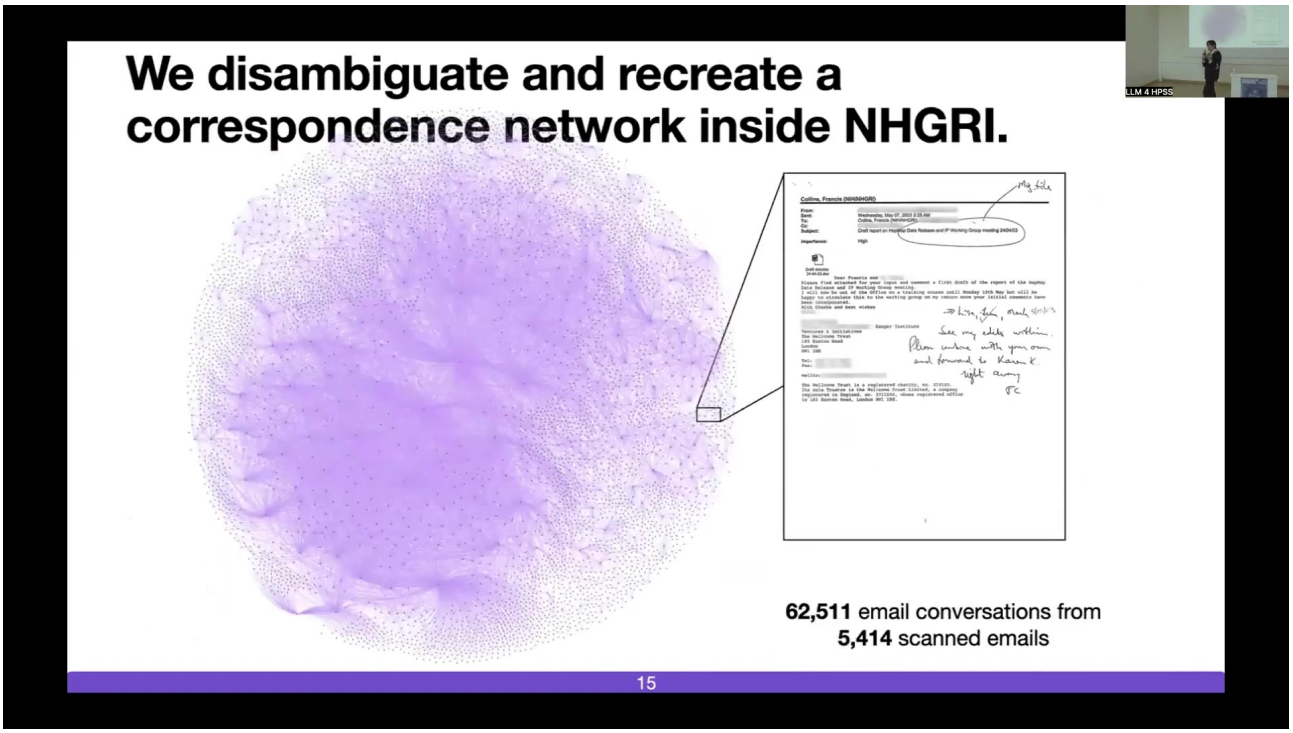

Figure 19.15: Slide 13

The research includes portfolio analysis focused on modeling funding decisions. A specific case study examines the decisions made regarding organism sequencing after the completion of the Human Genome Project. The problem involved scientists and leadership within the funding agency determining which organismal community's genome should be sequenced next. This decision-making process required allocating limited funding among various competing proposals advocating for the sequencing of different organisms.

19.17 Computational Model for Funding Decisions: Features

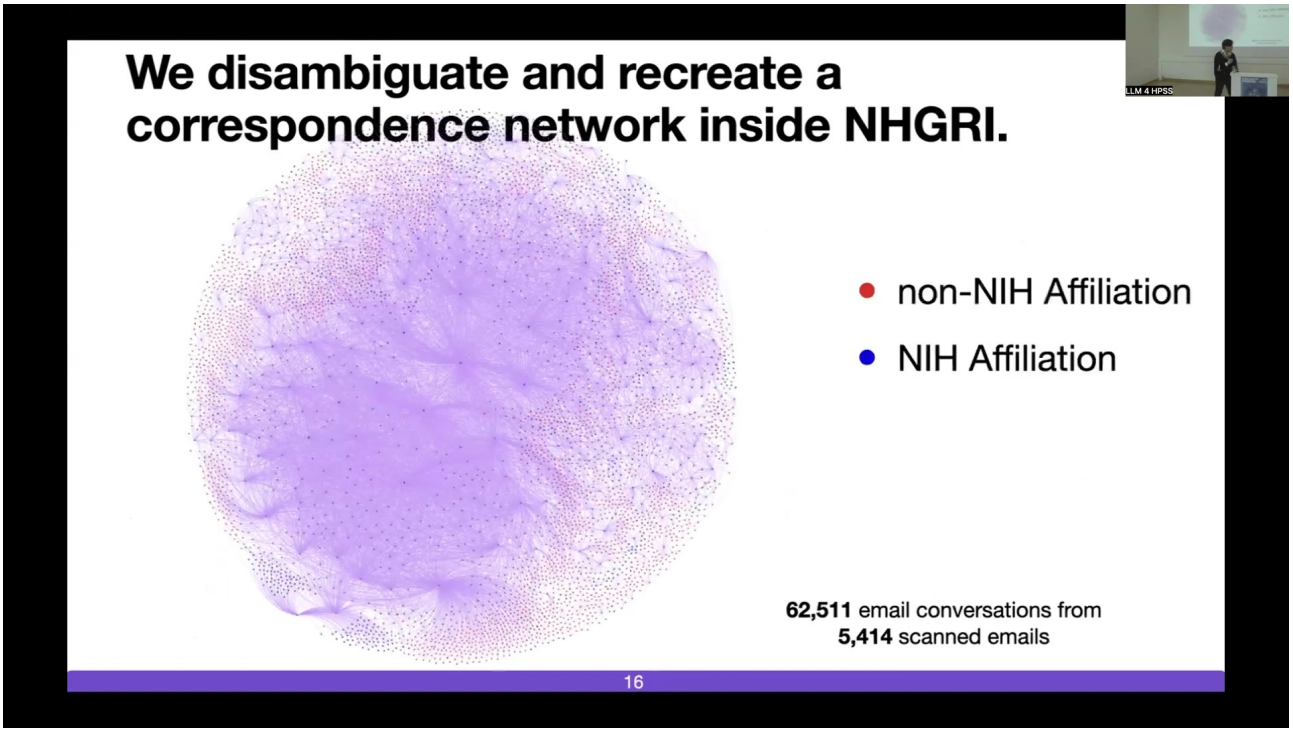

Figure 19.16: Slide 14

A computational model is developed using machine learning to recapitulate the funding decisions made for organism sequencing projects. The model incorporates various features categorized as Biological, Project, Reputation, and Linguistic.

Biological features include the genetic distance of the proposed organism to already known sequenced model organisms and the organism's genome size.

Project features encompass characteristics of the proposal and the proposing team, such as the team size, the time elapsed since the first submission of the proposal, the gender equity within the proposing team, whether the proposal is standalone or part of a larger initiative, and if the proposal was written internally within the agency.

Reputation features assess the standing of the individuals and community involved, including the H-index of the proposal authors, the size of the research community focused on the specific organism, the centrality rank of the proposers within the NHGRI network, and indicators of broader community support for sequencing the organism.

Linguistic features analyze the text of the proposals, examining aspects such as the level of argumentation and repetitiveness.

19.18 Computational Model Performance and Feature Informativeness

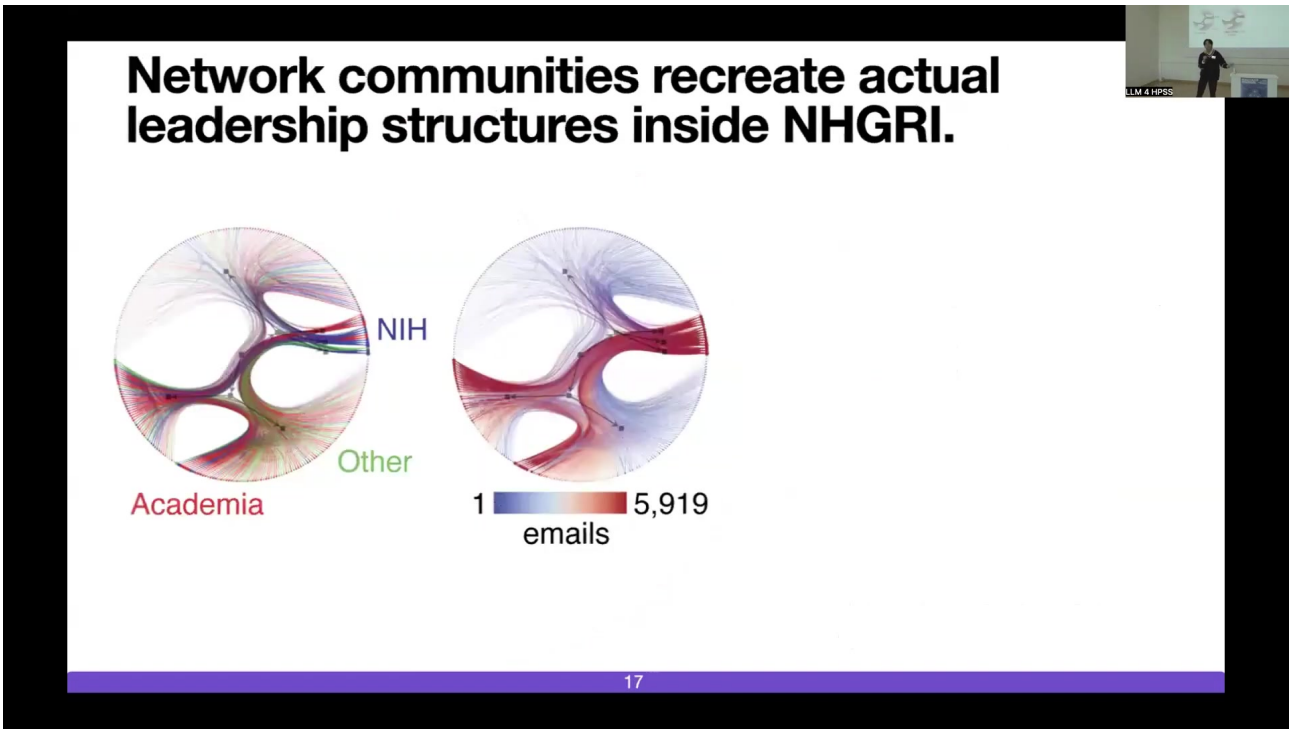

Figure 19.17: Slide 15

The performance of the computational model for predicting funding decisions is evaluated using ROC curves. The model demonstrates that all categories of features are informative for predicting funding success. The Biological features achieve an ROC AUC of 0.76 ± 0.05, Project features achieve 0.83 ± 0.04, Reputation features achieve 0.87 ± 0.04, and Linguistic features achieve 0.85 ± 0.04. When all features are combined, the model achieves the highest predictive performance with an ROC AUC of 0.94 ± 0.03, significantly outperforming individual feature categories and a random classifier (ROC AUC 0.5). The informativeness of these features enables further analysis to understand how they influence the funding decision models.

19.19 Feature Interpretability Analysis

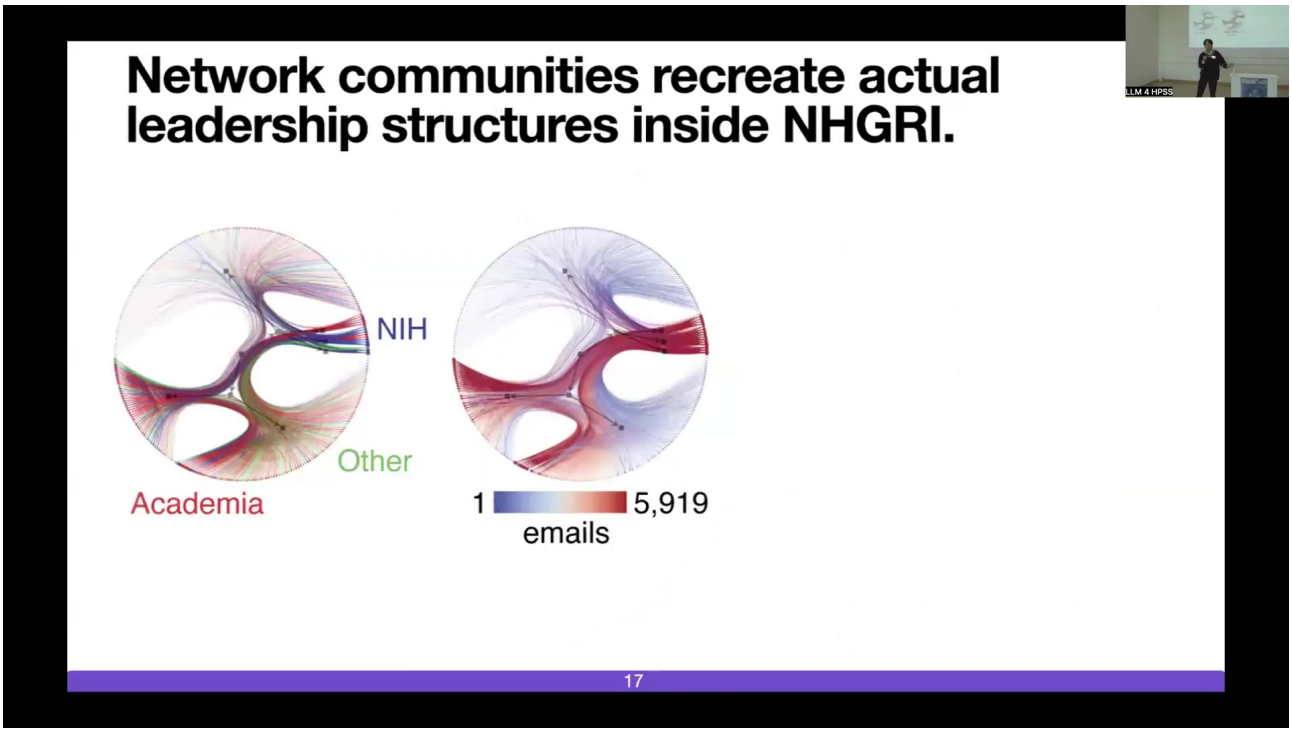

Figure 19.18: Slide 15

Feature interpretability techniques are employed to analyze the computational model. The purpose is to understand how individual features contribute to or inform the model's prediction of funding success. This approach relates to methods for explaining individual predictions, particularly in contexts where features may be dependent. Relevant work in this area, focusing on more accurate approximations to Shapley values for dependent features, is cited, specifically Aas, K., Jullum, M. & Løland, A. (2021) in *Artificial Intelligence*. A diagram visually represents features that are more likely to contribute positively to funding success prediction and those less likely to contribute positively.

19.20 Finding: Matthew Effect in Funding Decisions

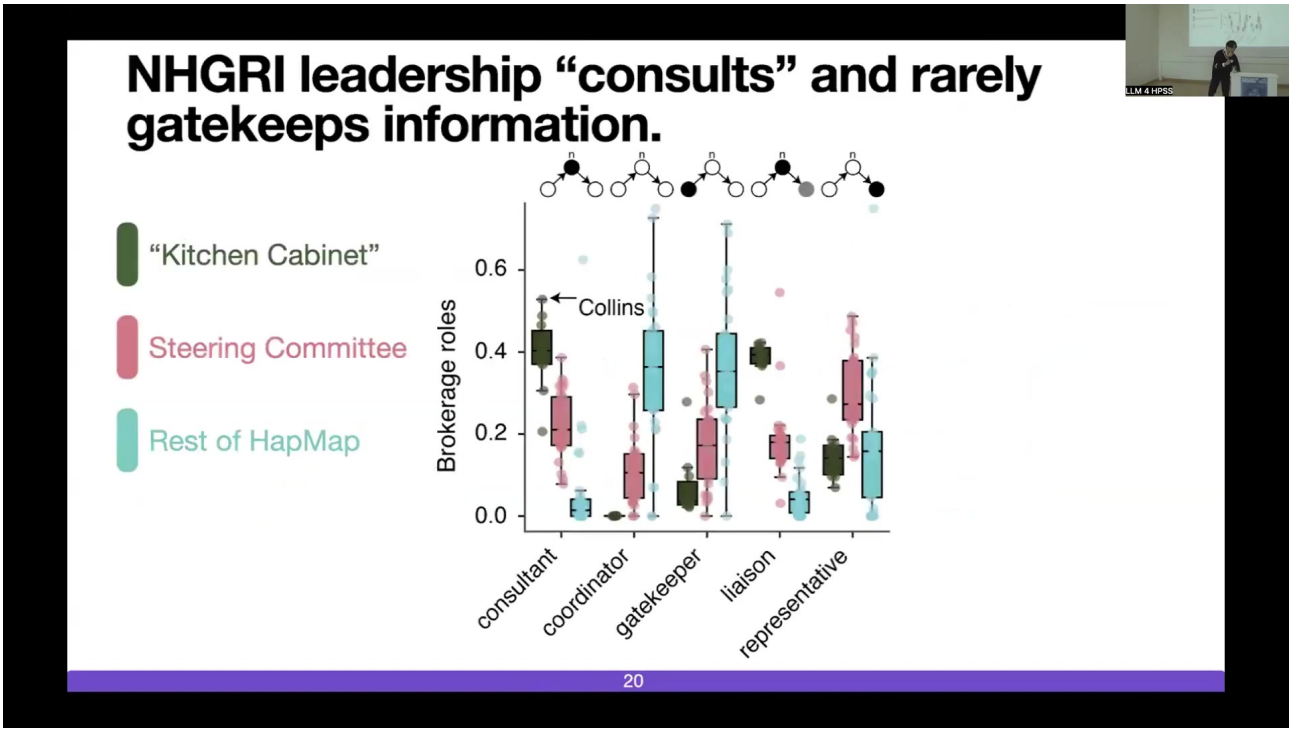

Figure 19.19: Slide 16

Analysis using feature interpretability reveals evidence of the Matthew Effect at play in the funding decisions. The Matthew Effect is observed in two key correlations: a higher maximum H-index among the proposal authors is associated with a greater likelihood of the proposal being approved for funding, and a larger size of the research community focused on the specific organism also correlates with a higher probability of funding success.

These correlations are visualized in scatter plots. The left plot shows an upward trend between the maximum H-index on the x-axis and a measure of contribution to funding success on the y-axis. The right plot similarly shows an upward trend between the logarithm of community size on the x-axis and the contribution to funding success on the y-axis. This confirms the expectation that funding agencies, aiming to maximize downstream impact in areas like clinical applications and technology, tend to favor proposals from more established researchers (higher H-index) and larger, more active communities. The findings align with existing understanding of the Matthew effect within the context of science funding.

19.21 Synthesis and Broader Applications

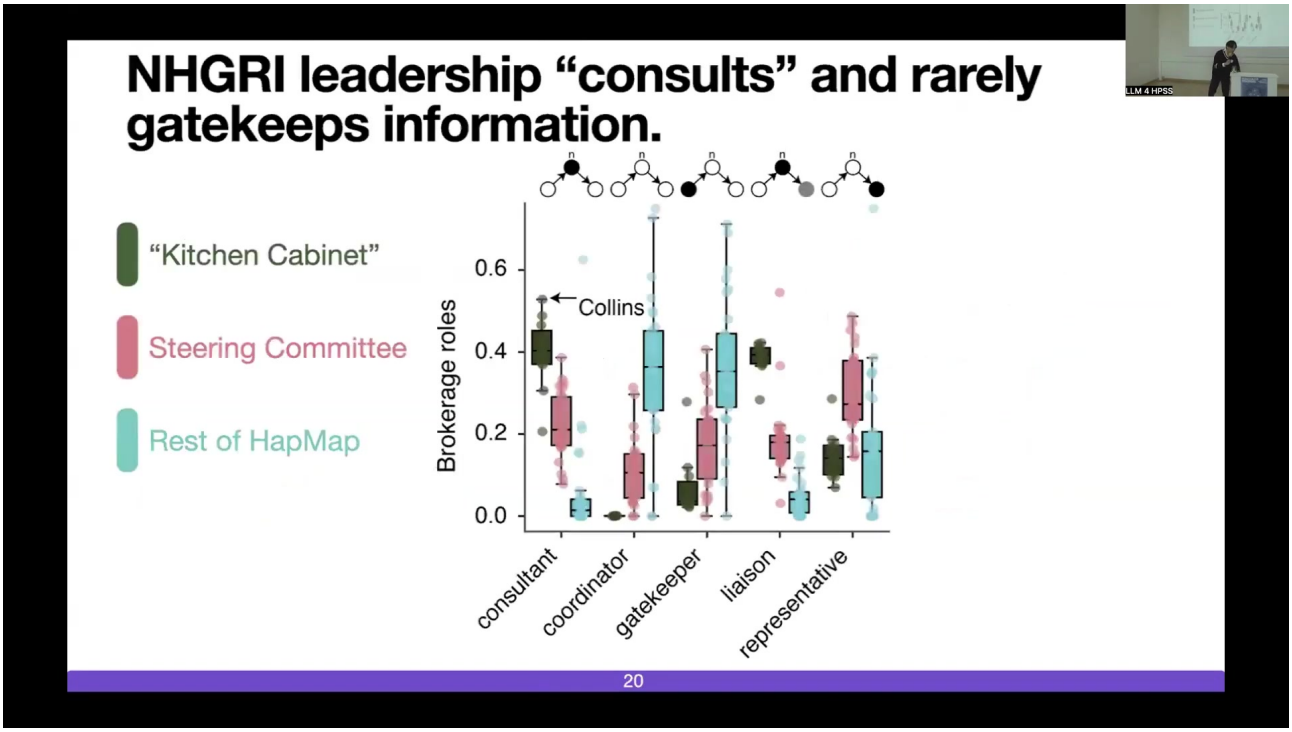

Figure 19.20: Slide 16

The work synthesizes the approach by demonstrating the potential achieved through combining born-physical archives with computational tools. This methodology is part of a broader initiative involving multiple partners beyond NHGRI, including custodians of federal court records from the United States and the EarthScope Consortium, which manages seismogram data. The core process involves translating data from these diverse sources using robust algorithms and cyber infrastructure.

The processed data and subsequent analysis have multiple applications. They can be used to inform policy decisions, increase the accessibility of previously locked-away data, and facilitate answering complex scientific questions. A diagram illustrates this process, showing Data and Metadata sources (NHGRI, federal court records, EarthScope Consortium) feeding into Knowledge Creation steps (Page stream segmentation, Handwriting extraction, Entity disambiguation, Layout modeling), which then enable Application Use (Scientific questions, Policy decisions, Data accessibility).

19.22 Importance of Preserving Born-Physical Archives

Figure 19.21: Slide 17

Born-physical archives hold valuable data that is currently contained in forms such as shipping containers. These physical archives face risks of neglect and are vulnerable to damage. The research underscores the critical importance of preserving this data to ensure its availability for future generations of scholars and scientists. The work highlights the necessity of efforts dedicated to the preservation and accessibility of these historical records.

19.23 Consortium and Call for Collaboration

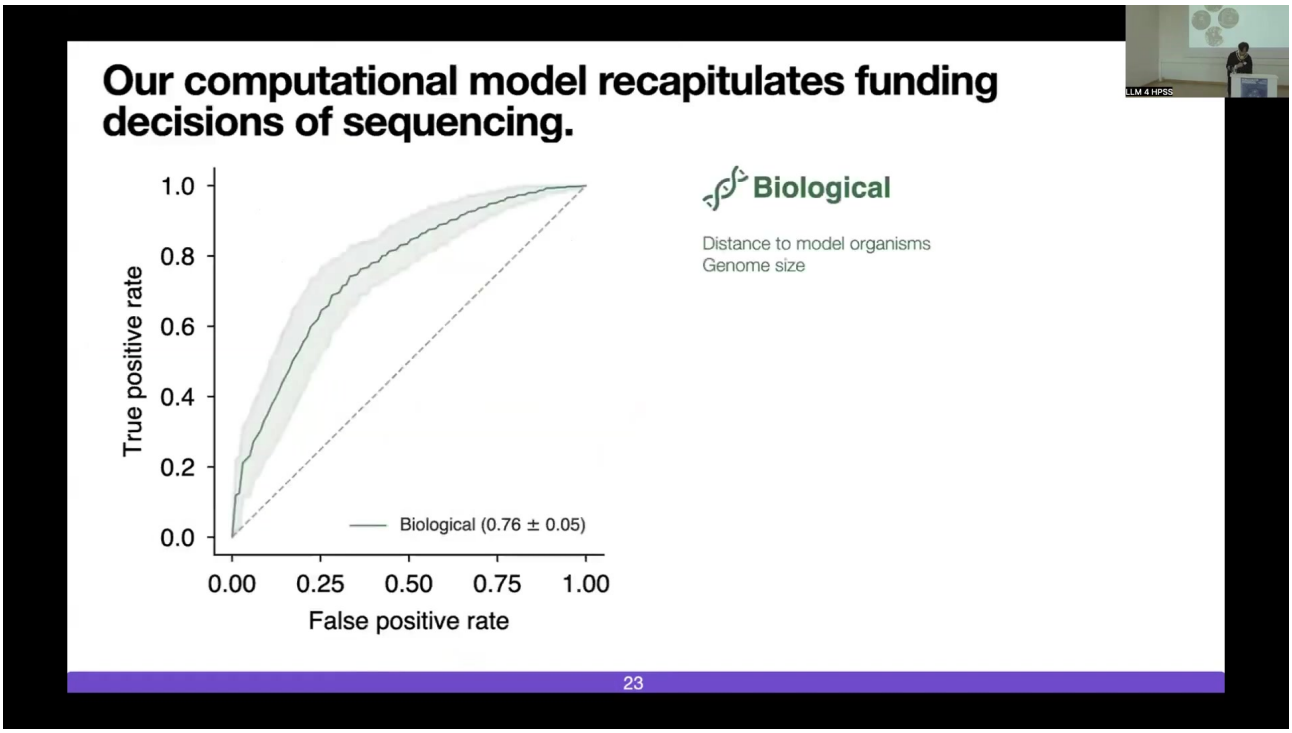

Figure 19.22: Slide 18

The project is part of a larger consortium named "Born Physical, Studied Digitally." The consortium is actively seeking to engage testers, partners, and users to collaborate on its initiatives. A call for collaboration is extended to individuals and institutions interested in working with the consortium.

A specific note is made regarding the recent status of NHGRI, which was among the agencies proposed for dissolution in the past year. The presentation argues that NHGRI stands as one of the most innovative funding agencies in the history of science and emphasizes that the data contained within its archive holds substantial potential for answering important scientific questions.

Chapter 20

From Source to Structure: Extracting Knowledge Graphs with LLMs

The project focuses on extracting structured knowledge graphs from unstructured historical and biographical sources using Large Language Models (LLMs) as part of a processing pipeline. The primary objective is to enable structured querying of these previously computationally inaccessible sources, such as printed biographical dictionaries. The problem addressed is the lack of inherent structure in many valuable historical sources, preventing complex analytical queries. The proposed solutio…

20.1 Overview

The project focuses on extracting structured knowledge graphs from unstructured historical and biographical sources using *Large Language Models* (*LLMs*) as part of a processing pipeline. The primary objective is to enable structured querying of these previously computationally inaccessible sources, such as printed biographical dictionaries.

The problem addressed is the lack of inherent structure in many valuable historical sources, which prevents complex analytical queries. The proposed solution is a two-stage pipeline. Stage 1 involves *Open Information Extraction* (*OIE*) using an *LLM* to extract subject-predicate-object triples from the text. This is followed by validation and refinement using an *LLM* ensemble acting as an adversary.

This stage includes a human-in-the-loop evaluation against domain expert extractions to assess quality using classical performance metrics. Stage 2 focuses on structuring the extracted triples into a knowledge graph. This stage is driven by research questions, defined as competency questions, which guide the creation of a domain-specific ontology.

LLMs are used to draft both competency questions and the ontology, with human experts providing final refinement. Entity disambiguation is performed, including resolution to *Wikidata* instances. The structured data, along with metadata (source, scoring, chunk), is encoded into an *RDF star graph*. The pipeline includes a final validation step before export.

Case studies include Zielinski's Polish biographical dictionary (1930) and the "Who was who in the GDR" reference work. Initial experiments demonstrate the ability to build networks (e.g., editors and authors) and conduct structured analyses (e.g., correlation between state awards and political affiliation/roles).

Key challenges identified are improving entity disambiguation and establishing proper benchmarking. Future work involves refining and completing the pipeline, systematic comparison with other graph extraction tools (*Neo4j graph*

builder, *Microsoft graph rack*), developing graph *RAG* systems for natural language querying, and building multi-layered networks for deeper analysis. The approach emphasizes task decomposition, data/research question-driven structuring, and verifiability through human intervention at key points.

20.2 Introduction: Extracting Structure from Unstructured Sources

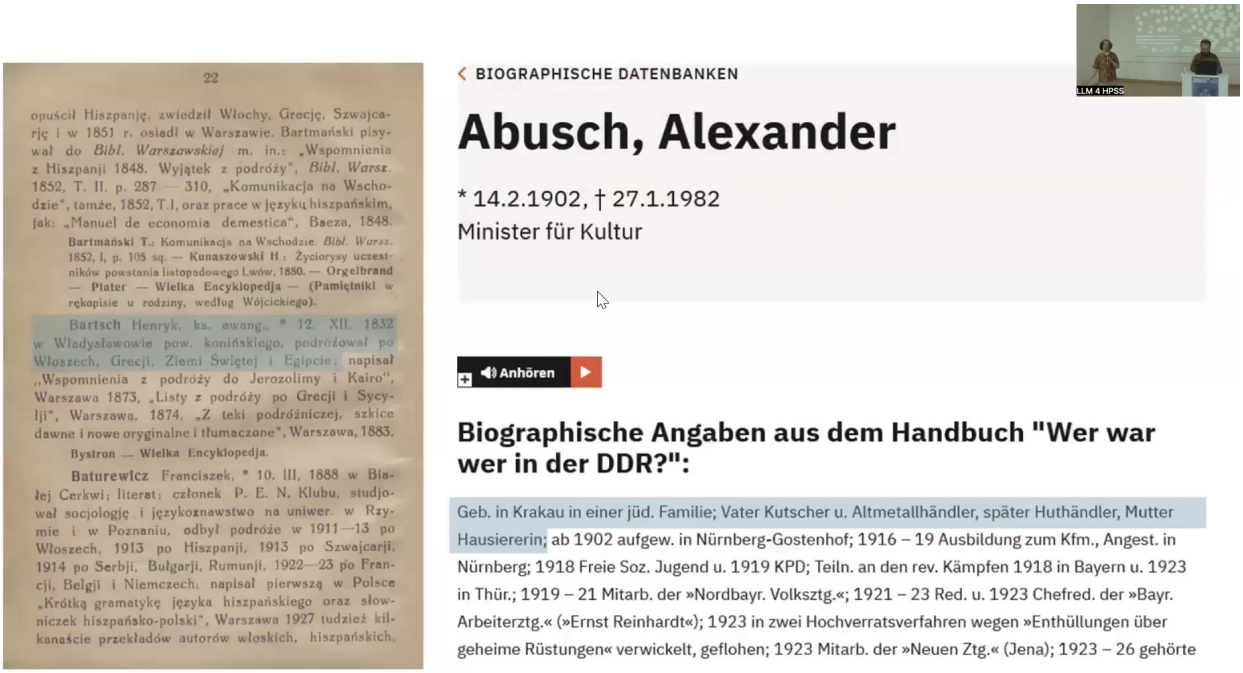

Figure 20.1: Slide 01

The project aims to extract knowledge graphs from source material using *Large Language Models* (*LLMs*). The primary focus is on accessing and utilizing new types of sources for research that are currently computationally inaccessible due to their unstructured nature. While historical, philosophical, and social science (*HPSS*) research often utilizes structured data sources such as publication databases or email archives, a significant amount of valuable information resides in unstructured formats, particularly printed books and biographical dictionaries. The core problem addressed is the inability to perform structured queries on these unstructured sources.

LLMs offer the potential to impose structure on this unstructured data. Historically, efforts like the *Get Grass* project attempted to address the computational accessibility of printed books. The current project specifically targets biographical sources, which are rich in detailed information about individuals but lack inherent structure. This absence of structure prevents researchers from asking complex, structured questions beyond simple facts like birth dates or work locations.

The goal is to enable queries about how professions formed networks over specific periods, how individuals migrating between locations contributed to the spread of ideas, or the specific roles of editors within a corpus in disseminating knowledge. The proposed solution involves using *LLMs* to construct knowledge graphs from this unstructured data in a controllable manner. A knowledge graph represents information as entities (such as persons, places, countries, or works) which become nodes in the graph. Relationships identified between these entities in the source material are represented as edges connecting the nodes. This structure allows for sophisticated structured querying.

The *Neo4j* graph database is used for representation and querying of the resulting knowledge graphs. The approach positions the *LLM* as one component within a larger processing pipeline, emphasizing its utility for specific tasks rather than seeking a universally perfect model. An example source snippet, an entry about the evangelical priest Henrik Bartsch born in 1832 who traveled and wrote books, illustrates the type of material processed. Traditional *Natural Language Processing* (*NLP*) approaches, such as *NLTK*, are often insufficient to extract the full contextual richness from such entries. The desired output is structured data in the form of statements or triples, capturing details like profession, birth date, birth place, and travel destinations from the text.

20.3 Two-Stage Pipeline: Stage 1 - Open Information Extraction

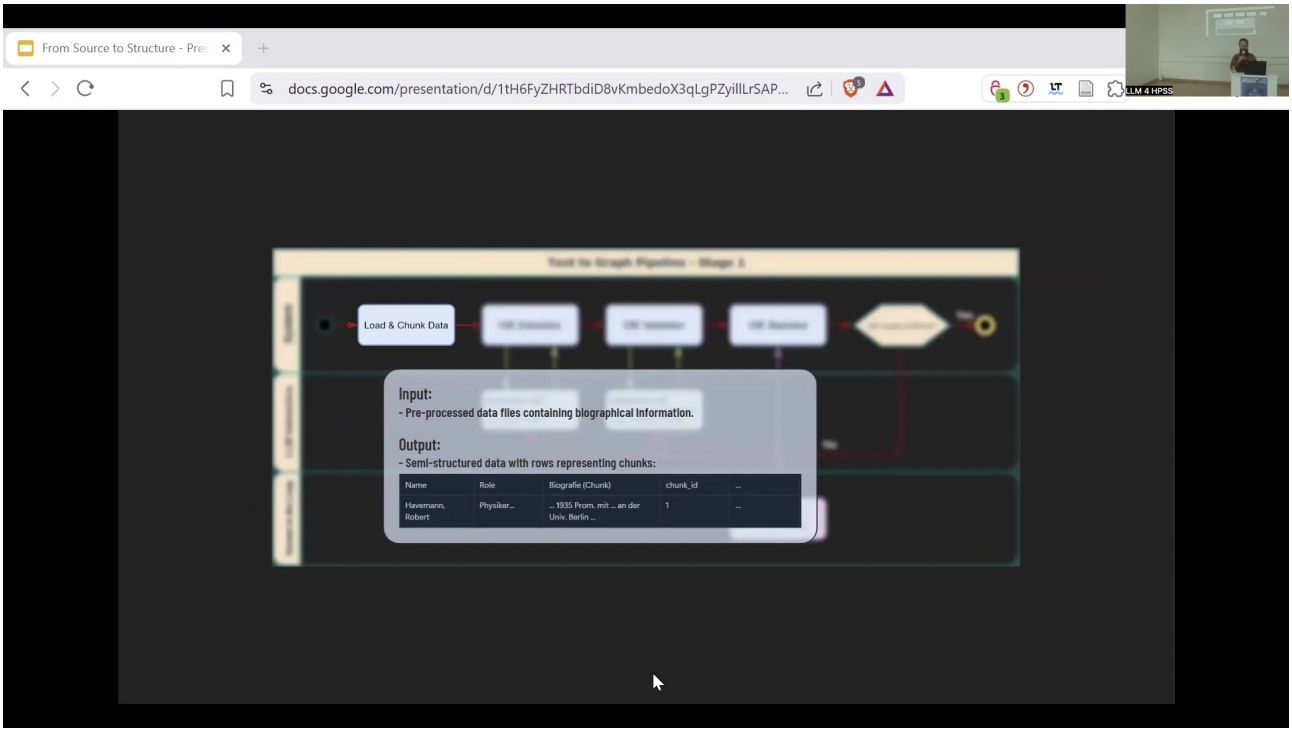

Figure 20.2: Slide 08

The process of transforming unstructured text into a structured knowledge graph is implemented as a two-stage pipeline. The first stage is responsible for extracting statements from the input text, while the second stage constructs the knowledge graph from these extracted statements, adhering to defined rules to ensure consistency and clear categorization.

The pipeline is built upon several core principles:

- It employs task decomposition, breaking down the complex process into smaller, controllable, and verifiable steps.
- It is data and research question driven, meaning the final structure of the knowledge graph is explicitly guided by the specific research aims of the project, rather than being solely dictated by a predefined ontology.
- Verifiability is integrated through human-in-the-loop steps at critical points in the process.

The input data for the pipeline often originates from messy, unstructured sources that typically require *Optical Character Recognition* (*OCR*) or scraping. Following these initial steps, the data undergoes preprocessing to achieve a semi-structured format before entering the pipeline.

The first central step of Stage 1 is *Open Information Extraction* (*OIE*). This step utilizes a *Large Language Model* (*LLM*) to extract all subject-predicate-object triples it can identify within the text, without relying on any preset categories or schemas. *OIE* is a rapidly evolving research area, with a significant shift towards using *LLMs* for this task.

The second step in Stage 1 involves *LLM* ensembling. A second *LLM* model is employed to validate and refine the initial statements extracted by the first model. This second model is specifically prompted to act as an adversary to the first, critically evaluating its output, correcting errors, identifying any missed triples, and assigning a confidence score to the extraction. This ensembling approach significantly enhances the quality of the extracted statements.

The final part of Stage 1 is evaluation. A sample of the validated output is evaluated against a corresponding sample created independently by domain experts. Classical performance metrics are calculated to quantify the quality of the extraction. This evaluation step represents the first instance of human-in-the-loop intervention, where domain experts judge the quality. A decision point is established: if the quality score meets a predefined threshold, the process moves to Stage 2; otherwise, Stage 1 is refined. The determination of what constitutes a "good enough" score is dependent on the specific use case and the characteristics of the dataset being processed.

20.4 Two-Stage Pipeline: Stage 2 - Knowledge Graph Structuring

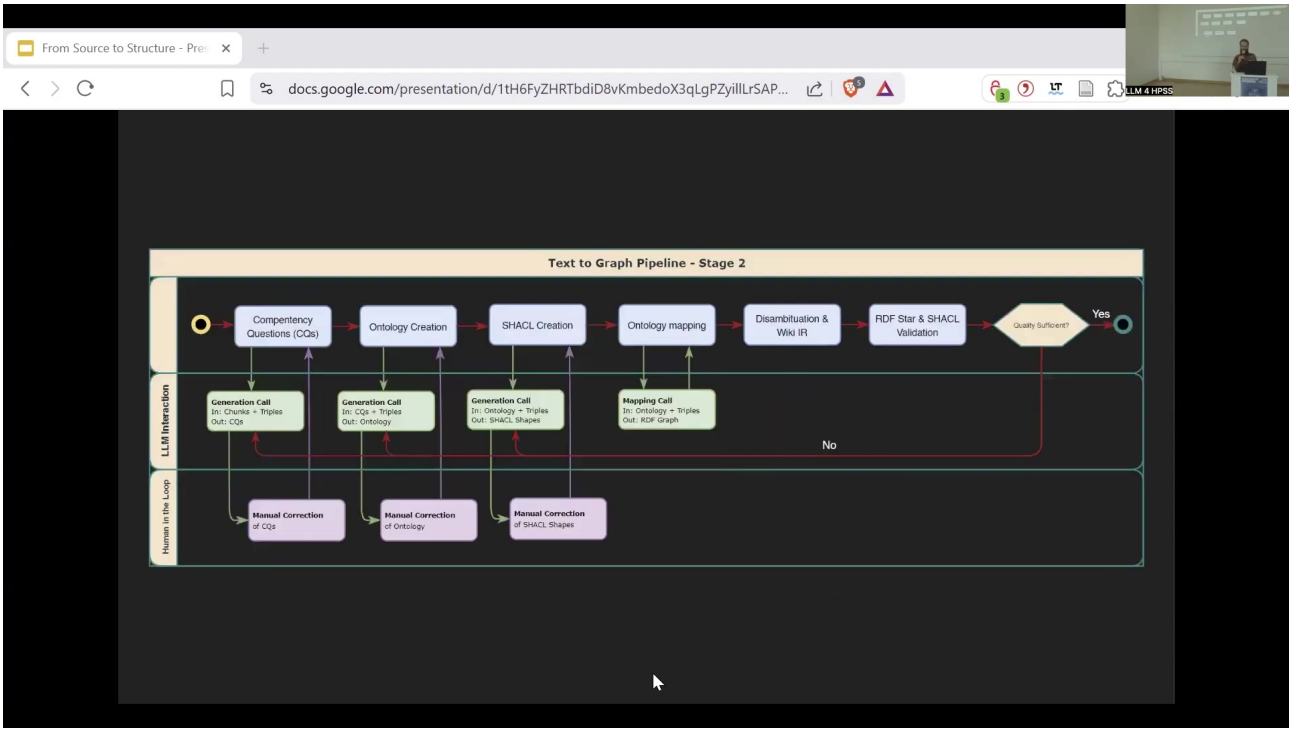

Figure 20.3: Slide 11

Following the extraction of statements in Stage 1, Stage 2 of the pipeline focuses on imposing further structure on this knowledge to form a knowledge graph.

The first step in Stage 2 involves defining competency questions. These are specific questions that the final knowledge graph is designed to answer. The purpose of starting with competency questions is to ensure that the structure of the knowledge graph is tailored to the specific research questions driving the project, making it research-driven rather than solely dependent on a predefined, potentially overly broad, ontology. A reasoning model is utilized to draft an initial set

of 20 to 30 competency questions. This draft is then refined and finalized by a human domain expert, incorporating a crucial human-in-the-loop step.

Building upon the competency questions and the extracted triples, the next step is to construct the ontology for the knowledge graph. A reasoning model drafts this ontology, which is subsequently finalized and corrected by a domain expert, again involving human oversight.

The final step in Stage 2 encompasses several tasks: entity disambiguation, data encoding, and metadata inclusion. Entity disambiguation involves resolving variations in names or references to the same entity, such as mapping “Humboldt Uni Berlin” to its standardized form “Humboldt Universität zu Berlin”. Entities are also resolved to external, standardized instances, such as those found in *Wikidata*. The data is then encoded, and relevant metadata is included. This metadata comprises the original source data, the scoring data generated during Stage 1’s evaluation, and the initial text chunk from which the triples were extracted. The output format for the structured data is an *RDF star graph*.

A final validation step is performed on the constructed knowledge graph. The resulting graph can then be exported in various formats depending on the intended use. Options include exporting it for network analysis into graph databases like *Neo4j* or storing it as a triple store to facilitate subsequent reasoning tasks. The pipeline is designed to handle the time dimension inherent in many biographical sources; if a time stamp is available for a statement in the source text, it is extracted along with the triple and modeled within the *RDF star graph*, which is crucial for analyzing developments and changes over time.

20.5 Use Cases and Applications

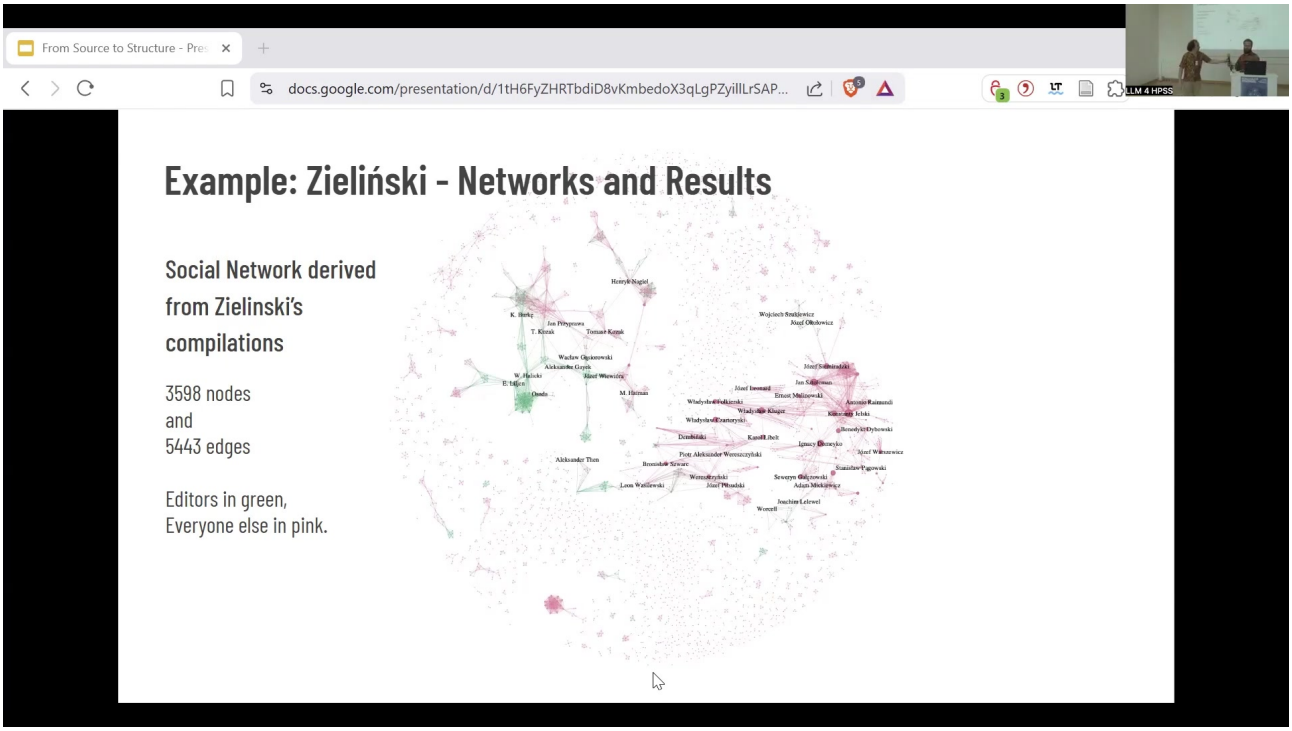

Figure 20.4: Slide 15

The primary motivation for this pipeline is to provide a controlled method for converting unstructured data into structured data, thereby enabling researchers to ask overarching questions that were previously impossible.

One key use case involves Zielinski's Polish biographical dictionary, a source compiled in 1930 as part of nation-building efforts during the formation of the Polish nation. This dictionary lists Poles who traveled abroad for exploration and other purposes. Although a PDF version is available, the printed format prevents structured querying. The pipeline enables researchers to ask specific questions about this corpus, such as how the migration of individuals facilitated the introduction of new ideas and fostered innovation in different locations, or the specific role played by editors listed in the dictionary in the spread of knowledge. An example result is a network graph, compiled by Alex Kay, visualizing the relationships between editors (represented as green nodes) and authors (represented as pink nodes). This type of network analysis, including the calculation of centrality measures and analysis across time, is not feasible manually from the printed PDF and provides novel information about the corpus.

Another use case is the biographical reference work "Who was who in the GDR". This source documents approximately 4,000 prominent figures from East German history, including politicians, dissidents, scientists, and artists. First published in the early 1990s, it was digitized and made available online in the 2000s, allowing for text searches but not structured queries. The source is noted to have a bias towards including individuals based on their fame. An example of a structured question that can be explored after processing this source with the pipeline is identifying the differences between individuals who received state awards and those who did not, specifically concerning their political affiliations and roles within the state apparatus.

An initial experiment using an early version of the pipeline, applied to 1,000 randomly sampled biographies from this source (not the final validated set), yielded preliminary findings. It indicated strong correlations between state awards such as the *Karl Marx Orden* and *Held der DDR* with affiliation to the *Socialist Unity Party of Germany* (*SED*) and holding political power. In contrast, the *Nationalpreis* showed a weaker link to high positions compared to individuals who did not receive any award. This type of analysis is highly relevant for *HPSS* research as it allows for a structural investigation into the interconnection between science and politics. It enables researchers to examine how factors like education, political affiliation, or positions of power varied across different generations or cohorts of scientists, providing insights into the dynamics of this relationship over time.

20.6 Conclusion, Challenges, and Future Work

In summary, the project facilitates a transition from viewing biographical sources as collections of isolated entities to enabling complex structural queries across the data.

The main challenges currently faced include improving the accuracy and robustness of entity disambiguation and establishing proper benchmarking methodologies to systematically evaluate the pipeline's performance.

Immediate future work involves refining and completing the current pipeline, which is presently considered a proof of concept. A key next step is to systematically compare its performance against other existing graph extraction pipelines, such as the *Neo4j graph builder* and the *Microsoft graph rack*.

Looking further ahead, future perspectives include utilizing the constructed knowledge graph in conjunction with the initially extracted text chunks to develop a graph *Retrieval Augmented Generation* (*RAG*) system. The objective of this graph *RAG* system is to allow users to query the entire dataset using natural language. Additionally, the project aims to develop methods for building multi-layered networks from the knowledge graph to support deeper structural analysis of the relationships within the data.

Chapter 21

References

Chapter 22

References

{.unlisted}