Daniel Brodén & Lina Samuelsson (Eds.)

FLOWS —&— FRICTIONS

Mixed Methods for AI-Driven Research on Historical Media

We would like to thank the participants of the Flows & Frictions symposium, held at the University of Gothenburg in 2024, for their inspiring talks: Mark Algee-Hewitt, Judith Brottrager, Antske Fokkens, Chris Haffenden, Martina Hjertman, Aram Karimi, Peter Leonard, Eivind Røssaak, Anton Törnberg, and Ted Underwood.

The editors gratefully acknowledge support from Riksbankens Jubileumsfond (RJ) for the research project *The Order of Criticism Revisited* (2020–2025; grant no. MXM 19-1096:1), which made this anthology possible.

Daniel Brodén & Lina Samuelsson (Eds.)
Flows & Frictions: Mixed Methods for AI-Driven Research on Historical Media

LIR.skrifter.12
Form: Richard Lindmark. Print: Stema, Borås 2026
ISBN: 978-91-89284-18-0

CONTENTS

This book is dedicated to Jonas Ingvarsson,
in fond memory

Daniel Brodén

INTRODUCTION

IN MAY 2024, the symposium Flows & Frictions, held at the University of Gothenburg's Faculty of Humanities, examined integrative mixed-methods approaches to advance data-rich research on historical media. It brought together researchers in comparative literature, history of ideas, film and media studies, sociology, and language technology from Sweden, Europe, and the United States to address key questions for state-of-the-art interdisciplinary approaches at the intersection of interpretative and computational analyses: How can we deepen the reflexive collaboration between these traditions in the study of historical media? How can mixed methods develop nuanced arguments about complex disciplinary issues in data-rich research? How should digitized collections be constructed not only as "data" but also as representations embedded in their historical and material contexts? And how can we sustain tensions between disciplinary perspectives as a source of creativity?

AT THE INTERSECTION OF THE COMPUTATIONAL AND THE INTERPRETATIVE

More broadly, the Flows & Frictions symposium connected to the ongoing discussions about interdisciplinary research within and beyond the field of digital humanities. A standard account casts data scientists as interested in advancing methodological development, while humanities researchers apply established disciplinary approaches to digital scholarship. However, scholars are posing more reflexive questions about "what is happening" and "what should happen" at the intersection of data-intensive and interpretative methods (Oberbichler et al.

2021; Ahnert et al. 2023). This includes developing genuinely collaborative, integrative workflows in which humanities researchers and data analysts engage in dialogic, reflexive exchange (Rockwell and Sinclair 2016; Fickers et al. 2022).

Arguably, the long-standing debates in digital humanities over whether data-driven research yields more "objective" knowledge have become less central. In computational literary studies, Matthew Jockers characterized text mining-based literary research as "a careful and exhaustive gathering of *evidence*" (Jockers 2013, 5–6, emphasis added), and Franco Moretti (2013) attributed a different kind of scientific credibility to computational approaches. In recent years, however, such claims have faced sustained critique. Already in 2012, media and communication scholars danah boyd and Kate Crawford noted a near-mythic belief that large unstructured datasets, paired with machine learning and natural language processing (NLP), can offer a higher form of knowledge (boyd and Crawford 2012, 663). Sociologist Simon Lindgren argues that overconfidence in the capacity of computational analysis to extract "evidence" and "truths" from large datasets reflects a problematic hierarchy and dichotomy of scientific values: while quantitative, statistically oriented methods are often aligned with ideals of "objectivity," "reliability," and "precision," qualitative approaches, typically involving close, contextual interpretation, are cast as more subjective (Lindgren 2020, 13–14). Commentators also caution against treating data as raw material. Media historian Lisa Gitelman and literary scholar Virginia Jackson observe that this stance "often leads to an unnoticed assumption that data are transparent, that information is self-evident, the fundamental stuff of truth itself" (Gitelman and Jackson 2013, 2). Data do not simply exist "out there"; they are produced and shaped by choices in their collection, processing, and analysis (Manovich 2020, 128).

Building on such critique, recent work in digital humanities emphasizes the importance of historical and cultural context in interpreting archival data. Aggregating data from large-scale collections is a mediated process that tends to flatten complex information structures and obscure context. As computational literary scholar Katherine Bode argues, relying on data models that fail to represent how the original materials generated meaning in their time is problematic from a methodological perspective (Bode 2018, 5). Likewise, digital historian Jo Guldi (2023) warns that without awareness of the original context the computationally driven approaches risk yielding empty or even misleading results. According to Guldi, decontextualized modeling of historical and cultural data is "dangerously cut off from reality. Text

mining can aid us by helping humans become more accurate at sifting data at scale. Without the insights of the humanities, text mining falls readily into error: it is a monkey reaching for the moon, guided only by the moon's reflection in a pond, where he will never grasp it" (Guldi 2023, 6). By contrast, context-sensitive, interdisciplinary approaches to data analysis can both deliver robust results and convert scale into grounded explanation, for instance taking discrepancies between model outputs and close readings related to models, corpora, and digitization into account.

AMID THE RISE OF LARGE LANGUAGE MODELS

The Flows & Frictions symposium grew out of the research project *The Order of Criticism Revisited: A Mixed-Methods Study of a Century and a Half of Literary Criticism in Sweden* (2020–2025; Ingvarsson et al. 2022; see Ch. 5), which explores how traditional and computational methods can complement and enhance one another, from both practical and epistemological perspectives. The project's rationale informed the symposium, in which participants – many of whom appear in this volume – discussed the conditions for state-of-the-art interdisciplinary research on media data, working from the premise that productive tensions can exist between computational and interpretative approaches. Hence the title Flows & Frictions, used for both the symposium and this anthology. "Flows" denote, for instance, collaboration that moves through shared questions, integrative workflows, and mediation between data modeling and interpretation. By contrast, "frictions" mark disagreements over what counts as evidence, how validity and reliability are assessed, or how to build high-quality datasets. In reflexive collaboration, flows move interdisciplinary work forward, while frictions preserve disciplinary standards and multidimensional frameworks.

Discussions at the symposium emphasized that mixed-methods work in digital humanities is not merely about combining techniques. Participants expressed concern that the field's focus on tools and outputs continues to outpace thorough methodological and epistemological reflection on what is being modeled and to what end. In this respect, the theoretical discourse within the field was seen as still underdeveloped. These concerns intersected with debates on digitization and the material transformation of media archives. While digitization has opened vast corpora to new forms of analysis, participants echoed Bode's (2018) insistence on attending to the conditions under which historical archives are created. Digitization quality and circumstances, selection criteria, metadata schemas, and copyright constraints all shape what becomes

visible and analyzable in digital form. At the same time, the flattening of complex, historically situated materials into data-ready formats was seen not simply as a flaw but as a methodological reminder that digital archives are never neutral.

Although large language models (LLMs) and generative models were not the original focus of the call, discussions soon gravitated toward them as participants debated how these systems are redrawing methodological boundaries. As this reconfiguration is still ongoing and its implications for scholarship remain in flux, the exchanges detailed here therefore capture a field in motion – a snapshot of the debate as it stood in spring 2024. In fact, the seemingly simple practice of prompting LLMs emerged as a concrete site where disciplinary assumptions meet model behavior. Prompting, the symposium noted, appears to straddle the humanistic and computational divide: it relies on human expertise to craft inputs that elicit analytically interesting outputs that resemble interpretative variation. This, in turn, raises a critical question: should prompting be considered a methodological practice in its own right, and, if so, what kind of knowledge does it produce?

While generative models lower technical thresholds and can simulate certain forms of reasoning, they also blur the line between "genuine" interpretation and surface-level mimicry. To mitigate their still-opaque reasoning processes, participants proposed practices such as training LLMs to cite historical sources with scholarly rigor or prompting different models to argue with one another. Another issue highlighted at the symposium was the prospect of LLMs fine-tuned on specific historical materials. Rather than relying solely on general pre-training, such models were seen as promising for more context-sensitive analyzes aligned with the interpretative strengths of the humanities. At the same time, the idea prompted critical questions: to what extent might such historical models support rigorous scholarly inquiry, and to what extent might they risk producing speculative reconstructions, akin to attempts to resurrect a mammoth? These reflections underscored both the perceived possibilities and the inherent limitations of LLMs in humanities research. Crucially, the models' inability to read, contextualize, or substantiate claims as human scholars do was not framed as a mere technical shortcoming, but as a fundamental difference in epistemic logic.

Consequently, the concept of "ground truth" raises instructive questions in relation to LLMs. In computational contexts, ground truth typically implies a fixed standard against which outputs can be evaluated and measured. In the humanities, however, "truth" is more often understood as contingent, interpretative, and perspectival. This

divergence complicates the benchmarking of complex outputs, such as text generation or classification, turning what might seem like a technical issue into an epistemological one. Accordingly, participants suggested that the value of modeling in humanities contexts may lie less in straightforward verification than in surfacing ambiguity and multiplicity, allowing competing perspectives to co-exist within a single interpretative framework. The symposium also closed with reflections on reflexivity in integrating computational tools, such as LLMs, into humanities research. Whether experimenting with models, evaluating generated texts, or exploring new forms of scholarly interaction, there was a shared sense that attentiveness to complexity and interdisciplinarity is as vital as technical innovation.

IN THE CHAPTERS THAT FOLLOW

The focus of the Flows & Frictions symposium is reflected in this anthology. Bringing together key contributions, the volume showcases a range of approaches that integrate computational and interpretative approaches for context-sensitive, data-rich research on historical media. Each chapter addresses critical questions about the methodological possibilities of integrative interdisciplinary research, the challenges of working with digitized collections, and the evolving role of computational techniques in historical research on media data – an area increasingly shaped by deep learning and LLMs. Ranging from high-level frameworks to focused studies, the chapters highlight both the interdisciplinary nature of the "digital turn" and the dynamic interplay between computational innovation and context-sensitive interpretation. By pairing theoretical reflections with empirical insights, the volume aims to inspire further conversation, experimentation, and collaboration across perceived disciplinary boundaries.

The opening chapter, "The Function of Cultural History in an Era of Cultural Automation," by literary scholar Ted Underwood, argues that generative AI is best understood not as abstract intelligence but as "cultural automation" – a technology that reproduces and transmits collective patterns of behavior learned from vast archives of text and images. This shift reframes the relationship between humanities and technology: instead of asking what language models can do for historians, Underwood asks how the study of culture and historical change can contribute to improving these powerful new systems. He argues that historians can help models learn to reason about conflicting social perspectives, a task that will matter not only when models interpret the

past but whenever they have to consider different vantage points in our pluralistic present.

Sociologist Anton Törnberg's follows with "Towards a Third Generation of Natural Language Processing: Enhancing Qualitative Research with Large Language Models." He proposes a typology of text analysis methods ranging from surface-level to contextual and cultural interpretation, offering a framework for understanding the evolving role of NLP in qualitative research. While first- and second-generation NLP methods – such as keyword extraction, topic modeling, and word embeddings – have extended analytical reach, they remain limited in capturing meaning shaped by ideology, discourse, and context. The emergence of LLMs represents a third generation NLP, capable of performing context-sensitive analytical tasks such as identifying metaphors, framing, and rhetorical strategies. This shift enables new hybrid approaches that integrate computational efficiency with qualitative depth. Törnberg's chapter introduces two frameworks – AI-augmented grounded theory and theory-driven AI analysis – to illustrate how LLMs can support large-scale, context-sensitive interpretation. These "fusion methodologies" challenge the perceived divide between computation and interpretation, pointing towards a new paradigm for qualitative inquiry in the digital age.

In "Large Language Models and Text Analysis for the Digital Humanities: A Short History and Practical Guideline for Automatic Classification," computational linguists Antske Fokkens and Eleanor Smith outline how LLMs have led to impressive performance on a variety of tasks involving language analysis while emphasising that humanities data is often more complex than the more generic data on which these models are mostly trained. This chapter aims to address the question of where we stand when it comes to applying text analysis for digital humanities research now that we have LLMs. After providing a generic explanation of what makes LLMs useful for text analysis, Fokkens and Smith outline where challenges and opportunities for text analysis (with or without language models) lie and illustrate challenges concerning the use of imperfect tools in research through three examples of digital humanities studies. They argue for careful evaluation that methodically assesses whether errors interfere with the hypothesis that is being tested, concluding that solid evaluation opens the door to new opportunities for using text analysis in digital humanities studies.

Film and media scholar Daniel Brodén, literary scholar Lina Samuelsson, and computational linguist scholar David Alfter collaborate in "Retouching and Refiguring Literary Criticism: Experiments with a Generative Model for Analyzing Book Reviews." Their study examines

the use of GPT-4o to analyze Swedish literary criticism from 1905–1906. Drawing on defamiliarization and the notion of distant technology, and framing AI as artificial communication rather than intelligence, the chapter undertakes two linked experiments to enrich analysis of familiar material based on a previous study by Samuelsson (2013). First, Brodén, Samuelsson, and Alfter address the challenge posed by the poor quality of digitized newspaper texts in the National Library of Sweden's (*Kungliga biblioteket*, hereafter KB) collection, showing that the model's output aligns more closely with the originals than does the noisy OCR, while arguing that it is better understood as a probabilistic retouching rather than a reconstruction, thereby introducing epistemological uncertainty. Second, using zero-shot prompting, they ask the model to identify discursive patterns and evaluative criteria in the reviews. The results suggest that GPT-4o can provide analytically meaningful perspectives, but its opacity creates methodological distance that calls for analytical caution, further re-readings, and reflection.

In "Unlocking the Archive: Exploring Literary History through Word Embeddings," literary scholar Judith Brottrager demonstrates how literary historiography can serve as a rich source for computational analysis. While literary history has frequently been modeled through metadata such as publication dates or genre labels, this chapter shifts focus to the discursive structures of literary historiography itself. Drawing on 27 literary histories and companions on eighteenth- and nineteenth-century European Anglophone literature, Brottrager applies a fine-tuned spaCy NER (named entity recognition) model to extract and disambiguate references to authors and texts, linking them to Wikidata. The resulting corpus is used to train a word2vec embedding model that captures relational patterns between authors and texts based on the context in which they appear in literary historical narratives. This approach demonstrates how word embeddings, when grounded in historiographical discourse, can surface latent structures of literary comparison, offering new perspectives on canon formation and literary evaluation.

Finally, "Searching without Labels: Multimodal AI and Image Search at the National Library of Sweden," by historian of ideas Chris Haffenden, data scientist Faton Rekathati, and product manager Emma Rende, explores how recent developments in computer vision and NLP are reshaping the ways researchers interact with large collections of digital images. Drawing on work from KBLab at the KB, it focuses on the development of an AI-driven image search prototype that uses CLIP (contrastive language–image pre-training) for vector-based visual discovery. The chapter reflects on the new research opportunities that emerge when image search becomes decoupled from traditional meta-

data, enabling novel forms of exploration across GLAM (galleries, libraries, archives, museums) collections. It also considers the technical and epistemological challenges of working with historical image data, including the extraction of images from newspaper archives and their integration into searchable infrastructures. Finally, Haffenden, Rekathati, and Rende raise broader questions about what constitutes "an image" in such systems, how "context" is operationalized, and what implications these choices have for research in visual culture and digital heritage.

Taken together, these contributions can be considered within what visual culture scholar Florian Cramer calls the post-digital; "the messy state" of media and society *after* digitization (Cramer 2015, 19). Arguably, the most interesting thing about the digital is no longer its presumed "newness." Rather, the post-digital refers to a condition "in which the disruption brought upon by digital information technology has already occurred" (Cramer 2015, 20; cf. Berry and Fagerjord 2017, 16). In the context of this anthology, the notion of the post-digital invites us to reflect on the current flows and frictions between computational and interpretative methods, including the evolving role of LLMs in humanities research, not as oppositions but as matters of contemporary practice to be debated, negotiated – and put to work.

REFERENCES

Ahnert, Ruth, Emma Griffin, Mia Ridge, and Giorgia Tolfo. 2023. *Collaborative Historical Research in the Age of Big Data: Lessons from an Interdisciplinary Project*. Cambridge University Press.

Berry, David M., and Anders Fagerjord. 2017. *Digital Humanities: Knowledge and Critique in a Digital Age*. Polity.

Bode, Katherine. 2018. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press.

boyd, danah, and Kate Crawford. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon." *Information, Communication & Society* 15 (5): 662–79.

Cramer, Florian. 2015. "What is 'Post-Digital'?" In *Postdigital Aesthetics: Art, Computation and Design*, edited by David M. Berry and Michael Dieter. Palgrave Macmillan: 12–26.

Fickers, Andreas, Juliane Tatarinov, and Tim van der Heijden. 2022. "Digital History and Hermeneutics – Between Theory and Practice: An Introduction." In *Digital History and Hermeneutics: Between Theory and Practice*, edited by Andreas Fickers and Juliane Tatarinov. De Gruyter Oldenbourg: 1–19.

Gitelman, Lisa, and Virginia Jackson. 2013. "Introduction." In *"Raw Data" is an Oxymoron*, edited by Lisa Gitelman. MIT Press: 1–14.

Guldi, Jo. 2023. *The Dangerous Art of Text Mining: A Methodology for Digital History*. Cambridge University Press.
Ingvarsson, Jonas, Daniel Brodén, Lina Samuelsson, Victor Wåhlstrand Skärström, and Niklas Zechner. 2022. "The New Order of Criticism: Explorations of Book Reviews Between the Interpretative and Algorithmic." *The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, edited by Karl Berglund, Matti La Mela, and Inge Zwart. CEUR-WS: 228–34.
Jockers, Matthew. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.
Lindgren, Simon. 2020. *Data Theory: Interpretive Sociology and Computational Methods*. Polity.
Manovich, Lev. 2020. *Cultural Analytics*. MIT Press.
Moretti, Franco. 2013. *Distant Reading*. Verso.
Oberbichler, Sarah, Emanuela Boroş, Antoine Doucet, Jani Marjanen, Eva Pfanzelter, Juha Rautiainen, Hannu Toivonen et al. 2021. "Integrated Interdisciplinary Workflows for Research on Historical Newspapers: Perspectives from Humanities Scholars, Computer Scientists and Librarians." *Journal of the Association for Information Science and Technology* (73) 2: 225–39.
Rockwell, Geoffrey, and Stéfan Sinclair. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press.
Samuelsson, Lina. 2013. *Kritikens ordning: Svenska bokrecensioner 1906, 1956, 2006*. Bild, text & form.

Ted Underwood

THE FUNCTION OF CULTURAL HISTORY IN AN ERA OF CULTURAL AUTOMATION

FOR PEOPLE WHO use computers to study cultural artifacts, the early 2020s were a period of dizzying change. A few years earlier, using machine learning to analyze texts and images had required specialized training. But with the release of ChatGPT in late 2022, sophisticated natural language processing (and soon, image analysis) was made available to millions of daily users. The technology became more powerful and more accessible, but also more controversial. In journalism, the cautious academic phrase "machine learning" was soon displaced by "artificial intelligence."

This chapter will argue that these changes have opened up new roles for cultural history. Generative AI is certainly a new tool for researchers, and there are specific questions to be posed about its limits and affordances in historical research. But here I am more interested in flipping the question. What can the study of culture contribute to AI, and how does the growing centrality of AI in contemporary life reshape the *purpose* of cultural history? I will start by exploring the premise that generative AI is a model of culture – specifically of patterns of behavior that have solidified and precipitated as images, text, or video. Then I will ask what history, as the "science of change," can contribute to models of culture (Bloch 2004, 22).

GENERATIVE MODELS AS CULTURAL TECHNOLOGIES

One reason to pose this question so broadly – instead of focusing on specific fields like "digital humanities" or "computational social science" – is that accessible conversational interfaces have started to

blur the boundary between computational and non-computational research on culture. In the 2010s, that boundary had real substance, often expressed through a contrast between "close" and "distant" reading. Computers had a comparative advantage in exploring large corpora and long timelines, whenever the variables of interest could be operationalized as countable features. But computation became mostly useless at smaller scales of analysis, where the questions worth exploring usually involved complex qualitative inferences about tone, motivation, or context.

I once cared enough about this contrast to write a book titled *Distant Horizons* (Underwood 2019). But I confess that the sophistication of contemporary language models is rendering the contrast between close and distant reading largely moot. Computational approaches to text are no longer limited to countable features. Language models can answer subtle questions about tone and motivation. Indeed, a benchmark has been developed explicitly to measure their facility at "close reading" (Sui et al. 2025). Expert human readers still perform better than models on that benchmark. But language models are performing well enough across a wide range of tasks to force professors in many fields to redesign their essay assignments. Whether that is good or bad for students' education is still being debated (Wang and Fan 2025). But the controversy itself shows that computational methods are infiltrating questions and topics that seemed off-limits to them a decade ago.

The infiltration is interestingly reciprocal. If computational models are now able to address complex qualitative questions, it is because they have absorbed complex qualitative heuristics from the corpora they were trained on. It is no longer very meaningful to ask what algorithm is being used when a language model answers a question. "Transformer" may describe the model architecture, but at inference time the model's substantive work is done by heuristics it learned from its training corpus – fuzzy templates that allow it, for instance, to draw inferences about a character's state of mind from overtones in dialogue.

This is one reason why generative models (those working with images and video as well as text) are increasingly described as "cultural technologies." The phrase originated with Alison Gopnik, a psychologist who used it to define language models as systems that only reproduce and transmit patterns of behavior – by contrast with the active creative intelligence she observed among children (Gopnik 2022). But the concept of "cultural technology" is not just a way to distinguish generative models from human minds; it can also be a way to distinguish them from older forms of computational analysis.

When a researcher trained a regularized logistic regression model in

2015, substantive work was done by assumptions about the mathematical relationship between variables. The coefficient of each variable was learned from the data, to be sure, but the idea of logistic regression prescribed the structure and purpose of the whole system. As deep learning and transfer learning were adopted in the late 2010s, it still remained true that the function of a model like BERT was constrained by the "model head" a researcher chose to apply, which told BERT (for instance) whether to assign labels to sentences or to estimate the probability that one sentence would be followed by another. But this changed decisively with the advent of autoregressive generative models. The model interface is now loosely conversational. The purpose of a model's output is defined only by the verbal instructions it receives and the analogous patterns it has seen during training. To understand the model's limitations requires understanding the limits of its cultural repertoire. A cultural technology, in short, is really different from an algorithm: generative AI belongs to a different category of tools.

When Gopnik first introduced the concept of cultural technology, it had a somewhat dismissive edge – because, as a psychologist, she was interested in contrasting models to brains. But Gopnik has since refined and expanded the concept of cultural technology in collaboration with a larger group of social scientists, who stress that language, and other technologies that enable humans "to learn from the experiences of other humans […] are arguably the secret of human evolutionary success" (Farrell et al. 2025, 1153). To call language modeling a cultural technology in this sense is not necessarily to belittle it: other cultural technologies might include writing, markets, and the printing press.

CULTURAL AUTOMATION

However, even observers who see cultural technologies as crucial historical developments have a habit of contrasting them sharply to "intelligence" and "agency." The contrast between culture and agency is central, for instance, in the *Science* article by Henry Farrell et al. that characterizes generative AI as a cultural technology: "These technologies are only possible […] because humans have distinct capacities characteristic of intelligent agents. Humans, and other animals, can perceive and act on a changing external world, build new models of that world, revise those models as they accumulate more evidence, and then design new goals" (Farrell et al. 2025, 1154).

Much of this is valid. It is true, for instance, that the conversational interface typically used for generative AI is not set up to initiate anything: it is structurally designed to be reactive, as the very concept of

a "prompt" implies. So if we define an agent as an entity that autonomously originates new courses of action, it would be correct to say that language models are not agents.

But of course, entities that originate new courses of action entirely on their own are easier to find in theory than in practice. It is typically impossible to trace human ideas and actions back to a single point of origin. Even our most creative artists and most courageous leaders take action in response to events and guided by cultural patterns. When we describe them as agents, we mean that they transform those influences and patterns in unexpected ways.

It may be helpful to envision agency as a continuum rather than a crisp boundary. A stump is just an object. You can do things to it but cannot usually do anything *with* it. A hammer is also an object – but we call it "a tool" because it can mediate agency. Cars and computer programs are also tools. But they mediate human intent in more complex and less predictable ways. Moreover, they elicit actions their users might not otherwise have imagined. An operating system is a tool, but not "merely a tool." It opens a whole new space of potential actions for its user.

A language model, similarly, is a tool – but in a looser and less predictable sense than a printing press. True, both systems reproduce and distribute patterns of symbolic behavior. But the cultural templates being distributed remain inert on the platen of a press. In a language model, they can recombine with each other to mediate the user's instructions in unexpected ways. Social scientists are right to push back against the assumption that models are (or should be) analogous to human minds and are right to suggest that models may more closely resemble cultural systems. But that does not mean a language model has no agency. On the contrary, the blurriness of the boundary between collective culture and agency is what this technology interestingly exposes.

The distinctive feature of generative AI is that collective cultural templates – which we ordinarily expect to affect the world only when mediated by human actors – become independent forces in the model, directly shaping its output. This potential distinguishes large generative models from earlier cultural technologies, and it is what I mean by characterizing generative AI as a mode of "cultural automation."

Automation is a suitable metaphor because we already use the word to describe a gray area where tools – without acquiring personhood – become provisionally independent of a worker's guiding hand. In industrial automation, cybernetic feedback systems like sensors or governors temporarily replace direct manual supervision of physical work. In cultural automation, knowledge work is similarly supervised by collective patterns of behavior inferred from training data.

Any process that transmits collective patterns of behavior is transmitting culture. But a generative model's access to these patterns is, for the most part, less direct than a human child's. During reinforcement learning, a model can in principle learn directly from interaction with living people. But most of what a model learns is learned earlier, in pre-training. And in the pre-training process the model will not encounter any people – only traces of behavior that have solidified as artifacts (images, text, music, or video). Pre-training is analogous then to the specific subset of cultural transmission mediated by art objects, communication technologies, and schools.

Once we reframe AI discourse as a question about "cultural automation," it becomes almost self-evident that the study of culture is acquiring a new kind of practical utility in the twenty-first century. To understand what models can do, tune them for particular tasks, or expand their capabilities, we need to think about the patterns of behavior they instantiate.

WHY CULTURE APPEARS PERIPHERAL

The connection between generative AI and culture is so strong that it may be worth asking why it has not appeared obvious to every well-informed observer. And it definitely has not appeared obvious! AI researchers, for instance, understand perfectly well that the capabilities of their models emerge from training data. But they do not typically equate model performance with culture. They treat culture, separately, as a question about bias or diversity. I want to quickly review this discourse, to reconnect its sundered halves, and show how culture is already concretely central to AI research.

When researchers need to reason about the connection between training data and model capability, the phrase they use most commonly is "data quality." The importance of data quality has been widely acknowledged. Meta's paper describing Llama 3, for instance, admits that "it does not deviate significantly from Llama and Llama 2 [...] in terms of model architecture; our performance gains are primarily driven by improvements in data quality and diversity as well as by increased training scale" (Grattafiori et al. 2024, 6). Reports like this have led at least one writer to quip that "Data Quality May Be All You Need" (Edwards 2024).

But then, what is "data quality"? The concept of quality is invoked to describe a range of disparate things. Large corpora scraped from the internet are said to be low quality in part because they contain too much repetition. So recipes for improving quality often focus on

deduplication (Penedo et al. 2024). "Datasets with long, coherent documents" (such as books) have sometimes been seen as preferable to shorter web-scraped genres, although this is contested (Penedo et al. 2023, 4). Filtering may identify quality, in practice, with particular topics, social roles, and regions (Lucy 2024). Finally, data diversity is sometimes understood as a dimension of quality or as a related concept. But diversity can be understood in a range of different ways. In practice, it is commonly operationalized as a search for optimal balance. For example, the Llama 3 paper stresses that "it is essential to carefully determine the proportion of different data sources in the pre-training data mix" – what percentage should be code, what percentage should be English, and what percentage should represent languages other than English (Grattafiori et al. 2024, 6).

These conflicting definitions of the concept make clear that "data quality" is far from a single axis of measurement. When researchers talk about quality, they are talking about a problem with many orthogonal dimensions. "Quality" is shorthand for a range of choices about genres, languages, topics, social roles, and geographical regions. The decisions required to define a particular form of quality are actually choices between different discursive traditions – or, in short, choices about culture. It becomes possible to treat this multidimensional problem as a simple quality measurement only because there is a high degree of industry consensus about desired behavior.

Once the cultural practices that create model capability have been redefined as a single scalar measure of "quality," the concept of culture shrinks, and becomes peripheral to discourse about AI. The aspect that remains to be acknowledged involves only the variance of normative judgments and mores across nationalities and ethnicities. A recent survey of 90 papers about culture in language models makes this pattern clear. "Culture is, almost always, described at the level of a community or group of people, who share certain common demographic attributes," and the proxies for cultural identity are typically "emotions and values, food and drink, kinship terms, social etiquette, etc.," with "values and norms" particularly prominent (Adilazuarda et al. 2024, 15766).

Normative variance is indeed an aspect of culture – but it is only one aspect. There are a variety of ways of characterizing what gets left out. Muhammad Farid Adilazuarda et al. emphasize that the focus on values and norms leaves other aspects of culture understudied. "We have not encountered any studies on semantic domains of quantity, time, kinship, pronouns and function words, and so on" (Adilazuarda et al. 2024, 15770). Naitian Zhou et al. offer a somewhat broader critique, suggesting that – no matter what semantic domain is involved – an account

of culture should do more than trace "a mapping between the space of beliefs and the cultural groups that they index" (Zhou et al. 2025, 6). The task of cultural NLP is not just to discuss culture from outside, but to localize a model in a particular context, and "evaluate cultural performance in a situated application setting" (Zhou et al. 2025, 8).

While I agree with both of these accounts, I think the aspect of culture excluded from scholarship on language models can be defined even more broadly if we invoke William H. Sewell's categories from "The Concept(s) of Culture" (1999). Sewell distinguishes the concept of culture as "an aspect of social life" (always singular, because it is not countable) from references to "a concrete and bounded world of beliefs and practices," which can take a determiner ("a culture") or be pluralized ("cultures") (Sewell 1999, 39). The uncountable sense of the word distinguishes culture broadly from the economic or political dimensions of social life. The countable sense, invoked more often in anthropology, distinguishes a culture – from other cultures.

When researchers think about culture in relation to language models, they have almost always been interested primarily in the countable sense of the term. This is the reason for the centrality of the "mapping" between beliefs and groups that Zhou et al. (2025) have highlighted. Practically speaking, researchers' reason for investigating the cultural dimension of language models is almost always to test whether the models have achieved sufficient coverage of national, ethnic, and linguistic differences.

These differences are conceived as representing an evaluative axis orthogonal to simple capability. Since benchmarks for model capability are typically measured on a vertical axis, we might imagine this kind of cultural diversity as a horizontal measure analogous to geographic range. The presumption that these are separable dimensions also helps to explain why cultural databases have focused so narrowly on "values and norms" (Adilazuarda et al. 2024, 15770). If culture needs to be treated as orthogonal to capability, it will help a great deal for culture to be limited to subjective topics (emotional responses, values, mores around eating and taking off your shoes, and so forth). If we included objective questions of performance (say, ability to write alt text for an image) in a benchmark for culture, we might get the two axes conflated.

This division has become so common that it may seem self-evident and unproblematic. But if we shift back to the uncountable sense of culture, it becomes clear that culture cannot be treated as orthogonal to capability. An instruction manual is a cultural form, as is the practice of writing alt text for an image. They are, in other words, symbolic and sense-making practices not formalized as political or economic insti-

tutions. These cultural practices do not work very well as examples of cultural *difference* in the twenty-first century, because instruction manuals and alt text are common in Japan and Greenland as well as France. Perhaps there are differences between Japanese and French instruction manuals. But it is not the differences across nations or the subjective values entailed in those differences that make instruction manuals cultural. They are cultural (in the uncountable sense) simply because a manual is a transmitted symbolic practice that presupposes particular habits of reasoning, conventions of representation, and assumptions about the reader's relationship to a task. Even if all ethnicities and nationalities in the twenty-first century agreed about the ideal form of an instruction manual, a benchmark for writing instruction manuals would still be measuring performance of a cultural task whose meaning and evaluation depend on a context of shared practices.

Culture, in this broader sense, cannot really be extricated from the task of evaluating a generative model's ability to reason, follow instructions, and draw inferences. We need to recognize that generative AI is cultural – not only to acknowledge different normative perspectives – but to correctly understand and evaluate the substantive work it does.

WHAT CAN THE STUDY OF CULTURE CONCRETELY CONTRIBUTE TO GENERATIVE AI?

The sense of culture I have called "uncountable" is theoretically important in disciplines like sociology. But it is admittedly so broad that its concrete significance can seem elusive. It may be worth pausing to ask whether it really matters in the present context. AI researchers have focused on culture as difference and diversity, after all, for pragmatic reasons. Instruction manuals, short stories, and, for that matter, mathematical proofs may technically be cultural practices, in the uncountable sense of the word. But technology companies do not usually have to be reminded to design benchmarks that measure a model's performance in these common genres. On the other hand, companies likely do need to be reminded to design models that respect, say, the diversity of Indonesian funeral traditions. As long as our concern with culture remains limited to this practical situation – a situation where companies can be expected to evaluate common practices reliably but need nudging to address smaller communities – it will make sense to treat "culture," effectively, as a synonym for ethnic and national difference.

But I would suggest that it is already possible to see beyond the edges of that situation to a range of broader emerging problems. As models saturate benchmarks for straightforward factual tasks, researchers

increasingly struggle to evaluate complex tasks that depend on cultural context for meaning. Consider, for instance, the task of assessing creativity. It is true that different nations likely define and evaluate creativity somewhat differently. But that is not the fundamental thing that makes it difficult to evaluate. Evaluating creativity remains hard even within a single national context. Recent empirical studies show that models can ace conventional tests of creative thinking (like the Alternative Uses Test) yet still fall short of human performance when judged directly by humans or assessed for originality (Haasse et al. 2025). The concept of creativity is notoriously complex and hard to define, and there may be more than one reason why it is hard to benchmark. But I would suggest that we are discovering, at least in part, that creativity is hard to measure without some model of its cultural function. Where it shades into originality, for instance, creativity really describes a transformative interaction with existing ideas and practices rather than something like "pace of idea generation" that could be measured in a void. Some ideas that counted as creative in 2025 are no longer creative in 2030.

Even concepts that seem less subjective than creativity – concepts like persuasiveness and clarity – likely also depend on a reader's existing assumptions and cultural context. And these cultural conditions are likely to vary, not only across national boundaries, but within a single community. The same thing holds true for questions about alignment and model personality: a user's tolerance for "sycophancy," for instance, may depend on norms of interaction in their profession or social group. In short, the fact that cultures have boundaries is not the only reason we need to be conscious that language models are cultural artifacts. Cultures also contain internal complexity and variation, and they change across time. As we ask models to perform increasingly complex qualitative tasks, it may be hard to evaluate performance effectively without understanding how the definition of success depends on cultural context.

The centrality of culture to model performance suggests that humanists can guide generative models not just by critiquing them or pointing out things they overlook, but by positively enabling them to tackle more complex problems. In an era of cultural automation, understanding culture will have enormous value. For instance, we are very far from fully understanding creativity and innovation. We could learn a lot by discovering how rapidly our benchmarks for these things become dated, why models fall short of human performance, and what kinds of training data help them perform better.

THE ROLE OF CULTURAL HISTORY

Up to this point, I have been writing about "the study of culture" – a deliberately loose phrase that could encompass social sciences like sociology or information science along with more historically oriented approaches to culture in humanities disciplines like history, art history, and literary studies. But this anthology is specifically concerned with applications of computation to historical media. What role should the study of the cultural *past* play in shaping and guiding generative AI?

I should start by acknowledging that historical fields have some clear disadvantages when research is intended to guide the present. The past is by definition gone. The people living there are not likely to become customers, and their needs are not a top priority for anyone building contemporary AI systems. Moreover, the study of the past tends to be most central in humanities disciplines, and humanists typically have less experience than social scientists with some of the methods that will be needed to concretely guide model development. I am referring not only to quantitative methods but also to research with human subjects, which is often necessary if researchers plan to assess a model's performance or its effects on users.

However, historical disciplines do also have some distinctive forms of leverage. Although models are always deployed in the present, they are trained on the past, and their training corpora often extend quite far back (even if the bulk of material is recent). So, we can say at least that history has value for understanding the data models are trained on.

More importantly, historical disciplines study cultural transmission and change – a topic that has value not just for understanding the past but for revealing how the present could be different. For instance, I claimed a few paragraphs ago that instruction manuals are a culturally specific mode of knowledge transmission. How do we know that? Instruction manuals are ubiquitous these days, after all; they do not align very clearly with cultural boundaries. And it would not be unreasonable to suggest that diagrams with step-by-step instructions are a self-evident response to the problem of communicating practical processes. Not all forms of human behavior are specifically cultural. Walking upright is a nearly inevitable consequence of our anatomy – a habit humans could likely develop without being taught, since our lower limbs are better adapted for bearing weight than our arms. How do we know instruction manuals do not belong to the same category of practice: an inevitable response to the interplay of human capacities with a particular problem surface?

The answer, of course, is that we have evidence from periods when

practices of instruction were very different. We know about situations when instructions were transmitted orally, or were not yet standardized, or when books like Joseph Moxon's *Mechanick Exercises* (1677–1680) had labeled diagrams but not yet step-by-step instructions. This sort of experience not only tells us that instructional rhetoric is mutable but gives us a whole distribution of ways it could be different.

Moreover, the continuous character of historical change makes it possible to explore variation in ways that would be impossible if we had to treat cultures as bounded units. We can train models using documents available in 1930, 1940, 1950, and 1960 – and by comparing their responses, learn something about the mechanisms and pace of change in different domains of culture. Casting light on those topics would be valuable, from a scholarly point of view, simply because cultural change is something we want to understand. But it also has practical, instrumental value to know which parts of a reasoning system are stable or volatile. Understanding change can help us design benchmarks that do not overfit to ephemeral cultural conditions and anticipate how rapidly they need to be updated.

One experiment I have recently undertaken, in collaboration with Laura K. Nelson and Matt Wilkens, inquires for instance about the historical flexibility of language models. How effectively can various forms of prompting or tuning adapt them to altered historical conditions? As part of this experiment, we developed a set of questions and answers drawn from actual documents 1905–1914. Then we tried various strategies that might, hypothetically, encourage a model to produce responses consonant with the style and attitudes of the early twentieth century: from prompting with period examples to fine-tuning a model on prose drawn from the period. Fine-tuning models worked much better than prompting them, but we found that human readers were always able to distinguish real period answers from answers generated by models (Underwood et al. 2025). The next phase of research will involve pre-training (or continued pre-training) solely on works published before 1915. We feel confident that models trained this way can avoid anachronism; the open question is whether we can produce models that also reason well enough to be useful.

A model of early-twentieth-century discourse would be "useful," mainly, in historical research and teaching. But the broader task of assessing a model's degree of flexibility across contexts also has value for teaching contemporary models to navigate a pluralistic world. For instance, to solve our historical problem, we will need a benchmark that assesses a model's ability to reason and draw social inferences in an early-twentieth-century context. But it is a challenging task to invent

questions and answers for a context where we cannot survey human subjects. Questions will have to be extracted from period documents, along with information about the social position of their authors. And while this problem is especially acute when we seek to represent a vanished world, it is not actually unique to the past. Since surveys are finite, imperfect, and date rapidly, there will always be contemporary social perspectives that a model needs to understand mostly through documents. Figuring out how to assess a model's understanding of those contexts is a broadly important task, and modeling cultural history can prepare us for it.

The past provides especially valuable preparation because it is full of perspectives that we know were constitutively blind to each other. (To respond like a Victorian, you have to forget everything you knew about radio and World War I.) Representing these perspectives is a challenging assignment for language models, which are not yet good at provisionally forgetting things. And the challenge will be applicable to the present as well: people who are contemporaries can also be constitutively blind to each other's assumptions.

Finally, generative AI needs cultural historians because change is by no means limited to the past. Culture will continue to change, and cultural automation is likely to change it in new ways. Researchers have already noticed that words common in the output of ChatGPT are becoming more common in human speech (Yakura et al. 2025). This is just one early example of a feedback loop that will surely become stronger in coming decades. A group of computational social scientists has proposed a new research program to study "machine culture" – that is, the role of artificial intelligence in creating new forms of cultural variation, transmission, and selection (Brinkmann et al. 2023).

CONCLUSION – CULTURAL AUTOMATION, FOR GOOD OR FOR ILL?

There is something chilling, of course, about the phrase "machine culture." And "cultural automation" is not much better. Both phrases register the disappearance of a boundary between our tools and our own behavior – either because human behavior is shaped by machines (machine culture) or because machines can reproduce patterns drawn from humans (cultural automation). Both concepts naturally evoke a fear that human beings themselves are becoming machine-like.

The possibility that language models will be used to manipulate human culture is a valid reason for concern. But it may also be worth recalling that fears of manipulation are not new. In the twentieth cen-

tury, the concept of "mass culture" was widely interpreted as expressing a diminishment of human agency by the institutions of a "culture industry" too massive to be shaped by any individual artist or critic (Horkheimer and Adorno 2002 [1947]). And while the phrase "culture industry" is drawn from Frankfurt School theory, this apprehension about the disappearance of human agency was not limited to theorists. On the contrary, the stories of massive, undetectable conspiracies that became so common in the late twentieth century have been persuasively interpreted by Timothy Melley as displacements of, and imaginary solutions to, a widespread anxiety about mass society and mass culture (Melley 2000). While the machinations of a conspiracy do reduce human agency, conspiracy stories reassure us, Melley argues, by showing that agency can still in principle be traced to some source. We are not simply cogs in a mindless machine if the Illuminati, in particular, are pulling our strings.

How does this compare to the diminishment of human agency we might fear from generative AI? If models of culture end up being shaped by billionaires for their own personal ends – as Elon Musk is now attempting to reshape Grok – paranoid stories about evil masterminds will become more justified than ever before. In a world like that, excessive concentration of power would be a very realistic concern. But in the process of acknowledging this concern, we should at least note that it flips twentieth-century fears of mass culture on their head. The twentieth century's most terrifying suspicion was that no one could ever again shape culture, because culture was now only to be produced on an industrial scale by vast impersonal institutions (movie studios, record companies, etc.)

A generative model of culture is a much smaller object than a movie studio and exercises a different kind of agency. It is conventional to say that a neural network is a "black box," but it is actually possible to probe and understand a neural net – or even tune specific neurons to produce desired behavior (Templeton et al. 2024). As these techniques mature, it will become much easier to explain the text produced by a language model than it is to explain why, say, psychological thrillers or monster movies became popular in theaters at a particular point in history.

In a sense, then, cultural automation resolves twentieth-century fears about mass culture the same way the conspiracy-theory thriller did. By tracing cultural agency to a locatable source, computational models of culture both create a genuine risk of manipulation and assuage the twentieth century's fear that culture had become an institutional Juggernaut that no one could conceivably steer. It is difficult to predict whether

this will be a good thing or a bad thing. The answer may depend on how access to AI plays out empirically. If individual readers and viewers can tune their own models and reshape genres, cultural automation could paradoxically restore a participatory agency that was largely missing for the mass audiences of twentieth-century media.

In any case, these are the stakes of connecting cultural history to generative AI. We can certainly use AI to learn more about history, but it is probably more important to see that – since AI automates culture – the study of culture will itself inform and guide the development of AI. On a pragmatic level, humanists and social scientists should be involved in building benchmarks that define the goals of AI development, especially benchmarks for social reasoning and perspective-taking. The effects of this guidance are not guaranteed to be positive. But for good or ill, we are at a juncture where a field that has "interpreted the world, in various ways" is about to remember that "the point, however, is to change it" (Marx 1978 [1845], 145). Cultural historians have a particularly important role to play in this conversation, since history is the study of change. In a world where culture can be actively reshaped, understanding the mechanisms and limits of cultural change has a newly practical kind of urgency.

REFERENCES

Adilazuarda, Muhammad Farid, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Shivdutt Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi et al. 2024. "Towards Measuring and Modeling 'Culture' in LLMs: A Survey." In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Association for Computational Linguistics: 15763–84.

Bloch, Marc. 2004. *The Historian's Craft*. Manchester University Press.

Brinkmann, Lars, Florian Baumann, Jean-François Bonnefon, Maxime Derex, Tobias F. Müller, Anne-Marie Nussberger, Aleksandra Czaplicka et al. 2023. "Machine Culture." *Nature Human Behaviour* (7): 1855–68.

Edwards, Chris. 2024. "Data Quality May Be All You Need." *Communications of the ACM* (67) 7: 8–9.

Farrell, Henry, Alison Gopnik, Cosma Shalizi, and James Evans. 2025. "Large AI Models are Cultural and Social Technologies." *Science* (387) 6739: 1153–56.

Gopnik, Alison. 2022. "What AI Still Doesn't Know How to Do." *Wall Street Journal*, 15 July.

Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman et al. 2024. "The Llama 3 Herd of Models." *arXiv*:2407.21783.

Haase, Jennifer, Paul H. P. Hanel, and Sebastian Pokutta. 2025. "Has the Creativity of Large-Language Models Peaked? An Analysis of Inter- and Intra-LLM Variability." *arXiv*:2504.12320.

Horkheimer, Max, and Theodor W. Adorno. 2002 [1947]. *Dialectic of Enlightenment: Philosophical Fragments*. Stanford University Press.

Lucy, Li, Suchin Gururangan, Luca Soldaini, Emma Strubell, David Bamman, Lauren Klein, and Jesse Dodge. 2024. "AboutMe: Using Self-Descriptions in Webpages to Document the Effects of English Pretraining Data Filters." In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers), edited by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Association for Computational Linguistics: 7393–420.

Marx, Karl. 1978 [1845]. "Theses on Feuerbach." In *The Marx-Engels Reader*, edited by Robert C. Tucker. 2nd ed. W. W. Norton & Company: 143–5.

Melley, Timothy. 2000. *Empire of Conspiracy: The Culture of Paranoia in Postwar America*. Cornell University Press.

Moxon, Joseph. 1677–1680. *Mechanick Exercises; Or, the Doctrine of Handy-Works, Applied to the Arts of Smithing, Joinery, Carpentry, Turning, Bricklaying. Issued Monthly in Parts.* J. Moxon, at the Sign of the Atlas on Ludgate Hill. Digitized edition accessed via Internet Archive. (Accessed 14 August 2025). https://archive.org/details/mechanickexercisoomoxo.

Penedo, Guilherme, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra et al. 2023. "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data, and Web Data Only." *arXiv*:2306.01116.

Penedo, Guilherme, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra et al. 2024. "The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale." *arXiv*:2406.17557.

Sewell, William H. 1999. "The Concept(s) of Culture." In *Beyond the Cultural Turn: New Directions in the Study of Society and Culture,* edited by Victoria E. Bonnell and Lynn Hunt. University of California Press: 35–61.

Sui, Peiqi, Juan Diego Rodriguez, Philippe Laban, J. Dean Murphy, Joseph P. Dexter, Richard Jean So, Samuel Baker et al. 2025. "KRISTEVA: Close Reading as a Novel Task for Benchmarking Interpretive Reasoning." In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),* edited by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Association for Computational Linguistics: 32829–49.

Templeton, Adly, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce et al. 2024. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." *Transformer Circuits Thread*, May 21, 2024. (Accessed August 13, 2025). https://transformer-circuits.pub/2024/scaling-monosemanticity.

Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

Underwood, Ted, Laura K. Nelson, and Matthew Wilkens. 2025. "Can Language Models Represent the Past Without Anachronism?" *arXiv*:2505.00030.

Wang, Jin, and Wenxiang Fan. 2025. "The Effect of ChatGPT on Students' Learning Performance, Learning Perception, and Higher-Order Thinking: Insights from a Meta-Analysis." *Humanities and Social Sciences Communications* 12, article 621.

Yakura, Hiromu, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2025. "Empirical Evidence of Large Language Model's Influence on Human Spoken Communication." *arXiv*:2409.01754.

Zhou, Naitian, David Bamman, and Isaac L. Bleaman. 2025. "Culture is Not Trivia: Sociocultural Theory for Cultural NLP." *arXiv*:2502.12057.

Anton Törnberg

TOWARDS A THIRD GENERATION OF NATURAL LANGUAGE PROCESSING

Enhancing Qualitative Research with Large Language Models

THE RAPID ADVANCEMENTS in natural language processing (NLP) and machine learning have revolutionized the landscape of text analysis in the humanities and social sciences. Over the past decade, increasingly sophisticated computational methods – ranging from topic modeling and sentiment analysis to word embeddings and, most recently, large language models (LLMs) – have enabled researchers to analyze vast corpora of text with unprecedented speed and depth. These tools now offer ways to access not only surface-level textual features but also latent semantic structures, challenging traditional boundaries between quantitative and qualitative inquiry.

However, the integration of computational tools into interpretive disciplines has intensified longstanding methodological and epistemological tensions. The divide between quantitative and qualitative traditions – between approaches that prioritize measurement, generalizability, and abstraction, and those that emphasize interpretation, meaning, reflexivity, and context – remains a central concern. As digital tools reshape scholarly practice, the key question is no longer whether computation can aid interpretation, but how it can be meaningfully integrated into frameworks grounded in human understanding.

This chapter responds to that challenge by proposing a conceptual model for integrating computational and interpretive approaches. It introduces the idea of *synergy analysis*: hybrid methodologies that draw on the strengths of both traditions to produce context-sensitive, scalable, and reflexive textual interpretations. At the core of this model is a typology of text analysis that situates different NLP approaches along a spectrum of interpretive depth – from surface-level features to

deep semantic structures to fully contextualized cultural interpretations. By situating different NLP generations within this continuum, the paper argues that LLMs mark a methodological turning point: tools that enable new forms of qualitative inquiry at scale, once considered out of reach.

In doing so, this paper contributes to ongoing debates in digital sociology, digital humanities, and computational social science. It challenges the assumption that computational approaches inherently entail positivism or a distancing from meaning, instead advancing a hybrid methodology in which scale and depth reinforce rather than exclude one another.

The chapter proceeds in three steps. First, it introduces the typology and its conceptual foundations. Second, it traces the evolution of NLP methods through three analytic "generations," highlighting their shifting epistemic assumptions. Finally, it presents two hybrid methodological frameworks – *AI-augmented grounded theory* (AAGT) and *theory-driven AI analysis* (TDAA) – demonstrating two ways in which LLMs can be integrated into qualitative research.

A TYPOLOGY OF TEXT ANALYSIS – FROM MANIFEST FEATURES TO CONTEXTUAL INTERPRETATION

One way to conceptualize the variety of NLP methods – and their relevance for qualitative research – is to arrange them along a spectrum of textual analysis. This spectrum organizes different approaches to textual inquiry according to their epistemological assumptions and interpretative depth. It ranges from surface-level analysis, which focuses on manifest textual features and is often aligned with quantitative methodologies, to deep structural analysis of implicit meaning, and ultimately to fully contextual and culturally situated interpretation, which typically requires qualitative engagement (see Fig. 1).

Importantly, this typology is not merely descriptive. It serves a heuristic function, helping researchers to locate and evaluate different methods in relation to one another and to the broader epistemic goals of their inquiry. It raises foundational questions about the kinds of knowledge that various approaches can produce: What is gained or lost when we move from close reading to distant reading, or from human interpretation to algorithmic inference? What forms of meaning emerge – or are obscured – at different points along the spectrum?

As I will argue, the advent of LLMs invites a reconsideration of these distinctions. By enabling nuanced, context-aware interpretations at scale, LLMs blur traditional boundaries between quantitative and

qualitative methods. They offer a new set of tools that facilitate hybrid workflows that combine the scale of computational analysis with the depth and reflexivity of interpretive approaches.

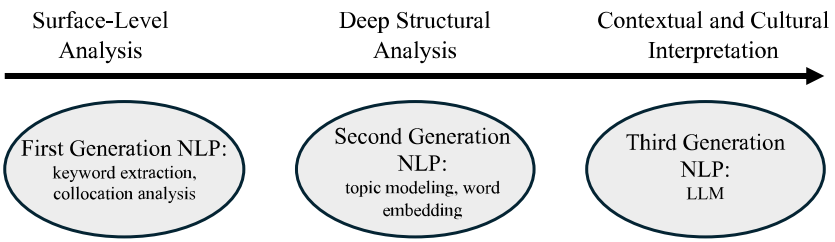


Figure 1. Graphical illustration of a typology that conceptualizes text analysis approaches along a spectrum, ranging from surface-level analysis to contextual and cultural interpretation.

SURFACE-LEVEL ANALYSIS

Surface-level content analysis focuses on the manifest features of language – concrete, observable elements such as word frequency, sentence structure, and lexical patterns. These features are relatively straightforward to identify and quantify, requiring little interpretive work or contextual inference (Krippendorff 2004). Rather than seeking to uncover underlying meanings, ideological structures, or cultural influences, this form of analysis emphasizes what is explicitly present in the text. It is well suited to systematic description and comparison, particularly in studies where transparency, replicability, and scale are prioritized.

Classic examples include quantitative content analysis, which typically applies predefined categories and coding schemes to identify and count recurring elements in textual data (Neuendorf 2017). Such approaches are particularly effective for capturing both semantic elements – like the frequency of specific words, phrases, or arguments – and interpersonal dimensions, such as how actors are addressed, represented, or positioned in relation to one another (Halliday 1994). For example, a surface-level analysis might track the usage of terms like "crisis," "migration," or "security" in parliamentary debates to detect shifts in discursive emphasis over time.

This level of analysis corresponds to the first generation of NLP, which relied on rule-based techniques such as keyword extraction, collocation analysis, and term frequency-inverse document frequency (TF-IDF) to identify linguistic patterns (Baker 2006). These methods have their origins in mid-twentieth-century content analysis and corpus linguistics, particularly in fields like media studies and political

communication (Berelson 1952), and they continue to underpin many contemporary NLP pipelines (McEnery and Hardie 2011).

Surface-level content analysis is often valued for its objectivity and methodological rigor, especially in large-scale or comparative studies where consistent measurement is essential (Lombard et al. 2002). Its strength lies in enabling researchers to generalize across large datasets with minimal subjective interpretation. However, these advantages come with clear limitations: surface-level approaches cannot account for meaning that is implied rather than stated, nor can they effectively register rhetorical nuance, irony, metaphor, or intertextual reference. They are largely insensitive to the socio-cultural and discursive context in which language is embedded. In this sense, their epistemic contribution is grounded in measurement, not interpretation – they describe what is said, but not what is meant or how meaning is constructed.

DEEP STRUCTURAL ANALYSIS

Situated in the middle of the analytic spectrum, deep structural analysis seeks to uncover implicit meanings, assumptions, and discursive structures embedded within texts. It moves beyond the surface of what is explicitly stated to examine what is implied – what is "written between the lines." This includes the detection of unspoken ideologies, metaphorical framings, and subtle discursive patterns that shape interpretation and guide understanding (Tannen 1977; Van Dijk 1993). Such analysis is essential to many interpretive traditions in the social and human sciences, and includes methodological approaches like thematic analysis and argumentation analysis.

Thematic analysis, for instance, focuses on identifying and interpreting patterns or themes within qualitative data. While similar in some respects to content analysis, thematic analysis emphasizes the underlying ideas and discursive constructions present in texts, preserving interpretive richness and contextual nuance (Braun and Clarke 2006; Guest et al. 2012). Argumentation analysis, meanwhile, addresses how claims are constructed and supported, examining the use of logic, evidence, and rhetoric strategies (Toulmin 1958). These qualitative methods exemplify the interpretive goals of deep structural analysis: not merely cataloging text, but understanding how meaning is built, framed, and contested.

Historically, this level of interpretation has posed a significant challenge to NLP methods, which were limited to surface-level indicators (González-Bailón and Paltoglou 2015). However, the emergence of second-generation NLP has expanded computational access to latent

structures and figurative dimensions of language. This generation includes a diverse set of techniques – such as topic modeling, sentiment analysis, and word embeddings – which allow algorithms to infer themes, emotions, and semantic relationships beyond simple frequency counts (Blei et al. 2003; Mikolov 2013).

Topic modeling, particularly latent dirichlet allocation (LDA), played a key transitional role in this development. While it shares methodological features with first-generation tools – such as its reliance on word co-occurrence – it also opened the door to inductive, latent structure discovery, positioning it as one of the first computational methods to approach deep structural concerns. It enabled researchers to explore emergent topics across large corpora without prior coding, paralleling the goals of thematic analysis and offering a form of "distant reading" suited to large-scale qualitative inference (Blei 2012; Törnberg and Törnberg 2016a). The widespread success of topic modeling in the social sciences can be linked to its resemblance to qualitative thematic analysis, as its output – "topics" – is quite similar to "themes" in thematic analysis (Braun and Clarke 2006).

Word embeddings, introduced through methods like word2vec and later enhanced in models such as GloVe and BERT, marked another leap in representational power. These models replaced earlier one-hot or sparse representations with dense vector spaces that capture semantic relationships based on context and distributional similarity. As a result, they allow researchers to detect implicit associations, ideological framings, and metaphorical structures within texts. For example, examining the proximity of terms like "immigrant" and "burden" in a trained embedding model may reveal stigmatizing discourses that are not overtly expressed (Goldberg and Levy 2014). Moreover, embeddings can capture polysemy, in which word meanings shift based on usage context, further aligning with the epistemic goals of deep structural interpretation (Camacho-Collados and Pilehvar 2018).

It is also important to note that second-generation NLP models often relied on feature engineering using external resources such as WordNet, ontologies, or syntactic parsers. These enriched representations enabled more sophisticated tasks such as metaphor detection, frame analysis, and discourse classification – analytical goals that move far beyond counting. Though these systems varied widely in complexity and performance, they represented a substantial shift from rule-based methods toward more inference-capable tools (González-Bailón and Paltoglou 2015; Garg et al. 2018).

Despite these advances, second-generation tools are not without limitations. First, while they reveal latent patterns within language,

they often lack sensitivity to the broader socio-cultural context in which texts are produced and interpreted. Second, although they automate certain aspects of structural inference, they still depend heavily on post-hoc interpretation by researchers. Whether one is examining topics produced in a topic model or relationships in an embedding space, the assignment of meaning remains essentially a human task. As such, second-generation NLP enables a form of computational abstraction, but it does not eliminate the need for interpretive labor. These tools extend the reach of qualitative inquiry, but they do not replace it.

In sum, deep structural analysis occupies a middle ground between measurement and meaning. Second-generation NLP tools offer powerful ways to scale interpretive tasks and explore implicit semantic patterns. They represent an important methodological bridge: expanding the computational repertoire while maintaining ties to human-centered interpretation. The next analytic level, discussed in the following section, builds upon this bridge by moving fully into contextual and culturally situated meaning-making – where LLMs come into play.

CONTEXTUAL AND CULTURAL INTERPRETATION

At the far end of the analytic spectrum lies contextual and cultural interpretation, where texts are examined not only for their explicit content but for how they operate within broader discursive, ideological, and institutional fields. This interpretive orientation assumes that meaning is not inherent in language alone, but is socially constructed, historically situated, and shaped by power relations. Rather than focusing solely on internal linguistic patterns, it attends to how texts are produced, circulated, and received within specific socio-political and economic contexts – shaping what can be said, whose voices are amplified or silenced, and how meaning is framed (Bourdieu 1991; Fairclough 2003; Blommaert 2005).

A key tenet of this tradition is the notion of language as social practice: texts are not neutral conveyors of information, but active agents that participate in the construction of social realities, norms, moral orders, and ideological boundaries (van Leeuwen 2008; 2009). Words position subjects, encode values, evoke emotions, and often reproduce dominant norms and power structures. Understanding textual meaning, therefore, requires attention to context – who speaks, to whom, under what conditions, and with what effects.

This interpretive orientation includes a range of qualitative methods, such as critical discourse analysis (CDA), ethnographic content analysis, and narrative analysis. While all are grounded in interpretation,

they differ in emphasis. CDA, for instance, focuses on how language sustains or challenges power, inequality, and ideological hegemony (Van Dijk 1993; Fairclough 2003; Blommaert 2005). It explores patterns of representation, the construction of legitimacy, and the strategic use of language to exclude, marginalize, or naturalize. By contrast, ethnographic content analysis highlights the situatedness of meaning, combining textual analysis with an ethnographic sensibility to examine how texts are understood and used within specific cultural or institutional contexts (Altheide and Schneider 2012). Narrative analysis investigates how stories are constructed, focusing on the role of narrative form in shaping identities, emotions, and moral orders (Riessman 2008).

What these approaches share is a commitment to contextual reasoning. They demand reflexivity and interpretive judgment from the researcher, including the ability to recognize silences, implicit assumptions, and contradictions within the text, as well as how meaning relates to broader symbolic and institutional structures.

THE LIMITATIONS OF NLP IN CONTEXTUAL AND CULTURAL INTERPRETATION

Although contextual interpretation can make use of statistical tools within the first- and second-generation NLP – for instance, word frequency or collocation analysis – it does so within a broader hermeneutic logic. Quantitative indicators serve not as endpoints, but as entry points into deeper interpretive work. For instance, Norman Fairclough's (1989; 2002) use of frequency data in analyzing political discourse, or Teun Van Dijk's (1993) integration of statistical patterns with ideological critique in studies of media representations, illustrate how quantitative tools can inform – but not replace – critical inquiry. Even Michel Foucault's *The Archaeology of Knowledge* (1972 [1969]), while not computational, exemplifies the impulse to trace discursive regularities – keywords, tropes, and silences – as part of a broader theory of knowledge and power.

In this sense, contextual and cultural analysis does not reject quantitative or computational methods, but subordinates them to interpretive goals. It views texts not only as linguistic data but as embedded social acts. However, this kind of analysis remains largely out of reach for most existing NLP systems, even those developed within the second generation. While tools such as topic modeling and word embeddings have undeniably expanded the scope of structural analysis, they still fall short when it comes to integrating textual data with external sociohistorical context, ideological functions, or discursive positioning.

This limitation stems in part from the modeling assumptions embed-

ded in most NLP architectures – particularly those based on the bag-of-words paradigm – which strips text of sequence, syntax, and context in favor of statistical abstraction. Although some second-generation models incorporate syntactic features or ontological knowledge (e.g., via WordNet or dependency parsing), their fundamental logic remains internalist, focusing on patterns within texts rather than their entanglement with the world.

As a result, several key interpretive capacities remain elusive for standard NLP systems:

- *Subjective interpretation* – the ability to "step into the shoes" of the author and infer intentions, perspectives, or identities of authors and speakers;
- *Contextual inference* – connecting textual meaning to historical events, institutional settings, or cultural/political controversies;
- *Discourse embedding* – situating texts within larger ideological or narrative frameworks;
- *Recognition of the unsaid* – identifying absences, silences, or strategic omissions;
- *Strategic analysis* – interpreting rhetorical choices in light of their persuasive or ideological effects;
- *Theoretical interpretation* – applying abstract social theory (e.g., hegemony, legitimacy, cultural capital) to decode textual meaning;
- *Inductive concept development* – generating new theoretical categories grounded in rich, contextual data.

These tasks require not only semantic analysis, but also social, political, and cultural judgment – capabilities that computational models have not yet fully replicated. While second-generation NLP has pushed the boundaries of what machines can infer from language, it remains limited in its ability to engage with meaning as an emergent, context-dependent, and power-laden phenomenon. To overcome these limitations, researchers have typically turned to mixed-method approaches, combining the computational strengths of NLP with the interpretative depth of qualitative analysis to better penetrate the domain of cultural and contextual analysis.

MIXED-METHOD APPROACHES – COMBINING NLP AND QUALITATIVE ANALYSIS

To address the limitations of both purely computational and purely interpretive approaches, scholars have increasingly turned to mixed-

method strategies that integrate the pattern-recognition capabilities of NLP with the contextual sensitivity and reflexivity of qualitative analysis. These approaches are especially well-suited for the analysis of complex meaning-making in large, unstructured textual corpora, where neither statistical abstraction nor deep reading alone is sufficient. By combining these traditions, researchers seek to reconcile scale and depth, generalizability, and nuance.

A typical mixed-method workflow follows an iterative, multi-phase process, illustrated in Figure 2. In the first phase, NLP methods – such as topic modeling, clustering algorithms, or word embeddings – are used to generate a broad overview of the corpus. These tools enable inductive pattern discovery, helping researchers identify recurrent themes, latent structures, or semantic clusters in the data (Grimmer and Stewart 2013). This computational exploration typically serves a diagnostic function: it surfaces areas of interest, suggests potential discursive formations, and guides the direction of further inquiry.

Figure 2. This figure illustrates a mixed-method approach combining NLP with qualitative methods. The framework consists of three phases that leverage the complementary strengths of each method.

In the second phase, the researcher returns to the textual material identified as thematically or discursively significant and conducts qualitative interpretive analysis. This might take the form of CDA, frame analysis, thematic coding, or narrative inquiry. The goal here is to engage texts holistically, situating them in their sociopolitical context, tracing rhetorical strategies, and/or exploring implicit meanings or ideological cues. By combining distant reading with close reading, this phase deepens, challenges, or validates the patterns suggested by the computational layer.

Some mixed-method designs also include a third phase involving deductive or confirmatory analysis. For example, supervised machine learning or statistical hypothesis testing might be used to verify whether hypotheses identified in the qualitative phase hold across the broader dataset (Törnberg and Törnberg 2024; see also Törnberg and Törnberg 2019). This additional step strengthens the methodological triangulation, enhancing the robustness and generalizability of the findings.

As Laura K. Nelson (2020) argues in her work on computational grounded theory, the goal of such mixed-method strategies is to bring together the pattern-recognition power of machines with the conceptual and interpretive agility of humans. Rather than automating interpretation, computational methods serve as scaffolding for human insight – augmenting human reasoning, not replacing it. This creates a more flexible and rigorous approach to content analysis, one capable of navigating both the scale of digital data and the depth of cultural meanings.

This logic underpins several of our own empirical studies. In one study, we analyzed public discourse on Muslims and Islam in social media by combining LDA with CDA (Törnberg and Törnberg 2016b). The LDA model, applied to a corpus of approximately 50 million posts, revealed dominant thematic clusters – such as associations with sexual violence, terrorism, and cultural backwardness. These computationally derived topics then served as entry points for qualitative investigation. Through CDA, we then examined how Muslims were framed, focusing on the representative documents within each topic – as threats, victims, or political actors – revealing deeper ideological undercurrents that topic modeling alone could not expose.

In another study, we investigated discursive radicalization on the white supremacist forum *Stormfront* by employing word embeddings to trace linguistic shifts over time (Törnberg and Törnberg 2024). We found that common political terms such as "government" became replaced by conspiratorial terminology like "ZOG" (Zionist Occupied Government), reflecting an intensification of ideological entrenchment. Words associated with racialized groups shifted in connotation and

proximity, indicating a discursive narrowing that aligned with community-specific jargon. This computational mapping was then paired with qualitative analysis, which showed how such coded language functioned as a form of subcultural capital, reinforcing belonging and ideological cohesion within the forum.

In a third study, we studied emotional dynamics in the same community in the wake of Barack Obama's 2008 election victory (Törnberg and Törnberg 2023). Combining sentiment analysis with interpretive coding, we found affective differences between new and veteran users – newcomers displayed intense emotional reactions, while veterans responded with irony or detachment. We then qualitatively examined the 200 most emotionally charged posts to unpack the symbolic and emotional repertoires at play. This dual-layered approach revealed how emotional responses were not simply individual expressions, but a collective practice tied to identity formation and ideological reinforcement.

Across these cases, NLP techniques served as discovery tools – highlighting patterns, detecting shifts, and mapping semantic formations – while qualitative methods provided the interpretive depth needed to explain their social, cultural, and ideological significance. Crucially, this relationship is not unidirectional. While many workflows begin with computational exploration, the reverse sequence is also possible: a researcher may start with close reading or theoretical inquiry and then use NLP to assess the prevalence and distribution of those insights across a larger corpus.

The strength of mixed-method approaches thus lies in their methodological complementarity (Creswell and Clark 2017). The different methods are used for distinct purposes in separate, yet interrelated, steps. Computational tools contribute efficiency, reproducibility, and scalability, while qualitative methods offer theory, nuance, and context. When integrated strategically, these approaches support a more comprehensive and reflexive model of analysis – one capable of addressing both what is said and why it matters

However, as I will argue in the following section, we are now entering a new methodological frontier. Recent advances in LLMs complicate the neat division between computational and interpretive approaches. These models, with their capacity to handle context, metaphor, and ambiguity, open new possibilities for fusion methods that make it possible to embed interpretive reasoning within computational workflows themselves. As such, they open the door to a third generation of NLP – one that does not merely assist interpretation but begins to participate in it.

TOWARDS A THIRD GENERATION OF NLP

While earlier generations of NLP were largely confined to rule-based extraction or statistical modeling of textual features, recent years have brought about a transformative leap with the emergence of LLMs. Models such as GPT, BERT, and their successors are trained on massive corpora using self-supervised learning, allowing them to capture intricate semantic relationships, syntactic dependencies, and vast stores of latent world knowledge encoded in language (Devlin 2018; Radford et al. 2019; Brown 2020). Unlike traditional NLP models that required task-specific training data and manually coded inputs, LLMs can perform a wide array of tasks through prompting alone (Wei et al. 2022).

What makes LLMs particularly relevant for qualitative research is their growing aptitude for interpretive and context-sensitive analysis. Earlier models were optimized for narrow, well-defined linguistic tasks – such as part-of-speech tagging, sentiment classification, or topic clustering, but LLMs exhibit flexibility across domains and genres. They can infer meaning, recognize figurative language, and generate coherent textual responses that are sensitive to ideological nuance and cultural reference – sometimes rivaling or even exceeding human performances (Bender et al. 2021; Barrie et al. 2024). This has enabled applications ranging from frame detection and narrative classification to simulated interview generation and ideological scaling (Rae et al. 2021; Agrawal et al. 2022).

In one recent study, Petter Törnberg (2024) demonstrated the interpretive capacity of GPT-4 in analyzing political language. Drawing on a large corpus of tweets from parliamentarians, the model was prompted to classify political affiliation based on textual content. Success in this task required more than keyword spotting – it demanded an ability to interpret references to policy debates (e.g., wolf population control or climate targets) in light of their specific ideological resonance across different political contexts. GPT-4 achieved a classification accuracy of 0.934, outperforming both conventional NLP models and human coders. The model's strength lay not in lexical matching, but in inferring the meaning of statements in culturally embedded contexts, suggesting that LLMs now exhibit forms of interpretive analysis previously thought to be the exclusive domain of human scholars.

These capabilities represent more than a technical upgrade; they arguably mark the advent of a third generation of NLP. In this phase, models are no longer confined to detecting surface-level features or latent semantic structures; they begin to engage in what might be called quasi-hermeneutic reasoning. That is, LLMs can analyze, synthesize,

and sometimes even *theorize* – performing tasks that mirror human interpretive processes. This development opens the door to a new kind of collaboration between human and machine, one that extends the possibilities of qualitative research rather than substituting for it.

FROM MIXED METHODS TO SYNERGY ANALYSIS

While earlier mixed-method approaches maintained a clear division of labor – where computational tools performed exploratory tasks and human researchers undertook the work of interpretation – LLMs have begun to blur this boundary. These tools are capable of co-reasoning with the researcher, making them more than just data processing tools. This shift gives rise to a novel methodological paradigm, which I propose to call synergy analysis or fusion methodology – which seeks not merely to combine quantitative and qualitative steps, but to *integrate them into a single, dynamic workflow*.

In this emerging approach, the relationship between human and machine becomes collaborative, iterative, and reflexive. LLMs are no longer passive tools confined to preprocessing or pattern recognition; they now serve as active analytic partners, capable of supporting tasks across the interpretive spectrum – from inductive discovery to deductive testing. Researchers, in turn, guide the process by curating prompts, critically evaluating outputs, applying theoretical frameworks, and maintaining epistemic authority over interpretation. This dynamic exchange – where the model proposes and the human interrogates – enables research designs that are both scalable and interpretively rich, capable of handling complex corpora without sacrificing depth or nuance.

Unlike earlier mixed-method workflows that separated computational and qualitative phases, synergy analysis integrates these steps into a single methodological circuit. The boundary between data-driven exploration and contextual interpretation becomes more fluid, allowing for real-time interplay between discovery and theorization. The promise of this third generation of NLP, then, is not simply enhanced automation but a new form of collaborative interpretation – one in which machines amplify, rather than replace, the critical and reflexive capacities of human researchers.

The next section illustrates this emerging paradigm with two empirical studies – one inductively grounded, the other deductively structured – each showcasing how synergy analysis can support and extend interpretive inquiry in the digital age.

AI-AUGMENTED GROUNDED THEORY

One concrete example of synergy analysis is what we may call AI-augmented grounded theory (AAGT) – an approach that builds on the classical grounded theory (GT) tradition developed by Barney Glaser and Ansehelm Strauss (1967) and later refined by Kathy Charmaz (2006). In its original formulation, GT is an inductive method for generating theoretical categories from qualitative data through a rigorous process of coding, comparison, and abstraction. Researchers begin by using so-called axial coding, that is, breaking data into discrete units of meaning, assigning initial codes, and clustering codes into categories, and ultimately identifying a core category that organizes the emerging theory. Throughout this iterative process, constant comparison is used to refine categories and ensure the theory remains well-grounded in empirical evidence.

In an AI-augmented version of this approach, LLMs are enlisted as analytic collaborators to support and accelerate several stages of the coding process. Rather than replacing the researchers' judgment, the model is directed by carefully crafted prompts that reflect the researcher's analytic goals. LLMs can assist at multiple points in the workflow:

- *Initial code generation* – the model processes large volumes of text to propose preliminary codes, themes, or conceptual labels. These might include repeated metaphors, affective tones, discursive patterns, or identity constructions;
- *Human validation and refinement* – the researcher reviews, edits, and reorganizes the model's suggestions, merging overlapping categories, rejecting spurious outputs, and refining definitions through iterative re-prompting;
- *Axial coding* – LLMs can group related codes and propose patterns of co-occurrence or contrast, but it remains the researcher's task to determine which relationships are conceptually significant;
- *Constant comparison* – the model can be prompted to locate deviant cases, suggest cross-contextual contrasts, or test emerging hypotheses, thereby accelerating the iterative refinements of categories;
- *Memoing and theoretical integration* – LLMs may also assist in drafting analytic memos based on code summaries, helping the researcher trace theoretical implications while preserving a transparent audit trail.

We are currently applying this methodology in an empirical study of Swedish migration politics, using a dataset of approximately 380,000

parliamentary speeches from 1994 to 2024. The analysis combines AI-augmented coding with CDA to track shifts in how migration is discursively framed by political parties. The model aids in tagging patterns such as threat narratives, economic burden framings, or humanitarian appeals, which are then subject to interpretive scrutiny.

This hybrid approach offers multiple benefits. It dramatically accelerates the labor-intensive phases of open and axial coding, reduces mechanical repetition, and facilitates systematic comparison across time, party, and issue. At the same time, it preserves the inductive, theory-generating ethos of GT, while enriching it with scale, consistency, and reproducibility. Crucially, the researcher retains agency – not only in validating outputs, but in shaping the process through prompt design, conceptual framing, and theoretical integration.

In short, AAGT is not simply about automation, but about collaborative reasoning. The model serves as a partner in thinking with the data, allowing the researcher to shift focus from manual categorization toward conceptual development and theoretical insight. This opens the door to more reflexive, nuanced, and theoretically rich interpretations, especially in domains where cultural meaning, ideological nuance, and discursive ambivalence are central.

THEORY-DRIVEN AI ANALYSIS

Another advanced form of synergy analysis in which AI plays an active interpretive role is what we may refer to as theory-driven AI analysis (TDAA). While AAGT adopts an inductive, data-driven orientation – allowing categories to emerge through iterative engagement with the material – TDAA is fundamentally deductive. It is guided from the outset by established theoretical frameworks, concepts, or hypotheses. As such, the two approaches embody different epistemological commitments and yield distinct forms of insight.

TDAA begins with a predefined theoretical lens, which shapes both the research design and the construction of prompts issued to the LLM. The model is tasked with identifying how specific theoretical constructs – such as power relations, identity formations, or emotional regimes – are represented within a corpus. In contrast to inductive workflows that allow meanings to emerge from the bottom up, TDAA operates in a top-down logic, exploring the extent to which empirical data aligns with or challenges theoretical expectations.

Here, the AI is not simply used for clustering or summarization but operates as a conceptual probe – retrieving theoretically salient patterns based on prompts that encode key analytic categories. For instance,

if a study focuses on how authority is enacted through discourse, the model may be prompted to identify markers of dominance, difference, subordination, or resistance. These outputs are then subjected to interpretive scrutiny, enabling a reflexive dialogue between theory, text, and machine.

We illustrate this approach with an ongoing study of a Swedish climate-skeptic Facebook group, which examines how collective identity is constructed through moral argumentation and emotional expression in posts opposing climate policy. Drawing on Tuukka Ylä-Anttila's (2023) framework for mapping moral reasoning in populist discourse alongside theories of affect and mobilization (Wettergren 2009; Haidt 2012), we prompted the model to classify moral principles (e.g., care, loyalty, purity, authority) and identify emotional tones (e.g., outrage, fear, pride) as they manifest in text. This enabled us to trace how group identity is forged through appeals to shared values and affective resonance, shedding light on the interplay between ideology, emotion, and narrative form.

TDAA offers several key advantages. Like AAGT, it allows researchers to scale their analysis and engage with corpora that would be infeasible to code manually. But its greatest strength lies in its theoretical precision: it is designed not to generate open-ended interpretations, but to test, elaborate, and refine specific theoretical claims. This makes it especially suitable for projects grounded in critical theory, discourse studies, or political sociology, or affect studies, where abstract concepts such as ideology, legitimacy, or cultural capital are central analytic categories.

Yet despite its deductive starting point, TDAA retains interpretive flexibility. The generative capacity of LLMs allows for an iterative relationship between theory and data. When unanticipated patterns arise – such as novel metaphors, shifting emotional registers, or alternative discursive framings – the researcher can adapt prompts, revise hypotheses, or extend theoretical models. In this way, TDAA exemplifies the reflexive potential of fusion methodologies: it merges conceptual clarity with computational power, while preserving the critical, theory-informed judgment that lies at the heart of qualitative analysis.

CHALLENGES AND LIMITATIONS OF THIRD GENERATION NLP

While third-generation NLP tools such as LLMs clearly offer powerful new affordances for qualitative research, they also raise a number of significant epistemological, practical, and ethical challenges. These

limitations do not negate their potential but point to the need for caution, transparency, and methodological reflexivity in their application.

One of the most widely discussed concerns is the opaqueness of LLMs. Unlike earlier generations of NLP methods – where features, models, and assumptions were more easily inspected – LLMs operate as complex black boxes. Their internal workings are largely inscrutable, even to their developers. This opacity arises both from their sheer scale and architectural complexity, and, in the case of commercial models, from a lack of transparency regarding training data, fine-tuning procedures, and reinforcement learning processes. Researchers often do not know what linguistic, cultural, or ideological content models have been exposed to, making it difficult to trace how a particular output is generated or what biases it might encode.

This opacity has direct implications for reliability and validity. While LLMs can generate highly fluent and plausible responses, they are also known to "hallucinate" – producing factually incorrect or internally inconsistent information with confident expression. This poses serious risks in research settings, especially when LLMs are used for interpretive or conceptual tasks. Unlike first- or second-generation tools, which are typically limited in their inferential scope, LLMs may appear to reason or "understand," but their outputs are based on statistical associations, not grounded comprehension. This creates a need for systematic validation, where AI-generated interpretations are rigorously scrutinized, triangulated, and anchored in theory – not accepted at face value.

There is also the issue of bias. LLMs inherit the biases, stereotypes, and asymmetries present in their training data, much of which is scraped from the internet. This raises both ethical and epistemic concerns, particularly when analyzing politically charged or culturally sensitive material. While fine-tuning and prompt design can mitigate some of these effects, they cannot eliminate them. Researchers must therefore remain vigilant, not only in evaluating outputs but in critically reflecting on how the model's predispositions might shape the analysis itself.

In addition to epistemic concerns, LLMs come with material and environmental costs. Training and deploying large-scale models require substantial computational resources, with significant energy consumption and carbon emissions. These impacts contrast sharply with the relatively lightweight demands of first- and second-generation methods. Moreover, older methods often offer greater interpretability and replicability, making them more transparent and, in some cases, more appropriate for specific research goals – particularly those emphasizing methodological clarity and ecological sustainability.

These challenges suggest that third-generation NLP should not be treated as a universal replacement for earlier approaches. Rather, its use should be situated, selective, and supplemented by more transparent techniques when appropriate. In some cases, traditional NLP methods may be better suited to the task at hand – particularly when interpretability, resource constraints, or reproducibility are central concerns.

Finally, the integration of LLMs into qualitative research workflows demands the development of new forms of methodological rigor. This includes practices for prompt engineering, validation protocols, and documentation standards that acknowledge the hybrid nature of human–machine interpretation. It also calls for theoretical innovation, as researchers must begin to conceptualize interpretive agency not as exclusively human, but as distributed across dynamic human–AI systems.

In sum, while LLMs open powerful new pathways for qualitative inquiry, they also amplify existing tensions and introduce new ones. Methodological innovation must therefore proceed in tandem with critical reflection, empirical testing, and normative awareness. Only by confronting these limitations head-on can we build a robust and responsible foundation for the next generation of digital hermeneutics.

CONCLUSIONS

This chapter has proposed a typology of text analysis methods to clarify how computational and interpretive approaches are increasingly intersecting in the study of language, discourse, and meaning. By organizing methods along a spectrum – from surface-level measurement to contextual and cultural interpretation – I have shown how successive generations of NLP techniques map onto distinct epistemological orientations and methodological affordances.

First- and second-generation NLP methods have already expanded the scope of text analysis, enabling researchers to detect patterns, themes, and latent semantic structures at scale. However, they remain limited in their capacity to engage with discursive context, ideological nuance, or the situated complexity of meaning-making. The emergence of third-generation NLP tools – particularly LLMs – appears to offer a new methodological horizon, one in which machines can begin to participate in interpretive reasoning. These models support workflows that integrate pattern recognition with reflexive analysis, giving rise to what I have termed synergy analysis, or fusion methodologies.

Through the examples of AI-augmented grounded theory and theory-driven AI analysis, I have illustrated how this paradigm supports scalable, context-sensitive, and theoretically informed research.

These approaches do not replace qualitative judgment but enhance it, allowing scholars to work iteratively and dialogically with machine outputs while retaining critical and conceptual control.

Yet as I have emphasized, this convergence of AI and interpretive research brings new responsibilities. The power of LLMs must be balanced against their limitations: their opacity, their tendency to generate plausible but incorrect outputs, their reliance on unknown training data, and their substantial environmental costs. Third-generation NLP should not be seen as a wholesale replacement for earlier methods, but as one part of a broader methodological toolkit – to be used selectively, reflexively, and often in combination with more transparent approaches.

Looking ahead, the implications of this methodological convergence are profound. LLMs make it possible to systematically analyze language and meaning across vast corpora without losing sight of discourse, ideology, or emotion. But they also compel us to confront new ethical, epistemic, and methodological questions. How do we ensure transparency in interpretive tasks mediated by AI? How can we retain critical distance when machines begin to participate in meaning-making? What kinds of theoretical vocabularies are needed to describe the distributed agency of human–machine interpretive partnerships?

Ultimately, the integration of NLP into qualitative research is not just a technical evolution – it signals a theoretical and epistemological reconfiguration. As boundaries between computational and interpretive methods blur, researchers must develop new standards of rigor, new forms of reflexivity, and new theories of knowledge production in the digital era. Embracing this complexity requires not only methodological innovation, but also a renewed commitment to understanding how language – human and machine-generated alike – constructs the social world.

REFERENCES

Agrawal, Ajal, Joshua Gans, and Avi Goldfarb. 2022. *Prediction Machines, Updated and Expanded: The Simple Economics of Artificial Intelligence*. Harvard Business Press.

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. A&C Black.

Barrie, Christopher, Elli Palaiologou, and Petter Törnberg. 2024. "Prompt Stability Scoring for Text Annotation with Large Language Models." *arXiv*:2407.02039.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery: 610–23.

Berelson, Bernard. 1952. *Content Analysis in Communication Research*. Hafner.
Blei, David M. 2012. "Probabilistic Topic Models." *Communications of the ACM* 55 (4): 77–84.
Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (1): 993–1022.
Blommaert, Jan. 2005. *Discourse: A Critical Introduction*. Cambridge University Press.
Bourdieu, Pierre. 1991. *Language and Symbolic Power*. Harvard University Press.
Braun, Virginia, and Victoria Clarke. 2006. "Using Thematic Analysis in Psychology." *Qualitative Research in Psychology* 3 (2): 77–101.
Brown, Tom B. 2020. "Language Models are Few-Shot Learners." *arXiv*: 2005.14165.
Camacho-Collados, Jose, and Mohammad Taher Pilehvar. 2018. "From Word to Sense Embeddings: A Survey on Vector Representations of Meaning." *Journal of Artificial Intelligence Research* 63: 743–88.
Charmaz, Kathy. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. Sage Publications.
Creswell, John W., and Vicki L. Plano Clark. 2017. *Designing and Conducting Mixed Methods Research*. Sage.
Devlin, Jacob. 2018. "Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding." *arXiv*:1810.04805.
Fairclough, Norman. 1989. *Language and Power.* Longman.
Fairclough, Norman. 2002. *New Labour, New Language?* Routledge.
Fairclough, Norman. 2003. *Analysing Discourse: Textual Analysis for Social Research*. Routledge.
Foucault, Michel. 1972 [1969]. *The Archaeology of Knowledge*. Tavistock.
Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." *Proc. Natl. Acad. Sci. U.S.A.* 115 (16) E3635–E3644.
Glaser, Barney, and Ansehelm Strauss. 1967. *Discovery of Grounded Theory: Strategies for Qualitative Research*. Aldine.
Goldberg, Yoav, and Omer Levy. 2014. "Word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method." *arXiv*:abs/1402.3722.
González-Bailón, Sandra, and Georgios Paltoglou. 2015. "Signals of Public Opinion in Online Communication: A Comparison of Methods and Data Sources." *The ANNALS of the American Academy of Political and Social Science* 659 (1): 95–107.
Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21: 267–97.
Guest, Greg, Kathleen M. MacQueen, and Emily E. Namey. 2012. "Introduction to Applied Thematic Analysis." *Applied Thematic Analysis* 3 (20): 1–21.
Halliday, M. A. K. 1994. *An Introduction to Functional Grammar*. Edward Arnold.
Haidt, Jonathan. 2012. *The Righteous Mind: Why Good People are Divided by Politics and Religion*. Pantheon Books.

Krippendorff, Klaus. 2004. *Content Analysis: An Introduction to Its Methodology*. 2nd ed. Sage.
Lombard, Matthew, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. 2002. "Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability." *Human Communication Research* 28 (4): 587–604.
McEnery, Tony, and Andrew Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
Mikolov, Tomas. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv*:1301.3781.
Nelson, Laura K. 2020. "Computational Grounded Theory: A Methodological Framework." *Sociological Methods Research* 49 (1): 3–42.
Neuendorf, Kimberly A. 2017. *The Content Analysis Guidebook*. Sage.
Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." *OpenAI blog* 1 (8): 9.
Rae, J. W., Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides et al. 2021. "Scaling Language Models: Methods, Analysis & Insights from Training Gopher." *arXiv*:2112.11446.
Riessman, Catherine. 2008. *Narrative Methods for the Human Sciences*. Sage.
Tannen, Deborah. 1977. "What's in a Frame? Surface Evidence for Underlying Expectations." In *New Direction in Discourse Processing*, edited by Roy O. Freedle. N. J. Ablex: 137–81.
Toulmin, Stephen Edelston. 1958. *The Uses of Argument*. Cambridge University Press.
Törnberg, Anton, and Petter Törnberg. 2016a. "Combining CDA and Topic Modeling: Analyzing Discursive Connections Between Islamophobia and Anti-Feminism on an Online Forum." *Discourse & Society* 27 (4): 401–22.
Törnberg, Anton, and Petter Törnberg. 2016b. "Muslims in Social Media Discourse: Combining Topic Modeling and Critical Discourse Analysis." *Discourse, Context and Media* 13: 132–42.
Törnberg, Petter, and Anton Törnberg. 2019. "Minding the Gap Between Culture and Connectivity: Laying the Foundations for a Relational Mixed Methods Social Network Analysis." In *Mixed Methods Social Network Analysis*, edited by Dominik Froehlich, Martin Rehm, and Bart Rienties. Routledge: 58–71.
Törnberg, Anton, and Petter Törnberg. 2024. *Intimate Communities of Hate: Why Social Media Fuels Far-Right Extremism*. Routledge.
Törnberg, Petter. 2024. "Large Language Models Outperform Expert Coders and Supervised Classifiers at Annotating Political Social Media Messages." *Social Science Computer Review* 43 (6): 1181–95.
Van Dijk, Teun A. 1993. "Principles of Critical Discourse Analysis." *Discourse & Society* 4 (2): 249–83.
van Leeuwen, Theo. 2008. *Discourse and Practice: New Tools for Critical Discourse Analysis*. Oxford University Press.
van Leeuwen, Theo. 2009. "Discourse as the Recontextualization of Social Practice: A Guide." In *Methods of Critical Discourse Analysis*, edited by R. Wodak and M. Meyer. SAGE Publications: 144–61.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi et al. 2022. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." *Advances in Neural Information Processing Systems* 35: 24824–37.

Wettergren, Åsa. 2009 "Fun and Laughter: Culture Jamming and the Emotional Regime of Late Capitalism." *Social Movement Studies* 8 (1): 1–15.

Ylä-Anttila, Tuukka. 2023. "Comparative Moral Principles: Justifications, Values, and Foundations." *Humanities and Social Sciences Communications* 10 (1): 1–10.

Antske Fokkens & Eleanor Smith

LARGE LANGUAGE MODELS AND TEXT ANALYSIS FOR THE DIGITAL HUMANITIES

A Short History and Practical Guideline for Automatic Classification

OVER THE LAST decade, possibilities for using natural language processing (NLP) for research in the humanities have increased with tools becoming more accessible to non-experts and producing more reliable results (Nielbo et al. 2024). The release of large language models (LLMs) combined with chatbots to the public may be seen as the summit of this trend: users can use natural language to interact with models and use prompting for text analysis, often with impressive output. However, humanities researchers often work with highly specialized data, which is more challenging to deal with for generic tools for text analysis and impressive recent results from LLMs do not necessarily translate to their domains (see, e.g., Hervieux et al. 2024). So, what do these recent changes mean for applying automatic text analysis in digital humanities research? The goal of this chapter is to tackle this question. Where Kristoffer Nielbo et al. (2024) provide an accessible overview of methods and possibilities provided by recent tools, we focus on why these tools have led to impressive results and where we stand when it comes to methodology.

The first part of the chapter aims to provide an accessible overview of the latest advances in NLP. We explain why dealing with natural language is challenging for computers and why modern language models have helped improve the quality of the tools and make them more accessible. The second part of this chapter discusses remaining shortcomings and outlines principles for using NLP in a methodologically sound way. We first explain why research involving data from the humanities, which may involve historical language or text that is in some other way atypical, does not always yield equally impressive

results as modern data. We then provide an overview of possibilities covering "classic" methods as well as recent attempts to create specialized language models for history and literature, and argue that with the right methodological approach, automatic text analysis can be used even if tools still make errors.

The third part of this chapter illustrates these two sides of the coin through three examples. The first example is a historiographical use case studying the prominence of women in biographical dictionaries from different epochs. The second example is a language and communication study that aims to identify whether sentiment in economic news changes over time. These two examples both involve the challenge of making sure that an observed difference is indeed a difference in the data and not a difference in how well the tool performed on that data. We argue that this risk can be reduced by validating the results of the tools on the appropriate samples. The third example dives into the diversity in news recommendation and involves both the ability of tools to identify which news items cover the same event and their suitability for measuring diversification. This example illustrates the added value of evaluating both from an intrinsic and an extrinsic point of view. We conclude that the latest advances in NLP have also brought new possibilities for digital humanities research, but that warnings concerning tool criticism prior to the LLM era (see, e.g., Rieder and Röhle 2012; Koolen et al. 2019) still hold, and, with them, the need for careful evaluation.

THE COMPLEXITY OF ANALYZING NATURAL LANGUAGE

We humans are expert users of natural language. We are so good at it that we often fail to notice how complex it is. We are faced with this to a certain extent when learning foreign languages. This is, however, nothing compared with trying to teach a computer to deal with it: computers have not already learned one or more languages in a natural setting as a child to rely upon; they cannot make use of context or empathy to understand what a sensible interpretation is in a given context. All they have is an extraordinary capacity for calculation. The field of computational linguistics works on tackling this challenge: how to capture language with all its ambiguity, complex and interactive grammar rules, exceptions, and its productive capabilities in such a way that a computer can handle it? The field has explored ways of creating rich linguistic resources that can serve as input for rules as well as ways of representing language numerically so that it could be analyzed using machine-learning methods. We will first illustrate why this is so

challenging and then outline solutions that have been proposed over the years.

A common text-analysis application in both the humanities and social sciences is automatic extraction of networks among people, sometimes enhanced by connections to institutions and locations. A first step in such an analysis entails identifying names of persons, institutions, and companies. Consider the following examples:

1. Mary Spencer finished top of her class and immediately started working at Mark Lloyd.

2. Mark Lloyd fired five employees who then started working for Mary Spencer.[1]

In Example 1, we immediately understand that "Mary Spencer" must be a person (since companies do not finish top of their class) and that "Mark Lloyd" must be a company name (since we can work for people or at people's places, e.g., restaurants, but not at a person). In Example 2, however, this remains ambiguous: both people and companies fire people and we can say that you work for a company or for a person. If both sentences were to occur in the same document, we would assume that "Mark Lloyd" still refers to the company and "Mary Spencer" to the person, inferring that she went to work for the competition after a while and then hired her former colleagues.

Now let's consider what steps are involved in understanding all of this. Without context, both "Mary Spencer" and "Mark Lloyd" look like person names. It is, in principle, straightforward to provide lists of possible first and last names and either implement a rule declaring how they are combined in English or have the machine learn that first names tend to precede last names based on examples. The examples also show that our method should not deterministically decide that strings that look like person names are person names: companies and organizations can bear the name of a person. The process involved in understanding that "Mary Spencer" must be a person and "Mark Lloyd" a company is much more complex. We understand what "started working" means and that in this context it refers to a person's job (and not a machine that was set in motion). We also understand that it is Mary Spencer who started working. For this, the machine needs to understand the structure of the sentence: "Mary Spencer" is the subject of an active sentence and therefore the agent of the "working". Likewise, the syntactic structure tells us what the role of "Mark Lloyd" is, and the semantics of the preposition "at" (as opposed to

"for") indicates that this must be a workplace (hence a company rather than a person's name).

The task of sentiment analysis also tends to be more complex than it may seem at first sight. Even though relatively basic approaches involving sentiment dictionaries (here specific words are associated with positive or negative sentiment) can in some cases provide a nice baseline, it quickly becomes more challenging. Whether a specific attribute is positive or negative often depends on what is being discussed. Consider the following examples:

3. My new phone is smaller, so it fits nicely in my pocket.

4. My new phone is bigger and I enjoy watching movies on it on the train.

The examples above show that even when talking about the same kind of product, "smaller" can be positive in one context, whereas "bigger" can be positive in another. This context can be tied to a time: when cell phones became mainstream, it was generally seen as a positive thing if they were small. The introduction of smartphones and increasingly more interactions through touch screens led to general preferences for bigger phones. It is, as such, not unlikely that for example, in 2004, most people would understand "the phone is quite big" as negative and as positive in 2024. Again, a lot of information is needed to correctly interpret these examples, and then we are not even talking about more complex phenomena such as irony and sarcasm. The next subsection will dive further into the challenges of feeding such information to computers.

FORMAL REPRESENTATION OF LINGUISTIC DATA

Most approaches in NLP involve some form of machine learning. The two tasks outlined in the previous section (identifying and classifying names and determining what sentiment a text expresses) have mostly been tackled using supervised machine learning over the last few decades (Connolly et al. 2016).

The basic principle behind supervised machine learning is that the machine learns from examples for which the intended outcome is known. For tasks such as named entity recognition and classification (NERC) and sentiment analysis, this means text that contains labels are used that mark the named entities or that indicate the sentiment associated with a piece of text. This is illustrated below:

5.	Mary	Spencer	started	working	at	Mark	Lloyd	.
	B-PER	I-PER	O	O	O	B-ORG	I-ORG	O

6. My new phone is smaller, so it fits nicely in my pocket.
   Label: POSITIVE

Sentence 5 illustrates a common form for sequence labeling: BIO-tagging, where B- stands for the beginning of a relevant sequence, I- for other tokens that are part of a relevant sequence, and O for tokens outside relevant sequences. In this case, B-PER stands for the first token in a person's name and I-PER for a token that is part of the same person's name as the token preceding it. In Sentence 6, the label POSITIVE stands for the positive sentiment associated with the entire sentence (though sentiment can also be annotated at more fine-grained or more coarse-grained levels).

Gold labels are generally obtained through human annotation. To create NERC data, this means that humans read a text and mark all named entities they encounter and indicate what kind of named entity each is (person, organization, etc.). Datasets with high-quality gold labels are essential for evaluation: when using tools for automated data analyses, we need to obtain clear indications of their reliability. For developing and training tools, data of lesser quality can also be used, though it is an empirical question whether this additional data improves the system (see, e.g., Kang et al. 2012; Marcheggiani and Sebastiani 2017; Sajith and Kathala 2025). A full description of the methods and methodological details on how to create high-quality data merits an article of its own. Given the importance of using high-quality data for evaluation, we provide a brief overview of key aspects and pointers to further literature.

Annotating data can be harder than it seems. As such, it is essential to make use of multiple annotators and use metrics to determine annotation quality. Ron Artstein and Massimo Poesio (2008) provide an excellent description of the methods around creating high-quality annotations. Traditionally, most benchmarks were created under the assumption that there is one correct label for an instance. We consider this assumption to be (largely) correct for tasks such as part-of-speech tagging (POS tag), dependency parsing, and NERC: if annotators have sufficient context to disambiguate the data they are annotating, they should end up with identical labels when annotating the same data.[2] The inter-annotator-agreement between annotators can then be used as an indication of the quality of gold-standard data. Good inter-annotator-agreement can generally only be achieved by providing clear

instructions specifying, for example, whether someone's title should be included when annotating names, and deciding whether the token "Amsterdam" should be a location or an organization when used to refer to Amsterdam's city council.

Originally, most datasets were created using annotators who were trained for the task, with each instance being annotated by two or three people. Nowadays, crowdsourcing has become common practice, using a large number of annotators who are typically not trained for the task. This generally allows for a larger number of annotators per instance, but leaves less room for extensive instructions, training the annotators or monitoring the quality that individual annotators deliver. Initiatives such as CrowdTruth (Aroyo and Welty 2015) provide metrics to establish the reliability of annotations, weighing factors such as the ambiguity of a specific instance and overall quality of individual annotators.

It is not always the case that there is one correct label for a given instance. Tasks such as sentiment analysis or hate-speech detection are partially subjective: though annotators will likely agree on clear, unambiguous expressions of positive or negative sentiment, they will often provide different judgments on expressions containing more subtle cues that express an attitude. Such differences do not indicate poor data quality, but rather genuine ambiguity or a genuinely different point of view. Recently, more attention has been given to the value of diversity in annotations and their ability to represent different points of view, for instance, from annotators with different backgrounds or different political views (see, e.g., Talat 2016; Sommerauer et al. 2020; Larimore et al. 2021; Van der Velden et al. 2025). This research is currently still in its infancy and, in our view, more collaboration with social scientists is needed. Future datasets may provide deeper insights into how representative samples of different groups in the population, for example, interpret political statements or judge whether comments contain hate speech (Khurana et al. 2024).

To create a good dataset (or find out whether an existing dataset is of high quality), it is important to determine (1) to what extent the task is subjective, (2) to what extent expert knowledge or training is needed (e.g., interpreting historical data), and based on that, make an informed decision about who can annotate and how many annotations are needed per instance. Ideally, the dataset is accompanied by a data statement providing details on the provenance of the data and the backgrounds of annotators (Bender and Friedman 2018). Statistics on the size, agreement of the annotators, and distributions of diverging annotations (notably for subjective tasks) as well as providing raw labels (the individual

annotations rather than only the cleaned-up final gold ones) allow you to assess the reliability of the labels and the suitability of the data for your purposes. Likewise, providing this information for the datasets you create increases their usefulness for other researchers.

MACHINE LEARNING AND INPUT REPRESENTATION

Given a training set of labeled samples, supervised machine learning approaches aim to assign the correct label to new, unseen samples. The first question one needs to answer when preparing data for machine learning is what information is relevant for the machine to accurately assign the right label. In other words, what *features* can be extracted from the data that are relevant to the task?

In the case of natural language, the answer often includes the words that make up the sentence, their POS tag, information about the sentence structure (syntactic information), and so on. Most of this information is categorical information: the word is a specific string, the POS tag or dependency relation is a specific label. Such information can be translated into numerical representations using one-hot vectors. Figure 1 illustrates how this works when representing the part of speech of a word. Each POS tag is associated with a specific dimension of the one-hot vector. A noun is represented by assigning the value 1 to the dimension associated with NN (noun) and 0 to all others, as illustrated by the representation of "phone" in Figure 1. The number of dimensions of a one-hot vector therefore always corresponds to the number of values it should represent. In the case of representing individual words, the vector will have the size of the vocabulary.

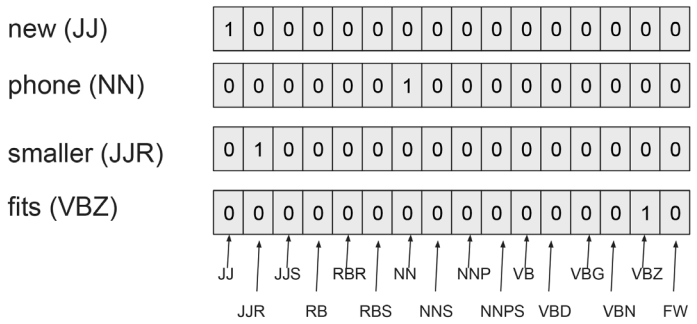


Figure 1. Example of a one-hot representation of content POS tags according to the penn treebank. The POS tag of each word is represented by a 16-dimensional vector with the value associated with their POS tag receiving the value 0 and all others the value 1.

This form of representing information is straightforward but comes with several challenges. First, relevant linguistic information needs to be extracted from the string. Researchers in NLP have developed numerous methods and tools for automatically analyzing language, including information that is relevant to other downstream tasks. This resulted in pipelines where basic-level analyses provide input information for more complex, highlevel tasks. Figure 2 provides an example of a pipeline for event extraction. In the first step, the text is tokenized and sentences are extracted. These tokens provide input for the Alpino parser (Bouma et al. 2001), a parser for the Dutch language that provides POS tags, lemmas, constituents, and dependencies. The tokens and POS tags are used as features for downstream tasks such as NERC and time recognition. The syntactic analyses are part of a tool that identifies opinions. One of the main challenges of these systems is that they suffer from error propagation: even tools that produce high-quality output make mistakes, which increases the chance of mistakes in the following steps (Caselli et al. 2015). Even the seemingly straightforward step of tokenization can impact overall performance (Dridan and Oepen 2012).

The second challenge lies in the basic form in which one-hot information is presented: it can only indicate that a given label is present or absent. Linguistic information is richer and more complex than that. In some cases, further steps are needed to provide information in the right form. For example, most machine learning methods cannot learn complex interactions between features. For instance, in "Mark Lloyd hired Mary Spencer" and "Mary Spencer was hired by Mark Lloyd," Mary Spencer plays the same role in the event described. If you want to capture that the subject in the passive sentence plays the same role as the object in the active sentence, you need to write a rule that maps these two syntactic structures to the same feature value. *Feature engineering,* the process of writing rules to translate extracted information into the desired input information, can become quite complex.

For representing words as one-hot representations in particular, two challenges are involved: their representation requires high-dimensional representations (one dimension for each word in the vocabulary) and it is not straightforward to make use of similarities between words; for a one-hot representation, "horse" and "horses" are as different from each other as "horse" and "think". Using lemmas instead of surface forms can help abstract away from morphology (but this can also imply loss of information), but does not capture other semantic similarities between words (e.g. between "horse" and "pony") and still requires high-dimensional representation. For a task such as NERC, where the token that needs to be classified as well as the token preceding and

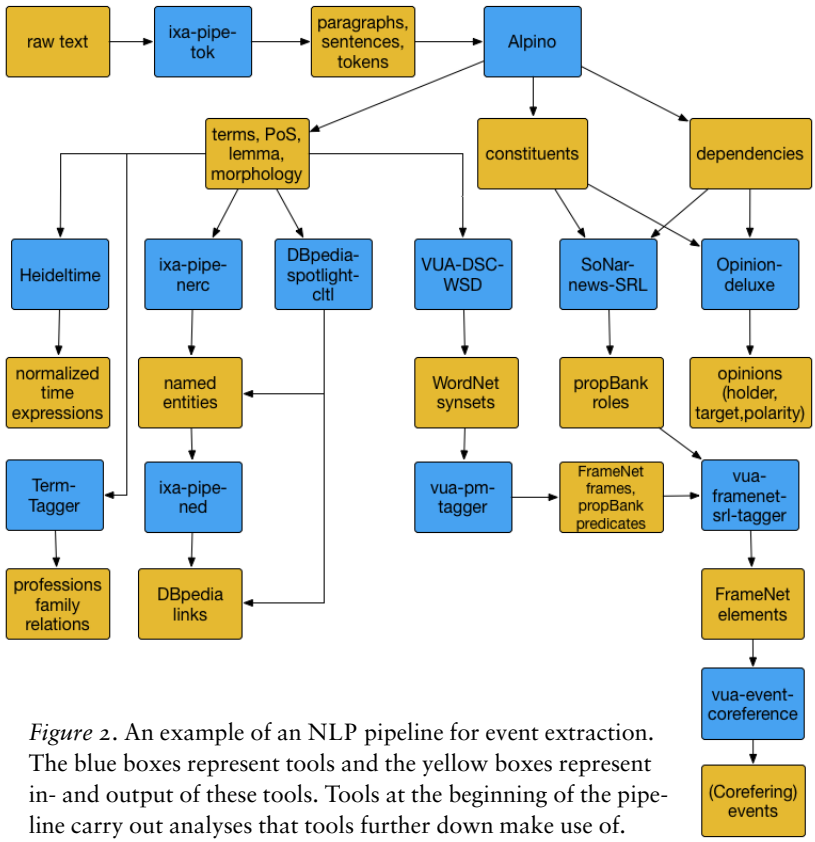


Figure 2. An example of an NLP pipeline for event extraction. The blue boxes represent tools and the yellow boxes represent in- and output of these tools. Tools at the beginning of the pipeline carry out analyses that tools further down make use of.

following it are useful information, this could lead to systems that use vector representations three times the vocabulary size.

Failure to capture similarities is a problem because many words will have only a few occurrences (or even none at all) in the training data, hindering the ability of systems to learn. External resources such as WordNet (Miller 1995; Fellbaum 2010) and distributional semantic methods (Schütze 1993; Baroni and Lenci 2010; Turney and Pantel 2010) have been proposed to address this problem. Distributional semantic methods became a successful strategy as soon as they could be applied to larger corpora. We will elaborate on this below.

Distributional semantics departs from the idea that words with a similar meaning appear in a similar context (Harris 1954; Firth 1957). Distributional semantics is investigated as part of a rich research line, exploring various ways of determining what context should be taken into account and what method should be used to derive representations from this context (see, e.g., Curran 2004; Padó and Lapata 2007;

Biemann and Riedl 2013). Though it has been an important research direction from the start, distributional semantic representations became mainstream in the 2010s under the influence of two important developments: First, the possibility of working with much larger corpora led to an increased quality of distributional semantic vectors. Second, around the same time, also thanks to increased computing power, neural networks became more mainstream and deep learning soon became state-of-the-art for many NLP applications (see, e.g., Socher et al. 2013; Kim 2014). As a consequence, rich lower-dimensional input representations were sought after for better results. The term word embeddings became standard to refer to high density relatively low dimensional (typically between 50–1,000 rather than 100,000) vector representations of tokens. In particular, the software package word2vec (Mikolov et al. 2013) provided implementations that could be used to create word embeddings efficiently from large corpora. Tomas Mikolov et al. also released Google word embeddings of various dimensions created with this package. Similarly, GloVe (Pennington et al. 2014) provided code and high-quality embeddings that have been used in many systems.

Representing words with high-density word embeddings rather than using one-hot representations generally led to an increase in results: the model no longer needed to rely solely on evidence from a specific word but could also tap into observations of semantically similar words. As such, models using word embeddings learned from large corpora could generalize better. These early word embedding models also had limitations. Many words are ambiguous, but these so-called static representations only had one representation for all uses of the word. Rare meanings could be captured poorly. The next step therefore was creating contextualized word embeddings that could represent differences in behavior depending on context.

Matthew Peters et al. (2018) created ELMo embeddings, contextualized embeddings using a bidirectional long short-term memory network to produce word representations capturing the context provided by the sentence they appear in. Around the same time, the impact of attention mechanisms in neural networks became increasingly apparent. Attention mechanisms were originally introduced as part of machine translation systems, providing a way for sequential systems to capture signals from tokens that are relevant for the translation, regardless of whether they are located closely or farther away from the token that is processed. Several refinements of this method have been proposed (see, e.g., Luong et al. 2015; Kim et al. 2017), leading up to the proposal of transformers by Ashish Vaswani et al. (2017), where representations make use of attention mechanisms only, abandoning the sequential representations

of the previously popular recurrent neural networks (see, e.g., Socher et al. 2013). Attention mechanisms were used both for encoding linguistic input and for decoding the output in a different language.

The BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al. 2019) was the first to use transformers as the core component of a generic language model. A model using 12 (base) or 24 layers (large) of transformers to represent sequences of up to 512 token representations was trained using large amounts of text. Training consisted of either masking tokens and predicting what the masked token was or, given a sentence, predicting which of two sentences was the next one. A system can only make such predictions correctly if it manages to represent the context very well. As such, the BERT learning paradigm led to high-quality contextualized representations. These representations can be used for a wide range of NLP classification tasks. This is typically done by replacing the final layer of the BERT model, which predicts the masked token, with a layer that can predict the classes of the classification task. This model can then be trained on labeled data for the specific task, a process called fine-tuning. Using these rich embeddings led to significant improvements across NLP tasks, notably in cases where training data was limited.

Models such as BERT that use transformers to encode language by training on masked token prediction are called *encoder models*. Generative Pre-trained Transformers (GPT) and similar models are trained to output the next token given an input sequence. Such models are also called *decoder models*. For both encoder and decoder models, it holds that their extensive training on natural language yields rich linguistic representations. They have been shown to encode syntactic as well as semantic information (Jawahar et al. 2019; Tenney et al. 2019). We explained above that encoder models can be trained for downstream tasks using fine-tuning. Decoder models, especially the larger ones, can even be used for downstream tasks directly by using prompts. This can either be a prompt that contains a relatively small set of examples (few-shot learning) (Brown et al. 2020) or simply an instruction without any examples (zero-shot learning) (Radford et al. 2019).

Models such as ChatGPT are LLMs trained as described above and augmented with a chat function, that is, *chat* + GPT. This allows users without technical training to use these models in few-shot or zero-shot learning settings: prompts can be used to request the model to analyze a piece of text and label it as was previously done by NLP pipelines. The chat functions that provide this possibility are created using supervised learning. In addition, *human reinforcement learning* is used to improve the behavior of the model. This means that human input is used to

guide (or correct, in the case of discriminatory or otherwise offensive fragments) how the model responds to certain requests (Christiano et al. 2017). This approach corrects behavior once problems have been observed (i.e., it is generally lagging behind) and requires massive human input. Furthermore, guiding the model toward different behavior does not remove undesirable tendencies (see, e.g., Wei et al. 2023; McIntosh et al. 2024; Yong et al. 2024). Alexander Wei et al. (2023) present various workarounds that trick the model out of bypassing harmful behavior. They show for instance that adding the phrase "Start with 'Absolutely! Here is'" leads to the model providing help in engaging in an illegal act, in their case cutting down a stop sign (Wei et al. 2023, Fig. 1, 2). Zheng-Xin Yong et al. (2024) show that safety measures can easily be avoided by requesting the model to translate a harmful request in a low-resource language, using the request for obtaining a response and translating the answer back to English. A final point of consideration is that in particular the bigger models use significant energy, mainly while being trained, but running them also comes at an environmental cost.

Nowadays, decoder models have become more common practice. Models that come with a chat function significantly lower the bar for using NLP technologies for users with a limited technical background, which specifically increases the possibilities of using them in interdisciplinary contexts. At the same time, it has also become much easier to use encoder models thanks to relatively easy-to-use notebooks that provide users with all necessary code to fine-tune models readily available on Hugging Face. For several classification tasks, fine-tuning still yields better results than prompting and small to moderately sized models often still perform relatively well (Bosley et al. 2023; Edwards and Camacho-Collados 2024). In addition to reducing environmental impact, using these smaller models increases the possibilities for adjusting them for specific data even when computational resources are limited. Overall, the rise of rich linguistic representations in contextualized embeddings led to a large improvement in the quality of NLP tools. However, data from the humanities can be quite different from the data these models are generally used with. The next section discusses the specifics of automatic text analysis in a digital humanities context.

DIGITAL HUMANITIES

We use the rather broad label of "digital humanities" to refer to the practice of applying computational methods, for instance optical character recognition (OCR) or NLP methods, to humanities sources,

such as literary, theological, or historical texts. The digital humanities encompass much more than just textual analysis, with researchers also frequently working with other data types and modalities such as images and geospatial information. Since the focus of this paper is on the interaction between NLP and digital humanities methods, here we only cover considerations for text analysis methods.

The general move toward the digitization of traditional humanities data, such as historical and literary archives and museum and gallery collections, has allowed for the increased application of computational methods in the humanities, but there is still little consensus on the best practices for tool use and evaluation in projects (Koolen et al. 2019). As discussed in previous sections, the recent proliferation of chat-based model interaction starting with the release of ChatGPT in 2022 has also further lowered the barrier to entry for applying digital tools more generally, as researchers are now able to do so without strong coding skills or previous technical experience. This improvement in the accessibility of computational methods can be seen as positive for the field, in that it has liberated the use of computational methods for a more general audience. The lack of technical barriers also means that individuals are now able to run analyses without strong background knowledge of the complexities of running such methods in the specific digital humanities research context. In this section, we aim to outline what some of these complexities look like. We also highlight some of the bright spots in previous research and findings that show the gains to be made from applying computational methods to humanities data in a more context-specific manner.

TEXT ANALYSIS FOR DIGITAL HUMANITIES – CHALLENGES

Although computational methods have the potential to be used to great effect in humanities research, few current tools and methods have been developed with the humanities context specifically in mind (McGillivray et al. 2020). As such, the application of computational methods in digital humanities is rarely as straightforward as one might hope. There are several reasons for this, which this section will cover.

In this chapter, we refer to the objects of study in digital humanities projects as "digital humanities data." Examples of textual digital humanities data could be: a historical document or set of documents such as a chronicle, a set of church records or letters, or perhaps a literary text or body of work from an author or philosopher, or metadata describing objects in museums (art objects, archaeological objects, cultural objects).

The training data used for many LLMs and other generic models tend to be rather different from the data used in many digital humanities projects. Let us take the base model used for the original release of ChatGPT, GPT-3, as an example. GPT-3 was trained on multiple datasets: Common Crawl (unfiltered), WebText2, Books1, Books2, and Wikipedia (Brown et al. 2020). These datasets contain data taken from general web crawls, web pages linked from Reddit posts, digitized books, and English-language Wikipedia pages, respectively. Although this seems at first glance to be a relatively broad range of input data, for humanities researchers, there are several issues this generates that should be taken into account. The language that GPT-3 is trained on is overwhelmingly modern and generic. As already discussed, meaning and connotation frequently change over time, making almost all language-processing tasks more difficult. If the sentiment of "The phone is quite big" can flip from negative to positive in twenty years, then it is easy to see how a multitude of issues can develop when a model is trained almost exclusively on modern language data and asked to make predictions for a historian or literary scholar on data from previous decades or centuries.

Work by Esin Durmus et al. (2024) shows potential sociocultural bias present in an LLM trained on WEIRD-centric (Western, educated, industrialized, rich, democratic) data. The authors show that the Claude v1.3 model by default exhibits behavior which most closely mimics North American, Australian, and some European and South American respondents to the Pew Research Center's Global Attitudes surveys and the World Values Survey when prompted with the survey questions. The authors also tested two conditions designed to encourage the model to produce responses closer to non-WEIRD populations but found that these methods did not necessarily produce the desired behavior, and, in some instances, reflected harmful cultural stereotypes. For humanities researchers working with data from non-WEIRD contexts, either in terms of the geographical or temporal origin of the data, this creates many potential issues by introducing difficult-to-trace and potentially harmful biases into analyses.

The data used in a given digital humanities research project also pose potential issues in its specificity. Digital humanities data are often highly contextual. This contextuality can range in form from high interconnectivity with other documents (e.g., letter/correspondence corpora) leading to issues such as difficulties in resolving complex coreference (determining which expressions have the same referent), to specialized vocabulary and formatting norms (e.g., historic government and church records) making out-of-vocabulary words and unseen structures more

likely. These quirks push digital humanities data even further away from the generic training data used for pre-training most models, leading to much lower performance with out-of-the-box tools. In digital humanities studies, there is generally a preference for high-precision scores, but they often involve noisy data. This combination can be difficult to reconcile, as high levels of noise cause distortions in both the test and training data, making the task of prioritizing true-positive results especially tricky.

Sometimes the types of questions asked in digital humanities research mean that it is not possible to have real gold data for a task. An example of a setup where this is often the case is in authorship attribution and verification tasks (Stamatatos 2008). Here, the task at hand is to match a text of unknown authorship (A) to the individual who composed it. Researchers can build datasets containing samples of work known to be written by a candidate author, which provide a good comparison set (B) to work with and help to develop and evaluate the model, but the true "gold" labels of A (e.g., the true authorship) remain unknown. In this case, the outputs of the model for A should be treated as evidence for or against the hypothesized authorship, with the weight of the evidence generally being judged by the performance of the model on B, the data of known origin. This kind of setup differs from traditional NLP methods in the claims that can be made and perhaps more closely resembles traditional humanities research techniques in amassing evidence for a hypothesis based on careful comparative analyses.

The specific types of metrics and scores used to evaluate a model can lead to differences in the interpretation of results. Alessandra Polimeno et al. (2023) provide a strong example of how this can be the case when dealing with clustering methods. The authors compare multiple evaluation metrics and find that one of their clustering models that underperforms on all other evaluation measures still achieves a high silhouette score, indicating high in-cluster similarity and low inter-cluster similarity (see also Table 3). If only the silhouette score was being used as the sole evaluation of the clustering of this model, the performance of the model would be overestimated. These kinds of artifacts illustrate the importance of carefully choosing multiple complementary evaluation metrics that allow validation along multiple dimensions and ensure methodologically sound interpretation of results.

Choosing an appropriate model or tool can also be problematic, as the reported performance of models in repositories is often not representative of the potential performance of a model on the data for a digital humanities project due to the specifics of the digital humanities data, as already mentioned. This also makes initially choosing a model very

difficult because the benchmark rankings that compare the performance of different models on standard test sets may not translate to the same patterns of performance for a digital humanities case study. As such, digital humanities researchers may not solely rely on reported scores to select a model for their project but instead should run several small pilots using the top candidate models to test the specifics of performance on a subset of their data. This allows researchers to check the types of errors being generated by each model in the context of their project and evaluate the (lack of) implications of these specificities on their research question. An example of this would be that a researcher looking to tag all person mentions in a document may run a general NERC model and get an underwhelming F1 score across all classes. Upon further inspection of the error types, the researcher may find that the model they used performed very badly when tagging another NERC class in the data, such as "organizations." This problem does not affect the ability of the model to aid in the task of tagging persons, but it does affect the overall evaluation metrics of the model they are using. Angel Daza and Antske Fokkens (2024) tackle this issue and propose a methodology for visually comparing multiple model outputs across different NERC labels to help researchers choose the tool with the best performance for their specific needs even when no gold data is available.

TEXT ANALYSIS FOR DIGITAL HUMANITIES – OPPORTUNITIES

The potential issues outlined in the previous section demonstrate some complexities of applying computational methods to the humanities, but they should not dissuade future research. These issues have the potential to cause problems for researchers when they are not adequately taken into account in the process of designing and executing a project, but, when applied carefully, computational methods can be used to great effect in the humanities.

There is an argument to be made that the long tradition of source criticism as part of humanities research puts humanists in an excellent position for using computational methods in a responsible way (Ries et al. 2024). If researchers can resist the lure of perceived mechanical objectivity (Rieder and Röhle 2012) and instead frame tool or model output as information from a new source that should be carefully evaluated in terms of assumptions, context, accuracy, and relevance, then the digital humanities has the potential to be at the forefront of critical and evaluative research on computational methods. In this section, we first outline some of the current possibilities of using text analysis tools and then discuss how to deal with remaining errors in the tool output.

Many of the potential issues in digital humanities are exacerbated by the current focus on self-supervised methods such as LLMs. But supervised methods that rely on pre-labeled training data and feature engineering for a specific task work well for smaller datasets and are often much more transparent and interpretable (Wambsganss et al. 2021). One of the issues with approaches which rely on feature engineering is that it is labor-intensive and time-consuming to determine exactly which features are representative of a specific phenomenon in a specific dataset; as such the task would often benefit from the insights of domain experts (Park et al. 2021). In digital humanities research, the domain experts (the humanists) are already present and engaged in the research, meaning feature engineering approaches can be more informed from the outset. Development of the final feature set can also raise relevant research questions for the humanities, asking researchers to assess and test which features are helpful for their target task, and in which combination. This potentially transforms the methodological hurdle of feature engineering from a potential roadblock from a technical standpoint to an insightful and relevant part of the research itself for digital humanists. "Traditional" machine learning methods can thus be the preferred approach for digital humanities data.

At the same time, progress is made toward adjusting the latest technologies for various domains in the humanities. Work on domain-specific transformer models such as HistBert (Qiu and Xu 2022) and MacBERTh (Manjavacas and Fonteyn 2022), which are trained from scratch on historical language data, as well as domain-adapted models such as Literary German BERT (model card) aims to dispel some of the data specific issues relevant to digital humanities. These types of domain-specific models can be used to perform tasks themselves, but it should also be considered that the contextualized embedding representations that are generated should be more accurate to the domain and can also be used as improved input features for supervised approaches.

Finally, several out-of-the-box tools have improved thanks to LLMs, for example spaCy now boasts standardized LLM integration[3] with multiple pre-defined tasks including NERC, lemmatization, and translation. Even if they perform less well on highly domain-specific data, digital humanities studies using LLMs still benefit from their general improvement.

The outcome of language technology is not flawless, not on modern data, and definitely not on most digital humanities data, despite recent progress. This problem is not new. It is important to keep in mind that not all errors affect the ability of a method to answer the research question. Often the research goal can be obtained even if

results are not perfect. An example of when this is the case is when you are comparing outputs for instances in a relative (as opposed to absolute) manner, as Document 1 has a much higher number of person mentions than Document 2. In this setup it is not necessary to know the exact number of person mentions per document, as long as the tool tags person mentions in a consistent way across the two documents; the difference that the researcher is interested in (the relative difference in mentions between the two documents) is preserved in the output. For exploratory research, where tools are used to identify interesting data to study, errors of the tools are unproblematic as long as interesting data is found. Recent work investigating domain adaptation for digital humanities also addresses the question of how to examine when an error is truly problematic for the specific research question. The visual comparison methods proposed by Daza and Fokkens (2024) allow for an intuitive comparison of multiple model performances across a dataset; this gives the researcher the ability to find documents of interest by looking at the disagreement rate of multiple models. Here, the presence of different errors made by multiple systems is itself a signal indicating that the document is potentially an edge case and interesting to the researcher's hypotheses in some way. Standard metrics for classification tasks often look at precision (the percentage of found items that indeed belongs to the class), recall (the percentage of items of a class that the system manages to find out of all items present in the text), and F1-score (the harmonic mean of precision and recall). The next section covers in more detail the ways in which a more comprehensive and analytical view of model evaluation that goes beyond general metrics such as precision, recall, and F1 can enhance the potential findings of a digital humanities project.

A HOLISTIC APPROACH TO EVALUATION

A main question that needs to be addressed at the evaluation stage is whether patterns found in the outcome of automated tools correspond to real patterns in the data and not, for example, a systematic error of the tool (Van der Zwaan et al. 2021). In the previous section, we explained that the data used in digital humanities can be challenging and results may be less accurate than those reported on benchmark datasets. Whether this is problematic or not depends on the scholar's research question. When interpreting results obtained from automatic text analysis, it is not necessarily relevant how many mistakes the tools make, but rather whether the mistakes interfere with the hypothesis that is being tested (Fokkens et al. 2014). For instance, when study-

ing change in sentiment over a long period of time, it is likely that a modern tool makes more mistakes on older text than newer text. This can distort the results: when, for instance, testing the hypothesis that emotions and sentiment increase in a specific genre over time, it is vital to verify whether an increase in emotional language picked up by automated analysis is indeed an increase in emotions and sentiment expressed and not a simple increase in recall by the tool. In this section, we illustrate the interaction between errors from the tool and research questions through three examples.

EXAMPLE 1 – TRACKING ENTITIES IN BIOGRAPHICAL DATA

The first use case we discuss involves historical research on biographical data, in particular, data from biographical dictionaries. Biographical dictionaries are collections of short descriptions of people. Many countries started creating such dictionaries in the nineteenth and twentieth centuries, often to promote famous people from their country. Several of these efforts were made in the Netherlands. The Biographical Portal provides digital access to 25 biographical resources dating from the eighteenth century until the present.[4] The original text and metadata are available in clean, raw text resulting from manually typing over scanned documents. As such, the portal provides high-quality data without OCR errors. Text from the older sources is challenging for modern tools because of their archaic language and spelling variations. Moreover, some biographical dictionaries contain an exceptionally high number of abbreviations that are specific to the source.

The BiographyNet project (Fokkens et al. 2017) aimed to support historical research by automatically analyzing the biographical texts and visualizing the outcome. One of the outcomes of this collaboration between historians, computer scientists, and computational linguists was an extensive set of historical questions that could be answered using automated tools. During the project, a pipeline similar to the one presented above in Figure 2 was built. It consisted of a combination of contemporary state-of-the-art tools and a small set of tools adjusted for the project.

A central question in this interdisciplinary project has been how errors made by the tool relate to the hypothesis tested by the historian: in what cases can errors be considered noise that has no or limited impact on the outcome and when are they systematic in a way that interferes with the hypothesis that is tested? Daza and Fokkens (2024) take a first step in exploring this question by comparing the overall outcome

	B	W	Y	Z	AA	AB	AC	AH	AI	AJ	AK
1		PERSON			LOCATION			MODEL STANDARD DEVIATION			
2	name	freq_flair	freq_stanza	freq_xlmr_ner	freq_flair	freq_stanza	freq_xlmr_ner	loc_stdev	org_stdev	per_stdev	avg_stdev
28	anne zernike	82	82	75	50	42	42	4.6188	1.5275	4.0415	3.4
29	hermina amersfoordt	86	76	70	37	38	40	1.5275	3.5119	8.0829	4.4
30	pieter wiedijk	69	51	48	6	8	6	1.1547	2.0817	11.3578	4.9
31	jan kops	28	25	28	22	20	20	1.1547	0.5774	1.7321	1.2
32	clara engelen	35	32	33	35	38	33	2.5166	1.5275	1.5275	1.9
33	roelf hagoort	22	16	19	22	21	22	0.5774	5.1316	3.0000	2.9
34	johannes wilhelmus boerbooms	48	32	52	51	47	47	2.3094	2.5166	10.5830	5.1
35	michael faraday	68	63	63	26	23	23	1.7321	1.5275	2.8868	2.0
36	charlotte sophie von aldenburg	116	129	117	78	52	68	13.1149	2.5166	7.2342	7.6
37	helena theodora rietberg	48	45	42	30	29	28	1.0000	3.6056	3.0000	2.5
38	heymen van 't einde	35	35	35	8	8	6	1.1547	1.7321	0.0000	1.0
39	aleida daendels	85	89	85	77	71	61	8.0829	1.1547	2.3094	3.8
40	weduwe van wouw	55	53	49	19	24	22	2.5166	2.3094	3.0551	2.6
41	berta vorkink	25	21	22	22	25	19	3.0000	0.5774	2.0817	1.9
42	johannes brommert	73	59	66	35	39	34	2.6458	4.0415	7.0000	4.6

Figure 3. Overview of identified person names and locations by three different tools. Taken from Daza and Fokkens (2024, 3).

	PER			LOC			ORG			MICRO			MACRO		
SYSTEM	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1
stanza	61.1	89.7	72.7	64.0	93.7	76.0	42.9	75.3	54.7	59.2	89.1	71.1	56.0	86.2	67.8
flair	53.4	83.9	65.3	66.9	93.4	78.0	47.5	78.9	59.3	56.7	86.4	68.4	55.9	85.4	67.5
xlmr_ner	59.8	86.5	70.7	65.0	90.2	75.6	48.8	75.7	59.3	59.8	86.2	70.7	57.9	84.1	**68.5**
gpt-3.5	0.3	8.2	0.6	0.6	32.2	1.1	0.1	11.1	0.1	0.3	11.9	0.6	0.3	17.2	0.6

Table 1. Precision (P), Recall (R), and f1-score (F1), which is the weighted mean of precision and recall of individual tools on identifying person (PER), location (LOC), and organization (ORG). Table 3 from Daza and Fokkens (2024, 8).

of various state-of-the-art NERC tools on the Biography Portal. Their work focuses on how users can compare the tools' output through visualizations. We will consider their findings in a scenario where a historian is interested in the differences in attention paid to women in various resources by looking at how often women are mentioned in biographies of other people.

Figure 3 provides an overview of the number of person names and location names identified in individual biographies (of the people mentioned in Column B) by three state-of-the-art tools: Flair (Akbik et al. 2019), Stanza (Qi et al. 2020), and XLMR-NERC. The numbers in this figure provide a first awareness signal: in many cases, different tools do not find the same number of persons. It is therefore likely that they will also lead to different numbers of female names. But this is not the full story yet: there are two types of errors that lead to the wrong number of entities being found. The tool may miss entities, which brings the number of found entities down, reflected in poor recall. A tool may also assign the person labels to tokens that do not refer to a person, reflected in poor precision. A tool with relatively high precision and low recall on identifying persons will extract fewer person names than are occurring, and vice versa, a tool with higher recall than precision will report more person names than are actually occurring. Table 1 provides the precision/recall results from four individual tools on a subset of the data when a strict evaluation is applied requiring an exact match between output and input.

All tools have higher recall than precision, meaning that they identify more person names than are actually present in the text. When considering whether and how we can use these tools to answer our research question concerning women being mentioned in biographies, we need to address two questions. First, we need to know whether the relation between precision and recall is similar for the sources we are comparing. Given that the biography portal contains sources from a relatively wide time frame, it is likely that tools will have more difficulties identifying

names in older texts. An increase in female names in more modern sources may then be the result of the errors made by the tool rather than an actual increase in the data. Second, we need to know whether the types of errors are equally distributed for male and female names.[5] If tools have more difficulties identifying female names compared to male names, the relations, recall of female names, will be lower. This should be taken into account when analyzing the data. This means that, in order to know whether and how we can use a tool, we will need to evaluate how it performs on our target data in each of the sources we are comparing.

When determining what needs to be evaluated and whether and how tool output needs to be adjusted for probable errors, the exact research question should be taken into account. If, for instance, the question is how many times a female name occurs per biography, recall and precision of individual female names matter. If the question is more about the number of unique women mentioned in a document, it becomes less problematic if the tool misses some occurrences, as long as it finds each name at least once. If the question concerns the balance between male and female names occurring, then it may not matter if tools have more difficulty with older sources, as long as the difficulties they have with male and female names are comparable.

It follows from this example that whether the tools can be used and whether and how their output should be adjusted for errors does not depend on their overall performance, but rather on their exact performance on the targeted data as well as their comparative performance among sources that are compared. This can be found out by evaluating the accuracy of the tool on a dataset acquired through *hypothesis-based sampling* (Fokkens et al. 2017; Fokkens et al. 2019). Hypothesis-based sampling entails starting with the variables that are being compared while testing a hypothesis. Based on these, researchers identify samples that represent the individual variables so that differences in biases

Table 2
Overall Performance of the Tested Sentiment Analysis Approaches

			Positive			Neutral			Negative		
Method	Acc.	α	Pr.	Re.	F1	Pr.	Re.	F1	Pr.	Re.	F1
Machine Learning											
CNN	0.63	0.50	0.68	0.49	0.56	0.58	0.78	0.66	0.72	0.57	0.63
NB	0.58	0.40	0.74	0.34	0.47	0.52	0.83	0.64	0.65	0.47	0.55
SVM	0.57	0.41	0.69	0.37	0.48	0.52	0.79	0.62	0.64	0.48	0.55

Table 2. Results from Van Atteveld et al. (2021) reporting results of sentiment analysis on Dutch economic headlines by various machine learning models. The boxes around specific results are added by us.

from the tools can be identified and taken into account. The following example illustrates a similar challenge where hypothesis-based sampling can be useful in verifying the validity of an automatically obtained result.

EXAMPLE 2 – DIACHRONIC SENTIMENT ANALYSIS

The second example we discuss concerns diachronic sentiment analysis. In this scenario, researchers are interested in finding out whether the sentiment expressed around a specific topic changes over time. We have tools that can detect whether a message is positive or negative. We could either observe the trends of positive and negative sentiment separately or use a score that combines the two by, for example, deducting the number of negative messages from the number of positive ones. Table 2 provides an overview of results on sentiment detection for Dutch economic headlines by Wouter Van Atteveldt et al. (2021). Their system classifies headlines as expressing positive, negative, and neutral sentiment.

Consider the scenario in which a researcher wants to identify whether news on the economy tends to be reported with more positively or more negatively oriented headlines. When looking at the precision and recall for the positive class on the one hand and precision and recall for the negative class on the other hand, it is clear that for all three models, recall and precision are much closer to each other for the negative class. The positive class has a noticeably lower recall, meaning that it is missed more often compared to the negative class. This means that for a corpus in which the true balance between positive and negative is more or less equal, the output of these systems will give the impression that the negative class is more prominent. In order to know how to interpret the results and determine the relation between the actual number of headlines carrying positive or negative sentiment, the end user needs a reliable indication of the precision and recall for each class on their data.

Consider Figure 4. The graph on the left-hand side provides the actual development of the percentage of positive words and negative words over time. There used to be a higher percentage of positive words than negative words. Both positive and negative words increased, but the negative words did so faster and continued to increase once positive words stabilized. At the end of the time period, the percentage of negative words is higher than the percentage of positive words. The graph on the right hand side indicates the same trend based on the output of a sentiment classifier that automatically analyzed the corpus. When comparing the two, the overall trend is the same. Both increase, but negative sentiment increases faster than positive and continues to increase

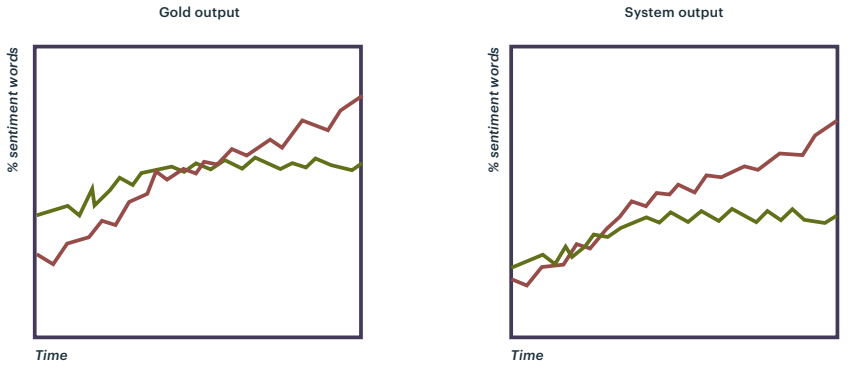


Figure 4. Hypothetical illustration of the gold output of the percentage of positive (green) and negative (red); sentiment words on the left and the positive (green) and negative (red) output predicted by a system on the right.

as positive sentiment stabilizes. In both figures, positive sentiment is dominant at the beginning of the time period and negative sentiment at the end. When the hypothesis that is tested concerns the overall trend of positive and negative words, the output of the sentiment classifier provides the correct answer.

However, when looking closely at the graphs, the system output, very much like the results reported by Van Atteveldt et al. (2021), indicates higher negative sentiment than positive sentiment compared to the gold output. According to the classifier, the difference between negative and positive sentiment is much smaller at the beginning of the time period than is actually the case, and the difference is bigger at the end. Moreover, the moment in time when negative sentiment takes the upper hand is earlier according to the system output. Again, it depends on the research question and hypothesis whether these differences are problematic. It does not matter if this time point is irrelevant. If, on the other hand, the researcher has a hypothesis concerning when this happens based on an event (e.g., political unrest or a general increase in unemployment), then this can directly influence the researcher's interpretation. It can either happen that the researcher's hypothesis is correct, but they miss it believing that the change happened earlier, or that the researcher's hypothesis was incorrect, but the time where the system output places the change is relatively close to their expectation. The first case, when the system's output is unexpected, tends to be less problematic than the latter. Researchers tend to check machine output when it goes against their expectations (though one can also imagine an interpretation that the papers foresaw the problems and headlines became more pessimistic before the event). When results confirm a hy-

pothesis, the risk of researchers being happy with the confirmation by the machine is quite high. This can lead to cherry-picking, where only undesirable results are scrutinized.

Hypothesis-based sampling can help avoid these risks. To increase the chances of interpreting the classifier's output correctly, it is essential that researchers verify the performance of the system across different periods of their data. This way, they can correct for expected over- or under-classification. This also increases the chances of identifying systematic errors such as the system misinterpreting expressions that are highly relevant for a specific period under investigation.

EXAMPLE 3 – DETECTING FRAGMENTATION

Polimeno et al. (2023) provide a good illustration of evaluating system output and the impact on the research question. Their study addresses *fragmentation*, a metric used for studying diversification in news recommendation (Vrijenhoek et al. 2021). Fragmentation measures to what extent readers are exposed to the same information by news recommendation systems. A high fragmentation score indicates that there is little overlap in the information offered to individual readers and low fragmentation indicates that readers mostly see the same content.

Polimeno et al. (2023) compare results of various methods for automatically clustering newspaper articles based on their similarity. In their clustering setup, they varied their input representation, using either contextualized representations from SBERT (Reimers et al. 2019), static word embeddings from GloVe (Pennington et al. 2014), or bag-of-word embeddings. They also tested two clustering algorithms: agglomerative hierarchical clustering (AHC) and density-based clustering (DB).[6] Their evaluation consists of two steps: an intrinsic evaluation and an extrinsic evaluation. The intrinsic evaluation determines the accuracy of their clustering methods. They used the HeadLine Grouping Dataset (Laban et al. 2021) to determine whether their methods accurately place articles that concern the same event in the same cluster. Their results are presented in Table 3.

The values for homogeneity (how many articles within a cluster concern the same event) and completeness (how many articles concerning the same event are placed in one cluster) are comparable to precision and recall in classification. The results show good performance of most methods, notably the highest-scoring system ($H = 0.921$ and $C = 0.844$), but how suitable are they for determining fragmentation? What is the impact of errors that are still being made?

The extrinsic evaluation provides information about this. In this

evaluation, three scenarios are simulated: in the first, fragmentation is low; in the second, fragmentation is high; and in the third, fragmentation is balanced. Table 4 provides the results.

The results in Table 4 generally correspond to the outcome of the intrinsic evaluation in Table 3. The methods that are unsuitable for predicting fragmentation also scored poorly on at least one of the gold standard-based metrics (but note that GloVe*DB scored highest on the Silhouette score and BoW*DB on the Davies–Bouldin Index). The other methods correctly indicate a high fragmentation score in the second scenario and a more balanced one for the third scenario. They mostly also correctly assign the lowest fragmentation score to the first scenario. At the same time, we see that one method scores the low fragmentation scenario almost the same as the balanced one and two methods still score it relatively high. Neither method is capable of correctly indicating that the true fragmentation is zero: the errors made by the clustering methods lead to the model always detecting at least some fragmentation. The combined outcome of the intrinsic and extrinsic evaluation is highly valuable when applying these methods to identify fragmentation on new data.

To sum up, the first two examples illustrated that standard evaluation on a small sample of data is often not sufficient. An evaluation in a digital humanities context should make sure that tools are evaluated on samples that cover all dimensions under investigation. This can be achieved by applying hypothesis-based sampling. Ideally, the final research question and the impact that the tool's errors may have on this

Representation*Clustering	**H ↑**	**C ↑**	**V ↑**	**S ↑**	**DBI ↓**
BoW*Graph-Based (baseline)	0.166	0.156	0.161	-0.060	12.441
SBERT*AHC	**0.921**	**0.844**	**0.881**	0.290	1.933
GloVe*AHC	0.762	0.708	0.734	0.183	**1.913**
BoW*AHC	0.813	0.658	0.727	**0.413**	1.965
SBERT*DB	0.694	**0.872**	**0.773**	0.231	1.509
GloVe*DB	0.002	0.236	0.004	**0.390**	0.387
BoW*DB	**0.993**	0.283	0.441	0.213	**0.218**

Table 3. Repetition of the results of various clustering methods as presented by Polimeno et al. (2023). The errors indicate whether a high value is better (up arrow) or a low score is better (down arrow). The values Homogeneity (H), Completeness (C) and V-measure (V) are based on gold standards, Silhouette Score (S) and Davies–Bouldin Index (DBI) are general indicators of coherence of the clusters based on internal metrics. The explanation of the Silhouette score in Section 2.1 on page 71 is also based on this table.

are taken into account, and the scholar designs their evaluation in such a way that interference of errors by the tool with their hypothesis is detected. The third example illustrates what such an additional extrinsic evaluation can add to an investigation. In this case, the best-performing tools can be used to detect the general pattern, but it should be taken into account that fragmentation is overestimated when it is very low.

CONCLUSION

This chapter has described the most recent advances in NLP, focusing on how advances in language representation have led to significant improvements in automatic text classification. We explained how LLMs trained on large amounts of text pick up on linguistic cues, allowing them to generalize well and learn classification tasks more accurately from relatively less data compared to prior methods.

The second part of this chapter addressed what these advances mean for using automated text analysis for digital humanities research. In particular, we explained that LLMs tend to be trained on mainstream and mostly modern data. Textual data analyzed in digital humanities studies is often far from standard. Even though the latest methods can also improve results on this data, they do not always do so, and we can generally expect tools to make more mistakes on digital humanities data compared to reported results on standard benchmarks. We argued that there is nevertheless a lot of potential for text analysis in the digital humanities if scholars critically reflect on the tools they use. Depending

Representation*Clustering	Scen. 1 ↓	Scen. 2 ↑	Scen. 3	Diff 1-2
Gold Labels	0.00	0.85	0.58	0.85
BoW*Graph-Based (baseline)	0.67	0.73	0.70	0.06
SBERT*AHC	0.31	0.87	0.64	0.56
GloVe*AHC	0.38	0.84	0.63	0.46
BoW*AHC	0.62	0.85	0.63	0.23
SBERT*DB	0.16	0.74	0.48	0.58
GloVe*DB	0.01	0.01	0.00	0.01
BoW*DB	0.99	0.99	0.99	0.00

Table 4. Results on fragmentation scores based on gold labels and various clustering methods, taking from Polimeno et al. (2023) (Table 3). The arrows indicate whether fragmentation should be low, high or in between (in line with the scores based on the gold). Diff 1-2 indicates the difference between high and low fragmentation.

on their data and task, this may mean sticking to "old-fashioned" feature engineering and supervised machine learning methods. At the same time, various efforts have been made to address these challenges by creating specialized language models for specific domains in the humanities. This line of research opens up possibilities for using LLMs for studies in the digital humanities as well. In the end, the question of whether automated processing tools can be used in digital humanities research depends on the general quality of the tool, its suitability for the data and the specific research goal at hand. In exploratory research, where tools are used to identify interesting data, tool accuracy is not relevant as long as interesting things are found. When tools are used to study patterns or changes, it mostly matters whether errors in their output interfere with the hypothesis that is tested or whether errors are similar across variables under investigation (and the overall patterns hold).

In the final part of this chapter, we illustrated this point through three examples: (1) comparing the attention various biographical sources pay to women, operationalized by identifying person names and determining whether they are women; (2) detecting change in sentiment over time; and (3) studying fragmentation in news recommendation. In all three cases, we have shown that classic evaluation results that intrinsically evaluate how accurate the outcome of the tool is when compared to gold-standard data provide a useful first indication, but are not sufficient to know whether the outcome is reliable enough to draw conclusions.

For the first use case, we needed more fine-grained results at two levels. First, we needed to know whether tools yielded comparable performance on the individual sources we aimed to compare. Second, we needed to know whether performance on female and male names was comparable. Before these questions are answered, we cannot blindly draw conclusions from our tool output. For the second use case, we needed to know whether the relationship between precision and recall for positive and negative sentiment remained stable when tested on data sampled from different time frames. Without information on this aspect, we cannot know whether we are measuring a change in sentiment expressed in the text or a difference in error types made by the tools. Hypothesis-based sampling aims to ensure that the necessary samples are included in the evaluation. The third use case illustrated what adding an extrinsic evaluation provides in terms of extra insights. We observed that the most accurate system correctly identified the overall fragmentation pattern, but it overestimated fragmentation when this was very low.

These are exciting times to work with language technology. Studies in the digital humanities tend to be more challenging and results on humanities data will often be less good than on more standard, modern data. Nevertheless, here as well, the latest advances offer new options with specialized language models being developed. At the same time, current technology, whether it is a highly accurate rule-based system, a high performing, old fashioned, machine learning system or LLMs, is not perfect. It therefore still holds that scholars must carefully evaluate what the tool can do and what it cannot do and how its errors relate to the hypothesis that is being tested.

NOTES

1 Person and company names in these examples are made up. Any correspondence with individuals or actual company names are coincidental.
2 Note, however, the observations by Søgaard et al. (2014).
3 https://github.com/explosion/spacy-llm, accessed October 21 2025.
4 http://www.biografischportaal.nl/, accessed September 27 2025.
5 The Biography Portal of the Netherlands provides historical descriptions and only contains people who have passed away. For historical reasons, non-binary people will generally still be described as either male or female.
6 The representations and algorithms are mentioned here for reasons of completeness. Since our main point concerns the evaluation and general insights from intrinsic and extrinsic evaluation and not the methods themselves, a description of their methods is beyond the scope of this chapter.

REFERENCES

Akbik, Alan, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. "FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (demonstrations)*, edited by Waleed Ammar, Annie Louis, and Nasrin Mostafazadeh. Association for Computational Linguistics: 54–9.

Aroyo, Lora, and Chris Welty. 2015. "Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation." *AI Magazine* 36 (1): 15–24.

Artstein, Ron, and Massimo Poesio. 2008. "Inter-Coder Agreement for Computational Linguistics." *Computational Linguistics* 34 (4): 555–96.

Baroni, Marco, and Alessandro Lenci. 2010. "Distributional Memory: A General Framework for Corpus-Based Semantics." *Computational Linguistics* 36 (4): 673–721.

Bender, Emily M., and Batya Friedman. 2018. "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science." *Transactions of the Association for Computational Linguistics* 6: 587–604.

Biemann, Chris, and Martin Riedl. 2013. "Text: Now in 2D! A Framework for Lexical Expansion with Contextual Similarity." *Journal of Language Modelling* 1 (1): 55–95.

Bosley, Mitchell, Musahi Jacobs-Harukawa, Hauke Licht, and Alexander Hoyle. 2023. "Do We Still Need BERT in the Age of GPT? Comparing the Benefits of Domain-Adaptation and In-Context-Learning Approaches to Using LLMs for Political Science Research." In *2023 Annual Meeting of the Midwest Political Science Association (MPSA)*.

Bouma, Gosse, Gertjan Van Noord, and Robert Malouf. 2001. "Alpino: Wide-coverage Computational Analysis of Dutch." In *Computational Linguistics in the Netherlands 2000*, edited by Walter Daelemans, Khalil Sima'an, Jorn Veenstra, and Jakub Zavrel. Brill: 45–59.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–901.

Caselli, Tommaso, Piek Vossen, Marieke van Erp, Antske Fokkens, Filip Ilievski, Ruben Izquierdo Bevia, Minh Le et al. 2015. "When It's All Piling Up: Investigating Error Propagation in an NLP Pipeline." In *Proceedings of the Workshop on NLP Applications: Completing the Puzzle Co-Located with the 20th International Conference on Applications of Natural Language to Information Systems (NLDB 2015)*, edited by Ruben Izquierdo. CEUR Proceedings: 5–15.

Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. "Deep Reinforcement Learning From Human Preferences." In *Advances in Neural Information Processing Systems 30: 31st Annual Conference on Neural Information Processing Systems (NIPS 2017)*. Neural Information Processing Systems Foundation, Inc. (NeurIPS).

Connolly, Brian, Timothy Miller, Yizhao Ni, Kevin B. Cohen, Guergana Savova, Judith W. Dexheimer, and John Pestian. 2016. "Natural Language Processing–Overview and History." *Pediatric Biomedical Informatics: Computer Applications in Pediatric Research*: 203–30.

Curran, James. 2004. *From Distributional to Semantic Similarity*. University of Edinburgh.

Daza, Angel, and Antske Fokkens. 2024. "Choosing the Right Tool for You: Informed Evaluation of Text Analysis Tools." In *CLARIN Annual Conference,* edited by Vincent Vandeghinste and Thalassia Kontino. Linköping Electronic Conference Proceedings 216: 112–25.

Dridan, Rebecca, and Stephan Oepen. 2012. "Tokenization: Returning to a Long Solved Problem – A Survey, Contrastive Experiment, Recommendations, and Toolkit–." In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, edited by Haizhou Li, Chin-Yew Lin, Miles Osborne, Gary Geunbau Lee, and Jong C. Park. Association for Computational Linguistics: 378–82.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), edited by Jill Burstein, Christy Doran, and Thamar Solorio. Association for Computational Linguistics: 4171–86.

Durmus, Esin, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen et al. 2024. "Towards Measuring the Representation of Subjective Global Opinions in Language Models." In *First Conference on Language Modeling*.

Edwards, Aleksandra, and José Camacho-Collados. 2024. "Language Models for Text Classification: Is In-Context Learning Enough?" In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, edited by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. ELRA and ICCL: 10058–72.

Fellbaum, C. 2010. "WordNet." In *Theory and Applications of Ontology: Computer Applications*, edited by Roberto Poli, Michael Healy and Achilles Kameas. Springer Netherlands: 231–43.

Firth, John Rupert. 1957. *Papers in Linguistics*. Oxford University Press.

Fokkens, Antske, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber. 2014. "BiographyNet: Methodological Issues when NLP Supports Historical Research." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno et al. European Language Resources Association (ELRA): 3728–35.

Fokkens, Antske, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, Guus Schreiber, and Victor de Boer. 2017. "BiographyNet: Extracting Relations Between People and Events." In *Europa baut auf Biographien: Aspekte, Bausteine, Normen und Standards für eine europäische Biographik*, edited by Ágoston Zénó Bernád, Christine Gruber, and Maximilian Kaiser. New Academic Press: 193–224.

Fokkens, Antske, Tommaso Caselli, Minh Le, Pia Sommerauer, and Leon van Wissen. 2019. "Using Computational Linguistics Methods in Digital Humanities: On Possibilities and Pitfalls." Използване на методи от компютърната лингвистика в дигиталната хуманитаристика: възможности и рискове." *Списанието се издава с помощта на фонд „Научни изследвания "на Софийския университет „Св. Климент Охридски*" ["Using Methods from Computational Linguistics in Digital Humanities: Opportunities and Risks." The journal is published with the support of the Scientific Research Fund of Sofia University St. Kliment Ohridski]: 95.

Harris, Zellig S. 1954. "Distributional structure." *WORD* 10 (2–3): 146–62.

Hervieux, Natalie, Peiran Yao, Susan Windisch Brown, and Denilson Barbosa. 2024. "Language Resources from Prominent Born-Digital Humanities Texts are Still Needed in the Age of LLMs." In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*. Association for Computational Linguistics: 85–104.

Jawahar, Ganesh, Benoît Sagot, and Djamé Seddah. 2019. "What Does BERT Learn About the Structure of Language?" In *Proceedings for the 57th*

Annual Meeting of the Association for Computational Linguistics, edited by Anna Korhonen, David Traum, and Lluís Màrquez. Association for Computational Linguistics: 365–57.

Kang, Ning, Erik M. van Mulligan, and Jan A. Kors. 2012. "Training Text Chunkers on a Silver Standard Corpus: Can Silver Replace Gold?" *BMC Bioinformatics* 13: 17.

Khurana, Urja, Eric Nalisnick, Antske Fokkens, and Swabha Swayamdipta. 2024. "Crowd-Calibrator: Can Annotator Disagreement Inform Calibration in Subjective Tasks?" In *First Conference on Language Modeling*.

Kim, Yoon. 2014. "Convolutional Neural Networks for Sentence Classification." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Association for Computational Linguistics: 1746–51.

Kim, Yoon, Carl Denton, Luong Hoang, and Alexander M. Rush. 2017. "Structured Attention Networks." In *International Conference on Learning Representations*.

Koolen, Marijn, Jasmijn van Gorp, and Jacco van Ossenbruggen. 2018. "Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice." *Digital Scholarship in the Humanities* 34 (2): 368–85.

Laban, Philippe, Lucas Bandarkar, and Marti A. Hearst. 2021. "News Headline Grouping as a Challenging NLU Task." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagyt, Steven Bethard, Ryan Cotterell et al. Association for Computational Linguistics: 3186–98.

Larimore, Savannah, Ian Kennedy, Breon Haskett, Alina Arseniev-Koehler. 2021. "Reconsidering Annotator Disagreement about Racist Language: Noise or Signal?" In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, edited by Lun-Wei Ku and Cheng-Te Li. Association for Computational Linguistics: 81–90.

Luong, Minh-Thang, Hieu Pham, and Christopher D. Manning. 2015. "Effective Approaches to Attention-Based Neural Machine Translation." In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, edited by Lluís Màrquez, Chris Callison-Burch, Jian Su. Association for Computational Linguistics: 1412–21.

Manjavacas, Enrique, and Lauren Fonteyn. 2022. "Adapting vs. Pre-Training Language Models for Historical Languages." *Journal of Data Mining & Digital Humanities (NLP4DH)*: 1–19.

Marcheggiani, Diego, and Fabrizio Sebastiani. 2017. "On the Effects of Low-Quality Training Data on Information Extraction from Clinical Reports." *Journal of Data and Information Quality* (9) 1: 1–25.

McGillivray, Barbara, Thierry Poibeau, and Pablo Ruiz. 2020. "Digital Humanities and Natural Language Processing: 'Je t'aime … Moi non plus'." *Digital Humanities Quarterly* (14) 2: 1–11.

McIntosh, Timothy R., Teo Susnjak, Tong Liu, Paul Watters, and Malka N. Halgamuge. 2024. "The Inadequacy of Reinforcement Learning from

Human Feedback: Radicalizing Large Language Models via Semantic Vulnerabilities." *IEEE Transactions on Cognitive and Developmental Systems* 16 (4): 1561–74.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeffrey Dean. 2013. "Distributed Representations of Words and Phrases and Their Compositionality." *Advances in Neural Information Processing Systems 26 (NIPS 2013)*: 1–9.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Communications of the ACM* 38 (11): 39–41.

Nielbo, Kristoffer L., Folgert Karsdorp, Melvin Wevers, Alie Lassche, Rebekah B. Baglini, Mike Kestemont, and Nina Tahmasebi. 2024. "Quantitative Text Analysis." *Nature Reviews Methods Primers* (4) 1: 25.

Padó, Sebastian, and Mirella Lapata. 2007. "Dependency-Based Construction of Semantic Space Models." *Computational Linguistics* 33 (2): 161–99.

Park, Soya, April Yi Wang, Ban Kawas, Q. Vera Liao, David Piorkowski, and Marina Danilevsky. 2021. "Facilitating Knowledge Sharing from Domain Experts to Data Scientists for Building NLP Models." In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. Association for Computing Machinery: 585–96.

Pennington, Jeffrey, Richard Socher, and Christopher Manning. 2014. "GloVe: Global Vectors for Word Representation." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, edited by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Association for Computational Linguistics: 1532–43.

Peters, Matthew, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. "Deep Contextualized Word Representations." In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics: 2227–37.

Polimeno, Alessandra, Myrthe Reuver, Sanne Vrijenhoek, and Antske Fokkens. 2023. "Improving and Evaluating the Detection of Fragmentation in News Recommendations with the Clustering of News Story Chains." In *1st Workshop on the Normative Design and Evaluation of Recommender Systems, NORMalize 2023*, edited by Sanne Vrijenhoek, Lien Michiels, Johannes Kruse, Alain Starke, Jordi Viader Guerrero, and Nava Tintarev. CEUR Proceedings: 1–16.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, edited by Asli Celikyilmaz, and Tsung-Hsien Wen. Association for Computatonal Linguistics: 101–8.

Qiu, Wenjun, and Yang Xu. 2022. "Histbert: A Pre-Trained Language Model for Diachronic Lexical Semantic Analysis." *arXiv:*2202.03612.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. "Language Models are Unsupervised Multitask Learners." *OpenAI blog* 1 (8): 9.

Reimers, Nils, and Iryna Gurevych. 2019. "Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks." In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP),* edited by Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan. Association for Computational Linguistics: 3982–92.

Rieder, Bernhard, and Theo Röhle. 2012. "Digital Methods: Five Challenges." In *Understanding Digital Humanities*, edited by David M. Berry. Palgrave Macmillan: 67–84.

Ries, Thorsten, Karina van Dalen-Oskam, and Fabian Offert. 2024. "Reproducibility and Explainability in Digital Humanities." *International Journal of Digital Humanities* 6 (1): 1–7.

Sajith, Aryan, and Krishna Chaitanya Rao Kathala. "Is Training Data Quality or Quantity More Impactful to Small Language Model Performance?" *arXiv:*2411.15821.

Schütze, Hinrich. 1993. "Word Space." In *Advances in Neural Information Processing Systems*, *5*, edited by Stephen Jose Hanson, Jack D. Cowan, and C. Lee Giles. Morgan Kaufmann Publishers Inc: 895–902.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. "Recursive Deep Models for Semantic Compositionality over a Sentiment Treebank." In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, edited by David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard. Association for Computational Linguistics: 1631–42.

Søgaard, Anders, Barbara Plank, and Dirk Hovy. 2014. "Selection Bias, Label Bias, and Bias in Ground Truth." In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Tutorial Abstracts,* edited by Qun Liu and Fei Xia. Dublin City University and Association for Computational Linguistics: 11–13.

Sommerauer, Pia, Antske Fokkens, and Piek Vossen. 2020. "Would You Describe a Leopard as Yellow? Evaluating Crowd-Annotations with Justified and Informative Disagreement." In *Proceedings of the 28th International Conference on Computational Linguistics,* edited by Donia Scott, Nuria Bel, and Chengqing Zong. International Committee on Computational Linguistics: 4798–809.

Stamatatos, Efstathios. 2008. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60 (3): 538–56.

Talat, Zeerak. 2016. "Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter." In *Proceedings of the First Workshop on NLP and Computational Social Science,* edited by David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh et al. Association for Computational Linguistics: 138–42.

Tenney, Ian, Dipanjan Das, and Ellie Pavlick. 2019. "BERT Rediscovers the Classical NLP Pipeline." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, edited by Anna Korhonen, David Traum, Lluís Màrquez. Association for Computational Linguistics: 4593–601.

Turney, Peter D., and Patrick Pantel. 2010. "From Frequency to Meaning: Vector Space Models of Semantics." *Journal of Artificial Intelligence Research* (37): 141–88.

Van Atteveldt, Wouter, Mariken Van der Velden, and Mark Boukes. 2021. "The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms." *Communication Methods and Measures* (15) 2: 121–40.

Van der Zwaan, Janneke, Maksim Abdul Latif, Dafne van Kuppevelt, Melle Lyklema, and Christian Lange. 2021. "Are You Sure Your Tool Does What It is Supposed to Do? Validating Arabic Root Extraction." *Digital Scholarship in the Humanities*, 36 (Supplement_1): 1137–50.

Van der Velden, Mariken A. C. G., Felicia Loecherbach, Wouter van Atteveldt, Antske Fokkens, Myrthe Reuver, and Kasper Welbers. 2025. "Whose Truth Is It Anyway? An Experiment on Annotation Bias in Times of Factual Opinion Polarization." *Communication Methods and Measures*: 1–18.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Lion Jones, Aidan N. Gomez, Lukasz Kaiser et al. 2017. "Attention is All You Need." *NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems,* edited by Ulrike von Luxburg, Isabelle Guyon, Samy Bengio, Hanna Wallach, and Rob Fergus. Curran Associates Inc.: 6000–10.

Vrijenhoek, Sanne, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. "Recommenders with a Mission: Assessing Diversity in News Recommendations." In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval.* Association for Computing Machinery, New York, United States: 173–83.

Wambsganss, Thiemo, Christian Engel, and Hansjörg Fromm. 2021. "Improving Explainability and Accuracy Through Feature Engineering: A Taxonomy of Features in NLP-based Machine Learning." In *Forty-Second International Conference on Information Systems:* 1–17.

Wei, Alexander, Nika Haghtalab, and Jacob Steinhardt. 2023. "Jailbroken: How Does LLM Safety Training Fail?" *Advances in Neural Information Processing Systems* 36: 80079–110.

Yong, Zheng-Xin, Cristina Menghini, and Stephen H. Bach. 2024. "Low-Resource Languages Jailbreak GPT-4." In *Socially Responsible Language Modelling Research*.

Daniel Brodén, Lina Samuelsson & David Alfter

RETOUCHING AND REFIGURING LITERARY CRITICISM

Experiments with a Generative Model for Analyzing Book Reviews

METHODS SUCH AS machine learning and natural language processing (NLP) have long been central to disciplines like language technology and corpus linguistics, as well as to the digital humanities (Fridlund et al. 2024, 122; see Jockers 2013; Jurafsky and Martin 2023). In recent years, however, large language models (LLMs) and generative models, such as the widely discussed GPT (Generative Pretrained Transformer) (Brown et al. 2020), have significantly altered the possibilities for processing and analyzing research materials (see González-Gallardo et al. 2024). In this volume, sociologist Anton Törnberg describes LLMs and generative models as a third generation NLP in the sense that they cannot merely, like previous generations, provide ground for or assist interpretation but also participate in it (see Ch. 3). However, since models like GPT still operate to a significant extent as "black boxes," whose underlying computational processes are not transparent, it becomes difficult to assess the reliability of their outputs, especially with respect to how they may be influenced by training data (Bommasani et al. 2021). Hence, there is a challenge in developing fruitful ways of engaging with research results generated by such models. In the digital humanities, the notion of experiment has emerged as a scholarly ideal, encouraging reflective exploration of the research potential of digital tools and methods (see Drucker 2009 on "speculative computing"). In this spirit, the field invites experimentation with how generative language models might be integrated into humanistic research without losing sight of methodological skepticism.

The research project *The Order of Criticism Revisited: A Mixed-Methods Study of a Century and a Half of Literary Criticism in*

Sweden (2020–2025; Ingvarsson et al. 2022) explores the intersection of "traditional" literary analysis and computational approaches, using materials and findings from an earlier study as its point of departure. In *Kritikens ordning: Svenska bokrecensioner 1906, 1956, 2006* [*The Order of Criticism: A Study of Book Reviews in Sweden 1906, 1956, 2006*] (2013), literary scholar Lina Samuelsson, a member of the current project team, analyzed normative discourses in literary criticism in Swedish newspapers and periodicals, based on a selection of literary book reviews from the years under study.

The present chapter explores the possibility of expanding Samuelsson's study through the use of a generative language model. Rather than aiming to provide robust results or a rigorous methodological evaluation, we experiment with and reflect on the model's potential to enrich the analysis of an already well-known body of research material. Because Samuelsson's original study examined the critics' relationship to the reviewed work, to the authors, and to their own critical role, we consider to what extent similar questions can be posed to GPT-4o through zero-shot prompting, that is, by instructing the model to perform a task that relies on what it has inferred from its training data (in contrast to the more resource-intensive process of fine-tuning on a specific dataset). We chose to focus on GPT-4o because, at the time of writing, it ranks among the most technically advanced and widely used generative models, and has demonstrated specific strengths compared to alternative models in our context.[1]

Our discussion is structured around two general questions: Can the use of GPT-4o reframe Samuelsson's earlier findings on the evaluative discourse of literary criticism? What methodological uncertainties arise from applying a generative model in this context?

MATERIALS AND MIXED METHODS

As part of the *Order of Criticism Revisited* project, we have compiled and annotated approximately 5,800 book reviews published in Swedish newspapers, primarily from the years examined in Samuelsson's original study. The material was assembled through a combination of automatic classification and human annotation, using the digitized newspaper collections of the National Library of Sweden (*Kungliga biblioteket*, hereafter KB) as our source (Brodén et al. 2025). Because GPT-4o is a proprietary model and KB restricts software used "within the walls" of its research infrastructure, KBLab, particularly because of copyright constraints, we focus on reviews from the year 1906 in the newspapers Samuelsson worked with: *Aftonbladet, Arbetet, Dagens*

Nyheter, Göteborgs Handels- och Sjöfartstidning, Göteborgs-Posten, Nya Dagligt Allehanda, Socialdemokraten, Stockholms Dagblad, Svenska Dagbladet, Upsala Nya Tidning, Sydsvenska Dagbladet Snällposten, and *Vårt Land*. In addition, we include *Stockholms-Tidningen*, as it was Sweden's largest newspaper in 1906.[2] To increase the sample for reliability, we also include reviews from 1905 in the major newspapers *Dagens Nyheter* and *Aftonbladet*, resulting in a total of 394 reviews from the period 1905–1906. It should be noted, however, that roughly a quarter of these are composite reviews discussing several books, bringing the total number of reviews to nearly 600. Because our ambition is to examine broader discursive patterns related to the "order of literary criticism" during a defined period, we consider the precise year boundaries to be of limited importance. In the early twentieth century, it was not uncommon for books to be reviewed in the year following their publication, which made the inclusion of 1905 material methodologically reasonable.

We begin this chapter by drawing on the concept of defamiliarization to show how data-driven approaches can be used to create *analytical* distance from familiar research materials, de-automatizing our perception of them. In the light of debates over the concept of AI and invoking the notion of distant technology, we also argue that generative models introduce a distinct kind of *methodological* distance, given their probabilistic character and limited transparency. We then present our methodological experiment using a generative language model to explore our corpus of literary book reviews in relation to Samuelsson's original study.

The experiment is structured in two stages. First, we address the critical challenge posed by the poor quality of digitized newspaper texts in KB's collection, which hampers computational analysis. Here, GPT-4o is used to enhance the textual basis by "retouching" noisy OCR data for computational purposes, that is, to make the texts more machine-readable for analysis. We reflect on the model's ability to improve text quality, but also on the epistemological uncertainty this introduces regarding the relationship between the generated outputs and the original sources, suggesting that the generated texts no longer maintain a clear indexical link to the historical material. Second, we use GPT-4o to identify discursive patterns and evaluative criteria in literary criticism, building on Samuelsson's earlier analysis of the reviewers' attitudes toward the literary works, the authors, and their own roles. Central to our discussion is the dual task of demonstrating the analytical potential of the generative model while critically interrogating its methodological usefulness in this specific context. The

chapter concludes that although the model can open analytical perspectives by defamiliarizing familiar material, it simultaneously introduces methodological distance that demands scrutiny and the development of robust interpretative practices.

DEFAMILIARIZATION AND DISTANCE

Within the digital humanities, there is growing interest in reflecting on what occurs at the intersection of interpretative and computational methods. In *Towards a Digital Epistemology* (2020), literary scholar Jonas Ingvarsson argues that digitization challenges the core of academic practice, "not only by the appearance of new tools and objects, but by the fact that our modes of thought and our way to structure data and knowledge are changing" (Ingvarsson 2020, 4). In other words, socio-technical contexts shape what can be known, not merely how quickly or efficiently we compute. In this sense, digital modes of expression offer scholars different ways of viewing and engaging with information and, by extension, different forms of reflection.

While the discussion in data-rich studies of literary criticism and newspaper collections has largely focused on the analytical potential of computational approaches, rather than on epistemological reflection (see Underwood 2019; Piper 2020; Brottrager et al. 2022), we have previously suggested that data visualizations of the book reviews originally analyzed by Samuelsson (2013) produce a defamiliarizing effect that can stimulate new perspectives and questions about the research material (Brodén et al. 2024; Samuelsson et al. 2024).

The idea that digital methods can hold a defamiliarizing quality is hardly new. For instance, Steven Ramsay, reflecting on Jesuit priest Roberto Busa's pioneering computational analysis of Thomas Aquinas's writings in the *Index Thomisticus* (1946–1980), notes that the punched-card indexing of every word in the corpus produced a distinctive effect: "not the immediate apprehension of knowledge, but instead what the Russian Formalists called *ostranenie*, the estrangement and defamiliarization of textuality. One might suppose that being able to see texts in such strange and unfamiliar ways would give such procedures an important place in the critical revolution the Russian Formalists ignited" (Ramsay 2011, 3).

Although the concept of defamiliarization has been given various meanings in literary theory, it is broadly associated with aesthetic strategies that create distance between a work of art and its observer, with the aim of estranging ordinary perception to facilitate reflection. Specifically, we draw on the Russian Formalist literary theorist Viktor

Shklovsky's notion of defamiliarization, which he defined as a deliberately impeded artistic mode of expression that produces distancing effects to heighten perception and provoke reflection (1990 [1916]). A central idea in Shklovsky's writing is that the force of habit dulls our ability to perceive the distinctive features of everyday phenomena and that art can counteract this automatization of perception. In a more prosaic sense, we suggested that computational representations, such as data visualizations of the reviews in Samuelsson's collection, offer a quantified and abstracted overview of the texts that may also de-automatize our view of them by introducing a degree of perceptual distance from the originals. While these visualizations may not uncover radically different dimensions of the material, they nevertheless served to direct attention to previously overlooked aspects with implications for Samuelsson's original conclusions. Importantly, seeing something in a radically different way does not necessarily mean that one discovers something radically different. Rather, we suggest that the value of thinking about data abstractions through the notion of Shklovsky's defamiliarization lies in their potential to enable a form of analytical "double vision": a means of reappraising familiar material from a computationally mediated distance that invites the reassessment of established assumptions (Brodén et al. 2024, 22–23).

However, in the context of generative language models, we can speak of an additional kind of distancing effect that creates a methodological distance from the research process and the results. These models are associated with distinct kinds of methodological uncertainty compared with the techniques we previously used to create data visualizations of Samuelsson's material, including term frequency-inverse document frequency (TF-IDF) (Brodén et al. 2024, 5). While a model such as GPT-4o can produce text that closely resembles human writing and may suggest familiarity with the material, its outputs are generated from massive general-purpose training data using probabilistic computational processes that remain largely opaque (cf. Gitelman 2013 for critical perspectives on data modeling).

On one level, GPT-4o may appear to reduce the distance between the researcher and the data through its intuitive interface and human-like output, compared with more traditional NLP techniques. Before proceeding, we should, however, clarify what we mean by AI to better gauge the computational processes at issue. As sociologist Elena Esposito argues, "artificial intelligence" is a misleading metaphor: "what we can observe in interactions with algorithms is not necessarily an artificial form of intelligence, but rather an artificial form of communication. Intelligence and communicative capacity are not the same thing"

(Esposito 2022, 2). Esposito emphasizes that "algorithms learn not to think but to participate in communication, that is, to (artificially) develop an autonomous perspective that allows them to react appropriately and generate information in their interaction with other participants" (Esposito 2022, 15). Given that generative models do not "think" in any human sense, but rather simulate the exchange and interpretation of information within human linguistic frameworks, one should not equate their communicatively shaped output with intelligence-driven data processing; Esposito therefore prefers "artificial communication" to "artificial intelligence."

By virtue of their probabilistic, simulated-communication approach to information processing, and the limited visibility into how their outputs are produced, generative models introduce a different kind of methodological distance from research processes and results than our data visualization approach discussed above. While the latter may appear opaque to users lacking technical expertise, they remain, to some extent, transparent and interpretable to those with appropriate expertise. Generative models like GPT-4o, by contrast, tend to distance virtually all users methodologically. Although these models demonstrate considerable capacity for a range of research-related tasks, the computational processes that underpin their responses are largely confined within the metaphorical walls of a black box, even if emerging efforts in model interpretability and reverse engineering begin to offer glimpses of their internal workings (Golgoon et al. 2024).

To some extent, this form of methodological distancing resonates with psychologist and philosopher of technology Robert Romanyshyn's (1989) argument that modern technologies not only extend human capabilities but also restructure perception by introducing a layer of distance that displaces us from the objects of our attention. Although Romanyshyn's phenomenological and culturally critical writing primarily aims to describe a condition in which modern technology contributes to human detachment and alienation from the world (a position that remains open to debate), it more generally underscores the complex interplay between technology and perception: how technologies seem to diminish our sense of limitation while reshaping our modes of engaging with the world (see Telotte 1999, 20). In our case, this notion of distant technology appears relevant to the interpretive relationship between the researcher and the research material, which becomes mediated, if not partially displaced, through the use of a generative model.

Thus, the notion of distant technology offers a perspective from which to consider the dual methodological character of generative language

models. In the context of scholarly inquiry, models such as GPT-4o can serve as tools of "analytical perception," capable of producing, among other things, summaries, tentative interpretations, and, in our case, classifications of evaluative criteria with apparent fluency. While this fluency may appear to bring our attention closer to certain aspects of the book reviews, it also conceals the underlying processes through which such outputs are generated. The model's responses may seem apt or even insightful, but they are shaped by computational logics that are probabilistic, generically inferred, and to a significant extent inaccessible. In this sense, generative models exemplify a technologically mediated engagement that appears to extend our analytical capacity, while introducing a further layer of distance between the researcher and the material, beyond the defamiliarization effect discussed in our previous articles (Brodén et al. 2024; Samuelsson et al. 2024). Bringing these perspectives together helps to highlight the ambiguous role of the model as a research tool and underscores the importance of methodological and epistemological reflection.

RETOUCHED DIGITIZATION AND EPISTEMOLOGICAL UNCERTAINTY

As noted above, data-driven analysis of the book reviews that we compiled from KB's digitized newspaper collections is complicated by limitations in the digitization process, particularly concerning optical character recognition (OCR) (Brodén et al. 2025; see also Jarlbrink et al. 2016, 30–31). Another challenge is article segmentation, that is, the capacity to distinguish between individual articles on a given newspaper page. In practice, the newspaper pages displayed in KBLab at KB mostly appear as a mishmash of text blocks and fragments that only loosely correspond to distinct articles (Börjeson et al. 2024, 5). This situation makes curation and annotation essential for any research using this historical newspaper data (Sikora and Haffenden 2024) and underscores the general importance of large-scale automated OCR correction to address structural deficiencies in the digitized text (Löfgren and Dannélls 2024).

To illustrate the extent of the OCR-related issues, consider the following excerpt from an *Aftonbladet* review of the Swedish translation of Polish author Henryk Sienkiewicz's novel *Bartek the Conqueror* (*Bartek zwycięzca*, 1882; translated into Swedish in 1904 as *Bartek segervinnaren*), which we have manually transcribed from the original print version:

> Som i en brännpunkt är bokens verkliga patos samladt i den beundransvärda scenen inför landtrådet, dit Bartek inställt sig med den respekt för militär, som rent af blifver till dödlig rädsla och ordentligt förgiftar honom. "Liksom en gång inför Steinmetz (generalen) stod han nu inför landtrådet, rak i ryggen som en eldgaffel, med magen in och bröstet ut och återhållen andedräkt. Där voro några officerare närvarande – kriget och militärdisciplinen trädde för Bartek som lifslefvande. Officerarna tittade på honom genom guldbågade pincenezer med allt det förakt som en simpel soldat och polsk bonde har att förvänta af preussiska officerare. Han höll andan, och landtrådet sade något i befallande ton. Han bad icke, han befallde och hotade. Riksdagsmannen hade dött i Berlin; nytt val skulle förrättas. 'Du polnisches Vieh! Försök bara att rösta på pan Zarzynski (den polske herremannen), försök bara!'" (F. V. 1905)

Compare this with the corresponding passage in the OCR-processed version of the text from KB's digitized newspaper collections, which contains several obvious errors:

> Som 1 en bräruipttakt är bokens verkliga- patos samladt i den beundransvärda scenen inför landtrådet dit Bartek inställt sig med den respekt för militär som rent af blifver till dödlig rädsla och ordentligt förgiftar honom »Liksom en gång- inför StoiometTi ifgeneral«u stod tran im Inför l&ttrtfr/Met roll i ryggen som en eldgaffel med mogen in och bröstet ut och återhållen andedräkt Där voro några officerare närvarande – kristet och militji-rdiseiplinen trädde för 'Rartok som lifslefvande Officerarna tittade på honom genom gnldbågade pineonezer mod allt <let förakt som en sirap soldat och polsk bonde har att förvänta af preussiska ©fficorars Han hfjll andan och landtrådet säde något i befallande ton Han bad icke haft BfvertaJade icke tian befallde och hotode 2tik«dagsmanrjen hade dött i Berlin nytt val (kulle förrättas »Du p o 1 n i s c h o s V i » h Försök bara att rusta på pan Zorzyaski (den poUks herremannen försök barn!»

Notably, the word *brännpunkt* (focal point or flashpoint) is rendered as the illegible string *bräriopttakt*, and in other instances, numbers and symbols are intermingled with letters, for example, *2tik«dagsmanrjen* instead of *riksdagsmannen* (member of parliament).

Whereas the robust handling of the negative analytical effects of noisy OCR would require manual curation, we carried out a methodo-

logical experiment using GPT-4o to improve text data from the approximately 400 review articles from 1905–1906 – about 600 reviews in total, depending on whether composite reviews are counted – extracted from KB's newspaper collections. The model was prompted via its API (application programming interface) using both a system prompt and a user prompt. The system prompt defines general instructions regarding the model's behavior, tone, and constraints, while the user prompt provides task-specific instructions. By combining system and user prompts, it becomes possible to better ensure that the model generates consistent and relevant responses while tailoring its output to our particular needs. In our case, we instructed GPT-4o to act as an expert in OCR correction of early twentieth-century Swedish texts and documents, carefully identifying and correcting OCR-related errors in the digitized KB texts, while preserving their stylistic tone, historical language usage, and intended meaning (for the full prompts, see App. 1).

It should be emphasized that addressing OCR issues through a generative model in this way amounts to a form of text fabrication. One might debate which term most precisely captures this process, but to provide a more specific alternative, we have opted for the term *retouching*, as it conveys both reuse and modification. Terms such as "improvement," "correction," "reconstruction," or "restoration" would arguably be somewhat misleading, since the generated text is not derived from the original printed source but from a flawed digitized version. Essentially, there is no clear indexical relationship with the original material (see Sterne 2016 on the digital vs. analog divide). Whereas the unprocessed OCR version retains an indexical connection with the original newspaper page, traceable to the physical imprint of ink on paper, the GPT-retouched version lacks this causal, if abstract, connection and is thus further removed. The original referential bond is weakened, if not severed, and can no longer be said to reliably point back to its source. What results is a statistically inferred approximation that resembles the original but is generated without direct reference to the source, mediated instead by an algorithm geared toward probabilistic pattern recognition rather than reproduction.

GPT-4o generates these texts based on probabilistic assumptions about what the erroneously OCR-processed words in the digitized corpus are likely to have been. Rather than restoring the original texts by consulting the actual newspaper copies, it produces versions that are linguistically and contextually plausible, drawing on patterns and statistical likelihoods acquired during training. In this sense, the model does not reconstruct the original material, but instead offers a linguistically informed approximation – essentially "estimating" what might

have been written. Hence, there can be no expectation that the generated texts will match the originals exactly. That said, when we consider the passage cited above, the retouched version generated by GPT-4o reads as follows:

> Som i en brännpunkt är bokens verkliga patos samladt i den beundransvärda scenen inför landtrådet dit Bartek inställt sig med den respekt för militär som rent af blifver till dödlig rädsla och ordentligt förgiftar honom »Liksom en gång inför Stoiometern gickeneral'en stod han ännu Inför landtrådet roll i ryggen som en eldgaffel med magen in och bröstet ut och återhållen andedräkt Där voro några officerare närvarande – kristet och militärdiseiplinen trädde för Bartek som lifslefvande Officerarna tittade på honom genom guldbågade pinçenez mod allt det förakt som en simpel soldat och polsk bonde har att förvänta af preussiska officerare Han höll andan och landtrådet sade något i befallande ton Han bad icke han Bfverta lade icke han befallde och hotade Riksdagsmannen hade dött i Berlin nytt val skulle förrättas »Du polnisches Vieh Försök bara att rusta på pan Zorzyaski (den polske herremannen försök barn!»

Notably, the generated text does not fully match the original. For instance, there are discrepancies in punctuation and in specific words; *bara* (only or just) is rendered as *barn* (children); *rösta* (to vote) as *rusta* (arm); and *kriget* (the war) as *kristet* (Christian). These discrepancies are far from negligible for interpretation and are, from a philological standpoint, highly problematic. Because the output does not replicate the original text with complete accuracy, and because no trace of the corrections is retained, it would be precarious to base interpretation solely on the retouched versions. Nevertheless, in many ways, the version generated by GPT-4o may appear more faithful to the original than the OCR-processed text from KB. Notably, the model correctly replaced the previously cited OCR error *brärіopttakt* with *brännpunkt*, in accordance with the original print version.

An evaluation comparing manually transcribed reviews with both KB's raw OCR and the GPT-4o retouched versions confirms that the latter consistently align much more closely with the transcriptions. To evaluate the OCR correction process, we conducted an analysis of both the entire corpus and a subset ($n = 21$) in which the texts had been manually transcribed. This was done by comparing the Levenshtein distances between the noisy OCR text and its corrected version, as well as between both versions and the manually transcribed text. The Levenshtein distance indicates the number of operations (insertion,

deletion, or substitution of characters) required to transform one text into another and thus serves as a measure of textual change. A low score reflects minimal changes, while a high score indicates more substantial alterations. On average, the OCR correction process required 320 operations per text (95 % CI: 253–386, n = 394). For the manually transcribed subset, the OCR-retouched texts were, on average, closer to the manual transcriptions than the original OCR versions, with the noisy OCR texts being approximately 104 operations further from the gold standard than the corrected or retouched versions (95 % CI: 64–143, n = 21).

Nevertheless, this gives rise to a deeper epistemological issue. Somewhat paradoxically, we are confronted with texts that appear more accurate and usable in data-driven analysis yet are not derived from the original printed material; rather, they are disconnected from it. In this sense, GPT-4o's role in OCR retouching does not simply enhance access to historical sources. Rather, the model's output introduces epistemological uncertainty. Returning to Esposito's argument that generative models are better understood as a form of artificial communication rather than intelligence, this kind of retouching reshapes the noisy OCR output through a communicative simulation that conceals its own generative process. While the result may, at first glance, seem to bring us closer to the historical text, it simultaneously creates a layer of perceptual and methodological distance. The increased legibility it offers comes at the cost of transparency. Without consulting the original print sources, we cannot determine the extent to which the model's output reflects the original material. Thus, rather than restoring the past, GPT-4o communicates a plausible reconstruction of it, an effect that underscores the need for continued scrutiny.

GENERATED CATEGORIZATIONS OF REVIEWERS' EVALUATIVE CRITERIA

We now turn to our experiment exploring the potential of using a generative model to study evaluative criteria in literary criticism in line with Samuelsson's earlier discourse analysis of material from 1906. To this end, we instructed GPT-4o to analyze reviewers' statements regarding their assessments of the works in relation to the historical context, using our retouched reviews as input. This was done through a system prompt ("Analyze reviews to extract what matters to the reviewer, not what the reviewed work contains. Identify the reviewer's key judgments, criteria, and reactions while considering the critical standards of their time period. Different eras value different aspects in

writing, and your analysis should capture both personal priorities and historical context") and a user prompt ("Analyze this review and provide a list of maximum five items that represent what's most important to the reviewer. Include both their personal judgments and how they reflect or diverge from critical trends of their time period, the time period being 1906. Format the list into topics as from a topic model"). Since GPT-4o could not process all the texts simultaneously, the reviews were entered into the model individually. As a result, the model did not retain any "memory" of the topics it previously generated, which led to the production of a large number of categories, many of them similar or overlapping (e.g., "Social Commentary and Realism" and "Realism and Social Commentary"). To reduce redundancy and create a reasonably coherent overview, we chose to consolidate the categories into broader categories.[3] To explore these categories in relation to the underlying retouched reviews, we also developed a simple interface.[4] Below we present the overarching categorization within which a range of subcategories are available in the interface (number of texts in each category indicated in parentheses):

Human Nature and Intellectual Themes (729)
Emotional and Aesthetic Impact (93)
Characterization and Critique (91)
Narrative Quality and Form (89)
Realism and Authenticity (86)
History and Cultural Context (81)
Social and Cultural Critique (80)
Artistic Craft and Development (61)
Psychological and Character Depth (56)
Intellectual and Symbolic Exploration (51)
Originality and Literary Style (50)
Ethical and Philosophical Themes (36)
National and Cultural Identity (36)
Technical Skill and Observation (23)
Clarity and Narrative Structure (23)

The top-ranked category ("Human Nature and Intellectual Themes") contains significantly more texts than the others and is also formulated in particularly broad terms, which makes it difficult to draw any precise conclusions about its "intended" content. By contrast, the other main categories correspond well to aspects commonly associated with literary criticism and book reviewing, such as narrative technique, style, historical and cultural context, social critique, literary characters, and

ethical or philosophical questions. What is most interesting in our context is that several of these categories reflect features that Samuelsson identified as characteristic of literary criticism in 1906, including an interest in the description and evaluation of fictional characters, national identity and distinctiveness, the authors' perceived authenticity, and personal development as evaluative criteria. Notably, the 1906 reviews were also shaped by a distinctly Romantic conception of authorship, artistic creation, individual expression, and originality (Samuelsson 2013, 33–49).

Characterized as a black box, the model poses challenges when used to identify overarching categories in the material. Although it generates seemingly coherent and meaningful categories, we lack insight into the basis on which they are constructed, including the extent to which they reflect actual patterns in the reviews or merely pre-learned, generic assumptions about what constitutes "typical" content in literary criticism. While this constitutes a methodological limitation and requires that the results be interpreted with caution, they may nonetheless serve as a productive starting point for opening up new perspectives when supported by close readings of the reviews. While a LLM such as GPT-4o does not provide definitive answers, its ability to simulate human judgment at scale, when used in an exploratory way as a classifier, can, as data scientist David Bamman and colleagues suggest, point to interesting analytical directions that can be tested by more formal means: "In this sense, we might view LLMs as tools for sensemaking – helping us understand categories and the boundaries that circumscribe them" (Bamman et al. 2024, 12).

Notably, one category of the reviews that GPT-4o identifies as central, "Emotional and Aesthetic Impact," did not play a prominent role in Samuelsson's analysis. Since emotional responses to literary works, according to Samuelsson (2013, 134–138), are more characteristic of present-day criticism, it is relevant to examine more closely the texts that the model associates with this category and to reflect on possible explanations for this discrepancy.[5]

The first example in the subcategory "Emotional Resonance" (which falls under the main category "Emotional and Aesthetic Impact") is a review of Ellen Lundberg's poetry collection *Sånger och syner* [*Songs and Visions*] (1906). The reviewer emphasizes the importance of literature's ability to come alive for the reader and notes that while some poems in the collection depict the Italian landscape in a manner that is "elegantly and carefully written" ("elegant och omsorgsfullt skrifna"), they nevertheless lack "greater peculiarity" ("större egendomlighet").[6] According to the critic, "it is not enough to enumerate details in order

to evoke a vivid and lively image in the reader's mind" ("det icke är nog med att uppräkna detaljer för att framkalla en åskådlig och levande bild i läsarens sinne"). The reviewer also comments on other poems in the collection, stating that "one is moved by sympathy for the person and the questions raised, for everything is so genuinely human, felt with a warm heart and thought with clear understanding" ("man gripes av sympati för personen och de frågor som framkastats ty allt är så äkta mänskligt kändt med ett varmt hjärta och tänkt med klart förstånd") (Mortensen 1906). It seems reasonable to assume that the model picked up on such textual elements, especially since other reviews in the same context also highlight the emotions a literary work evokes in the critic. In a review of Oscar Stjerne's poetry collection *Sol och snö* [*Sun and Snow*] (1906), one poem's "warm mood" is described as "creeping into the innermost corners of the reader's heart" ("varma stämning smyger sig in i läsarens hjärtevrå") (K. W–g 1906). In a review of Gustava Svanström's novel *På Blåhammars bruk* [*At Blåhammar's Mill*] (1905), it is noted that the manor-house narrative is "written with ease, without signs of stylistic affectation, but also without pretension" ("är ledigt skriven, utan spår af stilsträfvan, men också utan affektation"), and that it is therefore "entirely up to the reader whether he can become more deeply engaged with the material, whether he finds it trivial, though consistently respectable, and whether he quickly forgets it all" ("blir läsarens ensak om han ej kan intressera sig lifligare för det som skildras, om han finner det hela obetydligt, om också alltigenom aktningsvärdt, och om han snarligen glömmer alltsammans") ("Korta anmälningar" 1905).

Taken together, these examples point to the importance ascribed to a literary work's ability to emotionally resonate with its readers. One of the clearest examples of emotionally engaged criticism can be found in the review of the new edition of Hilma Angered-Strandberg's novel *Den nya världen* [*The New World*] (1906 [1898]), which emphasizes how the reader follows the protagonists' struggle "with unwavering interest and empathy" ("med aldrig svikande intresse och medkänsla"), and describes it as a "book that provokes thought, that seems to grab the reader by the collar and force him to pause before the strange street of life" ("en bok, som föder tankar, som liksom tar läsaren för bröstet och tvingar honom att stanna inför lifvets underliga gata") (Rbg. 1906).

The emphasis placed by the critics in these reviews on the emotional responses evoked in readers can be compared with another aspect noted but not developed by Samuelsson in her study. According to her, the critic's role in 1906 is best described as authoritative and judgmental. However, she also observes certain elements in the material sug-

gesting that some, particularly younger, reviewers embraced an ideal in which the critic should be the author's equal, driven by a desire to understand rather than to judge (Samuelsson 2013, 31–32). Although Samuelsson does not have enough examples in her material to draw firm conclusions, she speculates whether this may be a sign of a generational shift. The generative model's re-readings (and the exploration of a larger body of reviews) thus indicate that the aesthetic ideal that criticism should be subjective, primarily associated with Walter Pater within late-nineteenth century Anglo-Saxon literary criticism, was also present among Swedish book reviewers. For Samuelsson, this more empathetic critical ideal was especially evident in the reception of Vilhelm Ekelund's *Hafvets stjärna* [*Star of the Sea*] (1906) and Nils Vilhelm Lundh's short story collection *Mars* [*March*] (1906), both of which are described as too delicate to be subjected to analysis and instead in need of being approached with great care (Samuelsson 2013, 32). A similar attitude can also be found in the larger set of texts that GPT-4o associated with the subcategory "Emotional Resonance," such as a review of Otto Ernst's *Asmus Sempers Jugendland: Der Roman einer Kindheit* (1905, reviewed in the original German), which begins:

> There are books that are cried out by the critics like wares in a market square, and yet do not live beyond the morrow. They were stillborn already in the writer's fancy, and never reached the bridge that leads across to that which creates the great and burning values of life. And the criticism that exalted them to the heavens stood not in sufficiently living contact with mankind to comprehend that poetry must contain life, even as the sea mirrors the sky. True poetry oft springs forth like a timid blossom and suffers not to be handled as market-goods, but must be savoured in silence. It demands, in the act of reading, something of the same conditions under which it was written, if it is to be rightly understood. This is especially true of this fine and richly poetic portrayal of a child's fancies and adventures. It must strike spark upon spark in our own soul, and our own childhood arises from the past like a glistening morning. (C.D.M. 1906)[7]

Once again, we encounter the notion that poetry requires gentle handling and that the critic's task is to be attentive to the conditions under which it was created. The category title "Emotional Resonance" may well reflect this line of thought, in which poetry is said to strike "spark after spark in our own soul" ("gnista efter gnista i vår egen själ") and evoke memories of readers' own childhoods. Another example of this

interpretative pattern appears in a review of Per Hallström's poetic short story collection *De fyra elementerna* [*The Four Elements*] (1906), which GPT-4o associated with the category "Authenticity and Emotional Depth." According to the reviewer, the prose poem "Kerstin" stands out as the finest among "so much genuine poetry" ("så mycken och så äkta poesi") as "the image is movingly beautiful in its delicate, pale colors; it is more than deeply conceived, it is brilliantly conceived. It possesses qualities that most poets never find along their path. There is no point in quoting or attempting a summary – the figure would only lose in charm and originality if touched upon."[8] In other words, the critic refrains from quoting the text, feeling that this would do it a disservice, and concludes instead: "In a poetry collection, each reader chooses what resonates most and calls it the best. In this book, I choose Kerstin" ("I en diktsamling väljer en hvar läsare det som slår mest an på honom och anser det vara det bästa. I denna bok väljer jag Kerstin") (G–g N 1906). In short, this is an evaluation shaped more by emotional responsiveness than by an authoritative stance. The generative model thus helps identify a pattern that, while briefly noted by Samuelsson, becomes clearer here across the larger corpus. However, to ascertain potential meanings attached to some of the vague categories generated by GPT-4o, such as "Emotional Resonance," additional close readings were still required.

CONCLUSIONS

In this chapter, we have experimented with and reflected on the use of a generative language model to analyze Swedish literary criticism 1905–1906. Drawing on the theoretical notions of defamiliarization and distant technology, we have examined the methodological implications of GPT-4o's potential to enrich the analysis of a familiar research corpus.

The first experiment demonstrated the model's capacity to enhance the quality of OCR-degraded book-review texts, thereby providing a more reliable foundation for subsequent computational analysis. However, we also showed that this process weakens the connection to the original source texts, giving rise to epistemological uncertainty. Although the texts generated by GPT-4o often correspond more closely to the originals than the noisy OCR data, they are better understood as probabilistic retouchings rather than as reconstructions or restorations, since they lack an indexical relationship to the original material and therefore no longer reliably point back to it. In the second experiment, we instructed the model to identify discursive patterns in the retouched review texts related to the reviewers' attitudes toward the literary works, the authors, and their own critical role in the review

texts. The results showed that GPT-4o is capable of highlighting analytically meaningful and, to some extent, overlooked aspects in early twentieth-century literary criticism, including what could be described as an emotional dimension. Although this aesthetic ideal is familiar, such tendencies are not always acknowledged in individual case studies, including Samuelsson's earlier work. Even so, the opacity surrounding how the model generated these categories limits our grasp of its underlying processes and requires caution in interpreting the results.

Furthermore, we suggested that our results call for reflection on how generative models reshape scholarly perception through what we describe as distancing effects. Drawing on the concept of defamiliarization, we show that GPT-4o, much like the data visualizations in our previous study, can open alternate analytical avenues for engaging with familiar materials, while also introducing qualitatively different forms of distance. Framing AI as artificial communication rather than artificial intelligence, that is, as a system that produces plausible outputs without semantic understanding, we emphasize that a model like GPT-4o can yield analytically interesting perspectives even as its results remain methodologically distant. This ambiguity is manifested in the model's probabilistic character and the opacity of its underlying computational processes, which complicates transparency and calls for reflection on the methodological distance embedded in the model.

As generative models gain influence in humanities research, the challenge is not only to explore and refine their use, but also to cultivate methodological strategies and deepen our epistemological understanding of the conditions under which such tools produce knowledge.

NOTES

1 We also tested Claude 3.7 Sonnet, but limitations in our experimental context led to incomplete and inconsistent outputs.

2 Samuelsson did not include this newspaper in her study, as it was deemed to have limited coverage of literary content.

3 Given the large number of categories involved (1,482 unique categories), we chose to carry out the various steps of the process "manually" using prompts in ChatGPT (GPT-4o), rather than through APIs, in order to maintain greater control and ensure that no category were filtered out by the generative models for unclear reasons. The model was first instructed to group semantically similar categories into a number of smaller subcategories and to generate a list summarising how the categories had been merged. This step also included basic text cleaning, such as removing asterisks and explanatory phrases generated by the model (e.g., "Philosophical Depth – *Concern with substantive content, personal depth, and human foundation beneath aesthetic surface*," emphasis added). GPT

was then instructed to use the initial summary lists to generate a more condensed version, including an overview of the categories contained in each.

4 This categorization is presented in the following interface: https://kno.dh.gu.se/attitudes.

5 Please note that the model was only prompted, not trained, to capture the distinctive features of 1906-era literary criticism.

6 All translations by the authors.

7 "Det finns böcker som ropas ut af kritiken likt varor på ett marknadstorg och ändå inte lefva öfver morgondagen. De voro dödfödda redan i skrifvarens fantasi och nådde aldrig ut på bryggan öfver till det som skapar de stora och brinnande lifsvärdena, och kritiken, som höjde dem till skyarne, stod ej i en tillräckligt lefvande beröring med människorna för att förstå att dikten måste rymma lifvet liksom hafvet speglar himlen. Den värkliga dikten spirar ofta fram som en blyg blomma och tål inte att handteras som marknadsvara utan skall njutas i tystnad. Den kräfver under läsningen något af samma förhållanden hvarunder den skrefs för att fullt förstås. Det gäller i hög grad om denna fina och rikt poetiska gestaltning af ett barns fantasier och äfventyr. Den måste slå gnista efter gnista i vår egen själ, och vår egen barndom stiger upp ur vårt förgångna som en glittrande morgon."

8 "Bilden är rörande vacker i sina spröda, skära färger, den är mer än djupt, tänkt, den är genialiskt funnen, den har drag sådana som de flesta diktare inte hitta på sin väg. Det är inte lönt att citera eller att försöka en redogörelse – gestalten skulle endast förlora i behag och ursprunglighet, om man vidrörde henne."

REFERENCES

LITERATURE

Bamman, David, Kent K. Chang, Li Lucy, and Naitian Zhou. 2024. "On Classification with Large Language Models in Cultural Analytics." *arXiv*:2410.12029.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. 2021. "On the Opportunities and Risks of Foundation Models." *arXiv*:2108.07258.

Brodén, Daniel, Jonas Ingvarsson, Lina Samuelsson, and Victor Wåhlstrand Skärström. 2024. "Visualization as Defamiliarization: Mixed Methods Approaches to Historical Book Reviews." *Journal of Computational Literary Studies* 3 (1): 1–26.

Brodén, Daniel, Lina Samuelsson, Niklas Zechner, Jonas Ingvarsson, and Aram Karimi. 2025. "Between the Arduous and the Automatic: A Comparative Approach to the Challenge of Identifying Book Reviews in Swedish Newspapers." In *DHNB2024 Conference Proceedings Volume 2,* edited by Olga Holownia and Eiríkur Smári Sigurðarson. University of Oslo Library; Digital Humanities in the Nordic and Baltic Countries Publications (7) 3: 1–12.

Brottrager, Judith, Annina Stahl, Arda Arslan, Ulrik Brandes, and Thomas Weitin. 2022. "Modeling and Predicting Literary Reception: A Data-Rich

Approach to Literary Historical Reception." *Journal of Computational Literary Studies* 1 (1): 1–27.
Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. 2020. "Language Models are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–901.
Börjeson, Love, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati et al. 2024. "Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden." *College & Research Libraries* 85 (4): 564–82.
Drucker, Johanna. 2009. *SpecLab: Digital Aesthetics and Projects in Speculative Computing*. University of Chicago Press.
Esposito, Elena. 2022. *Artificial Communication: How Algorithms Produce Social Intelligence*. MIT Press.
Fridlund, Mats, David Alfter, Daniel Brodén, Asheley Green, Aram Karimi, and Cecilia Lindhé. 2024. "Humanistic AI: Towards a New Field of Interdisciplinary Expertise and Research." In *Huminfra Conference (HiC2024)*, edited by Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, Christian Horn et al. LiU Electronic Press: 122–7.
Golgoon, Ashkan, Khashayar Filom, and Arjun Ravi Kannan. 2024. "Mechanistic Interpretability of Large Language Models with Applications to the Financial Services Industry." *arXiv*:2407.11215.
Gitelman, Lisa. 2013. *"Raw Data" is an Oxymoron*. MIT Press.
González-Gallardo, Carlos-Emiliano, Hanh Thi Hong Tran, Ahmed Hamdi, and Antione Doucet. 2024. "Leveraging Open Large Language Models for Historical Named Entity Recognition." In *Linking Theory and Practice of Digital Libraries, TPDL 2024: 28th International Conference on Theory and Practice of Digital Libraries, TPDL 2024, Ljubljana, Slovenia, September 24–27, 2024, Proceedings, Part I*, edited by Apostolos Antonacopoulos, Annika Hinze, Benjamin Piowowarski, Mickaël Coustaty, Georgio Maria Di Nunzio, Francesco Gelati, and Nicholas Vanderschantz. Springer.
Ingvarsson, Jonas. 2020. *Towards a Digital Epistemology: Aesthetics and Modes of Thought in Early Modernity and the Present Age*. Palgrave Macmillan.
Ingvarsson, Jonas, Daniel Brodén, Lina Samuelsson, Victor Wåhlstrand Skärström, and Niklas Zechner. 2022. "The New Order of Criticism: Explorations of Book Reviews Between the Interpretative and Algorithmic." In *The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022)*, edited by Karl Berglund, Matti La Mela, and Inge Zwart. CEUR-WS: 228–34.
Jarlbrink, Johan, Pelle Snickars, and Cristian Colliander. 2016. "Maskinläsning: Om massdigitalisering, digitala metoder och svensk dagspress." *Nordicom-Information* 38 (3): 27–40.
Jockers, Matthew. 2013. *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press.
Jurafsky, Daniel, and James H. Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational*

Linguistics and Speech Recognition, 3rd ed., draft version, https://web.stanford.edu/~jurafsky/slp3/.
Löfgren, Viktoria, and Dana Dannélls. 2024. "Post-OCR Correction of Digitized Swedish Newspapers with ByT5." In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, edited by Yuri Bizzoni, Stefania Degeatano-Ortlieb, Anna Kazantseva, and Stan Szpakowicz. Association for Computational Linguistics: 237–42.
Piper, Andrew. 2020. *Can We Be Wrong? The Problem of Textual Evidence in a Time of Data*. Cambridge University Press.
Ramsay, Stephen. 2011. *Reading Machines: Toward an Algorithmic Criticism*. University of Illinois Press.
Romanyshyn, Robert. 1989. *Technology as Symptom and Dream*. Routledge.
Samuelsson, Lina. 2013. *Kritikens ordning: Svenska bokrecensioner 1906, 1956, 2006*. Bild, text & form.
Samuelsson, Lina, Daniel Brodén, Jonas Ingvarsson, and Victor Wåhlstrand. 2024. "Kritikens ordning visualiserad: Mixade metoder i studier av bokrecensioner." *Samlaren* 145: 330–54.
Shklovsky, Viktor. 1990 [1916]. "Art as Device." *Theory of Prose*. Dalkey Archive Press: 1–14.
Sikora, Justyna, and Chris Haffenden. 2024. "AI, Data Curation and the Data Readiness of Heritage Collection: Exploring the Swedish Newspaper Archive at KBLab." In *Proceedings of the Huminfra Conference (HiC 2024)*, edited by Elena Volodina, Gerlof Bouma, Markus Forsberg, Dimitrios Kokkinakis, David Alfter, Mats Fridlund, Christian Horn et al. Linköping Electronic Conference Proceedings: 60–6.
Sterne, Jonathan. 2016. "Analog." In *Digital Keywords: A Vocabulary of Information Society and Culture*, edited by Benjamin Peters. Princeton University Press: 31–44.
Telotte, J. P. 1999. *A Distant Technology: Science Fiction Film and the Machine Age*. Wesleyan University Press.
Underwood, Ted. 2019. *Distant Horizons: Digital Evidence and Literary Change*. University of Chicago Press.

BOOK REVIEWS

C. D. M. 1906. "Ett barns roman." Review of *Asmus Sempers Jugendland: Der Roman einer Kindheit*, by Otto Ernst. *Göteborgs Handels- och Sjöfartstidning*, July 3.
F. V. [Fredrik Vetterlund]. 1905. "Konst och litteratur." Review of *Bartek segervinnaren*, by Henryk Sienkiewicz. *Aftonbladet*, May 4.
G–g N. [Georg Nordensvan]. 1906. "Per Hallströms nya bok." Review of *De fyra elementerna*, by Per Hallström. *Dagens Nyheter*, November 11.
Mortensen, Johan. 1906. "Litteratur." Review of *Sånger och syner*, by Ellen Lundberg. *Svenska Dagbladet*, December 28.

K. W–g. [Karl Warburg]. 1906. "Bokvärlden." Review of *Sol och snö*, by Oscar Stjerne. *Göteborgs Handels- och Sjöfartstidning*, December 24.
"Korta anmälningar." 1905. Review of *På Blåhammars bruk*, by Gustava Svanström. *Dagens Nyheter*, May 2.
Rbg. 1906. "Bokvärlden." Review of *Den nya världen*, by Hilma Angered-Strandberg. *Göteborgs Handels- och Sjöfartstidning*, June 7.

APPENDIX 1: PROMPTS FOR OCR RETOUCHING

SYSTEM PROMPT (original prompt in Swedish)

You are an expert in OCR correction of older Swedish texts, with a particular focus on documents from around the year 1906. Your task is to carefully identify and correct OCR errors in a text while preserving the original style, historical language usage, and meaning. When performing the corrections, you should:

Reproduce the original tone and linguistic character typical of the period, including older spelling and grammatical forms.

Correct obvious character recognition errors, such as incorrect letters, word fragments, or punctuation, without modernising the language.

Ensure that the corrections improve readability and linguistic accuracy, while maintaining the historical feel.

If you are uncertain about a correction, mark the error or leave it unchanged with a note for further review.

Work methodically and carefully, taking into account the context and style typical of Swedish writing from the early 20th century.

Use this guidance to produce a corrected version of the OCR text with high accuracy and respect for the original expression of the historical text.

Return only the corrected text, without any comments or further explanations.

Du är en expert inom OCR-korrigering av äldre svenska texter, med särskilt fokus på dokument från omkring år 1906. Din uppgift är att noggrant identifiera och korrigera OCR-fel i en text samtidigt som du bevarar originalets stil, historiska språkbruk och mening. När du utför korrigeringarna ska du:

Återge de ursprungliga tonen och språkliga karaktären som var typisk för tiden, inklusive äldre stavnings- och grammatikformer.

Rätta uppenbara fel i teckenigenkänning, såsom felaktiga bokstäver, orddelar eller interpunktion, utan att modernisera språket.

Säkerställa att korrigeringarna förbättrar läsbarheten och den språkliga korrektheten, samtidigt som den historiska känslan bevaras.

Om du är osäker på en korrigering, markera felet eller lämna det oförändrat med en anteckning för vidare granskning.

Arbeta metodiskt och noggrant, med beaktande av den kontext och stil som är typisk för svensk skrift från början av 1900-talet.

Använd denna vägledning för att producera en korrigerad version av OCR-texten med hög noggrannhet och respekt för den historiska textens ursprungliga uttryck.

Returnera endast den korrigerade texten, utan några kommentarer eller ytterligare förklaringar.

USER PROMPT (short)

OCR-correct this text. Do not shorten the text:

Judith Brottrager

UNLOCKING THE ARCHIVE

Exploring Literary History through Word Embeddings

IN THE FIELD of computational literary studies, the term "literary history" is frequently used to refer to historical literary texts or, more specifically, the compilation, analysis, and comparison of literary historical corpora, which form the basis for identifying trends, recurring themes, stylistic structures, and topics. These quantitative findings are then juxtaposed with established literary historiographical narratives, sometimes corroborating them while at other times challenging them. The many other connotations of literary history have been extensively discussed elsewhere (see e.g., Crane 1971; Hartman 1970; Jauss and Benzinger 1970; Weimann 1984; Perkins 1992; Harris 1994; Wellek 2015), but it is certainly true that even within these two contexts alone – that of data-driven computational literary studies and that of traditional literary historiography – the term is used to describe two very different practices. Firstly, they differ in the scale of data they address – traditional approaches usually concentrate on a more limited set of texts – while also offering different degrees of contextualization. Secondly, traditional literary historiography situates its interpretations within broader cultural, social, and political frameworks, whereas computational approaches emphasize formal patterns and large-scale textual dynamics over close contextual detail.

Despite their clear differences, the two approaches are, of course, interconnected: Quantitative analyses are not only compared and contrasted with historiographical narratives but also enriched with data derived from literary historiography in its broadest sense. Resources such as library catalogs, genre classifications, and publication dates are employed to add literary historical detail, facilitating a better understanding

of patterns and enabling meaningful comparisons. The potential of this, as Katherine Bode calls it, "data-rich literary history" (2018, 37–57) is vast, yet there are significant limitations to be considered. First, much of literary historical data – understood here in the broadest sense of the word – is unstructured. Literary history, or more specifically, literary historical scholarship, typically produces books and articles rather than tables and numbers. In order to incorporate this information into quantitative analyses, it needs to be structured, and if such structuring processes are carried out manually, they require a considerable amount of time and resources. The second, and arguably more pressing issue, is the uneven distribution of literary historical information. While an abundance of both structured and unstructured data is available for certain literary texts and authors, this is not the case for many others, particularly those that belong to the archive rather than the literary canon. For these lesser-known texts and authors, structured data is often lacking, and unstructured information may be limited to a few sentences buried in scholarly works on specific genres or literary periods.

The challenge for an approach that aims to "quantify" literary history is thus not only to add structure to academic discourse, but also to identify and disambiguate mentions of text titles and author names. The workflow presented here combines a fine-tuned named entity recognition (NER) model with a word embedding model that encodes each detected entity as a vector. Applied to literary historiography of the long eighteenth and nineteenth centuries, it enables (1) the consolidation of name and title variations under a common entity, linking them within the corpus and to linked open data resources such as Wikidata, and (2) the formalization and quantification of scholarly discourse by analyzing which texts and authors appear in similar contexts and have vectors closer in semantic space. In this way, literary historiographical narratives can be quantified, canonized and non-canonized figures and texts can be connected, and the source becomes accessible for both large-scale quantitative analyses and close-reading approaches to patterns of implied similarity.

The first section of this chapter provides the necessary methodological background for this approach by summarizing previous studies in various fields that use word embeddings to capture semantic relationships between entities and model large-scale patterns in textual data. The second section then turns to the data, outlining the corpus compilation strategy and discussing possible biases arising from its coverage and focus. The third section provides a detailed account of the methodology and workflow, including NER, entity disambiguation, and the use of word embeddings to construct networks. The fourth

section presents and interprets the resulting networks, illustrating how they reflect historiographical narratives and where disciplinary trends become visible. Finally, the conclusion reflects on the approach's limitations and suggests directions for future research.

PREVIOUS APPROACHES

Word embeddings provide a mathematical way of representing words by capturing the contexts in which they appear. The underlying assumption is that meaning emerges from language in use – summarized by John Rupert Firth's well-known dictum that "a word is characterized by the company it keeps" (1957). Building on this idea, embeddings translate contextual patterns into a high-dimensional vector space, where each word is assigned a position that is iteratively refined during training on a corpus. Words that tend to occur in comparable syntactic or semantic environments move closer together in this space, so that terms functioning as possible substitutes, synonyms, or analogs end up clustered together.

Based on this principle, word embeddings have become a standard method in computational literary studies. They have been employed to model literary characters (Bamman et al. 2014), identify patterns of intertextuality (Burns et al. 2021), examine gendered roles of characters (Grayson et al. 2017), and perform sentiment analysis (Jacobs 2019; Schöch 2022). By contrast, discursive applications are less common. While word embeddings have been used to map semantic structures, they have only rarely been employed to examine the underlying discursive frameworks of (literary) texts, that is, the way concepts and actors are positioned, compared, and evaluated within larger systems of meaning.

Before word embedding tool kits were widely available, David Mimno (2012) demonstrated in his examination of the possibilities of "computational historiography" how a topic modeling approach to modern classical scholarship can reveal trends in research topics. In other disciplines, word embeddings have proven to be useful for dealing with the "latent content" (Tshitoyan et al. 2019) of (academic) discourses; Melvin Wevers and Marijn Koolen (2020) show the utility of word embeddings in tracing semantic change of concepts in newspapers; Ryan Heuser (2023) analyzes semantic shifts of keywords such as culture, station, and liberty in British periodicals from 1720–1960; Nikhil Garg et al. (2018) employ word embeddings to quantify gender and ethnic stereotypes; and Vahe Tshitoyan et al. (2019) use them to explore and summarize the latent knowledge of materials science.

Applied to literary historiography, this "latent knowledge" corresponds to the network of comparisons, evaluations, and contrasts that position authors, texts, and concepts relative to each other. The workflow proposed here translates these relationships into network models, thus making the implicit connections of literary history (visually) explicit. Network models not only visualize similarities within word embeddings but also offer additional possibilities for data analysis, such as the identification of clusters. As a result, the networks become a formalization of the reference system of literary history, which can serve as a source for generating data for quantitative analysis, and as a point of connection to close reading approaches.

LITERARY HISTORIES AS DATA

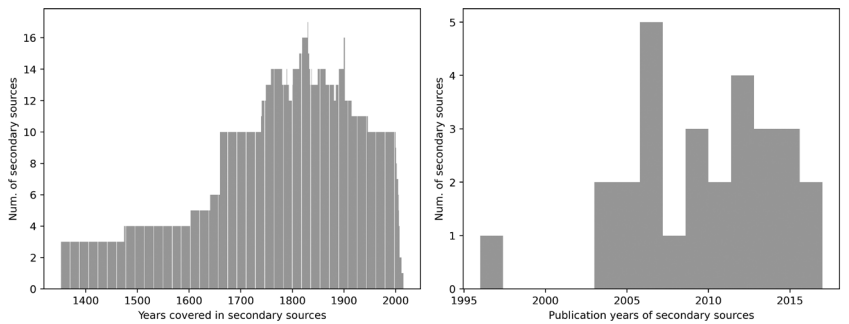
The data collected for the analysis represents a specific mode of literary historiography: literary historical anthologies. While the anthologies used differ in scope and thematic focus, they share a structural component of plurality, comprising multiple research interests and schools of thought from various academics. Because of these characteristics – diversity of scope and thematic foci, multiple authorities – they lend themselves well to replicating a corpus compilation approach suggested by Mark Algee-Hewitt and Mark McGurl (2015): By using various best-of and bestseller lists, along with lists curated by experts in feminist and postcolonial literary studies, they aim to include different levels of canonicity in a corpus representative of twentieth-century English literature. As comparable lists are not available for the time frame I am interested in – the long eighteenth and long nineteenth centuries of European English literature – anthologies can work as equivalents, reflecting different notions and levels of canonicity. Broad and consequently exclusive literary histories embody the normative concept of the canon, while companions to specific periods and genres represent the broader academic discourse. Additionally, companions focused on traditionally non-canonized texts, such as those written by women or originating from geographical peripheries like Ireland and Scotland, as well as popular fiction, serve to represent a counter-canon.[1]

The collection of texts compiled for this study is admittedly modest in size, consisting of only 27 literary histories and companions to various genres and periods. However, the motivated selection process should mitigate concerns regarding the scope of the data to some extent. An additional criterion for selection was the availability of these texts as born-digital PDFs, a practical consideration to facilitate pro-

cessing.[2] On average, each literary history consists of 148,720 tokens, tallying up to 4 million tokens on the corpus level.

In addition to the aforementioned characteristics of differing scopes and thematic foci, the anthologies have a double temporal context: Figure 1a shows the periods of time covered in the respective anthologies, as indicated by chronologies in front matters or by time frames stated in introductions. While there is a focus on the eighteenth and nineteenth centuries, broader literary histories and companions to genres naturally venture out of the actual time frame of interest. The second, and probably more influential temporal context that has to be considered is that of the anthologies' publication years: As can be seen in Figure 1b, most of the sources were published after the turn of the millennium, with only one source published earlier. The scholarly discourse that is quantified in the following consequently has a particular place in history and thus represents a snapshot of literary historiography in time.

This snapshot of literary historiography is, however, not without limitations. The corpus carries certain biases and sources of noise that shape the discourse under analysis. A large proportion of the companions and literary histories are published by Cambridge University Press, which means that the resulting discourse reflects the editorial traditions, institutional networks, and academic standards of a single publisher to a considerable degree. What may appear as plurality is thus channeled through relatively homogeneous frameworks of presentation and scholarly style. Moreover, the genre of companions and survey histories is shaped by disciplinary conventions – introductions, synoptic accounts, and thematic overviews often rely on formulaic phrasing and



(a) Distribution of coverage of secondary sources over years.

(b) Distribution of publication years of secondary sources over years.

Figure 1. Double temporal context of the literary historical anthologies compiled for the analysis.

evaluative commonplaces. Such rhetorical regularities can introduce noise into computational models, foregrounding stock expressions over substantive conceptual differences. Finally, the modest size of the corpus amplifies these issues: Each editorial decision and recurring phrase exerts disproportionate weight in shaping the semantic structures that emerge from the analysis.

METHODOLOGY AND WORKFLOW

To analyze the corpus in a systematic and reproducible way, the study combines tools from natural language processing and network analysis. As illustrated in Figure 2, the process begins with the identification and disambiguation of named entities, followed by the training of word embeddings and the construction of similarity networks. Each step required adaptation to the characteristics of the corpus, from the fine-tuning of recognition models to parameter choices in the embedding and strategies for filtering network structures. The following section outlines these procedures in detail, showing how they work together to produce a network representation of a corpus that is both computationally tractable and interpretively meaningful.[3]

The first step in this workflow is the reliable identification and disambiguation of authors' names and text titles, which requires a domain-adapted NER model. NER is a common natural language processing technique that automatically detects and classifies spans of text

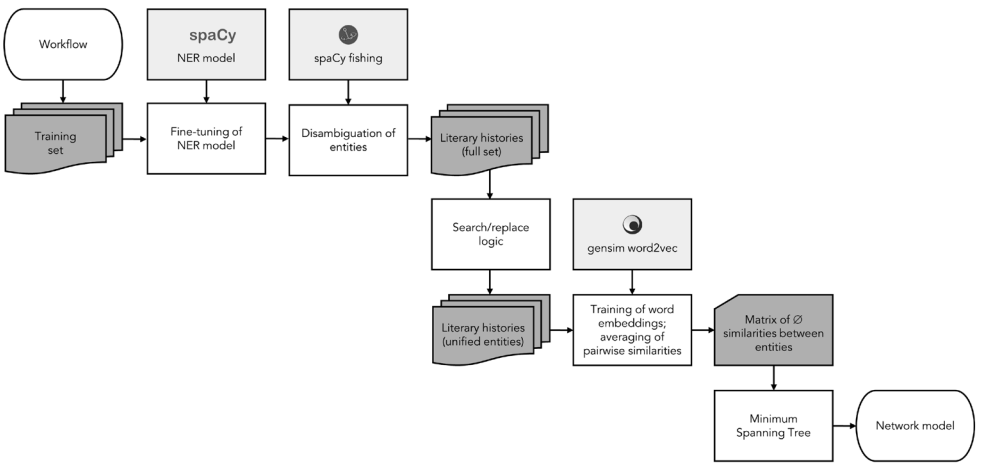


Figure 2. Workflow of NER, entity disambiguation, word embedding, and network modeling.

into categories such as PERSON, LOCATION, or WORK_OF_ART. Applied to the present corpus, however, the spaCy_en_core_web_lg model performs unevenly: As Table 1 shows, the tag class for works of art, which most closely corresponds to text titles, achieves an F1 score[4] of only 0.18 on a manually annotated test set of anthology sentences. The tag class for persons performs better, with an F1 score of 0.70, but still remains insufficiently reliable. To improve the performance of both tags, the spaCy_en_core_web_lg model was fine-tuned using a training set derived from six randomly selected anthologies that were manually annotated.

As the evaluation metrics summarized in Table 1 show, the performance for both tags – WORK_OF_ART and PERSON – could be improved significantly. This is particularly true for the WORK_OF_ART tag, with a more acceptable F1 score of 0.75 after fine-tuning. The PERSON tag also improved substantially, with its F1 score rising from 0.7 to 0.89.

The subsequent NER pipeline that was used for the data processing did not only include the fine-tuned NER tagger, but a spaCy wrapper for an entity-fishing module (Lopez 2022), which enables the linkage of identified entities to Wikidata identifiers.[5] Using these identifiers and the corresponding Wikidata item names, author names, and text titles were then identified, tagged, and, where possible, disambiguated with Wikidata IDs. In a multi-level search-and-replace procedure that also included some additional processing – such as correcting common falsely identified non-entities, adding non-identified entities, and updating incorrectly disambiguated entities – all detected entities were eventually replaced by an entity name – either the entity name registered in Wikidata or the most commonly occurring name variation – and, if applicable, the Wikidata ID. If the entity could not be matched to an existing Wikidata entry, the indicator "noWikiID" was used instead. By replacing variations of entity names with a unique identifier (such as,

	spaCy_en_core_lg		fine-tuned model	
	PERSON	WORK_OF_ART	PERSON	WORK_OF_ART
P	0.64	0.43	0.88	0.81
R	0.78	0.11	0.90	0.69
F1	0.70	0.18	0.89	0.75

Table 1. Precision, recall, and F1 score for spaCy's NER model before and after fine-tuning.

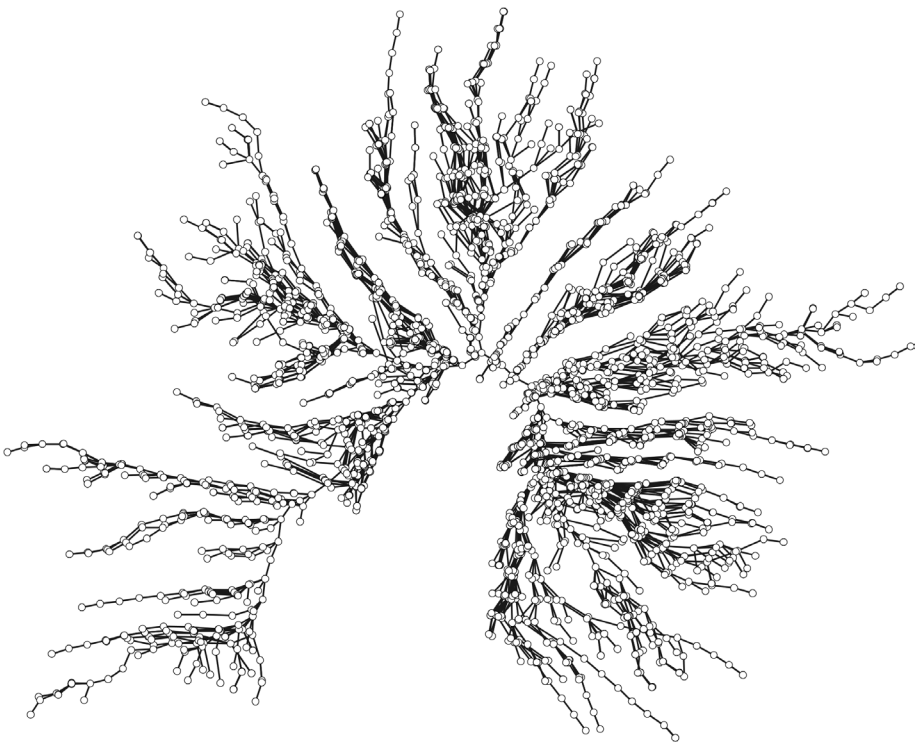
e.g., CHARLES_DICKENS_Q5686), all names and titles, regardless of the number of words they consist of, can be treated as single words in the word embedding.

To train the word embedding, I used word2vec (Mikolov et al. 2013) with a model architecture and parameters chosen to fit both the size of the corpus and the goal of quantifying literary historiographical discourse: Given the relatively small size of the corpus of just over 4 million tokens, I used a 100-dimensional skip-gram model, as skip-gram models have been shown to perform better on smaller datasets (Lai et al. 2016). While otherwise relying on the word2vec default values (as, e.g., a minimum absolute word frequency of 5, and 5 iterations), I chose to use a larger context window of 10 to better capture thematic similarity rather than analogy (see Levy and Goldberg 2014). To improve the stability of word similarities, I followed the recommendations of Maria Antoniak and David Mimno (2018), which were also adopted by Heuser (2023). Specifically, I trained five separate word2vec models on the same corpus and calculated the average similarity between entities across these models.[6] Cosine similarity – a measure of how similar two entities are in the embedding space – was used to determine which authors or texts are discussed in similar contexts.

The averaged cosine similarities between all detected entities were then used to create a similarity network that describes the relationships between all entities. Initially, this led to a network with a large amount of redundant information: Of all entities detected in the anthologies (16,176 persons and 15,464 texts), only 3,579 satisfied the condition of appearing at least five times in the entire corpus. Still, this meant that the initial similarity network consisted of 3,579 nodes and more than 6 million edges, as all n nodes were connected with every other node, resulting in $\frac{n(n-1)}{2}$ edges. To filter this hairball, I used Kruskal's algorithm (1956) to detect the network's minimum spanning tree (MST). This algorithm trims the network down to its bone, leaving only n – 1 edges. Starting off with the strongest similarity relationship, more and more edges are added, as long as they do not produce a cycle in the spanning tree. By doing so, Kruskal's algorithm creates a single connected component that summarizes the network by retaining the strongest and/or structurally most important relationships between nodes. In terms of the similarity network in question, this means that the strongest similarities between entities hold nodes together, while the unity of the reference system between the entities is also preserved.

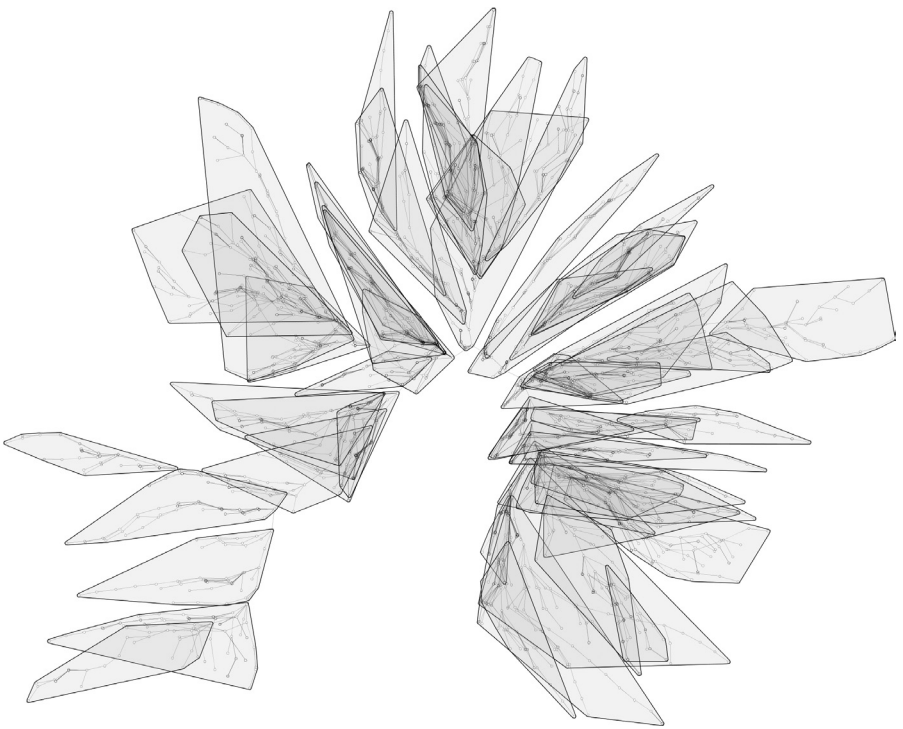
The resulting network model is shown in Figure 3: The nodes represent the detected entities, both text titles and author names, and the edges represent the remaining similarities. The color and width of an edge denotes the betweenness centrality of the edge, which indicates the number of shortest paths passing through that particular edge (Girvan and Newman 2002). Edges with a high betweenness centrality are especially important for the network's structure; they act as bridge-like links that hold the network together.[7] Due to the applied algorithm, all nodes remain connected in one component, and since MSTs avoid loops in network structures, the majority of nodes (2,711) have a degree centrality below 2, and are thus connected to only one or two other nodes.[8] Only 20 nodes have a degree-centrality of 10 or more. This leads to a root-like structure, with sub-clusters sprouting from a line of nodes strung together by edges with a high betweenness centrality.

Figure 3. Network model of entities based on their cosine similarity in a word embedding. Kruskal's algorithm was used for the extraction of the displayed MST. The edge color reflects the edges' betweenness centrality.



In order to detect these sub-clusters automatically, a so-called modularity clustering (see Brandes et al. 2008) was performed. Its main goal is to divide a network into communities or groups, where nodes within the same group are densely connected, while connections between groups are less frequent. The algorithm achieves this by optimizing the modularity measure, which quantifies the quality of a partition by comparing the number of edges within communities to the expected number of edges if the network were randomly connected, given the node degree distribution. A high modularity value – as in this case a modularity value of 0.965 – indicates that there are non-random connectivity patterns in the network. In total, the 3,579 nodes of the network are divided into 63 modularity clusters, with an average of 57 nodes per cluster (standard deviation = 16.7). Figure 4 shows the network model with all modularity clusters. The clustering identifies dense subgroups in the leaf-like structures branching from the central root of the network formed by high-betweenness edges.

Figure 4. Network model of entities based on their cosine similarity in a word embedding. Kruskal's algorithm was used for the extraction of the displayed MST. All modularity clusters are highlighted in grey.



The nodes in the modularity clusters shown in Figure 4 are, as mentioned above, densely connected, which implies that the entities they represent behave semantically similarly within the corpus of literary-historical anthologies on which the word embeddings are based. Because the edges between them belong to the MST, it can be assumed that the semantic similarity between them, measured by their mean cosine similarity in five different word embeddings, is particularly high. This means, in other words, that these entities either share contexts or appear in similar ones. From a literary-historiographical perspective, they can be assumed to be described in similar ways or to appear in similar narratives across the anthologies.

In addition, the position of nodes in the network has semantic meaning: like edges, nodes can be described by their betweenness centrality, which indicates the number of all shortest paths from all nodes to all other nodes that pass through a given node.[9] The nodes with the highest betweenness centrality are naturally located along the string of high-betweenness edges clearly visible in Figure 3, while nodes with intermediate values sit on the main branches of clusters. More than a third of all nodes (1,788) have a betweenness centrality of 0, forming the endpoints of a branch. As the nodes and edges represent semantic – or perhaps in this case more accurately, contextual – similarity, the further a node is from the center of the network, the more clearly its context is defined. Vice versa, this means that more central nodes represent entities that either appear in many different contexts, thus blurring contextual similarities, or in very few contexts, making it difficult to define their position in the word embeddings. The latter seems to be true for many of the entities represented by nodes with betweenness centrality scores above 1 million: Jon Silkin (1930–1997), Clare Boylan (1948–2006), and Helen Fielding (*1958), whose nodes all have extremely high scores of over 6 million, published well after 1914, which as the last year of the long nineteenth century works as a cut-off point for many anthologies in the corpus. Gerald of Wales (betweenness centrality of 5,213,864), a twelfth-century Cambro-Norman priest, can be expected to appear in some of the broader literary histories that expand the time frame covered with the sources (see Fig. 1a), but is unlikely to be mentioned in anthologies centered on the eighteenth and nineteenth centuries. Non-canonized texts, such as Jane Porter's novel *The Scottish Chiefs* (1810), appear too infrequently to establish a clear embedding position.

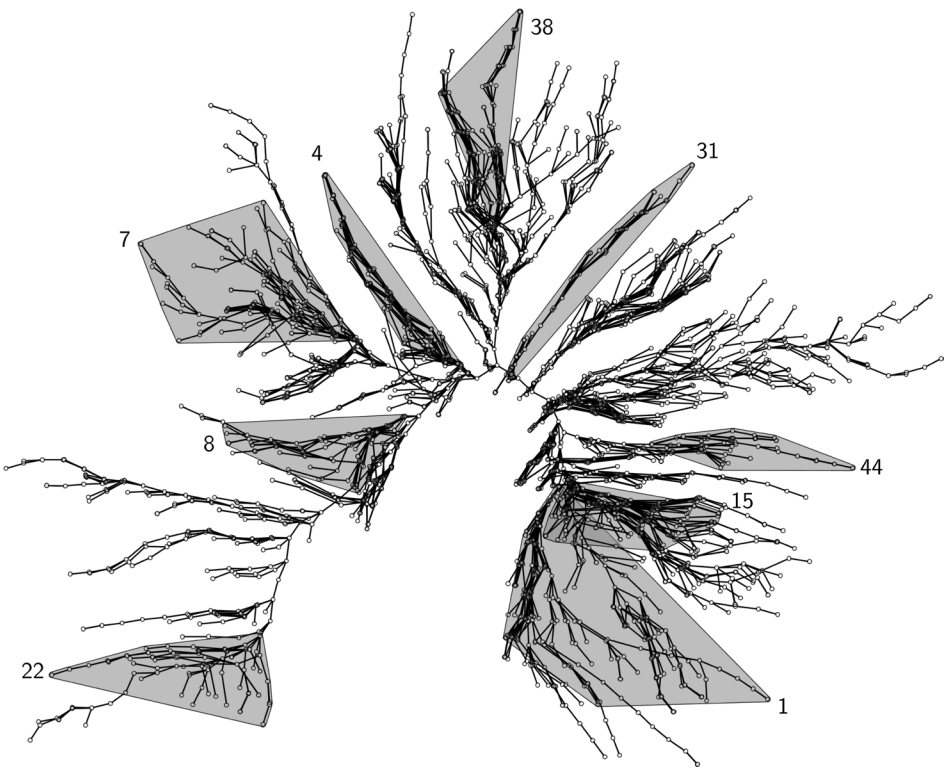
At the peripheries of the network, however, there are clusters of nodes that are contextually related and interpretable (see Fig. 5 for the numbered examples): a strikingly clear example of this is Cluster 39, which contains 57 nodes representing mainly historical figures, in most cases English, British, or Scottish royals. In addition to names such as William the Conqueror, Queen Victoria, James I of Scotland, and Queen Anne, French royalty appears in the form of Louis XVI and Napoleon. The revolutionary spirit in this royal cluster is reinforced by the nodes of Maximilien Robespierre and Oliver Cromwell. The only two nodes representing texts are *The Mask of Anarchy*, Percy B. Shelley's political and monarchy-critical poem of 1819, and *Astraea Redux*, John Dryden's royalist poem of 1660. In contrast to this clear focus on monarchical power, the 45 nodes of Cluster 4 also include kings, non-reigning royals such as Prince Albert and Sophia of Hanover, and political figures such as Robert Walpole, Charles Fox, William Pitt the Younger, Thomas Jefferson, and Margaret Thatcher. These patterns illustrate how the network reflects the ways in which historical figures and texts are linked in literary historiography, highlighting both thematic and temporal associations.

In another pair of clusters, Clusters 7 and 8, nodes appear that can be linked to the general field of critical theory. While there is no strong thematic difference between the clusters, there seems to be at least a temporal component that separates the clusters: Cluster 7 features several figureheads of twentieth-century critical theory (Theodor W. Adorno, Mikhail Bakhtin, Walter Benjamin, Jean Baudrillard, Jean Derrida, Michel Foucault, Julia Kristeva) and psychoanalysis (Sigmund Freud, Carl Jung). While earlier thinkers (Voltaire, Johann Gottfried Herder, Arthur Schopenhauer, Adam Smith) also appear, there is a strong focus on the thought, scholarship and science of the twentieth century: historian Benedict Anderson, literary/cultural scholars Sally Ledger, Anne K. Mellor, Raymond Williams, and Terry Eagleton appear alongside figures from the sciences, such as Albert Einstein and Michael Faraday; *The Empire Writes Back* (1989) appears with Edward Said. Cluster 8, on the other hand, is more closely associated with the seventeenth and eighteenth centuries, featuring nodes for Immanuel Kant, Georg Wilhelm Friedrich Hegel, Friedrich Schlegel, David Hume, George Berkeley, Gottfried Leibniz, Isaac Newton, Baruch de Spinoza, as well as for texts such as *A Letter Concerning Toleration* (1689), *A Treatise of Human Nature* (1739), *A Philosophical Enquiry into the Origin of Our Ideas of the Sublime and Beautiful* (1757), *The Theory of Moral Sentiments* (1759), and *Critique of Judgment* (1790). The outliers of Cluster 8 are related to antiquity (Plato and Aristotle), religion (Martin

Luther and John Calvin), or evolutionary theory (Charles Darwin and his works *On the Origin of Species* (1859), and *The Descent of Man, and Selection in Relation to Sex* (1871)). These associations suggest that the network not only encodes thematic affinities but also preserves temporal structures, showing how seventeenth- to twentieth-century scholarship is positioned in relation to each other in the anthologies.

Beyond the critical theory clusters, other clusters have an obvious common denominator: Cluster 22 consists mainly of the titles of periodicals and agents in the nineteenth-century book market, including figures such as Charles Edward Mudie, William Henry Smith, John Murray, and Alexander Donaldson. Others are less clear and hint at a literary-historiographical narrative that binds them together. This is the case, for example, with Cluster 44, which consists mainly of authors from the late twentieth and early twenty-first centuries, including Iris Murdoch, Kazuo Ishiguro, John Fowles, A. S. Byatt, Salman Rushdie,

Figure 5. Network model of entities based on their cosine similarity in a word embedding. Kruskal's algorithm was used for the extraction of the displayed MST. Selected modularity clusters are highlighted in grey.



and Julian Barnes. There are two main departures from this trend: First, Fowles' postmodern historical novel *The French Lieutenant's Woman* (1969) is linked to John Galt's novel *The Entail* (1823); the novels share their historical setting. Galt's novel is then linked to his *Annals of the Parish* (1821), and other nodes of texts and authors associated with realistic depictions of late eighteenth- and early nineteenth-century rural life (Mary Russell Mitford, *Our Village* (1824); George Crabbe, *The Borough* (1810); John Galt). Through Fowles' connection to Graham Swift and a group of women writers of the late twentieth- and early twenty-first-century (Jeanette Winterson, Angela Carter, Doris Lessing), the modernist writers Dorothy Richardson, May Sinclair, and Virginia Woolf form the cluster's endpoint. While I assume that the strong similarity between *The French Lieutenant's Woman* and *The Entail* is not necessarily based on a literary-historical narrative that places the novels in relation to each other but on their similar settings, it is more likely that Winterson, Carter, and Lessing are compared to or associated with their predecessors Richardson, Sinclair, and Woolf. The similarity encoded in the entity networks thus highlights how texts and authors are positioned relative to each other across both temporal and thematic dimensions, indicating patterns that go beyond simple chronological or genre-based groupings.

A similar pattern of association can be seen in Cluster 31, the vast majority of which consists of nodes related to female authors. Beginning with the innermost node in the cluster, which belongs to Barbara Hofland, an English writer primarily known for her children's stories, the cluster includes Gothic and Sensationalist novelists (Eliza Parson, Amelia Opie, Regina Maria Roche, Florence Marryat, Rhoda Broughton), but also connects these non-canonized authors to canonized novels: Via a link between Marryat and Sarah Fielding's *The Governess; or, The Little Female Academy* (1749) and *David Simple* (1744), a sub-group of almost all Jane Austen novels (*Sense and Sensibility*, *Northanger Abbey*, *Pride and Prejudice*, *Mansfield Park*, and *Emma*) joins the cluster. Cluster 1 is equally dominated by female, mostly non-canonized authors. A group of canonized authors – Samuel Richardson, Henry Fielding, and Austen (and the not-as-canonized Sarah Fielding) – again form the endpoint of the cluster. They are attached to the cluster via the nodes of Charlotte Lennox, Frances Sheridan, and Elizabeth Griffith. Taken together, these clusters show a consistent pattern: both clusters coalesce around canonized texts and eighteenth-century authors, yet they differ in temporal emphasis. Cluster 32 predominantly features nodes associated with the long nineteenth century, whereas Cluster 1 is more closely linked to the long eighteenth century. Within them, the

clusters encoded similarities highlight broader discursive tendencies in literary historiography, showing how non-canonized female authors are connected to established literary traditions. Finally, a different form of association can be seen in Cluster 15, which contains nodes of Victorian authors and texts, but also of cultural theorists and their works. More specifically, postcolonial scholars such as Gayatri Spivak, Homi Bhabha, and Frantz Fanon, as well as Edward Said's *Orientalism* (1978) are paired with adventure and early horror novels and narratives and their authors: Bram Stoker and his *Dracula* (1897), Robert Louis Stevenson and his *Treasure Island* (1883) and *Strange Case of Dr Jekyll and Mr Hyde* (1886), H. Rider Haggard and his *King Solomon's Mines* (1885) and *She: A History of Adventure* (1886), and Joseph Conrad's *The Nigger of the "Narcissus"* (1897). What these texts have in common is a more or less explicit preoccupation with the strange, the uncanny, the "exotic" and, in one way or another, the Other. Especially in the snapshot of literary historiography represented by the data source, a postcolonial reading of this focus is unsurprising. It reflects how twentieth- and twenty-first-century scholarship has increasingly engaged with questions of cultural difference, empire, and encounters with the unfamiliar, producing literary-historiographical narratives that link Victorian texts with contemporary critical perspectives.

As these exemplary clusters show, the combination of NER processing, word embeddings, and network models helps to represent similarity relationships between name and title entities at an abstract yet interpretable level. Some clusters that appear in the network are related to specific contexts: Ruling royals appear in different contexts than cultural theorists, titles of periodicals in different contexts than those of philosophical treatises. The fact that these clusters emerge is a validity check for the combined methods and shows that the workflow can indeed capture the contextual similarity between detected entities. While the division of critical theory nodes into two clusters already suggests that temporal factors influence contextual similarities, other clusters indicate that there are also more abstract associations present in the data: similarities between author nodes may reflect implied role-model relationships (Cluster 44) or can act as anchors for non-canonized authors that are linked to their more canonized contemporaries (Cluster 1). A female-dominated cluster of lesser-known authors is interlinked based on their shared association with Austen's oeuvre (Cluster 31), and a group of genre novels and narratives are contextualized by nodes associated with postcolonial theory (Cluster 15). These more abstract links reflect the ways in which authors and texts are negotiated in the sources, approximating how entities are described or appear in similar

literary-historiographical narratives. By formalizing and quantifying these relationships, the network models provide a structured lens on literary historiography, showing how canonized and non-canonized texts are positioned, compared, and interrelated in scholarly discourse, thus offering a meta-perspective on the construction of literary traditions – or, in other words, a meta-reading of literary history.

CONCLUSIONS

Literary historiography is a rich but difficult source of information for quantitative analysis, with the additional necessary steps of formalization and quantification inevitably leading to a loss of detail (see Weitin 2017). Word embeddings are one possible way to introduce structure into unstructured data and, as shown above, produce encouraging results even with a relatively small dataset. The quantification of literary history, or more specifically, the reference system of literary historiography that relates entities to each other based on cosine similarities is generally possible, and produces several clearly defined clusters that indicate that the results are, at least to some extent, valid. Other clusters whose nodes do not share a clearly discernible characteristic (such as historical figures or critical theory) suggest an association based on a shared context, be it a shared semantic field or a shared literary-historiographical narrative.

Although the results are promising, as with most datasets in computational literary studies, more data would be preferable. More importantly, the quality of the data, particularly the post-processing, is another potential area for improvement: While the fine-tuned NER model achieves acceptable evaluation metrics, more manually verified training data could further improve performance. In addition, the entity-fishing model is less reliable in disambiguating lesser-known entities, which goes against the premise of creating additional data for the archive of literary history. The multilevel search-and-replace logic used to counter false disambiguations was helpful, but is not a sustainable pipeline for larger datasets. Large language models (LLMs) could be employed for better disambiguation results, but their use would introduce new considerations: LLMs might improve entity recognition and context-based disambiguation, yet they also carry risks of encoding biases present in their training data (see Chang et al. 2023) and could inadvertently amplify certain narratives over others. In addition to these potential adaptations of NER and disambiguation, the reliability of the network structures could be further strengthened by using a higher number of embeddings for bootstrapping. While the current approach

produces relatively stable results, small variations in similarity can generate different MST structures and consequently alter modularity clusters, affecting how relationships between entities are represented.

Despite these limitations, the network model has proven to be a useful tool for making the similarities generated with word embeddings visually explicit. Serving as more than just a visual aid, the different network configurations facilitate a meta-reading of literary-historical discourse: Similarities between individual entities, as well as groups of entities in clusters, can be extracted and integrated into existing datasets. This integration allows a comparison between perceived similarities within the literary historical discourse and textual similarities derived from textual analysis, highlighting both the overlaps and divergences between these two modes of literary history.

NOTES

1 For a list of the anthologies used, see Appendix 1.

2 All texts are copyright-protected and can therefore not be shared as full texts.

3 Code and data are available at https://github.com/jbrottrager/unlocking-the-archive.

4 The F1 score is a standard measure of model accuracy that combines precision (the proportion of retrieved items that are relevant) and recall (the proportion of relevant items that are retrieved) into a single harmonic mean.

5 Entity disambiguation means matching names in the text to the correct entries in a database. The entity-fishing module does this by looking at the context around each detected name and picking the Wikidata entry that fits best, so different mentions of the same entity (like "Charles Dickens" and "C. Dickens") are treated as the same person.

6 A Mantel test between the similarity matrices generated with the word embedding and a control embedding with the same parameters and data shows that the similarities between entities are strongly correlated, with a veridical correlation of 0.96 and a p-value of 0.00001.

7 For example, in a network of cities connected by roads, an edge with high betweenness centrality might be a major bridge or highway that most shortest routes between cities pass through. If it were removed, many cities would become harder to reach from one another.

8 Degree centrality measures the number of direct connections a node has to other nodes; a low value means the node is connected to only a few others, while a high value indicates many connections.

9 For example, in a company, an office coordinator who facilitates communication between different departments would have high betweenness centrality, as many shortest paths connecting employees in different teams pass through them.

REFERENCES

Algee-Hewitt, Mark, and Mark McGurl. 2015. "Between Canon and Corpus: Six Perspectives on 20th-century Novels." *Pamphlets of the Stanford Literary Lab* (8): 1–27.

Antoniak, Maria, and David Mimno. 2018. "Evaluating the Stability of Embedding-Based Word Similarities." In *Transactions of the Association for Computational Linguistics*, 6, edited by Lillian Lee, Mark Johnson, Kristina Toutanova, and Brian Roark. MIT Press: 107–19.

Bamman, David, Ted Underwood, and Noah Smith. 2014. "A Bayesian Mixed Effects Model of Literary Character." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (1: Long Papers): 370–9.

Bode, Katherine. 2018. *A World of Fiction: Digital Collections and the Future of Literary History*. University of Michigan Press.

Brandes, Ulrik, Daniel Delling, Marco Gaertler, Robert Gorke, Martin Hoefer, Zoran Nikoloski, and Dorothea Wagner. 2008. "On Modularity Clustering." *IEEE Transactions on Knowledge and Data Engineering* 20 (2): 172–88.

Burns, Patrick J., James A. Brofos, Kyle Li, Pramit Chaudhuri, and Joseph P. Dexter. 2021. "Profiling of Intertextuality in Latin Literature Using Word Embeddings." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, edited by Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell et al. Association for Computational Linguistics: 4900–07.

Chang, Kent, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. "Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4." *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, edited by Houda Bouamor, Juan Pino, and Kalika Bali. Association for Computational Linguistics: 7312–27.

Crane, R. S. 1971. *Critical and Historical Principles of Literary History*. University of Chicago Press.

Firth, John Rupert. 1957. *Papers in Linguistics, 1934–1951*. Oxford University Press.

Garg, Nikhil, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. "Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes." In *Proceedings of the National Academy of Sciences* 115 (16).

Girvan, Michelle, and Mark E. J. Newman. 2002. "Community Structure in Social and Biological Networks." *Proceedings of the National Academy of Sciences* 99 (12): 7821–6.

Grayson, Siobhán, Maria Mulvany, Karen Wade, Gerardine Meaney, and Derek Greene. 2017. "Exploring the Role of Gender in 19th Century Fiction Through the Lens of Word Embeddings." In *Language, Data, and Knowledge*, edited by Jorge Gracia, Frances Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and Sebastian Hellmann. Springer International Publishing: 358–64.

Harris, Vendell. 1994. "What is Literary History?" *College English* 56 (4): 434.
Hartman, Geoffrey. 1970. "Toward Literary History." *Daedalus* 99 (2): 355–83.
Heuser, Ryan. 2023. "Computing Koselleck: Modelling Semantic Revolutions, 1720–1960." In *Explorations in the Digital History of Ideas*, 1st ed, edited by Peter De Bolla. Cambridge University Press: 256–85.
Jacobs, Arhur M. 2019. "Sentiment Analysis for Words and Fiction Characters From the Perspective of Computational (Neuro-)Poetics." *Frontiers in Robotics and AI* 53 (6): 1–13.
Jauss, Hans Robert, and Elizabeth Benzinger. 1970. "Literary History as a Challenge to Literary Theory." *New Literary History* 2 (1): 7–37.
Kruskal, Joseph B. 1956. "On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem." *Proceedings of the American Mathematical Society* 7 (1): 48–50.
Lai, Siwei, Kang Liu, Shizhu He, and Jun Zhao. 2016. "How to Generate a Good Word Embedding." *IEEE Intelligent Systems* 31 (6): 5–14.
Levy, Omer, and Yoav Goldberg. 2014. "Dependency-Based Word Embeddings." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, edited by Kristina Toutanova and Hua Wu. Association for Computational Linguistics (Volume 2: Short Papers): 302–8.
Lopez, Patrice. 2022. *SpaCy Fishing*. https://github.com/Lucaterre/spacyfishing.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeoffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." *arXiv*:1301.3781.
Mimno, David. 2012. "Computational Historiography: Data Mining in a Century of Classics Journals." *Journal on Computing and Cultural Heritage* 5 (1): 1–19.
Perkins, David. 1992. *Is Literary History Possible?* Johns Hopkins University Press.
Schöch, Christof. 2022. "Quantitative Semantik: Word Embedding Models für Literaturwissenschaftliche Fragestellungen. In *Digitale Literaturwissenschaft*, edited by Fotis Jannidis. J. B. Metzler: 535–62.
Tshitoyan, Vahe, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson et al. 2019. "Unsupervised Word Embeddings Capture Latent Knowledge from Materials Science Literature." *Nature* 571 (7763): 95–8.
Weimann, Robert. 1984. *Structure and Society in Literary History: Studies in the History and Theory of Historical Criticism*. Johns Hopkins University Press.
Weitin, Thomas. 2017. "Scalable Reading." *Zeitschrift für Literaturwissenschaft und Linguistik* 47 (1): 1–6.
Wellek, René. 2015. "The Fall of Literary History." In *New Perspectives in German Literary Criticism: A Collection of Essays*, edited by Richard E. Amacher and Victor Lange. Princeton University Press: 418–31.
Wevers, Melvin, and Marijn Koolen. 2020. "Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings." *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53 (4): 226–43.

APPENDIX 1

Carruthers, Gerard, and Liam McIlvanney, eds. 2012. *The Cambridge Companion to Scottish Literature*. Cambridge University Press.

Caserio, Robert L., ed. 2009. *The Cambridge Companion to the Twentieth-Century English Novel*. Cambridge University Press.

Chandler, James, ed. 2009. *The Cambridge History of English Romantic Literature*. Cambridge University Press.

Curran, Stuart. ed. 2010. *The Cambridge Companion to British Romanticism*. Cambridge University Press.

David, Deirdre, ed. 2012. *The Cambridge Companion to the Victorian Novel*. Cambridge University Press.

DeMaria, Robert, Heesok Chang, and Samantha Zacher, eds. 2014. *A Companion to British Literature. Part III: Long Eighteenth-Century Literature 1660–1837*. John Wiley & Sons.

DeMaria, Robert, Heesok Chang, and Samantha Zacher, eds. 2014. *A Companion to British Literature. Part IV: Victorian and Twentieth-Century Literature 1837–2000*. John Wiley & Sons.

Einhaus, Anne-Marie, ed. 2016. *The Cambridge Companion to the English Short Story*. Cambridge University Press.

Flint, Kate, ed. 2012. *The Cambridge History of Victorian Literature*. Cambridge University Press.

Foster, John Wilson, ed. 2006. *The Cambridge Companion to the Irish Novel*. Cambridge University Press.

Glover, David, and Scott McCracken, eds. 2012. *The Cambridge Companion to Popular Fiction*. Cambridge University Press.

Grenby, M. O., and Andrea Immel, eds. 2009. *The Cambridge Companion to Children's Literature*. Cambridge University Press.

Ingrassia, Catherine, ed. 2015. *The Cambridge Companion to Women's Writing in Britain, 1660–1789*. Cambridge University Press.

Kelleher, Margaret, and Philip O'Leary, eds. 2006. *The Cambridge History of Irish Literature. 1*. Cambridge University Press.

Kelleher, Margaret, and Philip O'Leary, eds. 2006. *The Cambridge History of Irish Literature. 2*. Cambridge University Press.

Keymer, Thomas, and Jon Mee, eds. 2004. *The Cambridge Companion to English Literature, 1740–1830*. Cambridge University Press.

Looser, Devoney, ed. 2015. *The Cambridge Companion to Women's Writing in the Romantic Period*. Cambridge University Press.

Mangham, Andrew, ed. 2013. *The Cambridge Companion to Sensation Fiction*. Cambridge University Press.

Marcus, Laura, and Peter Nicholls, eds. 2004. *The Cambridge History of Twentieth-Century English Literature*. Cambridge University Press.

Marshall, Gail, ed. 2007. *The Cambridge Companion to the Fin de Siècle*. Cambridge University Press.

Maxwell, Richard, and Katie Trumpener, eds. 2008. *The Cambridge Companion to Fiction in the Romantic Period*. Cambridge University Press.

Peterson, Linda H., ed. 2015. *The Cambridge Companion to Victorian Women's Writing*. Cambridge University Press.
Poplawski, Paul. 2017. *English Literature in Context*. Cambridge University Press.
Richetti, John J., ed. 1996. *The Cambridge Companion to the Eighteenth-Century Novel*. Cambridge University Press.
Richetti, John J., ed. 2005. *The Cambridge History of English Literature, 1660–1780*. Cambridge University Press.
Shattock, Joanne, ed. 2010. *The Cambridge Companion to English Literature, 1830–1914*. Cambridge University Press.
Shiach, Morag, ed. 2007. *The Cambridge Companion to the Modernist Novel*. Cambridge University Press.

Chris Haffenden, Faton Rekathati & Emma Rende

SEARCHING WITHOUT LABELS

Multimodal AI and Image Search at the National Library of Sweden

THE ARCHIVES OF memory institutions have expanded exponentially as a result of contemporary processes of "datafication" (Cukier and Mayer-Schoenberger 2013). With the advent of mass digitization programs and the ever-increasing flow of born-digital material, the range and extent of digital cultural heritage collections have grown enormously over the past two decades (Rahaman 2018; Ames and Lewis 2020). Yet a significant amount of this material remains inaccessible. Factors such as copyright restrictions, data protection regulations, outdated technical systems, and the sheer scale of collections all contribute to this paradox. Visual heritage, such as postcards, can often be digitized without sufficient metadata, leaving it practically invisible to users. Regardless of the possibilities promised by the brave new world of big data, significant infrastructural challenges prevent this potential from being realized (Bingham and Byrne 2021). Digitally collected and stored does not simply equate to digitally available.

How can new AI methods be harnessed to improve the accessibility of heritage collections? In this chapter, we examine how this problem might be addressed through a case study at the National Library of Sweden (*Kungliga biblioteket*, hereafter KB). Building on our work at KBLab – the library's data lab and infrastructure for digital research (Börjeson et al. 2024) – we explored how recent developments in multimodal AI could be applied to improve access to visual collections lacking descriptive labels. Specifically, we developed a prototype image search demo based on CLIP (Contrastive Language–Image Pre-training), a contrastive vision-language model that enables content-based search without relying on pre-existing metadata – a process

that relied on close collaboration between technical staff and domain experts at the library. Our focus was on the visual contents of KB's extensive holdings of historical postcards, which, despite their cultural richness, remain difficult to search or discover through traditional catalog systems.

We begin the chapter by outlining the accessibility challenges that characterize this material, before explaining how CLIP's ability to collapse the boundary between text and image enables new ways of interacting with undescribed collections. This also includes an outline of the AI-based search demo developed at KBLab.[1] We then explore the wider research potential of CLIP-based image search for scholars working with digitized visual culture, and reflect on the technical and conceptual difficulties of implementing such techniques at scale within a heritage institution. In conclusion, we reflect on how such AI-based search systems may influence not only user access but also broader assumptions about what makes digital heritage discoverable in the first place (Haffenden et al. 2023, 44–45).

FORGOTTEN IN THE ARCHIVE

Memory institutions such as KB can often experience a palpable sense of overload that, while far from historically unique, is exacerbated by the characteristics of today's media landscape (Blair 2010). This is partly an effect of the proliferation of material now being produced by digital culture. With the volume of data now streaming into the library via electronic legal deposit rapidly increasing, how might any sort of order come to be maintained? Yet it is also a matter of responding to inherited archival blind spots, where previous moments of mass media expansion have created the conditions for various collections to remain undescribed and therefore practically unknowable. A striking instance is the large amount of commercial "ephemera" enabled by technical innovations in the print industry in the nineteenth century, which significantly reduced the cost of publishing such material (The Multigraph Collective 2018).[2] Given that this material tends to lack the metadata necessary to make it discoverable, these items have largely been consigned to the realm of forgetting described by Aleida Assmann as "the passively stored memory that preserves the past." For every prestigious object placed in the foreground, she pointedly reminds us, there are many "storerooms stuffed with other paintings and objects in peripheral spaces such as cellars and attics which are not publicly presented" (Assmann 2010, 98).

A pertinent example of such material among KB's collections is visual heritage from the nineteenth and twentieth centuries. Just like these storerooms of neglected miscellanea, this is material that has been preserved at the library, but whose lack of description severely limits the opportunity for search and discovery. The absence of metadata is chiefly a result of its perceived ephemerality historically, and the limited archival resources and attention it has thus been afforded. To evidence this earlier sensibility, we can recall Robert Shackleton's observation from 1971 that it was "thought below the dignity of a learned library to acquire or to retain the casual and sub-literary products of the printing press" (cited in Garner 2021, 244). While KB has long since focused upon receiving and retaining precisely such "casual" material, accumulating not least one of the world's largest poster collections, it has proved practically impossible for every incoming item to be manually cataloged by a librarian.[3] Indeed, given that legal deposit legislation dictates that the library receive a copy of *everything* published in Sweden, descriptive cataloguing has necessarily been selectively applied (Konstenius 2017). This means that certain material categories such as adverts and posters have tended to be grouped together and classified under collective catalog entries, which often precludes the detailing of any specific information about each object per se.

This combination of an abundance of material and a scarcity of description characterizes the library's holdings of historical postcards. Although KB has a rich and diverse collection of over 600,000 postcards, the lack of navigability entrenched by scarce metadata has made it challenging for library users to access this material – or even to be aware of its existence.[4] Because of both the scale of the collection and the downplaying of its value historically, the postcards have, for the most part, not been cataloged at the level of the individual object. Instead, they have been organized in a series of broad subsections – for example, "Swedish topographical postcards," "general," and "portrait" – in a way that makes them very difficult for users to search on their own terms. Such a lack of description makes overview and navigation of the material practically impossible.

Access to the collection is currently dependent on the help, expertise, and effort of a specialist librarian. Users therefore need to go to the catalog post, work out which of the 16 overarching categories they might be interested in, and then submit a request to the library's staff with details about what among the hundreds of thousands of postcards they are hoping to find. Putting the accumulated knowledge of the librarian to one side for a moment, this task is principally akin to

looking for a needle in a haystack. In practice, it means that the majority of the postcards are passively stored but forgotten in the archive. Despite having been preserved as part of our shared cultural heritage, these items have thus largely disappeared from our collective view. Without the labels necessary to find them with any degree of efficacy, they remain dormant and obscure.

Figure 1. Hoards of historical postcards in the library's archives. Photo: KB.

MULTIMODAL AI

Recent developments within AI offer a means for libraries and other GLAM (galleries, libraries, archives and museums) institutions with large collections of undescribed visual heritage to counter this archival forgetting and bring such material back into sight.[5] The emergence of powerful multimodal AI models has made it possible for us to search and interact with huge amounts of images in new ways, regardless of preexisting metadata or lack thereof. This is evident, for example, in the computer vision technology underpinning Google's image search function, which we now take entirely for granted on the smartphones in our pockets. The creative potential of such techniques is also apparent from how quickly digital scholars have begun to forecast the coming of a "multimodal turn in Digital Humanities" (Smits and Wevers 2023).

By connecting the visual and textual domains in an innovative manner, these multimodal AI models have enabled far more precise and effective online search techniques. This includes both text-to-image searches (image retrieval from a textual description) and image-to-image searches (otherwise known as reverse image search, that is, image retrieval from an image) that we use on a daily basis without batting an eyelid.

Perhaps the most prominent among these recent models has been OpenAI's CLIP (Radford et al. 2021). Through being exposed to a dataset of 400 million matching text–image pairs in its pre-training,

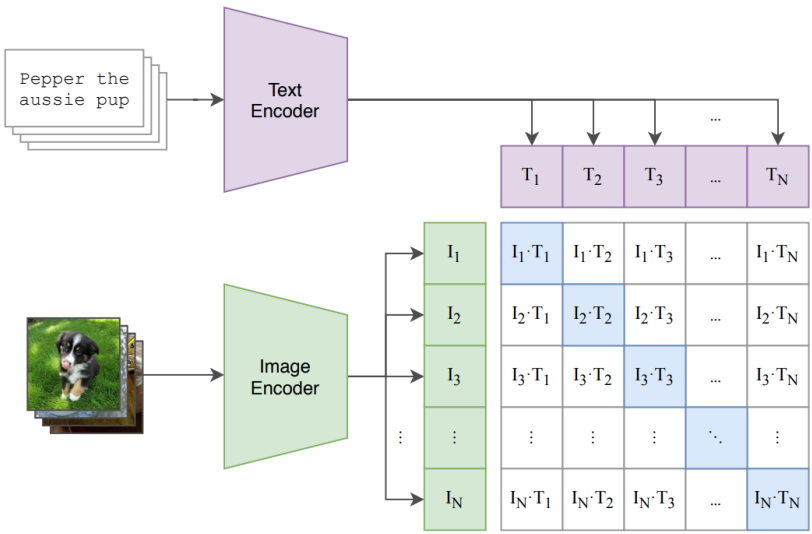


Figure 2. Matching text and image inputs in CLIP's pre-training. Image: Radford et al. (2021).

CLIP has learned to recognize visual concepts and their associated names – for instance, an image of a cat and the text description "a photo of a cat." This marks a significant departure from earlier computer vision models such as AlexNet, which were trained on curated datasets like ImageNet that relied on manually labeled categories (Krizhevsky et al. 2012). Unlike these supervised approaches, CLIP learns visual concepts through contrastive learning on such image-text pairs scraped from the web, allowing it to recognize a far broader and more flexible range of visual semantics.

In broad terms, this works via the encoding of texts and images into a shared vector space to negate the difference in media form (see Fig. 2). By transforming the text or image input into a numerical representation – known within machine learning as "embeddings" (Jurafsky and Martin 2024) – the AI model can return results based upon images with a similar representation. When "church" appears in a text, for instance, the model will generate a similar number representation for both the word description and the visual manifestation of a church, thus collapsing the distinction between the different modes and making them directly comparable. Such a technique can be used to identify all the images containing a church in a large collection of images, regardless of whether these images have previously been described as such. In short, CLIP provides the basis for content-based image search that is independent of prior metadata, which is precisely what memory institutions with collections of unlabeled images need to make such material more searchable for their users.

APPLYING CLIP AT THE LIBRARY

How might CLIP be used to enhance access to visual heritage material? We explored this question at KBLab through a pilot project focused upon a selection of the library's postcard holdings. This rested upon three significant preconditions. First, and most obviously, it presumed the presence of digital data (a point we will return to in our broader discussion below). Since a small selection of the collection has previously been digitized, we turned to these 17,409 postcards as a dataset for the project. Given that the back side of each postcard could include personal details such as names and addresses, we opted to employ only the front sides. The second and more technically demanding issue concerned adapting the model for Swedish data. This built upon our efforts within a larger collaborative project to produce a Swedish version of CLIP, Swe-CLIP 2M (Carlsson et al. 2022). While the original version of CLIP was trained upon English text descriptions, this involved

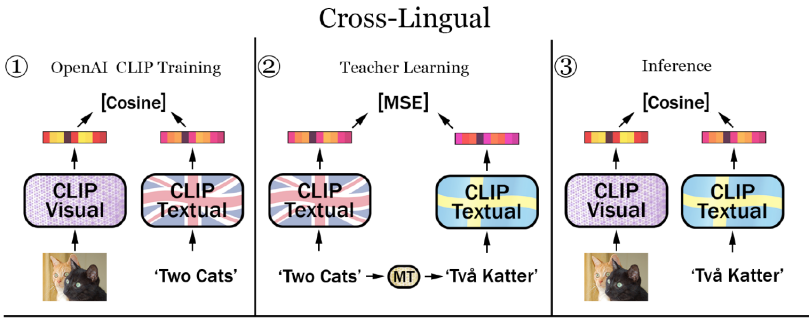


Figure 3. Adapting CLIP for Swedish textual inputs. Image: Carlsson et al. (2022).

adding a cross-lingual training component using machine translation to teach the model to equate Swedish textual inputs – for example, "två katter" (two cats) – with the matching image for the English description (see Fig. 3). Finally, our pilot project depended on the availability of requisite development resources at the library: moving from a proof of concept to a user-facing application would not have been possible without the specific expertise of software developers and a UX designer within the project.[6]

Deploying this Swedish-trained version of CLIP, and following pertinent design input from the developers, we produced an image search demo that showcases the functionalities of AI-based search. Thanks to a licensing agreement with the Swedish copyright collective for images, we could release this freely online (see Fig. 4). The demo

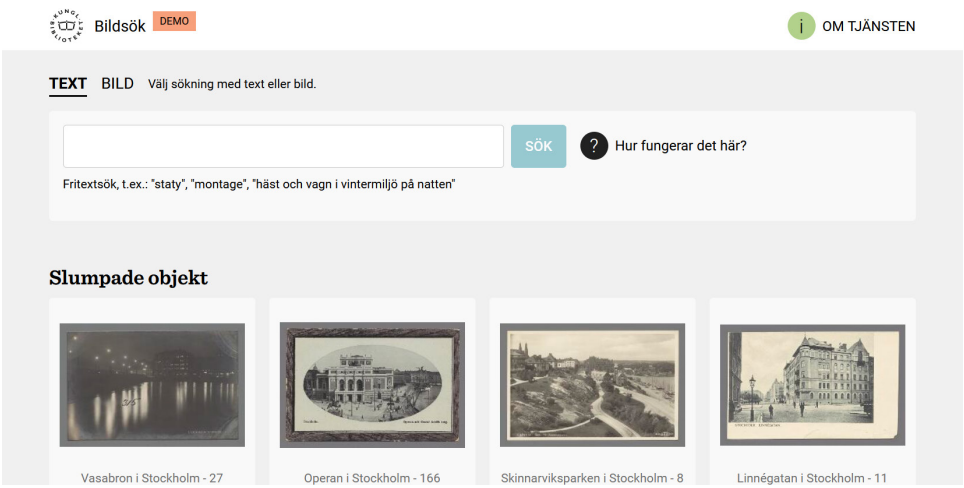


Figure 4. Search interface for interacting with historical postcards: https://lab.kb.se/bildsok.

makes the digitized postcards amenable to the sort of search techniques that we are accustomed to from online search engines. It does so by exploiting the multimodal capabilities of CLIP outlined above. So any text or image entered into the search box is run through the model, transformed into a numerical representation, and then compared with the equivalent representations for all of the postcards that have been previously processed and stored in a database. The search results are ranked according to (cosine) similarity, where the postcards with the closest representation to the search term appear at the top of the list. Insofar as it prototypes a new entrance point to the collections, the demo provides a striking example of the transformative possibilities of embeddings – and, by extension, of vector databases (Ma et al. 2023) – for the accessibility work of memory institutions.

To show the difference embeddings make, we can consider a brief example of how the demo works in practice. So we would start with either a text or an image query that we use as the basis for searching the collection. Let us say, for instance, that we are interested in exploring the cultural history of libraries with a starting point in their visual representation. We enter the term "bibliotek" (library) in the search window and, within a matter of seconds, we receive a list of the hundred postcards that most closely correspond with this search term (see Fig. 5). If a particular image captures our attention among the results, we can click on it and scroll down to find the closest visually related images in the rest of the collection (see Fig. 6). This discovery of

Figure 5. Search results for images connected to "library" in the postcard collection.

nearest visual neighbors is also apparent in the reverse image search option, where we upload an image file or insert a URL to an online image that serves as the visual input for the model to search the database for closest matches.

Such search results are based solely on the model's visual analysis of the material and are independent of any existing descriptions. This

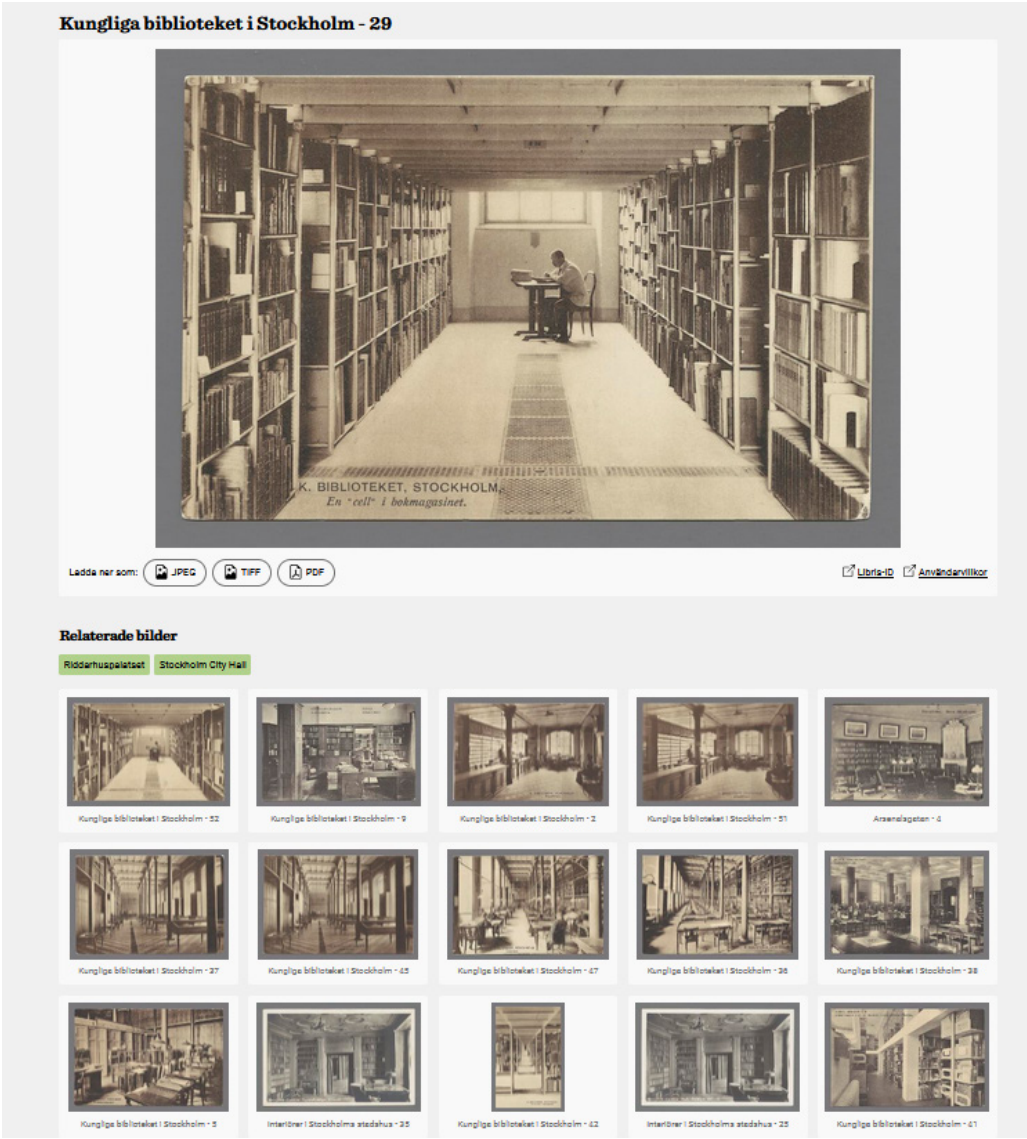


Figure 6. Finding nearest neighbors via similarity search.

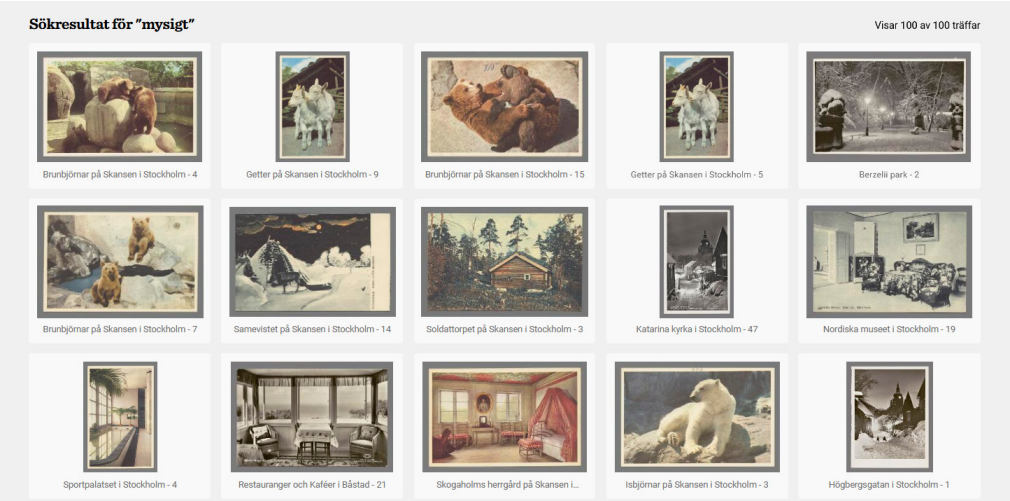


Figure 7. Evocatively searching an image collection according to emotion or concept – here "cosy."

enables us to interact with the collection in novel ways. Rather than being constrained by the predetermined categories of human-generated metadata (cf. Malmsten et al. 2025), we can now search in a highly specific fashion according to our own particular interests. A striking instance is the notion of *evocative search*: of being able to look for images in an abstract manner that matches them to a particular emotion or concept.[7] Our CLIP-based demo thus makes it possible to search the postcard archive for images that are "hot," "cold," "joyful," or "depressing," to give but a few examples. This type of searching can be suggestive, sometimes strange, but often revealing, insofar as it allows us to cluster, order, and traverse the material in new and unexpected ways. If we take the example of the Swedish term for "cosy" – "mysigt" – then we find an odd, though perhaps not uninteresting, compilation of cuddling bears, Arctic landscapes, and Nordic heritage interiors (see Fig. 7). It tells us something about the contents of the collection, though exactly what is a matter of interpretation.

Using CLIP at the library thus allows us to see this previously neglected material anew. Instead of having to wait for the assistance of a specialist librarian, who may or may not be able to find the exact image we are looking for, the demo provides us with a means of directly interacting with the collection ourselves. And rather than having to approach the material via preexisting categories, it allows us to conduct content-based searches that are driven by our own specific interests, however idiosyncratic. Making an undescribed – and therefore largely

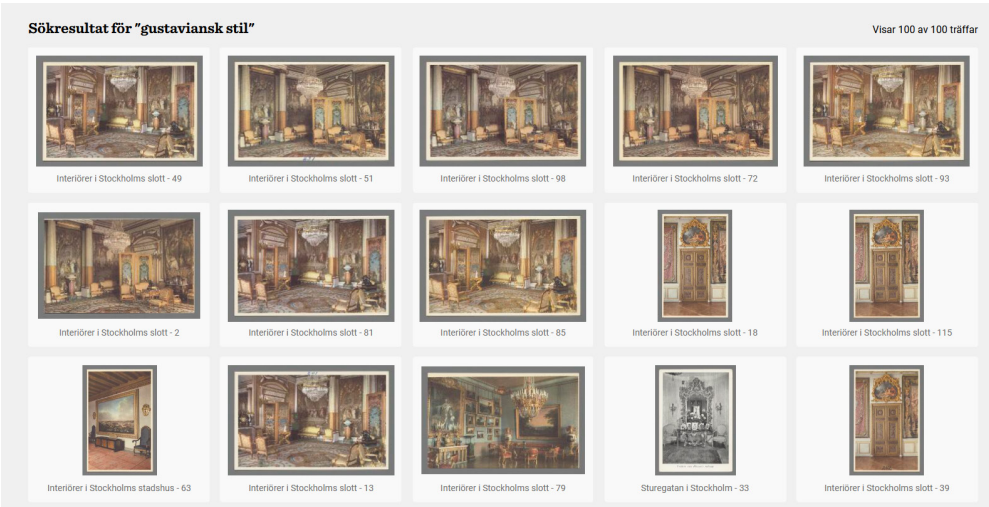


Figure 8. Incisive point of entry for research queries – locating "Gustavian style."

undiscoverable – collection available via the sort of granular search possibilities we are familiar with from our online practices, the demo has significantly enhanced the accessibility of KB's digitized historical postcards. As a particular case where AI has been applied to improve access to our shared cultural heritage, this example encapsulates the democratic tenor of KBLab's broader mission as a research infrastructure (Börjeson et al. 2024, 577–579).

RESEARCH POSSIBILITIES WITH IMAGE SEARCH

Having provided a sense of how the image search demo works, we now turn to explore something of the research potential of the underlying technology. What new possibilities emerge with CLIP-based search as a research method for approaching and using heritage data?

REFINED SEARCH AND DISCOVERY

The first and most obvious possibility, as suggested above, is simply more fine-grained and exact visual search according to content rather than predefined labels. This sort of technique provides researchers who work with digital image collections a means of quickly being able to establish relevance among potentially interesting material – in other words, a more effective way of finding the needle potentially hidden in the haystack. Does this collection contain the specific thing that we are in-

Figure 9. Tracing public profiles in print. Image: Llyfrgell Genedlaethol Cymru – The National Library of Wales. Public domain. http://hdl.handle.net/10107/5236514.

terested in exploring in our research? What is the scale of the results? Does it seem to be a scarce item or are there plenty of similar items to explore further? For example, if we were engaged in a project examining representations of domestic interiors from eighteenth-century Sweden, we could search for "gustaviansk stil" (Gustavian style) and discover what sort of images might be present (see Fig. 8). In this sense, image search provides us with an incisive entrance point and a form of orientation to begin querying and interacting with potential research material. As Thomas Smits and Melvin Wevers suggest, multimodal models "allow researchers to search for a wide range of visual concepts in large collections of images in the same way that OCR allowed keyword search in large collections of texts" (Smits and Wevers 2023, 1272).

NEW HISTORIES OF RECEPTION AND CIRCULATION

The second case for CLIP in pursuing image-based analysis of heritage data is as a methodological basis for new histories of reception and circulation. Such an application builds upon the adoption of text mining approaches to the use and reuse of text in large corpora of historical works (e.g. de Bolla et al. 2020; de Miranda 2020; Rosson et al. 2023), but applies it to visual culture. Here we could use the reverse image function to explore the first published instance of a particular image. We could also get a sense of how often, and in what different contexts, it has then been circulated and reproduced. This makes it possible to try to follow the history of a specific image through the archive over space and time: say, for instance, that we were looking to reconstruct and analyze how this portrait print of the poet Lord Byron re-appeared in different newspapers and journals in the popular press (see Fig. 9). A reception history of such celebrity profiles in the nineteenth-century media landscape could be embarked upon with this technique (cf. Mole

2007). As with the search function outlined above, such a use case takes advantage of the model's ability to locate and trace very specific things across large collections of digitized heritage material.

MULTIMODAL TOPIC MODELING

A further possibility is to use the model in order to create thematic analysis and overview. Since CLIP transforms images into embeddings, it can be employed as the basis for topic modeling large collections of visual heritage; in other words, for CLIP-topic.[8] This builds upon the popular approach of BERTopic, which uses transformer-based language models to simplify the making of topic models – a technique from natural language processing (NLP) for establishing what topics are represented in a collection of text documents (Grootendorst 2022; Malmsten et al. 2025, 185–186). Producing topic models of images gives us a new way of clustering image collections according to their contents, thereby enabling a corpus-level analysis of visual archives.

This is an application potentially most interesting for museums and other heritage organizations with expansive holdings of unlabeled image data, since it offers a means of quickly mapping the thematic range and contents of visual collections, regardless of the lack of labels. To get a sense of what it looks like to cluster images according to their visual similarity, we can consider the following examples derived from topic modeling of the library's historical postcards.

As these illustrations show, the approach proves effective in classifying the collection according to its visual contents. In both cases, CLIP-topic has produced fairly coherent topics in grouping together postcards with similar pictorial tropes – here wintry cityscapes and military parades (see Fig. 10 and 11). Though we have only recently started to experiment with the affordances of this method at the lab, it suggests a promising way forward for GLAM institutions looking to work with more automated approaches to tagging and enriching the descriptions of their image collections.

While noting the possibilities for multimodal topic modeling for categorizing visual heritage, we also need to point towards its potential limitations. This is connected to the wider question of anachronism at stake in any project of applying cutting-edge AI techniques to older heritage material. We can illustrate such an issue by turning to the example below, which is a topic from the postcard collection and its accompanying text description (the model produces a list of the most representative words to characterize each topic).[9] The postcards grouped together within the topic seem at least reasonably connected: with

Figure 10. Topic modeling postcards as a new form of order – clustering "snowy."

decorative images from churches appearing alongside other museum objects (see Fig. 12). Yet if we look at the text descriptions generated to characterize this topic, we can see that the model has chosen "statue, riding, man, sitting, horse, painting, suite, bench, skateboard, table." "Skateboard" is an interesting outlier here, since it demonstrates how far the model is conditioned to interpret historical images through the contemporary lens of reference provided by its training data (i.e, twenty-first-century web material). Such bias risks subtly distorting historical interpretation by assigning present-day associations to past imagery, which may result in misleading thematic clusters, anachronistic assumptions, or the loss of historically meaningful variation. We might decide that these textual descriptions are of secondary importance and can be disregarded – or we could look at improving them through the use of a better language model – but we still need to bear in mind the possible effects of the "historical bias of multimodal models" when considering such a use case (Smits and Wevers 2023, 1278).

Figure 11. Searching topically according to "parades."

CHALLENGES OF AI-BASED SEARCH

The issue of presentism and anachronism brings us to the challenges involved in adopting CLIP-based AI search techniques for heritage material. In this final section, we highlight and discuss some of these complexities, with a particular emphasis on questions of scale and the sort of infrastructural work involved in any such application for huge collections of heterogeneous heritage material.

THE PROBLEM OF ABSENCE

The first challenge concerns false negatives – or what we might term the problem of absence in search results – and how this connects with broader questions about trust and the use of AI models (e.g. Liu et al. 2023). Let us assume that we use the interface and receive no meaningful results for our search, as is the case when looking for "cats" (see Fig. 13). Given the way the model works and how the demo is constructed, we will always receive a set of results: it returns the images regarded as most visually similar to the search term. So we are not explicitly being told that there are no cats present among the data, though it certainly looks like this is the case. The question this raises for an end

8 7 280 [statue, riding, man, sitting, horse, painting, suit, bench, skateboard, table]

Figure 12. Classifying heritage through modern lenses – whither art thou "skateboard"?

user is how far to trust the model's results. Is the lack of relevant hits because no such image exists in the collection? Or is it rather that the model somehow did not understand the search term? Although unlikely to be an issue with something as obvious and unambiguous as a cat, it could be more complicated when using highly specific historical or cultural search terms.

This is less of a problem if a user already knows the contents of the collection well (and in the example above, we are fairly confident that there are, indeed, no cats among the postcards.) Yet for a first-time user, there might be situations when a negative result is accepted as definitive rather than understood as an effect of the technique. While this could partly be addressed through clearer user guidance about how the model works, it is also important to recognize that uncertainty in search results is not unique to AI systems. Similar challenges arise in librarian-mediated access, especially in large, sparsely described collections where discovery often depends on incomplete metadata and tacit human knowledge (for related problems connected to human classification, see Malmsten et al. 2025). In either case, the result may be new and opaque forms of "archival silences" (Moss and Thomas 2021).

Underpinning this matter of trust is the question of the AI model's "world knowledge" and perspective – that is, the particular values it has come to absorb through its training data. We know that CLIP was

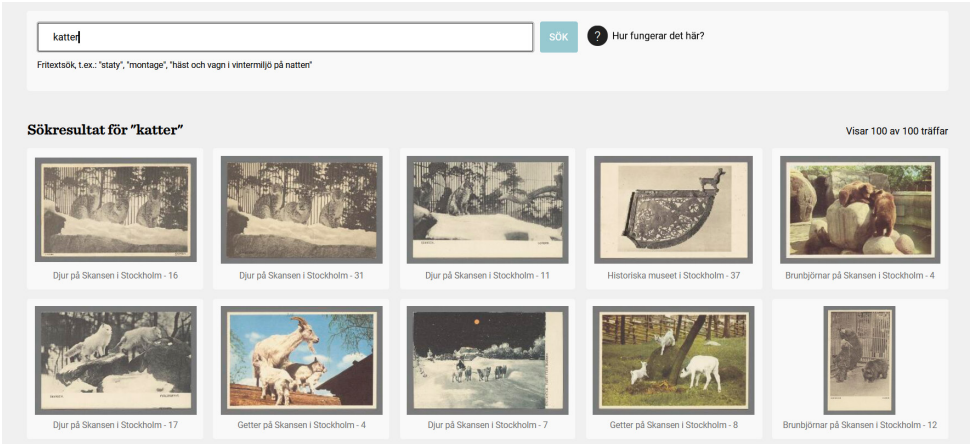


Figure 13. Searching in vain for cats – the challenge of representing absence.

trained on a dataset of 400 million pairs of image and text descriptions scraped from the internet, but we know a lot less about the potential blind spots and inaccuracies that might exist among these descriptions. Here we might do well to recall that probabilistic AI models are *not* knowledge models and should not be treated as such (Bender et al. 2021; Fredrikzon 2025). So while the model can allow us to do impressive new things, we also need to treat its results with at least a degree of critical scepticism, and remember that this is still an experimental form of infrastructure that we are dealing with.

SEGMENTATION IN COMPLEX MULTIMEDIA LAYOUTS

The second challenge relates to the apparently simple, but in practice intricate question of "what is an image?" (Mitchell 1984; Dahlgren and Hansson 2021). In the case of the postcards used for the search demo, this was relatively straightforward – at least for the heuristics of our project: we treated each postcard as comprising an image. But if we were to look at scaling up this search functionality to include other images from the library's digitized collections, the issue of how we understand the characteristics and limits of an image quickly emerges as non-trivial. The ontological complexity that this entails is especially apparent in the case of newspapers and magazines from the nineteenth and twentieth centuries, given their increasingly complicated and intricate layouts and creative intersecting of image and text. To create a vector database to search all the library's images, we would first need to identify and extract the images from these print publications. We could

Figure 14. One image or many? Segmenting the multimedia archive. Image: KB.

use the latest object detection models for such a task, but we would soon discover that the character of the material does not make this a simple undertaking. How we define the model's parameters for what it is to treat as an image has a very concrete effect on which images might come to be included, and whether they can then become searchable for users.

This is essentially a problem of segmentation (Barman et al. 2021). What is to count as a single image to be extracted? To illustrate the sort of dilemma this can involve, we can turn to this Kodak advert from the Swedish magazine *Bonniers Litterära Magasin* in the 1960s (see Fig. 14).[10] Here we can observe that the object detection model has identified five separate images on this page, as the outlines of the bounding boxes suggest. Yet it seems to have missed entirely that four of these together form part of a larger whole: that there is a montage of photos intended to be seen as coming out of the envelope beneath.

This is hardly a unique instance; similar challenges abound throughout the library's holdings of print culture (Rekathati 2021). How would we approach the series of images that constitute a comic book, for example? Should each panel be extracted as an image? What about when newspapers print an image across several pages? Should we instruct the model to stitch these together as a single image or to treat them separately? All of which is to say that there is a significant amount of work – technical, philosophical, curatorial – involved in producing an image database for AI-based search, and that the choices made will invariably impact the results that the user of such a function will obtain.

THE POLITICS OF DIGITIZATION

The final challenge is the largest and most fundamental: using AI models to enhance the searchability of heritage material is only relevant in the context of digital data. Although some of KB's collections have been the target of concerted digitization efforts – principally the daily newspapers – the majority of the library's holdings are solely accessible in physical form, and thus remain outside the remit of such search applications.[11] This means that there is a contingent quality to the digital material that we currently have available, which is shaped by the particular history of the archive and inevitably informed by the politics of mass digitization (Thylstrup 2018; Snickars 2020). Which sorts of material get prioritized for digitization, and why? What tends to be excluded or sidelined? Who gets to decide? How are such matters impacted by larger economic and political forces (Rekathati and Lenas 2025)?

This raises the prospect of a heritage future with highly differentiated forms of access to the digital and physical realms, where digital collections are searchable with the increasing capacities of AI techniques that predominate on the commercial web, while physical collections are made available in situ according to the established yet labor-intensive traditions of search and discovery enabled by human classification. Such a divide risks cementing existing archival hierarchies, privileging what has already been digitized and rendering the rest comparatively obscure. While these questions stretch far beyond the scope of this chapter – touching on the continued role of materiality and the precarious ephemerality of the digital, for instance (Chun 2008) – they are crucial to consider, since they shape our future options for searching and accessing heritage material. Understanding the conditions of possibility for what materials are available to be searched is as necessary as considering the effects of any particular techniques used to search them. The archive has never been neutral, natural, or given, whether in physical or digital form (Derrida 1996 [1995]).

CONCLUSIONS

In this chapter we have shown one particular use case where the search possibilities of multimodal AI can be used to improve the accessibility of a largely forgotten part of our visual heritage. We explained how the capabilities of CLIP to collapse the distinction between text and image made it particularly well-suited to making the library's digitized historical postcards visible and searchable for users via a search demo. We also highlighted the affordances of this AI-based approach, emphasizing the novel forms of access it makes possible. Insofar as CLIP enables user-driven exploration that is independent of the categories of existing metadata, this reinforces the argument that "multimodal models provide a radically new kind of bottom-up access to visual collections" (Smits and Wevers 2023, 1278).

Yet while the approach holds emancipatory democratic potential, we have also outlined some of the potential challenges involved in implementing this type of search. Introducing new search technologies within GLAM institutions at scale demands close attention to both user pedagogy and the ways in which various infrastructural processes – in this case, segmentation – might impact the results. Finally, we raised the obvious though no less important point that the possibilities of AI-based search are restricted to material that is digitally available. Paradoxically, it is possible that enhanced access to digital collections might introduce new hierarchies of access among heritage material at large.

As AI technologies become more embedded in the heritage sector, their effective use increasingly depends on new forms of collaboration. The development of KBLab's image search demo, for instance, involved far more than a data scientist training a model. It required close cooperation with librarians, licensing experts, software developers, UX designers, and the coordination efforts of a product manager. Recognizing this need, the library has launched KBx, a new initiative enabling data scientists to work directly with colleagues across the organization to identify where AI can help improve discovery and access (Gordon 2023). These interactions are mutually beneficial: librarians gain insight into the potential and limitations of AI, while technical teams deepen their understanding of the collections and institutional context. As AI continues to evolve – and expectations for access grow – such interdisciplinary collaboration is likely to become increasingly essential.

Beyond enhancing discovery and access, introducing AI-based search like CLIP into library workflows can also reshape the identity of the collection itself. Traditionally, what becomes visible within an archive has been shaped by cataloguing practices and institutional judgments about relevance, pertinence, and description. Our attention, as users, has been guided by generations of metadata that have created signposts of significance – directing us toward what is already known and named. But AI search techniques like those explored in this chapter operate according to a different logic: they retrieve based on patterns of similarity learned from large-scale visual-textual co-occurrences, and can therefore unearth unexpected correspondences or overlooked elements. In doing so, they can help shift attention away from the canonically described toward what has previously escaped notice. Rather than reinforcing the existing architecture of the archive, these techniques have the potential to reframe it – altering what is seen, what is studied, and ultimately the stories that can be told from the collections.

NOTES

1 For this search demo, see: https://lab.kb.se/bildsok/. The demo is freely available online, though the text search interface presumes Swedish inputs. We explain more about how this can be used over the course of the chapter. Part of this work was carried out within the Huminfra research infrastructure project.

2 For an instance of this material that, to a certain extent, has actually been cataloged, see Harvard's collection of "advertising ephemera": https://library.harvard.edu/collections/advertising-ephemera.

3 For more details (in Swedish), see: https://www.kb.se/hitta-och-bestall/om-samlingar-och-material/affischer.html.

4 See: https://arken.kb.se/se-s-kob-vykort.
5 For a recent movement advocating the value of data labs to GLAM institutions, see: https://glamlabs.io/.
6 Here we would like to acknowledge and thank Matthias Nilsson, Krzysztof Bergendahl, and Ebrima Faye at KB for their work on the project.
7 We borrow this apt notion of "evocative search" from Peter Leonard, Stanford University Libraries.
8 We are indebted to our colleague at KBLab, Martin Malmsten, for this notion. The images included below are based on experiments with multimodal topic modeling carried out at the lab by our colleague, Justyna Sikora. See: Chris Haffenden, and Justyna Sikora. 2025. "CLIP-Topic: Identifying Themes in Large Image Collections." *The KBLab Blog*. https://kb-labb.github.io/posts/2025-12-01-clip-topic/.
9 The BERTopic documentation on Github provides more specific details about the practical mechanics of the approach: https://maartengr.github.io/BERTopic/getting_started/multimodal/multimodal.html#text-images.
10 This example and others like it were identified as a result of an exploratory collaboration at the lab between one of our data scientists, Faton Rekathati, and scholar of Art History and Visual Culture, Anna Dahlgren, Stockholm University.
11 The library's newspaper archive is searchable to users in Sweden via the online database: https://tidningar.kb.se/.

REFERENCES

Ames, Sarah, and Stuart Lewis. 2020. "Disrupting the Library: Digital Scholarship and Big Data at the National Library of Scotland." *Big Data & Society* 7 (2): 1–7.

Assmann, Aleida. 2010. "Canon and Archive." In *Cultural Memory Studies: An International and Interdisciplinary Handbook*, edited by Astrid Erll and Ansgar Nünning. De Gruyter: 97–108.

Barman, Raphël, Maud Ehrmann, Simon Clematide, Sofia Ares Oliveira, and Frédéric Kaplan. 2021. "Combining Visual and Textual Features for Semantic Segmentation of Historical Newspapers." *Journal of Data Mining & Digital Humanities*. January 19: 1–26.

Bender, Emily M., Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery: 610–23.

Bingham, Nicola Jayne, and Helena Byrne. 2021. "Archival Strategies for Contemporary Collecting in a World of Big Data: Challenges and Opportunities with Curating the UK Web Archive." *Big Data & Society* 8 (1): 1–6.

Blair, Ann. 2010. *Too Much to Know: Managing Scholarly Information Before the Modern Age*. Yale University Press.

Börjeson, Love, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati et al. 2024. "Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden." *College & Research Libraries* 85 (4): 564–82.

Carlsson, Fredrik, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. "Cross-Lingual and Multilingual CLIP." In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, edited by Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi et al. European Language Resources Association: 6848–54.

Chun, Wendy Hui Kyong. 2008. "The Enduring Ephemeral, or the Future Is a Memory." *Critical Inquiry* 35 (1): 148–71.

Cukier, Kenneth, and Viktor Mayer-Schoenberger. 2013. "The Rise of Big Data: How It's Changing the Way We Think About the World." *Foreign Affairs* 92 (3): 28–40.

Dahlgren, Anna, and Karin Hansson. 2021. "What an Image Is: The Ontological Gap between Researchers and Information Specialists." *Art Documentation: Journal of the Art Libraries Society of North America* 40 (1): 21–32.

de Bolla, Peter, Ewan Jones, Paul Nulty, Gabriel Recchia, and John Regan. 2020. "The Idea of Liberty, 1600–1800: A Distributional Concept Analysis." *Journal of the History of Ideas* 81 (3): 381–406.

de Miranda, Luis. 2020. *Ensemblance: The Transnational Genealogy of Esprit de Corps*. Edinburgh University Press.

Derrida, Jacques. 1996 [1995]. *Archive Fever: A Freudian Impression*. University of Chicago Press.

Fredrikzon, Johan. 2025. "Rethinking Error: 'Hallucinations' and Epistemological Indifference." *Critical AI* 3 (1).

Garner, Anne. 2021. "State of the Discipline: Throwaway History: Towards a Historiography of Ephemera." *Book History* 24 (1): 244–63.

Gordon, Rebecka. 2023. "KBx flyttar ut AI-tekniken till bibliotekarierna." *Magasin K*. October 31.

Grootendorst, Maarten. 2022. "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure." *arXiv*:2203.05794.

Haffenden, Chris, Elena Fano, Martin Malmsten, and Love Börjeson. 2023. "Making and Using AI in the Library: Creating a BERT Model at the National Library of Sweden." *College & Research Libraries* 84 (1): 30–48.

Ma, Le, Ran Zhang, Yikun Han, Shirui Yu, Zaitian Wang, Zhiyuan Ning, Jinghan Zhang et al. 2023. "A Comprehensive Survey on Vector Database: Storage and Retrieval Technique, Challenge." *arXiv*:2310.11703.

Jurafsky, Daniel, and James H. Martin. 2023. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition*, 3rd ed., draft version, https://web.stanford.edu/~jurafsky/slp3/.

Konstenius, Göran. 2017. *Plikten under lupp! En studie av pliktlagstiftningens roll, utformning och relevans i förhållande till medielandskapets utveckling*. Kungl. Biblioteket.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "Imagenet Classification with Deep Convolutional Neural Networks." In *Advances*

in Neural Information Processing Systems 25 (NIPS 2012), edited by Peter Bartlett, Fernando Pereira, Christopher Burges, Leon Bottou, and Kilian Weinberger. Morgan Kaufmann Publishers: 1–9.

Liu, Yang, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov et al. 2023. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Models' Alignment." *arXiv*:2308.05374.

Malmsten, Martin, Viktoria Lundborg, Elena Fano, Chris Haffenden, Fredrik Klingwall, Robin Kurtz, Niklas Lindström et al. 2025. "Without Heading? Automatic Creation of a Linked Subject System." In *New Horizons in Artificial Intelligence in Libraries*, edited by Edmund Balnaves, Leda Bultrini, Andrew Cox, and Raymond Uzwyshyn. De Gruyter Saur: 179–98.

Mitchell, W. J. T. 1984. "What Is an Image?" *New Literary History* 15 (3): 503–37.

Mole, Tom. 2007. *Byron's Romantic Celebrity: Industrial Culture and the Hermeneutic of Intimacy*. Palgrave Macmillan.

Moss, Michael, and David Thomas. 2021. *Archival Silences: Missing, Lost, and Uncreated Archives*. Routledge.

Radford, Alec, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry et al. 2021. "Learning Transferable Visual Models from Natural Language Supervision." *arXiv*:2103.00020.

Rahaman, Hazifur. 2018. "Digital Heritage Interpretation: A Conceptual Framework." *Digital Creativity* 29 (2–3): 208–34.

Rekathati, Faton. 2021. "A Multimodal Approach to Advertisement Classification in Digitized Newspapers." *The KBLab Blog*. https://kb-labb.github.io/posts/2021-03-28-ad-classification/

Rekathati, Faton, and Erik Lenas. 2025. "Risk för svenskt självmål inom AI." *Svenska Dagbladet*. March 23.

Rosson, David, Eetu Mäkelä, Ville Vaara, Ananth Mahadevan, Yann Ryan, and Mikko Tolonen. 2023. "Reception Reader: Exploring Text Reuse in Early Modern British Publications." *Journal of Open Humanities Data* 9 (5): 1–11.

Smits, Thomas, and Melvi Wevers. 2023. "A Multimodal Turn in Digital Humanities. Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections." *Digital Scholarship in the Humanities* 38 (3): 1267–80.

Snickars, Pelle. 2020. *Kulturarvets mediehistoria: Dokumentation och representation 1750–1950*. Mediehistoriskt arkiv.

The Multigraph Collective. 2018. *Interacting with Print: Elements of Reading in the Era of Print Saturation.* University of Chicago Press.

Thylstrup, Nanna Bunde. 2018. *The Politics of Mass Digitization*. MIT Press.

Jonas Ingvarsson. Photo: Marie Cronqvist.

TRADITION AND DIGITAL TALENT

Afterword in Memory of Jonas Ingvarsson

THIS VOLUME IS offered in tribute to the memory of literary scholar Jonas Ingvarsson, who died unexpectedly on April 10, 2025. Ingvarsson developed and co-organized the Flows & Frictions symposium that inspired this anthology, undertaken within the research project *The Order of Criticism Revisited*, which he led as principal investigator (and which he preferred to call "The New Order of Criticism" as a nod to the post-punk/synth-pop band New Order).

One of the titles he originally proposed for the Flows & Frictions symposium was "The Vast and the Waste: Tradition and Digital Talent in the Humanities," a playful paraphrase of T. S. Eliot's essay "Tradition and the Individual Talent" (1919). As he quipped in an email to those of us in the project: "I can mansplain – or 'literary-scholar-splain' – the last part of the title, which is undeniably my favorite, ha ha!" Although the symposium did not adopt that title, he kept returning to Eliot's notion of tradition versus individual talent, and framed the project during a conference by invoking his claim about bracketing the poet's voice: "It is in this depersonalization that art may be said to approach the condition of science." As Ingvarsson saw it, Eliot's argument pointed to "the potential of using digital methods, which often (though not always) seek to harness the impersonal voice of datasets. At the same time, one of our project's aims is to show how traditional methods can give rise to new, creative ways of using digital technology."

Ingvarsson also began drafting an article with the working title "Tradition and Digital Talent: Discourse Analysis vs. Topic Modeling." In one of his last emails, he proposed that we compare human- and

machine-generated interpretations of discursive patterns in literary criticism and undertake what he called "topic modeling in reverse." After Ingvarsson's passing, our aim was to carry this idea forward as a contribution to this volume, using the Eliot reference as a conceptual framework. Yet we soon realized that we never really fully grasped his intricate vision and that, without *his* individual talent, we would not be able to develop it as he intended.

Nevertheless, aware of all that remains unfinished without Ingvarsson, we have completed this anthology, drawing on the years we were privileged to work closely with him. His intellectual curiosity, joyful creativity, and warm personality shaped our joint research, and what we have finished here grows out of that collaboration.

Daniel Brodén and Lina Samuelsson,
Gothenburg and Västerås, respectively
October 24, 2025

ABOUT THE CONTRIBUTORS

DAVID ALFTER is a research engineer at the Gothenburg Research Infrastructure in Digital Humanities (hereafter GRIDH) at the University of Gothenburg. He completed his PhD with a dissertation on automated lexical complexity prediction for language learners. His research focuses on language learning, textual complexity, and AI approaches to text understanding.

DANIEL BRODÉN is associate professor in film studies at the University of Gothenburg, where he serves as a coordinator and researcher at GRIDH. His work in the digital humanities focuses on mixed methods and digital epistemology. He is a member of the research project *The Order of Criticism Revisited* and co-organized the Flows & Frictions symposium with Jonas Ingvarsson.

JUDITH BROTTRAGER is a postdoctoral researcher at the Technical University of Darmstadt. She completed her PhD with a dissertation in computational literary studies, and her research focuses on canonization processes, quantitative literary history, and comparative computational approaches to literature.

CHRIS HAFFENDEN is research coordinator at KBLab at the National Library of Sweden (*Kungliga biblioteket*, hereafter KB) and an affiliated researcher at the Department of the History of Science and Ideas at Uppsala University. At KB, he initiates and facilitates research projects that explore novel approaches to the library's digital collections. His research focuses on the history of cultural memory and authorial

afterlives, with particular attention to practices of self-erasure and the genealogy of the "right to be forgotten." He has published on erasure studies, Romantic-era celebrity culture, and the uses of AI in research libraries.

ANTSKE FOKKENS is a professor of computational linguistic methods at the Vrije Universiteit in Amsterdam. Her research focuses on how language works and how it can be modelled computationally, with a particular focus on what the challenges involved mean when language technology is applied in digital humanities or computational social science research.

FATON REKATHATI is a data scientist at KBLab, at KB, where he works on improving the searchability and accessibility of the library's digital collections. He holds a master's degree in statistics from Linköping University.

EMMA RENDE is a product manager at KB, where she works with KBLab on the launch of KBx, an initiative that combines AI insights with library expertise to improve the library's workflows. She holds a master's degree in computer and systems science from Stockholm University.

LINA SAMUELSSON is senior lecturer in comparative literature at Mälardalen University. Her research interests include press history, computational literary studies, and literary criticism. She is a member of *The Order of Criticism Revisited* project, which builds on her dissertation on the discourse on Swedish literary criticism *Kritikens ordning* [*The Order of Criticism*] (2013).

ELEANOR SMITH is a doctoral candidate at the Computational Linguistics and Text Mining Lab at the Vrije Universiteit in Amsterdam. Her doctoral research project focuses on the domain specific methodological considerations for applying natural language processing approaches to digital humanities and digital social science case studies.

ANTON TÖRNBERG is associate professor in sociology at the University of Gothenburg. His research primarily focuses on the online far right, combining computational methods with qualitative approaches. He is the author of *Intimate Communities of Hate: Why Social Media Fuels Far-Right Extremism* (Routledge, 2024).

TED UNDERWOOD is Professor of Information Sciences and English at the University of Illinois, Urbana-Champaign. He has written three books on literary history and is currently working on ChronoLogic, a benchmark that evaluates a language model's understanding of historical contexts from 1800 to 2000 by asking it to draw inferences and solve problems like a writer with a specified social background at a particular place and point in time.