Doing science studies with large language models
An introduction

Arno Simons, Adrian Wüthrich, and Michael Zichert

Introduction

Large language models (LLMs) are rapidly becoming part of the everyday infrastructure of scholarly work. They can read, summarise, search across, and transform texts at a speed and scale that changes the way researchers find and engage with literature and sources. As with any powerful new technique, their uptake raises questions about reliability, transparency and appropriate use, and these questions are intensified by the opacity and commercial governance of many current systems. For researchers in the history, philosophy and sociology of science (HPSS), this creates a double task: to explore how LLMs can be integrated into their own research practices, and to critically assess and guide that integration. HPSS is well positioned to do both, given its reflexive stance and close engagement with science and technology, but this double role also creates a tension between optimistic exploration and cautionary, critical reflection. The chapters in this volume are written from within this tension.

The volume documents how the HPSS community thought and felt about the arrival of LLMs as research tools in mid 2025. It captures the ideas, doubts, hopes, and arguments sparked by the ERC-funded international workshop "Large Language Models for the History, Philosophy, and Sociology of Science", held in Berlin in April 2025[1], and develops them further through continued reflection and exchange. All contributions were drafted and revised in a shared online document, where authors could read, comment on and respond to each other's chapters as they wrote. Cross-references and explicit dialogue across chapters were thus part of the design rather than an afterthought. This framing invited openness, experimentation, and personal voice, and encouraged authors to build on their own styles, interests, and positions within or adjacent to HPSS. In doing so, the volume carried the workshop's lively, collaborative and intellectually diverse atmosphere into written form. While many authors had presented their work at the workshop,

1 See https://www.tu.berlin/hps-mod-sci/llms-for-hpss/workshop-llms-for-hpss for general information about the workshop and https://www.youtube.com/playlist?list=PL5rAX6ywmP7O_nT99Osd74uino78BJMVT for video recordings of the talks.

not all speakers are represented in this volume, and some contributors joined the project later without having taken part in the Berlin meeting.

Building on our survey paper (Simons et al., 2026), this volume is situated within a growing literature on LLMs as research tools in fields related to HPSS. Programmatic publications and reviews have appeared in scientometrics (Zhang et al., 2025a), (digital) humanities (Karjus, 2025; Lozić and Štular, 2023; Zhong et al., 2025), quantitative (Thapa et al., 2025; Ziems et al., 2024) and qualitative (Friese and Morgan, 2026) social sciences (see also Christou, 2025; Davidson and Karrell, 2025), and related fields (Varnum et al., 2024). Computer scientists are also increasingly calling for interpretative and humanistic perspectives in the development and evaluation of LLMs and explicitly invite collaboration with the humanities (Ehrlich and Hazzan, 2025; Hemment and Kommers, 2025). In parallel, work on LLMs for scientific research more generally is expanding (Binz et al., 2025; Luo et al., 2025; Ren et al., 2025), including efforts to adapt models to scientific languages and tasks (Ho et al., 2024; Zhang et al., 2024) and to support literature reviews and research synthesis (Han et al., 2024; Scherbakov et al., 2025; Zhuang et al., 2025).

In the remainder of this chapter we briefly sketch what current LLMs are, how they are trained and in which broad ways they can be used. We then outline major opportunities and challenges that LLMs pose for research in HPSS. The chapter closes with an overview of the structure of the volume and a short introduction to each contribution.

What are LLMs, how are they trained, and how can we use them?

By LLMs we mean neural language models based on the transformer neural network architecture (Vaswani et al., 2017).[2] Transformers combine ideas that had been developed in earlier neural language models, such as contextualised word embeddings, encoder and decoder modules, and attention mechanisms, in a way that scales to very large datasets and model sizes. Although the original transformer was introduced in machine translation, it has since become the standard architecture for many kinds of language and multimodal models (Yin et al., 2024).

In everyday use, people often equate LLMs with chat oriented, generative systems such as ChatGPT. Technically, these are one branch of transformer models, built around the decoder part of the architecture and trained with a simple next-token prediction objective. Given a sequence of tokens, the model learns to predict the next token in the sequence, where a token can be a word, a subword, or sometimes just a single character (Radford et al., 2018; 2019). During training and generation the model is only allowed to use its representation of the tokens to the left of the prediction point. This left to right

2 The term "large" in large language models is a moving target. In some discussions it is reserved for frontier-scale generative models with tens or hundreds of billions of parameters, roughly from the size of GPT-3 onward. In this introduction we use the term LLM in a broader sense to refer to all transformer-based language models, including encoder-only models like the original BERT whose parameter counts are "only" in the hundreds of millions, since these models also play an important role in HPSS workflows.

setup is what makes these models generative. At inference time they can be asked to continue a text, answer a question, or transform an input, one token at a time.

In contrast, encoder-based transformer models are trained differently. The paradigmatic example is BERT (Devlin et al., 2018). Instead of predicting the next token, BERT-style models learn to fill in masked tokens in a sentence or paragraph. During training, a certain proportion of tokens in the input is replaced by a special mask symbol and the model is asked to recover the original tokens. Crucially, it can use both left and right context to do so. This bidirectional attention lets encoder models take into account the entire surrounding sentence or passage when deciding which token fits best.[3]

Both kinds of models, decoder-based and encoder-based[4], learn internal vector representations for each token in context, often called contextualised word embeddings (CWEs). Roughly speaking, these are numerical encodings of how a word is used in a particular sentence, such that uses with similar meanings end up with similar vectors (Ethayarajh, 2019; Jawahar, 2019). In decoder models, these representations are geared towards predicting the next token, using only information from the past in a sentence. In encoder models, they are shaped by information from both sides of a token and are therefore particularly informative about local meaning and sentence level structure.

This distinction matters because it gives rise to different application types, also within HPSS (Simons et al., 2026). Decoder-based, generative LLMs are extremely flexible text generators and can also be used for structured extraction. The introduction of GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020) showed that training larger networks on larger datasets led to the emergence of unanticipated skills, such as few- and zero-shot learning. In response, the AI industry scaled up decoder models at high speed and added further post training steps, such as instruction tuning and reinforcement learning from human feedback (RLHF), to improve their ability to follow prompts (Ouyang et al., 2022). More recently, specialised prompting and training techniques for step-by-step, so-called chain-of-thought (CoT) "reasoning" (Wei et al., 2022) have further increased their performance on tasks that involve mathematical proofs, logic puzzles or multi-step argumentation. These efforts resulted in today's proprietary, energy intensive generative frontier models and their striking ability to solve many tasks through natural language prompting, including many tasks directly relevant to HPSS. While this success has concentrated attention and investment on very large decoder models, it has not made encoder models obsolete.

---

3 After BERT, researchers improved this encoder-style "masked token" approach in two main ways. RoBERTa (Liu et al., 2019) largely kept the same masked-word objective but showed that changes to the training recipe can make this kind of learning work better in practice. ELECTRA (Clark et al., 2020) introduced an alternative objective in which the model learns to detect tokens that have been replaced, offering a different and often more efficient route to learning strong bidirectional representations. DeBERTaV3 (He et al., 2022) builds on these ideas by pairing a strong encoder architecture with ELECTRA-style training, illustrating that post-BERT progress also came from rethinking the pretraining objective itself.

4 Some transformers are encoder–decoder hybrids, such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2020), combining bidirectional encoding with left-to-right generation. This blurs the split, but the roles still matter as components that can appear alone or together.

Encoder-based models with their context-rich CWEs still play a central role for HPSS, especially for tasks that concern word and sentence meaning, such as semantic similarity, topic exploration, or tracking conceptual change (Simons et al., 2026; in this volume, see Ahmadi, 2026; Simons, 2026a). They are often adapted in different ways. A common strategy is task specific fine-tuning for classification, sequence labelling or token tagging on relatively small labelled datasets. Another is to fine-tune them to produce sentence embeddings (Reimers and Gurevych, 2019), that is, single vectors that represent whole sentences or paragraphs and can be compared with similarity measures. Such embeddings underpin state-of-the-art text-clustering and topic-exploration tools (Grootendorst, 2022). They are also widely used in many retrieval-augmented generation (RAG) setups to index and search large corpora, so that an encoder model first retrieves the most relevant passages which are then handed to a generative model that composes an answer on top of them (Gao et al., 2024). In other words, BERT-like models quietly power the search and filtering stages that make many GPT-style systems look well informed.[5]

For many HPSS-related tasks it therefore makes sense to ask not only which single model type fits best, but how differently specialised models can be combined in task-specific pipelines. In many current systems, these pipelines are wrapped into so called "agents", in which a generative LLM acts as a controller that breaks a task into steps, calls external tools such as search engines, databases, code interpreters or encoder models, and integrates their outputs into a final answer (Huang et al., 2024; Ren et al., 2025; Schmidgall et al., 2025). Emerging multi-agent setups go one step further by coordinating several such agents in different roles, for example as planners, critics or domain specialists (Guo et al., 2024; Tran et al., 2025). LLMs are also increasingly used as judges, for instance to rate or compare candidate answers from different prompts or models (Li et al., 2025). Beyond model choice, HPSS scholars can experiment with prompting strategies, including CoT and prompt chaining (Sun et al., 2024), to decompose complex research questions into smaller steps that LLMs can handle more reliably. Many chapters in this volume, such as Danilova et al. (2026), Scharnhorst et al. (2026), and Schlattmann et al. (2026), illustrate such a division of labor, discussing the combination of encoder and decoder models or multiple prompt stages within a single workflow.

Opportunities

Natural language text is one of the most important data sources in HPSS research, which makes LLMs, with the improved context-sensitive semantic capacities and broad task performance sketched above, natural candidates for many kinds of text-related work.

5 On the encoder side, development continues beyond the original BERT generation. Some BERT-style models have been pushed to much larger parameter counts, including DeBERTa (He et al., 2022; 2021) and multilingual encoders such as XLM-R (Goyal et al., 2021). In addition, prompt-based methods like pattern-exploiting training (PET) show that masked language models can support few-shot classification by reframing labels as fill-in-the-blank decisions, offering an alternative to full task-specific fine-tuning (Clavié et al., 2025; Schick and Schütze, 2021; He et al., 2022). These newer encoder-focused lines of work are worth following to see which yield robust, meaning-sensitive tools that could become useful for HPSS.

HPSS scholars routinely engage with a wide range of genres, from scientific publications across virtually all fields and subfields, through transdisciplinary materials such as industry reports, policy papers and media coverage, to field notes, archival documents and interview transcripts. Modern LLMs can also process structured material, for example tables, databases or code, and increasingly multimodal inputs such as images, figures and video (Yin et al., 2024). Taken together, these capacities open up a broad landscape of possible applications, from large-scale corpus analysis to interactive exploratory reading. At the same time, they lead HPSS into largely uncharted territory, where questions about methodological fit, expertise, trust, and control over the tools become newly pressing. In this section we highlight some of the main opportunities that LLMs create for HPSS, before turning to the challenges and risks that accompany them in the next section.

A first opportunity is domain and task adaptability. Frontier generative models are strong multitask and in-context learners, so they can often be applied to very HPSS-specific problems simply by careful prompt design, for example to generate alternative readings of citation contexts (Simons, 2026b). At the same time, both decoder and encoder models can be adapted more deeply to particular domains by continued pretraining on targeted corpora, such as physics articles for tracing shifting meanings of key terms like "Planck" or "virtual" over time (Zichert et al., 2025; Zichert and Simons, 2026; Simons, 2026a). This kind of domain adaptation sharpens a model's sensitivity to local vocabularies, genres and histories of use, which is crucial for many HPSS questions. Together, prompt-based task adaptation and corpus-based domain adaptation make it possible to build task-specific pipelines in which different models do what they are best at. For example, Danilova et al. (2026) first use small decoder LLMs to generate structured historical texts, then use encoder models to check how diverse and historically plausible these texts are and to map out the main topics they cover, and finally treat this vetted synthetic corpus as training data for a future genre classifier, which could itself be an encoder-based model.

A second opportunity is the methodological tunability of LLMs. They can support multiple epistemic roles in HPSS, with adjustable levels of delegation, commitment, and evidential responsibility. Depending on how they are configured and validated in context, LLMs can be used to delegate well-specified tasks such as document classification, first-pass coding, or information extraction. In this mode, they return relatively stable outputs that can serve as inputs to later analysis or as deliverables in their own right. They thus scale up familiar close reading practices to large corpora (e.g., Alves et al., 2026; Boulanger, 2026; D'Alessandro, 2026). But the same architectures can also be positioned in lower-commitment, supportive roles, for example as conversational partners in the exploratory reading of particular sources (e.g., Hill, 2026) or entire archives (e.g., Scharnhorst et al., 2026), where outputs function as suggestions, contrasts, or "what if" scenarios. Importantly, the dimensions do not always align: models can be used to automate exploratory work (for example by generating candidate codes or hypotheses at scale), or to support more evidential tasks (for example by producing checkable extractions with quotes and pointers that the researcher verifies). Across projects, and sometimes across stages of a single project, this tunability lets researchers choose how much

agency to delegate, how strongly to treat outputs as results, and how explicitly evidential the model's contributions are supposed to be (cf. Simons et al., 2026).

A third opportunity lies in the ability of LLM-based methods to act as translation layers that improve interoperability across scales, methods, and heterogeneous materials. Because they can operate at different textual scales, from passages to full documents and large corpora, LLM-based methods facilitate the design of multi-level analyses that move between abstracts, sections, and full texts (cf. Malaterre and Lareau, 2026). In addition, they enable movement between fine-grained textual interpretation and aggregate analysis, allowing researchers to relate local readings and categories to broader distributions and patterns, including in mixed-method settings (e.g., Schlattmann et al., 2026). This interoperability also shows up in the ease with which LLM components can be inserted into otherwise familiar pipelines, for example embedding-based semantic expansion of key terms to broaden or refine term sets in conceptual and terminological studies (e.g., Aguilar-Valdez et al., 2026; Malaterre and Lareau, 2026), or the generation of synthetic or semi-synthetic training data to bootstrap genre classifiers or coding models when annotated examples are scarce (e.g., Danilova et al., 2026). Overall, the opportunity lies in using LLMs as mediators that make different scales of text, heterogeneous sources, and qualitative and quantitative modes of analysis interoperable, with modular insertion points that let researchers build those connections where a project most needs them. Finally, as multimodal systems mature, this kind of mediation can extend beyond text to figures, tables, diagrams, and other visual or audiovisual materials that are central to scientific communication and archival work (see Graßhoff, 2026).

A fourth opportunity concerns documentation, reproducibility, and shared infrastructure, especially for interpretative work. LLM-based analyses are often implemented through prompts, scripts, and configuration files that spell out, in a concrete, machine-readable way, what was asked of the model, which data were used, and how intermediate steps were wired together. Such files can be run, modified, and inspected by others. They also make prompts and pipelines visible as concrete objects of study in their own right. For earlier computational methods in HPSS, such as topic models or citation networks, this kind of scripted reproducibility was already possible, but it mostly applied to pattern-oriented analysis. LLMs extend this possibility into domains of interpretative work, for example qualitative coding, summarising or reconstructing arguments, that previously lived almost entirely in tacit practices and prose descriptions (e.g., Hitch, 2024; Morgan, 2023; Zhang et al., 2025b; but see Nguyen and Welch, 2025, for a sharp critique of applying LLMs to qualitative research). If their pipelines are carefully documented, and if the models are clearly version-pinned and run with fixed, deterministic decoding settings (for example with temperature set to zero), then even interpretative LLM-based studies might come much closer to the reproducibility standards of other computational methods. With proprietary frontier models, strict repeatability of outputs is harder to achieve (see the fourth challenge below), yet publishing prompts, code, and logs could still improve procedural transparency. We see an opportunity in turning such workflows into reusable models, shared datasets or shared prompt libraries tailored to particular sources, thereby creating a common infrastructure that others can adopt, adapt and extend, and that helps robust methods spread across projects, institutions and subfields of HPSS and beyond (cf. Boulanger, 2026; Scharnhorst et al., 2026; Simons, 2026b).

A fifth opportunity is the way LLMs push HPSS toward a more reflexive articulation of concepts and reasoning. To obtain useful behaviour from a model, researchers must specify tasks, categories, examples, and evaluation criteria with an unusual degree of explicitness. Writing and revising prompts, programming pipelines, curating training or few-shot examples and testing how sensitive outputs are to small changes all tend to surface tacit assumptions about conceptualization, context or relevance (e.g., Simons, 2026b; Liesegang and Gläser, 2026; Zichert and Simons, 2026). In this way, model steering and evaluation can double as methodological self-study, where the frictions, failures, and ambiguities encountered with the system help clarify where our own concepts are ill-suited, overloaded or contested, and where they are robust.

Challenges

The opportunities just outlined do not come for free. They are intertwined with a series of challenges and open questions that HPSS needs to confront if LLMs are to become responsible and intellectually productive parts of its research practices.

A first challenge concerns the trade-off between accessibility and literacy (Simons et al., 2026; see also Marx, 2024). Frontier models in polished interfaces make it very easy to run impressive-looking analyses without much technical knowledge, which lowers the barrier to entry but makes it tempting to rely on off-the-shelf behaviour for genuinely difficult HPSS tasks. Many of these tasks, however, require careful prompt design, task-specific evaluation and often some form of corpus adaptation or fine-tuning, which presupposes at least basic coding skills, familiarity with the tools and interfaces used to run and adapt models, and an understanding of how these systems fail. This creates a tension between accessible tools that many can use and more accurate, tailored workflows that only a subset of researchers can build and diagnose, and raises the question of how HPSS as a field can develop and share LLM literacy rather than leaving it to individual enthusiasts (cf. Meding and Daugh, 2026b; Wagner and Hermes, 2026; Vogl et al., 2026).

A second challenge is the limited temporal and contextual sensitivity of current models (see Part 2 of this volume, "Historicizing LLMs", which includes chapters by Büttner, 2026; Olival et al., 2026; Danilova et al., 2026; and Wolf, 2026). Most LLMs are not explicitly trained to track historical change in language, concepts or genres, yet this is central to the history of science and to historically-informed strands of philosophy and sociology of science. Today's models tend to blur or collapse historical distance, for example by projecting contemporary meanings into older texts or by smoothing over shifts in terminology that are precisely what HPSS wants to study. Addressing this may require not only new training regimes and datasets, but also architectural innovations such as explicit time encodings or attention mechanisms that respect temporal structure (cf. Büttner, 2026). HPSS scholars are well placed to formulate what genuine time awareness would mean in practice and should therefore take part in defining and testing such developments rather than treating them as purely technical questions left to others.

A third challenge concerns interpretative work and the absence of fixed ground truths. Much HPSS research deals with shifting categories, multiple plausible readings and situated judgments that cannot easily be reduced to single correct labels (cf. Alves

et al., 2026, who show how disputed labels and shifting boundaries complicate "ground truth" in discipline classification). This sits uneasily with standard machine-learning notions of training data and evaluation, which assume clear targets against which model performance can be scored. HPSS therefore needs to reflect on what it means to use LLMs in domains where disagreement and ambiguity are not bugs but core features of the object of study, and to experiment with evaluation practices that foreground qualitative appraisal, comparative reading and argumentation rather than simple accuracy metrics (e.g., Meding and Daugs, 2026a). In doing so, the field can connect to neighbouring interpretative disciplines in the humanities and social sciences that face similar issues (Friese and Morgan, 2026; Hitch, 2024; Morgan, 2023; Zhang et al., 2025b), and respond to emerging calls for a "qualitative turn" in contemporary AI (Hemment and Kommers, 2025). Such a turn treats model outputs, especially from LLMs, as cultural artefacts that are subject to the kinds of judgments and curatorial practices familiar from literary studies, art criticism or ethnography. At a time when some interpretative communities are highly skeptical of (Nguyen and Welch, 2025) or calling for bans on generative AI in their own domains (Jowsey et al., 2025) while major social-science and computer-science preprint servers are narrowing the space for interdisciplinary debates about LLMs from both ends (arXiv, 2025; SocOpen, 2025; see also Brainard, 2025), it becomes all the more important for HPSS to claim room within its own institutions for sustained, technically-informed discussion of these tools (cf. Simons et al., 2026).

A fourth, more general challenge follows from these first two points. If serious HPSS work with LLMs requires literacy, temporal sensitivity, and a nuanced view of interpretative judgment, then the field needs to co-shape its own tools (cf. Eberle, 2026). This includes defining benchmarks and task formulations that reflect HPSS-specific questions (e.g., Boulanger, 2026), building and curating domain-specific corpora, and finding ways to deal with "ground truth" scarcity in areas where categories are contested or annotations are expensive. It also involves organising prompt, code, and model sharing in ways that make successful workflows visible and reusable, in line with the fourth opportunity above, for example through shared repositories, community guidelines or small purpose-built platforms. A key part of this agenda is to work toward open-science solutions, wherever possible using or co-developing open models and datasets so that HPSS research does not depend too much on closed, evolving systems controlled by a few large providers (Valleriani, 2025). This will often require collaboration beyond HPSS, for instance co-developing general-purpose open models with other scientific communities, or working with neighbouring fields that face similar challenges, such as historical linguistics or digital humanities for historical languages, while focusing HPSS-internal efforts on more specialised models and datasets. Without such collective efforts, HPSS risks either importing ill-fitting tools and evaluation practices from other fields or fragmenting into isolated projects that cannot easily learn from each other.

A fifth challenge is how to engage with LLMs critically, neither demonising them nor using them naively (e.g., Meding and Daugs, 2026b). Beyond technical literacy (see the first challenge above), HPSS needs reflective frameworks for integrating these systems into established methodologies without letting them quietly deflect or erode those methods (e.g., Khutsishvili, 2026). Models can offer seemingly convincing shortcuts, for example quick summaries, instant codings or plausible-sounding explanations, that risk

nudging researchers away from slow reading, archival work or theory-driven interpretation. They can also reconfigure qualitative workflows in ways that shift or endanger the epistemic authority of researchers themselves, raising questions about who is recognised as the primary knower and interpreter in a project (Khutsishvili, 2026; Nguyen and Welch, 2025). At the same time, LLM-driven automation intervenes in very material struggles over scarce academic resources and workloads, fuelling justified concerns about the outsourcing or elimination of research and teaching tasks, especially for early-career and precariously employed scholars. The field will have to decide in what roles LLMs can serve as co-analysts, pattern-finders or sparring-partners, how to disclose and justify their use in publications, and how to resist the pressure to align research questions with what is easy for the models rather than what is substantively important.

Finally, HPSS faces the general challenges that come with LLMs, which are not specific to this field but still shape its choices (cf. Bender et al, 2021; Guest et al., 2025; Lang, 2026; Khutsishvili, 2026; Strubell et al., 2019). On the epistemic side, there is the black-box character of large models and their tendency to hallucinate or fabricate details in ways that are hard to detect and correct. On the structural side, there is dependence on a small number of large technology companies, their changing business models, and their entanglement with political and economic agendas that HPSS itself often studies. On the ethical side, there are issues of energy use and climate impact, the exploitation of human annotation labour, opaque data collection and privacy risks. Engaging with LLMs in HPSS therefore also means taking a position on these wider conditions, or at least making them visible as part of the cost and context of any research that relies on such systems.

Outline of the book

The volume is organised into six parts.

Part 1, "General challenges and limitations", gathers six chapters that set the conceptual and critical scene. **Sarah Lang**'s opening chapter maps key ethical and epistemic risks of bringing proprietary LLMs into humanities workflows, from opaque and biased training data and "open-washing" to exploitative data labour, environmental costs, and threats to research integrity such as hallucinations and paper mills. She critiques explainable AI as insufficient on its own and foregrounds dataset documentation and auditing, care-based and solidaristic approaches, and stronger institutional and funding requirements for accountable AI. **Kristina Khutsishvili** examines how LLMs fracture epistemic authority by shifting scientific authorship from the act of judgement toward the act of approval. Drawing on themes of recognition and "imagined communities", she argues that AI-enabled writing and synthesis can blur boundaries of belonging and reframe scientists as curators of outputs whose genesis is opaque. For HPSS, this both threatens interpretive identity through mimicry and increases its relevance as a source of epistemological orientation about trust, responsibility, and what counts as inquiry. **Oliver Eberle**, writing from both ML and HPSS perspectives, argues that reliable use of frontier and "reasoning" models in HPSS hinges on better evaluation and interpretability. He surveys approaches from behavioral benchmarks to mechanistic explanation

methods, highlights the lack of HPSS-specific datasets and ground-truthed evaluations, and calls for curated materials, culturally and temporally grounded benchmarks, and stronger HPSS participation in AI research and policy to shape trustworthy systems. **Holle Meding and Aurel Daugs** map five recurrent limitations of LLMs for historical scholarship, from hallucinations and chrono-insensitivity to informational presentism and Anglocentrism, alignment effects, and black-box opacity. They argue for a cautious, task-specific use where concepts, prompts, parameters, and evaluation criteria are made explicit, and they illustrate this with an evaluated named entity recognition (NER) workflow on German memory-discourse data that treats LLM outputs as checkable components rather than as autonomous historical narration. **Andreas Wagner and Jürgen Hermes** stage a dialogue about "encoded humanities" that pushes back against the one-size-fits-all, chat-centric drive toward generative systems. They argue that many humanities tasks are better served by encoder-based pipelines that support traceability, minimal computing, and reproducible workflows, while reserving generative models for carefully framed exploratory roles like hypothesis generation or cultural analytics under conditions of openness and critical verification. Finally, **William D'Alessandro** reports a case study in philosophy of mathematical practice, using Gemini to mine thousands of arXiv papers and assemble a large annotated dataset of mathematicians' explanatory aims. He finds that current LLMs can assist impressively with long-context retrieval and disciplined analytical comparison of existing views, yet tend to fail at producing genuinely novel philosophical theories. He argues that this gap is largely a training and evaluation problem: unlike math or coding, philosophical progress offers fewer repeatable, reinforcement-learning-friendly patterns and little consensus ground truth.

Part 2, "Historicizing LLMs", includes four chapters that turn explicitly to temporality. **Jochen Büttner** argues that most LLMs treat texts from many centuries as if they belonged to one undifferentiated "present", which blurs historical change and pushes outputs toward modern language and assumptions, making anachronism a structural risk for historical work. He surveys strategies for temporal conditioning, from period-focused pretraining and time-slicing to time tokens, time embeddings, temporal attention, and temporal RAG, and concludes that time grounding will likely be most feasible and useful in targeted tasks given the scale and metadata hurdles of building time-conditioned foundation models. **Fernanda Olival, Helena Freire Cameron, António Branco, and Renata Vieira** draw on their work with the 18th-century Portuguese Memórias Paroquiais to show how generative LLMs stumble on pluricentric language variation and uneven digital representation across Portuguese varieties, historical spelling and transcription/normalization constraints, and formally structured tasks like named entity recognition, where a Portuguese BERT-like full-context model outperforms generative models. They argue for tailored historical corpora, domain fine-tuning, and hybrid workflows that enforce formal constraints, with close human and multidisciplinary oversight. **Vera Danilova, Julia Reed, Andrew Burchell, Gijs Aangenendt, and Ylva Söderfeldt** test zero-shot synthetic genre generation for 20th-century patient-organisation medical periodicals in four European languages using small open-weight models. Combining historian evaluation with diversity metrics, they find reasonable genre mimicry but weak historical accuracy and strong year and digit preference biases, highlighting a diversity–hallucination trade-off and motivating a cautious view of "historical plausibility" and down-

stream reuse. **Jeffrey C. Wolf** closes the part with a reflection on where LLMs actually fit in multilingual, longue-durée digital history. Rather than chat-style generation, he argues that encoder embedding models are the workhorse for cross-lingual semantic matching and text reuse at scale, but only if they are evaluated and adapted to four hard constraints: strong multilingual coverage, OCR noise, modern out-of-domain bias, and semantic shift across centuries. He outlines the trade-offs in temporal modeling, motivates corpus-specific fine-tuning and preprocessing, and stresses interdisciplinary collaboration and transparent workflow design.

Part 3, "Linguistic change, markers, and differences", brings together six chapters that all use distributional approaches to study scientific language and its evolution. **Christophe Malaterre and Francis Lareau** propose "epistemic framings" as patterned ways disciplines articulate what counts as knowledge and how it is justified, and show how embeddings and LLMs can help detect and map these framings at scale. They argue for hybrid workflows that combine curated epistemic markers with embedding-based expansion, syntactic pattern mining for multiword expressions, and selective LLM annotation or labeling where full-corpus LLM use is still too costly. **Michael Zichert and Arno Simons** offer a programmatic and methodological chapter on computational conceptual history for HPSS. They situate LLM-based methods within a longer trajectory of computational attempts to model conceptual change and use this perspective to highlight the specific opportunities and challenges that LLMs introduce for this undertaking. They argue that contextualized embeddings and hybrid workflows expand how polysemy and change can be modelled, but that corpus construction, concept operationalization, evaluation, and interpretive responsibility remain the decisive bottlenecks, now sharpened by model choice, training data, and infrastructure. **Arno Simons**' chapter then provides a detailed case study. Focusing on the multi-sense term "Planck" in astrophysics and high-energy physics, he compares several general and domain-adapted BERT models, including his own Astro-HEP-BERT, showing how CWEs can disambiguate senses, induce sense clusters and track diachronic shifts in their relative prominence. The chapter demonstrates what highly specialised, domain-adapted models can do for HPSS-style conceptual history and offers practical guidance on model choice, domain adaptation, and validation. Using similar CWE-based methods, **Elaheh Sadat Ahmadi** turns to "codification", understood as the degree to which a field stabilises its terminology. Drawing on sociological debates about codification, she proposes a semantic uniformity score derived from contextual embeddings and applies it to astrophysics and sociology, comparing how consistently key terms are used in each field. The chapter not only reports clear differences in semantic uniformity but also reflects on limits and validation strategies, thus linking classic codification theory to contemporary distributional methods. **Diego Alves, Sergei Bagdasarov, Badr Mohammed Abdullah, and Stefania Degaetano-Ortlieb** use a fine-tuned Llama-3 model to classify Royal Society Corpus articles across more than three centuries (1665–1996) into a small, explicitly defined set of disciplines, treating disciplines as evolving linguistic registers. Through manual annotation, error analysis, and a close look at "out-of-list" labels, they show both strong practical performance and how misclassifications often stem from diachronic lexical and semantic change, shifting genre conventions, and genuinely fuzzy boundaries between disciplinary sublanguages. Using embeddings of

LLM-generated TLDR summaries, they visualise overlaps between disciplinary language varieties as a way to probe interdisciplinarity, and argue that discipline taxonomies are historically loaded instruments that may require period-sensitive refinement. Finally, **Sofia Aguilar-Valdez, Bách Phan-Tất, Dirk Speelman, Dirk Geeraerts, and Stefania Degaetano-Ortlieb** offer a deliberate counterpoint by not using LLMs for their core analysis, arguing that detecting signals of the late 18th-century shift from phlogiston to oxygen requires evidence traceable to a defined corpus. They instead combine LDA topic modeling and information-theoretic change measures with dependency-based slot–filler patterns around "air" and its modifiers, and then contrast this approach with ChatGPT and Grok to clarify what LLMs can supply as macro-narrative versus what they miss when the aim is inspectable, reproducible linguistic markers.

Part 4, "Knowledge graphs and topic modeling", consists of four chapters and moves from word- and phrase-level analysis to structured representations of scientific knowledge and discourse. **Christian Boulanger** shows how building a socio-legal knowledge graph for disciplinary history runs aground on missing and messy citation data in non-English, footnote-heavy law and humanities texts. To tackle this, he develops an open, TEI-annotated gold-standard corpus and sets up an LLM-based extraction and evaluation workflow using LLamore, benchmarking it against Grobid and highlighting trade-offs in accuracy, speed, and compute. He argues that opening bibliometrically neglected domains will require open evaluation data, shared standards for interchange, and collaborative infrastructure building. **Raphael Schlattmann, Aleksandra Kaye, and Malte Vogl** trace the development of a two-stage, human-in-the-loop pipeline that extracts and validates open triples from biographical lexicons and then turns them into an ontology-guided knowledge graph with linking and constraint checks. Using an interview format plus LLM "sparring partner" critiques, they make the iterative toolbuilding process, trade-offs between nuance and operationalisation, and the structuralist commitments of graph-based history unusually explicit. **Francis Lareau and Christophe Malaterre** compare topic modeling on titles, abstracts, and full texts in astrobiology journals, contrasting LDA with embedding-based BERTopic and evaluating models with diversity, coherence, recall, and topic-balance measures. They find that title-only models are generally too information-poor, that abstracts often hit a "Goldilocks" sweet spot with high coherence at far lower cost, and that full texts can improve recall while sometimes reducing coherence due to methodological noise. The chapter argues for goal-driven, and sometimes multi-level, use of different text structures rather than assuming that more text is always better. Finally, **Malte Vogl, Alexander von Schwerin, and Sabrina Kirschke** reflect on an interdisciplinary project that combines BERTopic's embedding-based topic modelling with a generative LLM used to produce concise, meaningful topic labels for qualitative narrative analysis of the German genetic engineering discourse. They show how LLM-assisted labeling plus interactive visualisations can help researchers orient themselves in a large, time-stamped corpus, while also surfacing heavy translation work between developer and users and an 'extra-dark' black box problem around the lack of best-practice norms, missing training, and reproducibility.

Part 5, "Retrieval-augmented generation (RAG)", gathers five chapters that focus on concrete LLM workflows in qualitative and archival research built around RAG architectures. In their second chapter, **Holle Meding and Aurel Daugs** present and critically

evaluate a modular RAG pipeline built on a 100,000+ article corpus from DER SPIEGEL, a major German weekly news magazine, showing how chunking, embeddings, and retrieval settings shape both semantic discovery and source-traceable responses. The pipeline was developed and piloted in a collaborative project with students, and it is evaluated via a pragmatic "silver standard", foregrounding retrieval quality as the key bottleneck, advocating prompts that force uncertainty when evidence is missing, and arguing that RAG only becomes reliable when paired with historian-led evaluation and source criticism. **Miira Hill** introduces the "data interview" as a reflexive, RAG-based approach to qualitative content analysis grounded in the Empirical Theory of Science (EToS), which treats knowledge as socially produced and stabilized through communicative practices, institutions, and infrastructures. Accordingly, LLM outputs are handled as situated, dialogical artifacts that require interpretation and validation rather than being taken as results. In a case study of German Facebook comments on migration and history, she shows how multi-interviewer prompting, corpus-grounding, and team-based coding can scale interpretation while surfacing persistent risks like misclassification, repetitive evidence, and normative "false balance" in sensitive content. **Andrea Scharnhorst, Han Yang, Jetze Touber, Kim Ferguson, Philipp Mayr, and Vyacheslav Tykhonov** situate RAG in the context of cultural-heritage and research-data infrastructures. Tracing an engineering journey from "archives for everyone" to "chatting with collections" on the Dataverse platform and in the MuseIT project, they show how local RAG chatbots that combine LLMs with knowledge-organisation systems can empower curators and users of cultural data, provided they are co-designed with galleries, libraries, archives, and museums and embedded in open, interoperable and commons-oriented infrastructures. **Gerd Graßhoff** introduces the AI-Reporter pipeline, which takes slides and recordings of academic talks and turns them into structured, publication-ready chapters. Using the opening lecture of our Berlin workshop as an example, he demonstrates how multimodal LLM pipelines can create a new genre of scientific communication that bridges live presentation and written text, while still relying on human editorial control for curation and quality. **Jeffrey C. Wolf**'s second contribution is an experiment in recursive reflection. Adapting an open-source NotebookLM RAG variant to run locally, he generates extended scholarly dialogues about his own workshop talk and uses them to explore what happens when we use LLMs to discuss LLMs in HPSS. The resulting conversations are technically impressive yet shallow, and he argues that such computational reflexivity, while playful and revealing, still falls short of the deeper, situated reflexivity that human scholars practice.

Part 6, "Citation context analysis (CCA)", includes three chapters that turn to an established method at the intersection of scientometrics and the sociology of science and explores how LLMs challenge and transform it. **Lydia Liesegang and Jochen Gläser** argue that LLM-based automation turns CCA itself back into an object of inquiry, because LLM support forces explicit rules that expose CCA's tacit assumptions. They challenge three assumptions: that nearby text reliably says something about the cited work, that coding schemes can ignore missing values, and that fixed sentence or paragraph windows match the real unit of analysis, the argument. They also warn about circular validation: if the "gold standard" annotations were produced using the same flawed rules, an LLM can score well by copying them while still misreading what the citation is doing. They

conclude that useful LLM support needs theory-driven CCA designs and evaluation that checks argumentative meaning, not just agreement with legacy labels. **Arno Simons, Hiba Arnaout, and Iryna Gurevych** propose a roadmap for "reconstructive" CCA that treats citation meaning as plural and inter-contextual, emerging across citing passages, cited sources, and patterns of uptake over time. They translate these aims into LLM-supported workflows from context detection and classification to citation–entity and citation–argument mining, similarity and clustering of citation framings, and chained, time-aware interpretative readings that surface alternatives while keeping human judgment, evidence visibility, and reproducible settings central. In doing so, they also sketch several forward-looking tasks that are still rare or largely unexplored in computational CCA, especially those aimed at tracing evolving meanings and argumentative linkages rather than refining evaluative metrics. Overall, they argue that LLMs can widen the scale and nuance of citation interpretation only when used reflexively as tools for inquiry rather than as mere substitutes for the hermeneutic work of researchers. In the final chapter, **Arno Simons** argues for "scaling in" thick, case-based citation interpretation, testing it on a canonical hard case from the CCA literature. He develops a two-stage GPT-5 workflow that pairs a stable surface label with cross-text checking and multiple, text-grounded interpretative hypotheses. A 2×3 prompt-variation study then shows how scaffolding and framing steer which plausible readings and vocabularies the model foregrounds, reinforcing that LLMs work best as inspectable co-analysts whose outputs remain curated and judged by humans.

In sum, the contributions in this volume probe and stretch traditional HPSS methodologies and invite readers to engage critically with a new item in their toolkit: LLMs. Taken together, they exemplify the openness and reflexivity toward LLM use that we hope this volume will foster.

Use of LLMs in the production of this volume

Parts of this introduction, as well as many of the chapters in this volume, were written with support from LLMs, such as OpenAI's ChatGPT. Any text produced with LLM assistance was reviewed, edited and, where necessary, substantially rewritten by the authors, who bear full responsibility for the published versions. Where particular uses of LLMs are methodologically relevant, they are documented in the respective chapters.

References

Aguilar-Valdez S, Phan-Tất B, Speelman D, et al. (2026) Discursive parallels of the chemical revolution. Topic modelling and distributional analysis. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

Ahmadi ES (2026) Exploring disciplinary differences in semantic uniformity. A computational approach to codification. In: Simons A, Wüthrich A, Zichert M, et al. (eds)

Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science. Bielefeld: transcript, part-3.

Alves D, Bagdasarov S, Abdullah BM, et al. (2026) Use of large language models in the classification of scientific texts into disciplines. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

arXiv (2025) Attention Authors: Updated Practice for Review Articles and Position Papers in arXiv CS Category – arXiv blog. Available at: https://blog.arxiv.org/2025/10/31/attention-authors-updated-practice-for-review-articles-and-position-papers-in-arxiv-cs-category/.

Bender EM, Gebru T, McMillan-Major A, et al. (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*, 2021, pp. 610–623.

Binz M, Alaniz S, Roskies A, et al. (2025) How should the advancement of large language models affect the practice of science? *Proceedings of the National Academy of Sciences* 122(5).

Boulanger C (2026) The potential of LLMs for constructing a socio-legal knowledge graph. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-4.

Brainard J (2025) As scientists explore AI-written text, journals hammer out policies. *Science*, 10 December. Available at: https://www.science.org/content/article/new-preprint-server-welcomes-papers-written-and-reviewed-ai.

Brown TB, Mann B, Ryder N, et al. (2020) Language Models are Few-Shot Learners. http://arxiv.org/abs/2005.14165.

Büttner J (2026) Why pursue temporally-grounded AI for historical disciplines, and what makes it so challenging? In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-2.

Christou PA (2025) *Artificial Intelligence (AI) in Social Research*. CABI.

Clark K, Luong M-T, Le QV, et al. (2020) Electra: Pre-training text encoders as discriminators rather than generators. http://arxiv.org/abs/2003.10555.

D'Alessandro W (2026) LLMs as philosophers: what can they do, and why aren't they better? In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Danilova VV, Reed J, Burchell A, et al. (2026) Zero-shot generation of synthetic historical data with LLMs. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-2.

Davidson T and Karell D (2025) Integrating Generative Artificial Intelligence into Social Science Research: Measurement, Prompting, and Simulation. *Sociological Methods & Research*.

Devlin J, Chang M-W, Lee K, et al. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. http://arxiv.org/abs/1810.04805.

Eberle O (2026) Grounding AI in humanistic inquiry. Interdisciplinary challenges for evaluation and interpretability. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Ehrlich N and Hazzan O (2025) The Converging Paths of Computer Science and the Humanities in the Age of GenAI – Communications of the ACM. In: *Artificial Intelligence and Machine Learning*. Available at: https://cacm.acm.org/blogcacm/the-converging-paths-of-computer-science-and-the-humanities-in-the-age-of-genai/.

Ethayarajh K (2019) How Contextual are Contextualized Word Representations? Comparing the Geometry of BERT, ELMo, and GPT-2 Embeddings. http://arxiv.org/abs/1909.00512.

Graßhoff G (2026) AI-Reporter: a path to a new genre of scientific communication. From presentation to publication through agentic LLMs. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-5.

Guest, O., Suarez, M., Müller, B., et al. (2025). Against the Uncritical Adoption of "AI" Technologies in Academia. https://doi.org/10.5281/zenodo.17065099.

Jawahar G, Sagot B and Seddah D (2019). What does BERT learn about the structure of language? *ACL 2019–57th Annual Meeting of the Association for Computational Linguistics*.

Friese S and Morgan D (eds) (2026) *Qualitative Data Analysis with Artificial Intelligence: Theory, Methods and Practice*. SAGE.

Gao Y, Xiong Y, Gao X, et al. (2024) Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997. http://arxiv.org/abs/2312.10997.

Grootendorst M (2022) BERTopic: Neural topic modeling with a class-based TF-IDF procedure. http://arxiv.org/abs/2203.05794.

Guo T, Chen X, Wang Y, et al. (2024) Large Language Model based Multi-Agents: A Survey of Progress and Challenges. http://arxiv.org/abs/2402.01680.

Han B, Susnjak T and Mathrani A (2024) Automating systematic literature reviews with retrieval-augmented generation: a comprehensive overview. *Applied Sciences* 14(19).

He P, Liu X, Gao J, et al. (2021) DeBERTa: Decoding-enhanced BERT with Disentangled Attention. http://arxiv.org/abs/2006.03654.

He P, Gao J and Chen W (2022) DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In: *The Eleventh International Conference on Learning Representations*, 29 September 2022. Available at: https://openreview.net/forum?id=sE7-XhLxHA.

Hemment D and Kommers C (2025) Doing AI Differently. Rethinking the foundations of AI via the humanities. The Alan Turing Institute, London, UK. Available at: https://www.turing.ac.uk/news/publications/doing-ai-differently.

Hill M (2026) The data interview. Reflexive integration of large language models in qualitative content analysis. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-5.

Hitch D (2024) Artificial Intelligence Augmented Qualitative Analysis: The Way of the Future? *Qualitative Health Research* 34(7): 595–606.

Lund BD, Wang T, Mannuru NR, et al. (2023) ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology* 74(5): 570–581.

Ho X, Nguyen AKD, Dao AT, et al. (2024) A Survey of Pre-trained Language Models for Processing Scientific Text. http://arxiv.org/abs/2401.17824.

Huang X, Liu W, Chen X, et al. (2024) Understanding the planning of LLM agents: A survey. http://arxiv.org/abs/2402.02716.

Jowsey T, Braun V, Clarke V, et al. (2025) We reject the use of generative artificial intelligence for reflexive qualitative research, SSRN Scholarly Paper. Available at: https://papers.ssrn.com/abstract=5676462.

Khutsishvili K (2026) AI and the scientist. On the fracture of epistemic authority. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Lang S (2026) Critical concerns for using LLMs in the (computational) humanities and beyond. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Lewis M, Liu Y, Goyal N, et al. (2019) BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. http://arxiv.org/abs/1910.13461.

Li D, Jiang B, Huang L, et al. (2025) From Generation to Judgment: Opportunities and Challenges of LLM-as-a-judge. In: *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (eds C Christodoulopoulos, T Chakraborty, C Rose, et al.), Suzhou, China, November 2025, pp. 2757–2791. Association for Computational Linguistics. Available at: https://aclanthology.org/2025.emnlp-main.138/.

Liesegang L and Gläser J (2026) Supporting citation context analysis with large language models raises questions that should have been asked 40 years ago. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Liu Y, Ott M, Goyal N, et al. (2019) RoBERTa: A Robustly Optimized BERT Pretraining Approach. http://arxiv.org/abs/1907.11692.

Lozić E and Štular B (2023) Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities. *Future Internet* 15(10).

Luo Z, Yang Z, Xu Z, et al. (2025) LLM4SR: A Survey on Large Language Models for Scientific Research. http://arxiv.org/abs/2501.04306.

Malaterre C and Lareau F (2026) Epistemic framings in science. Charting scientific knowledge with embeddings and LLMs. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

Marx V (2024) Quest for AI literacy. *Nature Methods* 21(8): 1412–1415.

Meding H and Daugs A (2026a) From RAGs to rich responses. Enhancing LLM reliability through retrieval-augmented generation. In: Simons A, Wüthrich A, Zichert M,

et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-5.

Meding H and Daugs A (2026b) On the use and limitations of large language models in historical scholarship. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Morgan DL (2023) Exploring the Use of Artificial Intelligence for Qualitative Data Analysis: The Case of ChatGPT. *International Journal of Qualitative Methods* 22: 16094069231211248.

Nguyen DC and Welch C (2025) Generative Artificial Intelligence in Qualitative Data Analysis: Analyzing—Or Just Chatting? *Organizational Research Methods*: 10944281251377154.

Olival F, Cameron HF, Branco A, et al. (2026) Generative LLMs and history research. Limitations for languages, periods, and tasks. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-2.

Radford A, Narasimhan K, Salimans T, et al. (2018) Improving Language Understanding by Generative Pre-Training. OpenAI. Available at: https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

Radford A, Wu J, Child R, et al. (2019) Language models are unsupervised multitask learners. OpenAI. Available at: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Reimers N and Gurevych I (2019) Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. http://arxiv.org/abs/1908.10084.

Ren S, Jian P, Ren Z, et al. (2025) Towards Scientific Intelligence: A Survey of LLM-based Scientific Agents. http://arxiv.org/abs/2503.24047.

Scharnhorst A, Yang H, Touber J, et al. (2026) Co-creation of AI technology, empowering curators of cultural heritage information and guarding research commons. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-5.

Scherbakov D, Hubig N, Jansari V, et al. (2024) The emergence of Large Language Models (LLM) as a tool in literature reviews: an LLM automated systematic review. http://arxiv.org/abs/2409.04600.

Schlattmann R, Kaye A and Vogl M (2026) From source to structure. Extracting knowledge graphs with LLMs. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-4.

Schmidgall S, Su Y, Wang Z, et al. (2025) Agent Laboratory: Using LLM Agents as Research Assistants. http://arxiv.org/abs/2501.04227.

Simons A (2026a) Meaning at the Planck scale? Contextualized word embeddings for doing history, philosophy, and sociology of science. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

Simons A (2026b) Scaling In, Not Up? Testing Thick Citation Context Analysis with GPT-5 and Fragile Prompts. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Simons A, Arnaout H and Gurevych I (2026) Reconstructive citation context analysis using large language models. A roadmap. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-6.

Simons A, Zichert M and Wüthrich A (2026) Large language models for history, philosophy, and sociology of science: Interpretive uses, methodological challenges, and critical perspectives. *Studies in History and Philosophy of Science* 117: 102151. https://doi.org/10.1016/j.shpsa.2026.102151.

SocOpen (2025) Moderation Policy. In: *SocOpen: Home of SocArXiv*. Available at: https://socopen.org/moderation-policy/.

Strubell E, Ganesh A and McCallum A (2019) Energy and Policy Considerations for Deep Learning in NLP. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (eds A Korhonen, D Traum, and L Màrquez), Florence, Italy, July 2019, pp. 3645–3650.

Thapa S, Shiwakoti S, Shah SB, et al. (2025) Large language models (LLM) in computational social science: prospects, current state, and challenges. *Social Network Analysis and Mining* 15(1): 4.

Tran K-T, Dao D, Nguyen M-D, et al. (2025) Multi-Agent Collaboration Mechanisms: A Survey of LLMs. http://arxiv.org/abs/2501.06322.

Valleriani M (2025) Large language models that power AI should be publicly owned. *The Guardian*, 26 May. Available at: https://www.theguardian.com/technology/2025/may/26/large-language-models-that-power-ai-should-be-publicly-owned.

Varnum MEW, Baumard N, Atari M, et al. (2024) Large Language Models based on historical text could offer informative tools for behavioral science. *Proceedings of the National Academy of Sciences* 121(42): e2407639121.

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. *Advances in neural information processing systems* 30.

Vogl M, von Schwerin A and Kirschke S (2026) Large language models in interdisciplinary research settings. A reflection. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-4.

Wagner A and Hermes J (2026) Encoded humanities, or: not everything has to be generative. A dialogue on AI tasks and roles. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-1.

Wei J, Wang X, Schuurmans D, et al. (2022) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *Advances in Neural Information Processing Systems* 35: 24824–24837.

Wolf JC (2026) LLMs and multilingual historical corpora in a digital history project. Reflections from the Berlin workshop. In: Simons A, Wüthrich A, Zichert M, et al. (eds)

Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science. Bielefeld: transcript, part-2.

Yin S, Fu C, Zhao S, et al. (2024) A Survey on Multimodal Large Language Models. http://arxiv.org/abs/2306.13549.

Zhang H, Wu C, Xie J, et al. (2025b) Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. *Computers in Human Behavior: Artificial Humans* 4: 100144.

Zhang Y, Chen X, Jin B, et al. (2024) A Comprehensive Survey of Scientific Large Language Models and Their Applications in Scientific Discovery. http://arxiv.org/abs/2406.10833.

Zhang Y, Zhang C and Kousha K (2025a) Editorial: artificial intelligence for scientometrics (Part I). *Scientometrics* 130(10): 5281–5284.

Zhong T, Yang Z, Liu Z, et al. (2025) Opportunities and Challenges of Large Language Models for Low-Resource Languages in Humanities Research. http://arxiv.org/abs/2412.04497.

Zhuang Z, Chen J, Xu H, et al. (2025) Large language models for automated scholarly paper review: A survey. *Information Fusion*: 103332.

Zichert M and Simons A (2026) From early digital methods to LLMs. Computational conceptual history of scientific concepts. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

Zichert M, Simons A and Wüthrich A (2025) Expanding conceptual histories: using contextualized word embeddings for the history and philosophy of the virtual particle concept. *Computational Humanities Research* 1: e16.

Ziems C, Held W, Shaikh O, et al. (2024) Can large language models transform computational social science? *Computational Linguistics* 50(1). MIT Press: 237–291.