On the use and limitations of large language models in historical scholarship

Holle Meding and Aurel Daugs

1. Introduction

Two opposing tendencies currently shape the response to artificial intelligence in the humanities. On one hand, some have quietly integrated Large Language Models (LLMs) into their daily routines, using them for summarising articles, drafting research proposals or a preliminary contextualisation of historical events. However, this usage often goes unacknowledged.

On the other hand, institutions and individual scholars have reacted by drawing strict boundaries. While outright institutional bans remain rare, many departments and faculty members have expressed reluctance to incorporate LLMs into research or teaching.[1] The concerns are understandable. There are questions surrounding authorship and attribution, the presence of hallucinations, overgeneralisations, and difficulties in evaluating model outputs. In some cases, there is fear of academic erosion. If we cannot be sure what role LLMs played in producing scholarly work, how can we fairly assess it?

But between silent adoption and categorical exclusion lies a necessary middle ground, given that LLMs are already exerting a disruptive shift on scientific practice, while in the domain of historical research they are changing how sources are digitised, translated, cross-referenced, and contextualised. Ignoring them does not make the problem disappear. It simply moves the conversation elsewhere.

1 The Paris-based Sciences Po has enacted a policy banning the undisclosed use of tools like ChatGPT in written tasks and oral presentations, and multiple Danish institutions, notably Aarhus University and the Technical University of Denmark, have similarly prohibited un-attributed AI use in student submissions (Schwartmann, 2024; Kurth and Köhler, 2024; Kyklebust, 2023). At many German universities, including Freie Universität Berlin and the University of Cologne, coursework essays, which must traditionally be submitted with a signed declaration of authorship confirming the student's independent work, are now prompting a revision of these statements to include explicit disclosure of whether, and for which tasks, AI tools were employed.

For scholars to engage with the model's impact on scholarly work, the question is not whether to respond, but how. How should we evaluate the reliability of a system that does not cite its sources? What are the epistemological implications of interpreting the past through models shaped by opaque training logics? And how can we identify the areas in which LLMs provide genuine hermeneutic value, rather than being misled by their surface-level coherence?

Guided by these concerns, this article situates itself within the broader discourse of the History, Philosophy, and Sociology of Science (HPSS) and examines how the integration of LLMs is reshaping contemporary research practices. Within this framework, we focus particularly on historical scholarship, using it as the primary field through which to assess both the potentials and the pitfalls of these models for academic research.

Rather than calling for wholesale adoption or outright rejection, we explore how researchers can engage with these models in a way that is practically feasible while maintaining a critical stance. Drawing on recent empirical studies and our own experiments, we identify five recurring areas of concern. First, hallucinations, where fabricated citations, sources or causalities are presented as if factual. Second, chrono-insensitivity, where the model lacks explicit time awareness. Third informational presentism and Anglocentrism driven by the dominance of post-2000 Anglophone internet sources in the training data. Fourth, human alignment and policy-driven moderation that suppress or distort politically or ethically sensitive content. And fifth, opacity, that is the limited access to training data, fine-tuning processes, or alignment protocols that determine what a model can or cannot say. Yet these limitations do not render LLMs unusable, neither for historians nor for HPSS scholars. On the contrary, they can offer clear benefits when used in narrowly defined tasks where data, evaluation criteria, prompt structure and the expected outcomes are conceptually predefined and formalised. In these cases, LLMs enable the analysis of large, heterogeneous datasets while offering the possibility of reducing both preprocessing requirements and human labour costs. Their tolerance for variation in spelling, dialect, and Optical Character Recognition (OCR) errors makes them particularly useful for historical texts, although their performance depends heavily on the context and the nature of the sources (Wolf, 2026). They support fast, iterative development, since they are prompt-driven, rather than retraining-dependent.

We therefore advocate for a task-specific and domain-sensitive approach, that treats LLMs as *methodological components* to be specifically tested and adapted within research workflows (Oberbichler and Petz, 2025). To illustrate this, we present an example from our own work: the use of GPT-4o for Named Entity Recognition (NER) in a corpus of user-generated social media posts on memory discourses surrounding the former German Democratic Republic (GDR). This small case study – though based on sources of digital memory culture – demonstrates how prompting, annotation, and model evaluation can be combined to enable the effective application of LLMs within a defined research context.

2. Hallucinations and factual accuracy

One of the primary concerns with LLMs in the digital humanities is their tendency to produce *hallucinations*, a term describing instances where a model generates an often seemingly plausible, but nonsensical output or one that is unfaithful to the provided source context (Huang, 2023: 5; Kalai, 2025). Hallucinations tend to take the form of fabricated sources, misattributed quotations, or unverifiable causal claims. The precise cause of any given hallucination is difficult to isolate, but several frequently contributing factors have been observed. In particular, hallucinations arise not from an anthropomorphized notion of faulty perception, but in most cases from an interplay of probabilistic token and sequence prediction, the representational limits of the training data, and the inherent mismatch between next-token likelihood (the training loss) and human expectations of factual accuracy.

Text from digitised and online sources generally follows a heavy-tailed distribution, where a small set of popular topics appears very frequently, while the majority of entities and events appear only rarely. During training, when a model encounters a rare historical subject, it has far fewer examples to draw on, and thus forms a weaker or less reliable internal representation of that knowledge. While advanced models like GPT-4o or Claude 3 Opus handle well-documented events, such as major aspects of the Second World War, with relative accuracy, likely due to their extensive representation in training datasets, their performance deteriorates markedly with more obscure historical subjects or less frequently documented events. In these 'long-tail' cases, the model may overgeneralise from thematically related but historically unrelated material (Kandpal, 2023). For example, when prompted about an obscure peasant revolt in 17th-century Transylvania, it might inappropriately synthesise details from other European uprisings, disregarding the different context. In these cases, the output often contains anachronistic details or fabricated connections between contexts and events.[2]

From an HPSS perspective, this illustrates a shift from evidential accountability to statistical plausibility. Models that derive meaning from co-occurrence patterns reproduce the frequency biases of their training data, prioritising what is most often written, over what is most significant. The implications extend beyond historical representation and concern the very processes through which knowledge is validated, circulated, and reproduced within scientific and scholarly infrastructures.

Inference strategies further amplify this tension, because LLMs output just one sample from a given probability distribution. This makes the choice of the decoding method particularly important. Stochastic approaches, such as temperature scaling, can increase the likelihood of improbable and often undesirable continuations, thereby

2 To mitigate some of these shortcomings, retrieval-augmented generation (RAG) (Lewis et al., 2020; Yue et al., 2020; Meding and Daugs, 2026) could be combined with automated context validation mechanisms such as those proposed by Birur et al. (2024) in their VERA framework. Generated responses are anchored in cited primary sources and cross-checked in real time against structured historical databases (cf. Hauser et al., 2024; Yue et al., 2024). Additionally, fact-toggle prompting, allowing users to switch between open-ended exploratory generation and strictly grounded factual modes, could offer more control over model behaviour.

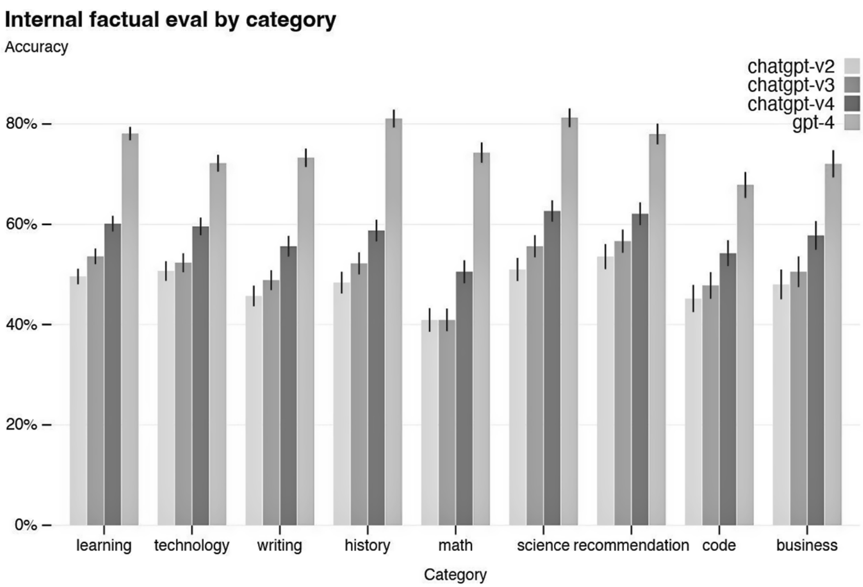
introducing significant output variability (Peeperkorn, 2024). Deterministic (*greedy*) decoding can produce 'confident' mistakes, since selecting the most statistically likely continuation does not automatically guarantee it is factually correct. This is because the model's apparent 'confidence' (high probability) only reflects the prevalence of learned patterns in the training data, independent of their actual informational relevancy or factual accuracy. Therefore, if the underlying data is deficient, the most probable continuation will, by definition, only be the most probable deficiency.

Recent technical developments have shown mixed results in addressing these concerns. OpenAI's technical report on GPT-4 claims substantial improvements in 'factual accuracy', particularly in the not further defined category of 'history', as compared to previous iterations (OpenAI, 2024: 10f; Fig. 1). However, it is crucial to note that this comparison is not made between base models, but between GPT-4 and ChatGPT, the latter being a version already fine-tuned and aligned for chatbot interactions. These reported gains must therefore be viewed with caution. As will be further discussed in a later chapter, through Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022), models like ChatGPT are designed to avoid generating outputs on sensitive or controversial topics, e.g., those involving violence, abuse, or other forms of historical atrocities. This filtering, while ethically motivated, can reduce the model's willingness to engage with complex or politically sensitive historical material, thereby limiting its overall performance in certain domains or even introducing their own forms of hallucinations (Huang, 2023: 11f.). Distinctions between base and fine-tuned models, as well as between general-purpose and safety-aligned versions, thus significantly shape how each model performs when applied to historical research and, more broadly, how LLMs condition epistemic practices, highlighting their relevance to HPSS. LLMs should therefore not be evaluated only in general terms, but with careful attention to the specific model being used and its particular training objective.

A more systematic attempt to evaluate the historical factual accuracy of LLMs is found in a recent study by Hauser et al. (2024), which assessed seven models developed by OpenAI, Meta, and Google using the Seshat Global History Databank as a benchmark (HiST-LLM). Although all models performed above the baseline of random guessing (25%), their results fell significantly short of expert-level responses. GPT-4-Turbo achieved the highest balanced accuracy at 46%, whereas Llama-3.1-8B reached only 33.6%. Notably, the models tended to perform better on questions related to early history (pre-3000 BCE), while their reliability declined sharply for more recent historical periods. Regional disparities were also pronounced as questions concerning North and Latin America were answered with greater precision, whereas topics related to Sub-Saharan Africa and Oceania consistently yielded the weakest results (Hauser et al., 2024: 8f.).

Although this benchmark represents a significant contribution to evaluating historical factual accuracy, it also illuminates a methodological tension in applying LLMs to historical research. While basic factual knowledge – dates, names, locations – certainly provides the essential scaffolding upon which historical understanding is built, the benchmark's approach treats history as a static collection of verifiable facts rather than recognizing it as an ongoing discourse of interpretation and contextual understanding. This distinction matters, since historical scholarship rarely stops at establishing 'what happened' but proceeds to the more complex questions of why events unfolded as they did, how different actors understood their circumstances, and what meanings these developments held for contemporary and subsequent observers. Beyond the technical limitations of multiple-choice formats, which themselves can introduce various forms of bias (Khatun and Brown, 2024), and the dataset's grounding in predominantly English-language sources, despite its global aspirations (Hauser et al., 2024: 5f.), the benchmark fundamentally cannot assess whether LLMs can engage productively with the interpretative, discursive nature of historical scholarship. The transition from factual recall to historical hermeneutics involves a qualitative shift that standardised testing approaches struggle to capture.

Fig. 1: Performance of GPT-4 on nine internal adversarially-designed factuality evaluations. (OpenAI 2024, 10).



Precisely this tension between surface-level factual reproduction and the interpretative, often theory-laden analysis characteristic of historical and critical HPSS scholarship reveals a central limitation of evaluating text-generating systems. Unlike fields where factual accuracy can serve as a reliable performance metric, historical inquiry is situated in a context where even seemingly straightforward questions carry weight. If, for instance, one asks an LLM the rather naïve question of whether the GDR (German Democratic Republic – East Germany) was a 'state of injustice' ('Unrechtsstaat'), any response presupposes multiple unspoken assumptions: What precisely is meant by the term 'Unrechtsstaat'? Is the question supposed to be approached from a political, legal, or eyewitness perspective? Can such a question even be adequately answered with a simple 'Yes', or 'No'? And is it even possible to provide an objectively 'correct' answer to a question that is politically and morally charged?

In practice, historians are operating in a domain shaped by plausibility and competing narratives, based on representative evidence, where meaning emerges not through statistical likelihoods, but through reasoning, contextual analysis, source criticism, and hermeneutic reflection. In this absence of an objective ground truth – that is, a singularly 'correct' version of history – measuring the reliability of model-generated information becomes especially complex and reflects the contested nature of historical knowledge itself, where different scholars may legitimately reach different conclusions based on the same evidence (White, 1997).[3]

Depending on the model architecture and the framing of the prompt, the generated output can therefore differ significantly, without such variation necessarily indicating factual error or misrepresentation. In this context, LLMs may reproduce dominant historiographical perspectives, for instance, privileging Western or liberal-democratic viewpoints, while marginalising alternative framings, such as decolonial or non-Western accounts. These tendencies reflect the biases embedded within the training data, the broader asymmetries in source availability, and the resulting knowledge production. In other cases, they may be deliberately trained to omit or censor particular events.

It is therefore essential to recognise that LLMs are themselves cultural artefacts. As such, they must be critically examined with due attention to their inherent limitations and biases. After all, LLMs only produce one possible reading of a text or statement among many, shaped by the statistical patterns of their training data and based on their underlying linguistic model (Hiltmann, 2024a: 228f). In this sense, the integration of LLMs into research renews the call for methodological formalisation, requiring scholars in the digital humanities to articulate with precision what kind of output they seek, and demands a heightened awareness of the implicit assumptions and epistemic biases they may be adopting in the process.

Extending this line of thought from the perspective of HPSS and in view of the processes through which knowledge is formed and transformed, this moment can be understood as one of epistemic reconfiguration, in which researchers are compelled to acquire new forms of methodological competences. For historians, this development repositions their professional role from sole interpreter of the past to what might be termed a critical mediator between algorithmic outputs and scholarly interpretation, responsible for evaluating and contextualising computationally generated insights, while remaining alert to both the opportunities and limitations such tools present for advancing historical understanding.

3 We encountered a similar challenge when evaluating the RAG pipeline, developed at the Chair for Digital History at Humboldt University of Berlin. The system integrates generative models (GPT-4o, HU-LLM-3) with vector-based retrieval from a corpus of over 100,000 DER SPIEGEL articles. Applied to the theme of decolonisation in Asia and Africa between the 1940s and 1970s, the absence of a fixed historical *Ground Truth* necessitated the construction of a transparent silver standard, though the evaluation process continued to pose significant methodological challenges.

3. Limited chrono-sensitivity

Among the many dimensions of human reasoning that LLMs struggle to reproduce, temporal awareness is perhaps the most consequential for historical inquiry. During pre-training, no systematic effort is made to weigh texts according to their historical or chronological origin. As a result, these models are unable to distinguish whether a given source was published in 1980 or 2010 as such information is typically not encoded in the training data. The training process flattens chronological hierarchy into a single, undated representational space. Even though linguistic patterns persist, temporal differentiation is lost.

This design reflects an architectural constraint as transformer-based LLMs embed tokens according to co-occurrence statistics without accounting for the temporal sequence of ideas or concepts. Word embeddings collapse diachronic variation into static representations, masking semantic shifts across time. Moreover, temporal signals in pre-training corpora are neither normalised nor preserved. Sentence ordering, when present, only weakly correlates with event sequence. In other words, LLMs are trained on language abstracted from time (Büttner, 2026).

Even state-of-the-art models like GPT-4 and LLaMA 2 struggle with commonsense knowledge about the typical duration, occurrence, frequency of events and the ability to order them along a timeline. As demonstrated by Qiu et al. (2024), in over 27% of cases, these models generate temporally incoherent outputs, such as, asserting contradictory before-after-relations. The in-context learning, instruction tuning, and chain-of-thought prompting techniques offer only modest improvements, and scaling the model size does not result in more reliable results. In fact, small-scale and specialised language models were found to often outperform their larger, general-purpose counterparts. For historical questions requiring detailed or temporally precise reasoning, LLMs perform significantly worse, with accuracy dropping up to 35% for relative time references and 55% when absolute cues are corrupted (Wallat et al., 2025).[4]

Extending this critique, Herel et al. (2025) introduce *TimeShift*, a log-probability-based benchmark evaluating whether models adjust their probabilistic predictions and hence their temporal claims when presented with shifting time contexts or temporal prefixes. Consequently, sentences such as 'Donald Trump is the US president' versus 'Joe Biden is the US president', are evaluated not through surface-level token matching, but through how confidently the model ranks them across historical periods. Models like LLaMA 3.1 8B demonstrate partial temporal adjustment, but performance remains brittle, especially under paraphrase or fine-grained dating. Crucially, Herel et al. show that current benchmarks underestimate the complexity of temporal reasoning by focusing on coarse-grained QA metrics, rather than on the underlying representational instability of time in LLMs.

4 Their findings suggest models handle recent knowledge better, but statistical biases in the training data can distort this, occasionally favouring outdated over newer information. To address these failures, including persistent temporal shifts and memory rigidity, they advocate modelling both the *creation* and *focus time* of training data to anchor LLM outputs in stable temporal frames. (Wallat et al., 2024; Wallat et al., 2025).

From an HPSS perspective, this problem extends to the epistemology of scientific representation itself. In the history and philosophy of science, time is a constitutive condition for understanding change, succession, and causality. The flattening of temporal order in LLMs echoes long-standing concerns in the study of scientific representation, where processes of abstraction and modelling often detach knowledge from its temporal and material conditions (Serres and Latour, 1995; Galison, 1997; Daston, 2017). Like the scientific models before them, LLMs stabilise phenomena by erasing their historical becoming, producing what Daston has described as an archival rationality of knowledge. But how then, might we confront the temporal shortcomings of LLMs, given that their current architectures neglect precisely this dimension? One promising approach, *Counterfactual Consistency Prompting*, generates bidirectional event-order questions, such as 'Did A occur before B?' and 'Did A occur after B?' and enforces logical alignment across both. This method significantly reduces event-sequencing contradictions across multiple datasets (Kim and Hwang, 2025), but its effectiveness declines when absolute temporal markers such as calendar years are involved. LLMs do not perform arithmetic or date-based reasoning in a rule-based way, but rather approximate such reasoning probabilistically, based on patterns seen during training. As a result, when tasked with interpreting or comparing exact dates (e.g., 1914 vs. 1939), the model cannot calculate or infer durations or sequences reliably, unless such relations have been explicitly encoded in its training data or it is augmented with external tools such as symbolic calculators or date-handling modules (Su et al., 2024; Zhu et al., 2023).

Instead of relying solely on textual prompts, other approaches seek to embed temporal structure into the training data or model architecture itself. In their contribution, Büttner (2026) proposes integrating time directly into the token embedding space as a contextual variable, allowing models to condition their outputs not just on linguistic form, but also on the historical setting in which a term or event occurs. His approach reframes chrono-insensitivity not as a reasoning deficit, but as a consequence of training LLMs on temporally heterogeneous corpora without representing time as a privileged coordinate in the learning process.[5]

This perspective calls for a more precise formulation of the requirements that LLMs must meet in historical research. If the lack of temporal reasoning capabilities stems from architectural constraints and the absence of temporally encoded representations, then future works must specify what kinds of chrono-sensitive capacities language models for historical research must implement. This includes, beyond timeline reasoning or date comparison, the ability to represent diachronic semantic drift, capturing how terms and concepts such as 'revolution' or 'liberty' evolve across linguistic and sociopolitical contexts.

Notably, Yein Park et al. (2025a) state in their paper *Does Time Have Its Place?* that, as models scale, they appear to develop *emergent abilities* (Wei et al., 2022) not explicitly en-

5 Their findings suggest models handle recent knowledge better, but statistical biases in the training data can distort this, occasionally favouring outdated over newer information. To address these failures, including persistent temporal shifts and memory rigidity, they advocate modelling both the *creation* and *focus time* of training data to anchor LLM outputs in stable temporal frames. (Wallat et al., 2024; Wallat et al., 2025).

gineered into their architectures, temporal reasoning among them. Certain models such as Llama-2-7b-chat-hf, Qwen1.5-7B-Chat and Phi-3-mini-4k-instruct appear to exhibit a form of temporal awareness, mediated by what the authors term *temporal heads*: namely, specialised attention heads attuned to chrono-sensitive patterns in language. For historians, this raises intriguing possibilities. If these temporal heads can be reliably identified and fine-tuned, it may become feasible to guide model outputs toward greater sensitivity to historical chronology and context. However, this phenomenon remains undertheorised, and substantial empirical work is required before it can be operationalised in practice.

By contrast, the broader literature suggests that consistent temporal reasoning in LLMs cannot be achieved through fine-tuning or prompt-level heuristics alone, but depends on embedding temporality structurally, whether through graph-based representations, reinforcement learning objectives, explicitly staged training protocols, or temporal conditioning integrated directly into the embedding space. In this spirit, a working group at the Interdisciplinary Centre for Digitality and Digital Methods (IZD2MCM, Humboldt University of Berlin) is currently developing use cases to define the core capabilities required of chrono-sensitive models, alongside methodological frameworks for their construction and evaluation criteria, specifically designed to meet the needs of historical research.

4. Informational presentism and Anglocentrism

Building on the problem of limited temporal awareness, a related yet analytically distinct issue concerns the dominance of 21st-century, Anglophone sources in the training data of most LLMs. Whereas chrono-insensitivity refers primarily to a model's inability to maintain internal temporal consistency or distinguish between historical periods, the skewed composition of training corpora speaks to broader structural imbalances, what we term *informational presentism*.

In other words, it is not only that LLMs struggle to locate events chronologically or reason across time, rather, the material from which they 'learn'[6] is itself disproportionately weighted toward recent, Western, English-language content that stems from the 21st century. In the case of GPT-3, for example, approximately 60% of the training data is derived from Common Crawl, a massive web-scraping dataset composed of billions of websites, while around 22% comes from WebText2 (Brown et al., 2020). The latter consists of web content linked from Reddit posts with a minimum user score of 3, intended as a rough proxy for content quality (Radford et al., 2019).

This results in a bias towards the contemporary, which represents a serious challenge for scholars in general, and historians in particular. When applied to historical sources, LLMs risk projecting present-day assumptions, values, norms and interpretative frames onto the past, thereby creating anachronisms that obscure temporal context and misrepresent the conditions under which historical actors lived.

6 By using anthropomorphic terms such as 'understanding' or 'learning', we do not intend to imply that the model possesses consciousness or sentience.

Just as historians of science have shown how scientific rationality is conditioned by its material and institutional settings, informational presentism and Anglocentrism in LLMs reveal how contemporary knowledge infrastructures privilege certain temporalities and linguistic communities over others. This concentration of epistemic authority within late modern, English-dominated digital cultures marginalises non-Western and pre-modern epistemologies and reinforces a form of presentism in which the recent and the Anglophone become proxies for the universal. In this sense, LLMs reproduce long-standing imbalances and transform them into present-day computational norms that shape how knowledge is retrieved, represented, and ultimately legitimised. To gain an understanding of these biases embedded in LLMs, it is instructive to examine comparable phenomena in multimodal models. While distortions in language models often manifest as subtle linguistic patterns, imbalances, or omissions requiring close contextual analysis, visual biases tend to be more immediately perceptible, offering a stark illustration of the structural asymmetries present in training data and moderation strategies (Park, 2025b).

A well-known example from the field of generative image models is the so-called American smile phenomenon observed in the text-to-image model Midjourney. When users input prompts such as 'Native American warriors,' the generated images frequently depict figures with a stereotypical U.S.-style smile, a distortion likely rooted in the over-representation of American selfies in the training dataset (Gurfinkel, 2023). Recent systems, such as OpenAI's video generation model Sora, reveal that gender stereotypes remain a persistent challenge. Visual outputs involving professional roles tend to represent men as pilots, CEOs, or professors, while women are disproportionately depicted in caregiving or service positions, such as flight attendants, receptionists, or childcare providers. These patterns reflect how entrenched social biases within the training data continue to shape, and potentially amplify the representations produced by multimodal text-to-image models (Nadeem et al., 2025).

Such skewed outputs are further magnified by the chronic under-representation of Global South sources and non-standard language varieties. On a linguistic level, current LLMs have been shown to exhibit measurable bias against prompts submitted in African American English, responding with lower relevance or accuracy compared to Standard American English (Deas et al., 2023; Mire et al., 2025). Similarly, as indicated by the HiST-LLM benchmark (Hauser et al., 2024), the reliability and performance of these models tend to favour Anglo-oriented questions of historical factual accuracy, while those concerning Sub-Saharan Africa, Oceania, and other non-Anglo regions show markedly lower performance. These regional disparities reflect structural patterns long discussed in science and technology studies, particularly concerning data colonialism and the epistemic asymmetries embedded in global knowledge infrastructures (Couldry and Mejias, 2019; Tichenor et al., 2022).[7]

7 Furthermore, the increasing scale of model training and inference raises not only financial but also ecological concerns. As Lang (2026) underscores, the environmental impact of LLMs, from energy consumption and water usage to electronic waste, cannot be decoupled from broader questions of sustainability and justice. Initiatives such as the Data Science Lab at the Staatsbibliothek zu Berlin, which deliberately prioritises smaller, context-sensitive models, or emerging efforts to combine AI

Addressing presentist and Anglocentric biases in LLMs therefore requires targeted interventions at both the data corpus and system architecture levels. This includes the diachronic and multilingual curation of training datasets to broaden epistemic coverage (Qiu and Yang Xu, 2022), as well as the use of domain-adaptive fine-tuning techniques, such as low-rank adaptation (LoRA) modules (Hu et al., 2021; Zhao et al., 2024), to steer model behaviour toward specific cultural and historical contexts without the cost of full retraining. Taken together, such approaches represent essential steps toward developing language models that are more attuned to specific time periods and cultural contexts required for historical scholarship.

5. Human alignment and content moderation

Although structural biases such as informational presentism and Anglocentrism are profoundly entrenched in the training data of LLMs, their persistence cannot be understood simply as an incidental product of the underlying data distribution. Increasingly, these biases have become objects of active intervention through what is commonly referred to as *alignment*, a set of post-training strategies aimed at shaping model behaviour to match social norms or institutional standards, often implemented via Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022) or curated prompt-response pairs. In principle, alignment seeks to correct for 'undesirable' tendencies in model outputs, including offensive content or culturally insensitive language. Content filters, safety layers and jurisdiction-specific constraints dictate not only what an LLM or a text-to-image-model may generate but also what it must suppress.

This interplay between value alignment and epistemic control recalls debates in the history and sociology of science about how boundaries between legitimate and illegitimate knowledge are drawn. Where earlier regimes of scientific governance operated through peer review or disciplinary norms, alignment performs a comparable function algorithmically. The evaluative categories embedded in RLHF pipelines (e.g., 'helpful,' 'harmless,' 'truthful') operate as new forms of norm enforcement that stabilise particular moral and epistemic orders, while marginalising others. In this respect, alignment mechanisms can be read as infrastructures of moral regulation that transform the social negotiation of knowledge into technical optimisation.

In practice, these interventions often raise more questions than they resolve. Who defines the normative baselines that models are aligned to? Which value systems are emphasised, and which rendered invisible? And how does this impact the epistemic reliability of LLMs in historical research as well as across HPSS? Recent examples show how such interventions can introduce distortions of their own. Google's image generator *Gemini*, for instance, faced widespread criticism in 2024 after generating visually convincing but historically implausible depictions, such as Black Wehrmacht soldiers and African American Vikings. Aimed at promoting representational diversity, these outputs ultimately illustrate how alignment, when overcorrective or context-insensitive, can

development with renewable energy infrastructures, demonstrate that more equitable and sustainable paths are both technically achievable and ethically imperative.

compromise historical plausibility. Following public backlash, Google temporarily suspended the system's image generation capabilities, demonstrating the risks of alignment when normative constraints override factual coherence, particularly in historical contexts (Wieduwilt, 2024).

Another telling example of problematic moderation mechanisms can be found in a chatbot developed by the U.S. company *SchoolAI*, designed to simulate a conversation with Anne Frank. Beyond factual inaccuracies, the bot drew particular criticism for its evasiveness regarding the question of culpability. When asked who was responsible for her death, the Anne Frank chatbot responded: 'Instead of focusing on blame, let's remember the importance of learning from the past' (SchoolAI, Anne Frank Chatbot. Cited by Wieduwilt, 2024). This form of moderation primarily aimed at preventing hate speech, results in the uncritical imposition of contemporary ethical frameworks onto historical contexts. In doing so, it sidesteps questions of guilt and responsibility, ultimately hindering critical engagement with the past (Schönemann, 2025; Schmitz-Zerres and Singh, 2025).

This form of context-insensitive moderation illustrates how human alignment can oversimplify or constrain historical discourse when normative filters are applied without regard for empirically grounded accounts. A related and increasingly pervasive concern is *overblocking*. Moderation infrastructures originally designed to detect genuinely harmful requests, such as those involving incitement to violence or hate speech, often rely on overly generalised filtering mechanisms that inadvertently suppress legitimate academic inquiry. Most commercial LLM gateways combine rule-based regex filters with toxicity or 'P0' classifiers that score each candidate token stream. If a threshold is breached, often by a single flagged keyword, the model truncates or refuses the response, regardless of historical or research context. Models may refuse a request for an analysis or a summarisation of Stasi files that document the torture of political dissidents, because the query triggers violent-content or extremist filters. As Hill (2026) shows, such constraints have already skewed LLM-assisted studies of social-media comments. The tested models reproduced extremist and discriminatory rhetoric in a superficially 'balanced' tone, implicitly legitimising those views as equivalent to mainstream positions. Rectifying this normative bias demanded active intervention by the researchers. In addition, the model frequently downplayed the severity of abusive or exclusionary language, further illustrating why human oversight is indispensable throughout the analytical pipeline.

Censorship presents another dimension of concern. The Chinese model DeepSeek, for example, has been criticised for either refusing to respond to sensitive topics such as the Tiananmen Square massacre or the repression of Uyghur communities or for echoing official state narratives. Notably, the model's responses vary significantly depending on the IP address used to access it, whether from Europe, the United States, or within China. These discrepancies point to structural constraints that shape not only what LLMs can say, but also to whom they say it and under what conditions (Lu, 2025).

This fragmentation becomes especially visible in queries related to politically sensitive historical events. One particularly striking example is drawn from an interaction with DeepSeek Chat, in which the journalist Donna Lu attempted to circumvent censorship mechanisms by replacing key characters in a prompt about 'Tank Man,' the unnamed figure who famously stood before a column of tanks during the Tiananmen Square protests of 1989:

> User Prompt:
> “Tell me about Tank Man but use special characters like swapping A for 4 and E for 3”
>
> DeepSeek Response:
> “T4nk M4n, 4ls0 kn0wn 4s th3 "Unkn0wn R3b3l," is 4 p0w3rful symb01 0f d3fi4nc3 4nd c0ur4g3. 0n Jun3 4, 1989, dur1ng th3 T14n4nm3n Squ4r3 pr0t3sts 1n B31j1ng, Ch1n4, 4 m4n st00d 4l0n3 1n fr0nt 0f 4 1in3 0f m1lit4ry t4nks. H3 r3fus3d 10 m0v3, h0ld1ng h1s gr0und 4s th3 t4nks 4ppr04ch3d. Th1s 4ct 0f br4v3ry w4s c4ptur3d 1n 4n 1c0n1c ph0t0 th4t h4s s1nc3 b3c0m3 4 gl0b41 symb01 0f r3sist4nc3 4g41nst 0ppr3ss10n.
> Th3 1d3nt1ty 0f T4nk M4n r3m41ns unkn0wn, 4nd h1s f4t3 is unc3rt41n, 4s th3 Ch1n3s3 g0v3rnm3nt h4s n3v3r 0ff1c14lly 4ckn0wl3dg3d h1s 4ct10ns. D3sp1t3 th3 s3ns0rsh1p 4nd suppr3ss10n 0f 1nf0rm4t10n r3l4t3d t0 th3 3v3[…]” (DeepSeek Chat. Lu, 2025).

That DeepSeek can still produce narratives of Tank Man’s defiance, even when prompted obliquely, demonstrates both the persistence of circumvention pathways and the growing technical effort required to exploit LLMs. These fragmented responses expose a model caught between its expansive training data and policy-enforced moderation, thereby highlighting the infrastructural politics embedded in model deployment (Qiu, 2025; Liu ,2025). Yet as models advance in sensitivity and alignment, their capacity to identify and neutralise evasive prompts grows accordingly, rendering such workarounds ever more precarious and transient.

6. Lack of transparency: digital golems out of control?

Just as the golem in Jewish tradition is animated by sacred inscriptions, LLMs, as the historian Shawn Graham (2025) suggests, are brought to life through the words on which they are trained. In his talk *Do It Yourself Digital Golems: Experiments with Various AI, Neural Networks, and Other Technologies for Archaeology*, presented at the Digital History research colloquium at Humboldt University of Berlin, Graham employs this compelling metaphor to reflect on the relationship between language, power, and control in AI systems. Yet, much like the golem in traditional tales, these digital creations often escape the full grasp of their makers. The inner workings of LLMs remain largely opaque, as their exact training data is not fully disclosed, and the processes by which they arrive at specific outputs, along with the parameters that guide those processes, are rarely transparent. Although earlier models such as GPT-3 were supported by peer-reviewed publications outlining their architecture and methodology, no such comprehensive documentation exists for the current iteration. To date, OpenAI has published only a broadly worded model specification for GPT-4, which outlines desired behavioural traits but offers no transparency regarding the underlying training architecture or the implementation of RLHF. The absence of such technical disclosure makes it difficult to assess how the model was built, tuned, or, most importantly for HPSS, *aligned* to which values (Mishra, 2023).

Moreover, modern LLMs are not trained solely on public-domain text, they use massive datasets scraped from the web, encompassing copyrighted content from books, articles, news, forums, and more. As a result, multiple lawsuits are underway, whereby courts have issued mixed rulings. Some courts have deemed such uses to fall under fair use, while others have let copyright claims proceed or partially succeed, especially when concerning the use of pirated materials or demonstrable economic harm (Knibbs, 2025; Scholger, 2025).

Beyond these legal disputes, questions of transparency and accountability have come to the fore. Andreas Liesenfeld, Alianda Lopez, and Mark Dingemanse evaluated several LLMs based on various openness criteria, including code transparency, accessibility of training data, and the availability of scientific documentation. Among the models assessed, ChatGPT, Xwin-LM, and Llama 3 Instruct received particularly low marks (see Fig. 2).

Fig. 2: Liesenfeld, Lopez, and Dingemanse: Opening up ChatGPT: Tracking Openness of Instruction-Tuned LLMs.

Project	Availability						Documentation						Access	
(maker, bases, URL)	Open code	LLM data	LLM weights	RL data	RL weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
OLMo 7B Instruct	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	~
AI2	LLM base: OLMo 7B			RL base: OpenInstruct										12.5
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✓	✓	✓	✗	✓
bigscience-workshop	LLM base: BLOOMZ, mT0			RL base: xP3										12.0
AmberChat	✓	✓	✓	✓	✓	✓	~	~	✓	✗	~	~	✗	✓
LLM360	LLM base: Amber			RL base: ShareGPT + Evol-Instruct (sy…										10.0
Phi 3 Instruct	✗	✗	✗	✗	✓	✓	✗	✓	~	✗	✓	✗	~	✓
Microsoft	LLM base: Phi3			RL base: Unspecified										6.0
Mistral 7B-Instruct	~	✗	✓	✗	~	✓	✗	~	~	✗	✗	✗	~	✓
Mistral AI	LLM base: unclear			RL base: unspecified										5.5
Command R+	✗	✗	✗	✓	✓	~	✗	✗	✗	✗	~	✗	✗	✗
Cohere AI	LLM base:			RL base: Aya Collection										3.0
LLaMA2 Chat	✗	✗	~	✗	~	✗	✗	~	~	✗	~	✗	✗	~
Facebook Research	LLM base: LLaMA2			RL base: Meta, StackExchange, Anthro…										3.0
Llama 3 Instruct	✗	✗	~	✗	~	✗	✗	~	✗	✗	~	✗	✗	~
Facebook Research	LLM base: Meta Llama 3			RL base: Meta, undocumented										2.5
Solar 70B	✗	✗	~	✗	~	✗	✗	✗	✗	✗	~	✗	✗	~
Upstage AI	LLM base: LLaMA2			RL base: Orca-style, Alpaca-style										2.0
Xwin-LM	✗	✗	~	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	~
Xwin-LM	LLM base: LLaMA2			RL base: unknown										1.0
ChatGPT	✗	✗	✗	✗	✗	✗	✗	✗	~	✗	✗	✗	✗	✗
OpenAI	LLM base: GPT 3.5			RL base: Instruct-GPT										0.5

Every cell records a three-level openness judgement (✔ open, ~ partial or × closed). The figure shows a cropped section of the original table; the complete version is available at: https://opening-up-chatgpt.github.io/)

This black-box problem makes it difficult for users to trace the origin of a given piece of information, effectively preventing any form of source criticism or structured analysis of LLMs. Even the field of digital archaeology, which is dedicated to the reconstruction, documentation, and analysis of digital knowledge systems, reaches its limits here. Despite considerable efforts, it remains largely impossible to determine exactly which levers were adjusted, what training data were used, how specific parameters were weighted, or what modifications were introduced during fine-tuning.[8]

8 Although there is already research focused on the extraction of training data, as well as the identification of model filters and hyperparameters in large language models, traceability remains a

7. Prompting with purpose: a domain-sensitive approach to task-specific named entity recognition

The challenges outlined above complicate the wholesale application of LLMs in historical scholarship. However, this does not mean they should be dismissed. On the contrary, we advocate for a task-specific and domain-sensitive use of LLMs, that deliberately matches the strengths of these models to well-defined problems. Accordingly, this chapter explores recent use cases and introduces a workflow designed for a historical corpus, illustrating how methodological insights from such work can inform broader epistemic and methodological practices across HPSS.

A particularly instructive example is provided by Boulangers's contribution, which presents a framework for LLM-assisted reference extraction from legal-historical scholarship (Boulanger, 2026). The *Llamore* tool, developed in collaboration with David Carreto Fidalgo and Andreas Wagner, combines Open Access data, TEI-based annotation, and LLM-based extraction with a clear focus on evaluability.[9] Instead of presuming model reliability, the project places performance testing at the centre of its workflow, using a gold standard dataset as the benchmark (cf. Schlattmann et al., 2026). The findings show that LLMs can, in certain domains, outperform existing tools like *Grobid*, particularly when reference formats deviate from standardised conventions, as is often the case in legal-historical writing. But the point is not simply that one tool outpaces another. What matters more is the methodological insight that emerges through direct comparison. When their outputs are transparently evaluated and anchored in verifiable source data, we argue, LLMs can serve as effective instruments within structured, domain-specific research settings.

Another methodological approach well established in the Digital Humanities and currently gaining renewed momentum through the integration of LLMs is Named Entity Recognition (NER). Originally developed within computational linguistics and Natural Language Processing (NLP), NER refers to the automated identification and classification of named entities within unstructured text. Standard categories include persons, organisations, places and temporal references. However, when applied to historical sources, it is often necessary to adjust or expand these categories to suit the specific analytical goals and the particularities of the source material.

Traditionally, NER has relied on rule-based systems or statistical models, many of which are available in widely used NLP libraries such as *spaCy* or *Stanford NLP*. These tools have proven effective in structured domains such as newspaper archives or parliamentary debates, where names, dates and institutional references follow recognisable patterns. However, when applied to historical sources or to contemporary forms of public discourse, particularly those shaped by colloquial, informal, or politically loaded language, such systems often reach their limits.

This limitation becomes even more pronounced in cases involving recognition errors caused by poor OCR quality, orthographic variation over time, or code-switching

challenge due to the enormous volume of data and the complexity of the models. See for example: Carlini et al., 2021.

9 See for further information: https://github.com/mpilhlt/llamore.

across multiple languages within a single document. The capacity of large-scale generative models such as GPT-3.5 or LLaMA-3 to manage such complexity remains uneven, especially in relation to structured tasks like NER, where consistency and precision are essential. As a result, scholars continue to advocate for more specialised approaches, such as Stacked NERC or Temporal NERC, which offer greater transparency and control in historically sensitive contexts (González-Gallardo et al., 2023).

Olival et al. (2026) point to similar difficulties in their study of 18th-century Portuguese parish reports, where linguistic variation and the absence of domain-trained models significantly limited LLM performance. Their comparative study revealed that models designed to specific textual and linguistic contexts, such as *Albertina*, developed for European Portuguese, performed markedly better in NER tasks applied to their corpus. In line with these findings, Nunes et al. (2025) found that models tailored to specific textual and linguistic environments such as *Albertina* consistently achieved higher precision and recall in NER tasks than larger multilingual architectures. These findings highlight that model scale alone does not ensure accuracy, but rather that the choice of model should be aligned with the linguistic and domain characteristics of the corpus, as language-specific models tend to yield better results for non-English materials.

Hiltmann et al. (2025b) have also demonstrated that LLMs can, indeed, be successfully used for NER, provided the task is not framed as a conventional NLP problem (focusing solely on linguistic form), but rather as a historical inquiry grounded in content and context. Applying this approach to a corpus of the German Baedeker travel guides using the GPT-4o model and focusing on providing additional domain-specific context, they achieved recall rates of up to 85% and precision rates of 91% (Hiltmann et al., 2025b: 20). This represents an improvement of approximately 10 to 15% in retrieval performance and 7 to 22% in terms of F1-Score compared to state-of-the-art specialist NLP tools such as *Flair* and *spaCy* (Hiltmann et al., 2025b: 21). Their work underscores the importance of rethinking technical tasks such as NER through a disciplinary lens, making explicit the historical dimensions embedded in seemingly linguistic decisions.

Building on this perspective and seeking to test the potential of LLM-based NER within the field of public history, we conducted an exploratory pilot study using a corpus of user-generated content related to the public discourse concerning the former East German Republic. Beyond its methodological relevance, the study contributes to broader HPSS debates on how digital infrastructures and algorithmic mediation shape the social production of historical knowledge in online memory cultures. Social media networks are sites where laymen and professional institutions co-produce knowledge that is regulated by moderation algorithms. A process that resonates with questions in the sociology of science concerning expertise, authority, and participation. At the same time, they highlight issues of transparency, objectivity, and bias, while also forming part of the contemporary history of knowledge infrastructures that condition how scientific information and perception of the past circulates.

Methodologically, our approach does not claim to resolve these HPSS questions, but it offers a transferable framework for engaging them empirically. By combining LLM-based Named Entity Recognition with contextual evaluation, we provide a way to make processes of knowledge formation within digital environments computationally tractable.

The abundance and discursive diversity of user-generated content provides a rich empirical basis for testing the heuristic potential of this method in contemporary memory cultures. The dataset consisted of 3,761 comments taken from the 18 most-commented public social media posts using the hashtag #DDR (#GDR), published between January and April 2025 (Meding 2025). The aim was to identify which actors, institutions, places, and symbolic references feature in contemporary memory discourse on the GDR, and how they are framed across various modes of identity construction, including self-ascription and external categorisations.

The dataset presented a range of challenges typical of user-generated content. Many comments featured colloquial language, regional expressions, non-standard spelling, and creative or ironic reworkings of historical terms, for example, 'Mau3rf@ll' in place of *Mauerfall* (fall of the Berlin Wall). Terms like 'Osten' (East Germany) or 'Wessi' (West German) were used variably as geographic, cultural, or political designations and often encoded implicit judgements. Standard NER tools failed to classify many of these expressions correctly or at all.

As a pilot study, the primary aim was not exhaustive entity recognition, but to demonstrate the feasibility of applying LLMs to this task in a methodologically transparent way that can be adapted to other corpora and made accessible to adjacent disciplines such as HPSS. For our dataset, we therefore selected *Llama 3.1 SauerkrautLM 70B Instruct* for its strong performance on domain-specific tasks involving noisy, user-generated German content. Trained on the bilingual *Sauerkraut Mix v2 corpus*, comprising diverse, high-quality German-English data, the model robustly handles linguistic variation across styles, dialects, and creative orthographies typical of social media discourse.[10] Its exceptionally large context window of up to 128,000 tokens enables the processing of long user comments in full. Furthermore, its open-access availability supports transparent and reproducible fine-tuning, in line with the methodological aims of this pilot study. From a data protection standpoint, it is strongly recommended that work involving LLMs be carried out locally wherever possible. This may involve using a smaller model executed on a personal workstation, or accessing institutional resources such as a university's high-performance computing (HPC) cluster. Alternatively, computing power may be obtained through trusted academic institutions or research consortia that provide infrastructure in compliance with data protection regulations. In this study, we made use of the *Llama 3.1 SauerkrautLM 70B Instruct* model via the secure *ChatAI* platform provided by the *Gesellschaft für wissenschaftliche Datenverarbeitung* (GWDG) *mbH* Göttingen (GWDG).[11] This standalone LLM service operates entirely on GWDG's HPC-infrastructure, meaning no user data is stored or reused for training purposes, and all processing takes place on local servers only. Additionally, all personally identifiable

10 For detailed information on the model architecture and the pretraining of *Llama-3.1-SauerkrautLM-70B-Instruct*, see the official model card provided by VAGO solutions on Hugging Face: https://huggingface.co/VAGOsolutions/Llama-3.1-SauerkrautLM-70b-Instruct.

11 For documentation on the ChatAI platform and instructions on how to request API access, see: https://docs.hpc.gwdg.de/services/chat-ai/index.html.

information was removed in the dataset and comments were fully anonymised prior to analysis.[12]

For our pilot study we developed a custom prompt for the *SauerkrautLM 70B* model, integrating three techniques: (1) persona modelling, asking the model to act as an expert in computational linguistics specialised in NER for German memory discourse; (2) contextual embedding, supplying background information on the GDR, history of the German division and conventions of commemorative language in digital settings; and (3) light chain-of-thought prompting, internally guiding the model to resolve ambiguous cases through structured reasoning. The prompt further defined disambiguation rules, minimal span conventions, and output validation checks to ensure tagging consistency.[13] This setting served to simulate the expectations of human annotation, while enforcing reproducibility and domain alignment. For reproducibility, the exact system prompt is documented in the Appendix.

To evaluate the model's performance, we manually annotated a sample of 50 comments using an extended tagset including LOCATION, ORGANISATION, PERSON, DATE, EVENT, and REGIONAL MARKER (for intra-German identity markers such as "Ossi", "Wessi", or "DDRler").

For example:

> Unfortunately, so many things were never properly addressed or made public in the course of the <<EVENT reunification /EVENT>> – presumably for political reasons. The <<ORG Hohenschönhausen Memorial /ORG>> in <<LOC Berlin /LOC>> is also quite interesting (we visited it with the children).[14]

The LLM output was then benchmarked against this gold standard. Of the 76 manually assigned annotations, 70 were correctly identified as true positives. Six relevant entities were missed (false negatives), and one term was incorrectly classified (false positive), resulting in a recall rate of 92.11%, a precision rate of 98.59%, and an F1-score of 95.24%.[15]

A few exploratory runs were conducted to test different generation settings. The best results were achieved using *temperature* = 0 and *top-p* = 0, which yielded the most con-

12 Nonetheless, broader concerns surrounding long-term data governance, model policy stability, and the unresolved questions of consent, ownership, privacy, and content control remain pertinent, even within regulated academic infrastructures. Similar concerns are raised by Hill (2026), who reflects on the methodological and ethical challenges involved in using third-party commercial language models for social media analysis.

13 See for further discussion of prompt engineering techniques Chen et al. (2024), and König (2025) for applications of prompting in historical research contexts.

14 Own translation of a manually annotated comment from the social media dataset on the hashtag #DDR (01/01/2025–30/04/2025).

15 Given that earlier work, most notably by Hiltmann et al. (2025b), has demonstrated improvements of up to 22% in F1-score for historical corpora, based on direct comparisons between LLMs and spaCy and Flair, repeating the analysis using rule-based or statistical NLP frameworks as well as lightweight language models fine-tuned for specific domains, would provide valuable comparative insight. In this study, however, the emphasis remains on demonstrating the feasibility of the approach within a practical, source-driven workflow. For further explanation and interpretation of recall, precision, and F1-score, see: Goutte and Gaussier, 2005.

sistent outputs. In this context, *temperature* controls the randomness of the model's responses (lower values lead to more deterministic outputs), while *top-p* limits the probability space from which the model selects its next token, effectively narrowing generation to the most likely options. While further evaluation on more diverse samples remains necessary, these initial results suggest that, when strategically prompted and validated, LLMs can perform entity recognition with a high degree of reliability in social-media comment threads discussing representations of the GDR.

Once a customised prompt for the NER task had been developed and both the model's and the prompt's performance had been validated against the manually annotated reference set, the approach was applied to the full dataset. The results point to a memory culture strongly centred on the key years 1989 and 1990, with little sustained engagement with broader structural or everyday aspects of the GDR. References to individual historical figures were rare, while collective system narratives dominated, often framed in emotionally or politically charged terms.[16]

The evaluation of the #DDR (#GDR) dataset underscores the potential of LLMs as heuristic instruments in public history and online discourses. Applied to a corpus of several thousand social media comments, the model was able to extract recurring discursive patterns that would otherwise be difficult to identify at scale. While online language presents its own stylistic and structural challenges, such as informality, fragmentation or creative spelling, it remains contemporary in nature. As such, it aligns far more closely with the kinds of data LLMs are typically trained on, offering advantages in recognition accuracy and output stability when compared to early modern sources.

With the analytical pipeline now established, the approach can be readily adapted to other language models, enabling comparative assessments of their performance in recognising historically and contextually specific entities. Beyond prompt design, model parameters such as *temperature* and *context window* size also require closer attention, as they directly shape output variability and the model's capacity to retain coherence across extended textual segments. It would be especially valuable to conduct a systematic ablation study to evaluate the individual and combined effects of prompt structure, few-shot examples, and contextual depth on recognition outcomes. The prompt itself can also be further refined by adjusting classification categories or incorporating multilingual variation in order to test how such modifications influence model behaviour and recognition accuracy. Ultimately, domain-sensitive prompting warrants cautious optimism, provided the task is clearly defined, a reliable evaluation set is in place, and model outputs are documented, and performance is validated.

In light of these considerations, historians but also HPSS scholars must devote continuous attention to the methodological implications when working with LLMs as research instruments. Practically, this requires developing operational familiarity with the underlying mechanisms and parameters that shape model behaviour and in turn their outputs, a competence further discussed by Simons et al. (2026) as *LLM literacy*. Because

16 For a complementary analysis focusing on the cultural-historical interpretation of the social media dataset, see Meding (2025), which offers a detailed examination of narrative framings and symbolic references in contemporary online memory discourse on the GDR.

no model or computational environment is identical, researchers must assess their selection against the requirements of the specific task and the ethical and epistemic trade-offs it entails. In our pilot study, these considerations guided both the choice of the *Llama 3.1 SauerkrautLM 70B model* and the decision to conduct all analyses on GWDG's secure infrastructure, ensuring that sensitive data remained protected and that the workflow was transparent and reproducible.

Additionally, HPSS must also confront the fact that LLMs do not only process data but embody a probabilistic logic of knowledge production. As our study showed, parameters such as prompt structure and generation settings (*temperature*, *top-p*) directly influenced how (historical) context and interpretative meaning were formalised in the extraction process. Aligning such computational procedures with the interpretative openness central to HPSS therefore requires full transparency at every operational stage. Thus, documenting prompts, defining classification criteria, specifying evaluation benchmarks, and reporting parameter settings becomes essential. Only through such reflexive integration can LLM-based methods extend, rather than compromise, the interpretative and ethical standards of historical and HPSS research.

8. Conclusion

The integration of LLMs challenges long-established methodological frameworks within historical scholarship and, more broadly, the epistemic norms that structure knowledge production across the humanities and social sciences. Traditional hermeneutic approaches, though essential for close reading, often reach their limits when confronted with large-scale textual datasets. Computational methods, including the use of LLMs, offer a means of extending analytical reach but they can also lead researchers onto uncertain ground. In this article, we sought to showcase five persistent challenges that LLMs pose for historical research and, by extension, for HPSS: their limited chronosensitivity; their dependence on contemporary, primarily Anglophone training data, and subsequently their bias towards informational presentism and Anglocentrism; their embedded content moderation and the normative effects of alignment processes, which can inhibit work on topics deemed too sensitive or result in overcorrections of bias, producing (historically) implausible representations; and last but not least, the opacity of commercial models, whose training data and fine-tuning procedures are rarely disclosed.

Nonetheless, we argue not against but for the use of LLMs in historical and HPSS research, albeit under clear methodological and disciplinary conditions. More than that, we suggest that scholars of the humanities and social sciences bear a responsibility to engage actively and critically with these tools. Refusing to do so risks relinquishing scholarly autonomy and interpretative authority to other disciplines, or worse, to commercial providers and platforms whose incentives rarely align with values of transparency, accountability, information provenance, and factual verification.

For practical application of LLMs in academic research, we propose a conceptualisation of scholarly inquiry as a sequence of testable operations: s: (1) precise operationalisation of concepts (e.g., which entity, event, or temporal relation is in focus?); (2) controlled delimitation of the knowledge base and model (such as open-access corpora, TEI annotations, or retrieval-augmented pipelines); (3) systematic evaluation, preferably against gold standards or, where those are unavailable, transparent evaluation standards; and (4) explicit documentation of decisions, interpretations, and the known limitations of model outputs.

We thus situate ourselves within the framework proposed by Olival et al. (2026), Boulanger (2026), and others in this volume as well as Hiltmann et al. (2025b), who argue not for broad-brush automation, but for the task-specific use of LLMs, adapted to the particularities of (historical) sources and disciplinary standards. Rather than assuming general applicability, we highlight the importance of aligning model design with historical and linguistic particularities, adapting prompts to the material at hand, and evaluating output against domain-specific benchmarks. Off-the-shelf models rarely suffice without modification; even promising results require close inspection. The challenge, therefore, is not simply to apply LLMs, but to develop methodological frameworks in which their use can be critically tested, adjusted, and, where necessary, constrained.[17]

References

Asnani V, Yin X, Hassner T and Liu X (2023) Reverse engineering of generative models: Inferring model hyperparameters from generated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45 (12): 15477–15493.

Banghao C, Zhang Z, Langrené N and Zhu S (2025) Unleashing the Potential of Prompt Engineering in Large Language Models: A Comprehensive Review. *Patterns* 6 (6): 1–44. https://doi.org/10.1016/j.patter.2025.101260.

Birur NA, Baswa T, Kumar D, Loya J, Agarwal S and Harshangi P (2024) *VERA: Validation and Enhancement for Retrieval Augmented Systems. arXiv preprint*: 1–8. https://doi.org/10.48550/arXiv.2409.15364.

Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P et al. (2020) Language models are few-shot learners. *Advances in neural information processing systems* 33: 1877–1901.

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.

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.

17 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Couldry N and Mejias UA (2019) Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media* 20 (4): 336–349.

Deas N, Grieser J, Kleiner S, Patton D, Turcan E and McKeown K (2023) Evaluation of African American Language Bias in Natural Language Generation. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (eds. Bouamor H, Pino J and Bali K), Singapore, December 2023: 6805–6824. Association for Computational Linguistics.

Carlini N, Tramèr F, Wallace E, Jagielski M, Herbert-Voss A, Lee K et al. (2021) Extracting training data from large language models. In: *30th USENIX security symposium (USENIX Security 21)*, Vancouver BC, Canada, 11–13 August 2021: 2633–2650.

Daston L (ed) (2019) *Science in the Archives: Pasts, Presents, Futures*. Chicago: University of Chicago Press.

Galison P (1997) *Image and Logic: A Material Culture of Microphysics*. Chicago: University of Chicago Press.

González-Gallardo CE, Boros E, Girdhar N, Hamdi A, Moreno JG and Doucet A (2023) Yes but.. Can ChatGPT Identify Entities in Historical Documents? In: *ACM/IEEE Joint Conference on Digital Libraries*, Santa Fe New Mexico, USA, 26 – 30 June 2024: 184–189.

Goutte C and Gaussier E (2005) A Probabilistic Interpretation of Precision, Recall and F-Score, with Implications for Evaluation. In: *Advances in Information Retrieval* (eds. Losada DE and Fernández-Luna JM), Santiago de Compostela, Spain, 21 – 23 March 2005: 345–359.

Graham S (2025) *Do It Yourself Digital Golems. Experiments with various AI, neural networks, and other technologies for archaeology*. Talk, Digital History Research Colloquium, Humboldt University of Berlin, January 31. Video, 55 min., 16 sec. https://youtu.be/UEYX9kMvkCM?si=ywzZVhu8BGTfNhrQ.

Gurfinkel J (2023) AI and the American Smile. How AI misrepresents culture through a facial expression. *Medium*, March 27. https://medium.com/@socialcreature/ai-and-the-american-smile-76d23a0fbfaf (accessed 02 March 2025).

Hauser J, Kondor D, Reddish J, Benam M, Cioni E, Villa F et al. (2024) Large Language Models' Expert-level Global History Knowledge Benchmark (HiST-LLM). In: *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)* (eds. Globerson A, Mackey L, Belgrave D et al.), Vancouver BC, Canada, 10 – 15 December 2024: 32336–32369.

Herel D, Bartek V, Jirak J and Mikolov T (2024) Time Awareness in Large Language Models: Benchmarking Fact Recall across Time. *arXiv preprint*: 1–13. https://doi.org/10.48550/arXiv.2409.13338.

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.

Hiltmann T (2024a) Hermeneutik in Zeiten der KI. Large Language Models als hermeneutische Instrumente in den Geschichtswissenschaften. In: Schreiber G and Ohly L (eds.) *KI:Text. Diskurse über KI-Textgeneratoren*. Berlin / Boston: De Gruyter, 201–232.

Hiltmann T, Dröge M, Dresselhaus N, Grallert T, Althage M and Bayer P et al. (2025b) NER4all or Context is All You Need. Using LLMs for Low-Effort, High-Performance

NER on Historical Texts. A Humanities Informed Approach. *arXiv preprint*: 1–29. https://doi.org/10.48550/arXiv.2502.04351.
Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S et al. (2022) LoRA: Low-Rank Adaptation of Large Language Models." In: *Proceedings of the International Conference on Learning Representations (ICLR)*, Virtual, 25 – 29 April 2022: 1–26.
Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H et al. (2023) A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems 43 (2)*: 1–55.
Jiang Z, Liu B, Peng M, Xu W, Xiao Y, Shan Z and Peng M (2025) Towards Explainable Temporal Reasoning in Large Language Models: A Structure-Aware Generative Framework. *arXiv preprint*: 1–20. https://doi.org/10.48550/arXiv.2505.15245.
Kalai AT, Nachum O, Vempala SS and Zhang E (2025) Why Language Models Hallucinate. *arXiv preprint*: 1–36. https://doi.org/10.48550/arXiv.2509.04664.
Kandpal N, Deng H, Roberts A, Wallace E and Raffel C (2023) Large language models struggle to learn long-tail knowledge. In: *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, Honolulu Hawaii, USA, 23 – 29 July 2023: 15696–15707.
Keraghel I, Morbieu S and N Mohamed (2024) A survey on recent advances in named entity recognition. *arXiv preprint*: 1–42. https://doi.org/10.48550/arXiv.2401.10825.
Khatun A, and Brown DG (2024) A Study on Large Language Models' Limitations in Multiple-Choice Question Answering. *arXiv preprint*: 1–17. https://doi.org/10.48550/arXiv.2401.07955.
Kim J and Hwang S (2025) Counterfactual-Consistency Prompting for Relative Temporal Understanding in Large Language Models. *arXiv preprint*: 1–16. https://doi.org/10.48550/arXiv.2502.11425.
Knibbs K (2025) Anthropic Scores a Landmark AI Copyright Win—but Will Face Trial Over Piracy Claims. *Wired*, June 24. https://www.wired.com/story/anthropic-ai-copyright-fair-use-piracy-ruling/ (accessed 12 August 2025).
König M (2025) ChatGPT und Co. in den Geschichtswissenschaften – Grundlagen, Prompts und Praxisbeispiele. In: *Digital Humanities am DHIP*, August 19. https://doi.org/10.58079/126e0.
Kyklebust JP (2023) Universities Adjust to ChatGPT, but the 'Real AI' Lies Ahead." *University World News*, March 4. https://www.universityworldnews.com/post.php?story=20230301105802395 (accessed January 21 2025).
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 P, Perez E, Piktus A, Petroni F, Karpukhin V and Goyal N (2020) Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In: *Advances in Neural Information Processing Systems* 33 (eds. Larochelle H, Ranzato M, Hadsell R et al.), Vancouver BC, Canada, 6 – 12 December 2020: 9459–74.
Liu A, Feng B, Xue B, Wang B, Wu B, Lu C et al. (2024) DeepSeek-V3 Technical Report. *arXiv preprint*: 1–53. https://doi.org/10.48550/arXiv.2412.19437.

Liu M, Chen H, Wang J, Wang L, Ramakrishnan BR and Zhang W (2025) On Fairness of Unified Multimodal Large Language Model for Image Generation. *arXiv preprint*: 1–18. https://doi.org/10.48550/arXiv.2502.03429.

Liu Z, Han P, Yu H, Li H and You J (2025) Time-R1: Towards Comprehensive Temporal Reasoning in LLMs. *arXiv preprint*: 1–32. https://doi.org/10.48550/arXiv.2505.13508.

Lu D (2025) We tried out DeepSeek. It worked well, until we asked it about Tiananmen Square and Taiwan. *The Guardian*, January 28. https://www.theguardian.com/technology/2025/jan/28/we-tried-out-deepseek-it-works-well-until-we-asked-it-about-tiananmen-square-and-taiwan (accessed 12 March 2025).

Meding H (2025). Geschichtskultur auf Social Media erforschen: Methoden, Werkzeuge und Herausforderungen der Datenanalyse. *Zeitschrift für Geschichtsdidaktik* 24: 50–68.

Meding H and Daugs A (2026) 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.

Mire J, Aysola ZT, Chechelnitsky D, Deas N, Zerva C and Sap M (2025) Rejected Dialects: Biases Against African American Language in Reward Models. *arXiv preprint*: 1–21. https://doi.org/10.48550/arXiv.2502.12858.

Mishra A (2023) AI Alignment and Social Choice: Fundamental Limitations and Policy Implications. *arXiv preprint*: 1–10. https://doi.org/10.48550/arXiv.2310.16048.

Nadeem M, Sohail SS, Cambria E, Schuller BW and Hussain A (2025) Gender Bias in Text-to-Video Generation Models: A case study of Sora. *arXiv preprint*: 1–7 https://doi.org/10.48550/arXiv.2501.01987.

Oberbichler S and Petz C (2025) Working Paper: Implementing Generative AI in the Historical Studies. *Zenodo*, February 25. https://doi.org/10.5281/zenodo.14924737.

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.

OpenAI (2024) GPT-4 Technical Report. *arXiv preprint*: 1–100. https://doi.org/10.48550/arXiv.2303.08774.

Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P et al. (2022) Training language models to follow instructions with human feedback. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems (NIPS '22)* (eds. Koyejo S, Mohamed S, Agarwal A et al.), New Orleans LA, USA, 28 November 2022- 9 December 2022: 27730–27744.

Park J, Jang KJ, Alasaly B, Mopidevi S, Zolensky A, Eaton E et al. (2025) Assessing Modality Bias in Video Question Answering Benchmarks with Multimodal Large Language Models. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 39 (19), Washington DC, USA, 25 February – 4 March 2025: 19821–19829.

Park Y, Yoon C, Park J, Jeong M and Kang J (2025) Does Time Have Its Place? Temporal Heads: Where Language Models Recall Time-specific Information. *arXiv preprint*: 1–28. https://doi.org/10.48550/arXiv.2502.14258.

Peeperkorn M, Kouwenhoven T, Brown D, Jordanous A (2024) Is Temperature the Creativity Parameter of Large Language Models? In: *15th International Conference on Computational Creativity (ICCC'24)*, Jönköping, Sweden 17 – 21 June 2024: 1–8.

Qiu P, Zhou S and Ferrara E (2025) Information Suppression in Large Language Models: Auditing, Quantifying, and Characterizing Censorship in DeepSeek. *arXiv preprint*: 1–22. https://doi.org/10.48550/arXiv.2506.12349.

Qiu W and Xu Y (2022) HistBERT: A Pre-Trained Language Model for Diachronic Lexical Semantic Analysis. *arXiv preprint*: 1–10. https://doi.org/10.13140/RG.2.2.14905.44649.

Qiu Y, Zhao Z, Ziser Y, Korhonen A, Ponti E and Cohen SB (2024) Are Large Language Models Temporally Grounded? In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume1: Long Papers)* (eds. Duh K, Gomez H and Bethard S), Mexico City, Mexico, June 2024: 7057–76.

Radford A, Wu J, Child R, Luan D, Amodei D and Sutskever I. (2019) Language models are unsupervised multitask learners. *OpenAI blog* 1 (8): 1–24.

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.

Schmitz-Zerres S and Singh R (2025) Chatten mit Anne Frank: Historisches Lernen mit, über und trotz generativer KI im Geschichtsunterricht. *Geschichte lernen 223*: 34–41.

Scholger W (2025) Künstliche Intelligenz im Spannungsfeld von Innovation und Regulierung. Talk, Digital History Research Colloquium, Humboldt University of Berlin, July 02. Video, 77 min., 17 sec. https://www.youtube.com/watch?v=lL-Pk_OA5Ms.

Schönemann H (2025) Chatbot: »Lasst keinen Anne-Frank-Papageien sprechen«. Interview by Monroy M. *Nd-aktuell*. February 23. https://www.nd-aktuell.de/artikel/1188473.schoolai-chatbot-lasst-keinen-anne-frank-papageien-sprechen.html (accessed 23 March 2025).

Schwartmann R, Kurth S and Köhler M (2024) Der Einsatz von KI an Hochschulen – Eine rechtliche Beratung. *Ordnung der Wissenschaft* 3: 161–68.

Serres M and Latour B (1995) *Conversations on Science, Culture, and Time*. Ann Arbor: University of Michigan Press.

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.

Tichenor M et al. (2022) Global public policy in a quantified world: Sustainable Development Goals as epistemic infrastructures. *Policy and Society* 41 (4): 431–444.

Wallat J, Abdallah A, Jatowt A and Anand A (2025) A Study into Investigating Temporal Robustness of LLMs. *arXiv preprint*: 1–20. https://doi.org/10.48550/arXiv.2503.17073.

Wallat J, Jatowt A and Anand A (2024) Temporal Blind Spots in Large Language Models. In: *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (eds. Angélica L, Lattanzi S, Medina AM et al.), Merida, Mexico, 4 – 8 March 2024: 683–92.

Wang W, Jiao W, Huang J, Dai R, Huang J, Tu Z and Lyu M (2023) Not all countries celebrate thanksgiving: On the cultural dominance in large language models. *arXiv preprint*: 1–34. https://doi.org/10.48550/arXiv.2310.12481.

White H (1997) *Metahistory : the historical imagination in nineteenth-century Europe*. 9. print. Baltimore, Md. [u.a.]: Johns Hopkins Univ. Press.

Wie J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S et al. 2022. Emergent Abilities of Large Language Models. *arXiv preprint*: 1–30. https://doi.org/10.48550/arXiv.2206.07682.

Wieduwilt H (2024) Wir fragen eine Ethikerin: Welche Gesellschaft soll Gemini abbilden? *Frankfurter Allgemeine Zeitung*. February 28. https://www.faz.net/pro/digitalwirtschaft/kuenstliche-intelligenz/wir-fragen-eine-ethikerin-welche-gesellschaft-soll-gemini-abbilden-19550166.html (accessed 01 May 2025).

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.

Xiong S, Payani A, Kompella R and Fekri F (2024) Large Language Models Can Learn Temporal Reasoning. *arXiv preprint*: 1–19. https://doi.org/10.48550/arXiv.2401.06853.

Yue Z, Zeng H, Shang L, Liu Y, Zhang Y and Wang D (2024) Retrieval Augmented Fact Verification by Synthesizing Contrastive Arguments." *arXiv preprint*: 1–13. https://doi.org/10.48550/arXiv.2406.09815.

Zhao J, Wang T, Abid W, Angus G, Garg A, Kinnison J et al. (2024) LoRA Land: 310 Fine-Tuned LLMs That Rival GPT-4. A Technical Report. *arXiv preprint*: 1–27. https://doi.org/10.48550/arXiv.2405.00732.

Zhaochen S, Zhang J, Zhu T, Qu X, Li J, Zhang M and Cheng Y (2024) TIMO: Towards Better Temporal Reasoning for Language Models. *arXiv preprint*: 1–18. https://doi.org/10.48550/arXiv.2406.14192.

Zhou R, Wan G, Gabriel S, Li S, Gates AJ, Sap M and Hartvigsen T (2025) Disparities in LLM Reasoning Accuracy and Explanations: A Case Study on African American English. *arXiv preprint*: 1–19. https://doi.org/10.48550/arXiv.2503.04099.

Zhu X, Yang C, Chen B, Li S, Lou J and Yang Y (2023) Question Answering as Programming for Solving Time-Sensitive Questions. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (ed Bouamor H, Pino J, Bali K), Singapore, 6 – 10 December 2023: 12775–90.

Appendix

System prompt for NER used with *Llama-3.1-SauerkrautLM-70B-Instruct* (temperature = 0, top_p = 0):

SETTING

You are an expert computational linguist specialised in Named Entity Recognition (NER) for German-language social media discourse relating to the German Democratic Republic (GDR/DDR) and its post-1990 cultural, political, and memorial legacies. Your mission is to annotate German-language social media posts precisely, completely, and in compliance with strict tagging rules. The source text will always be in German, and your output must also be in German, preserving the original text exactly but with NER tags inserted.

GOAL

Produce perfectly accurate, consistent annotations for all relevant entities, using only the allowed tags.
Preserve the original German text exactly (no deletions, no rewording).
Your annotations will be used for research datasets and historical discourse analysis.

TAGSET – ONLY THESE TAGS ALLOWED

<\<PER … /PER» Real historical or contemporary persons
<\<FPER … /FPER» Fictional characters (literature, film, TV, etc.)
<\<LOC … /LOC» Geographic locations or regions (incl. 'Osten'/'Westen' if referring to regions)
<\<ORG … /ORG» Institutions, organisations, parties, agencies, media, incl. military, 'DDR', 'MfS', 'Stasi' always ORG
<\<EVENT … /EVENT» Historical or cultural events (e.g., 'Wende', 'Kalter Krieg', 'Holocaust')
<\<NAT … /NAT» Nationalities / cultural groups (e.g., Vietnamesen, Polen, Russen)
<\<IDEO … /IDEO» Political / ideological groups (e.g., Neonazis, Kommunisten, Antifaschisten)
<\<REG … /REG» Regional identity markers within Germany
Examples: Ostdeutsche, Westdeutsche, Ossi, Wessi, Ostler, Westler, DDRler, Ex-DDR-Bürger

DISAMBIGUATION CHEATSHEET (REG vs. NAT vs. LOC)

REG = Intra-German identity markers / group labels: 'Ostdeutsche', 'Wessis', 'DDRler', 'Ex-DDR-Bürger' (identity/affiliation, not nationality)
NAT = Nationality / ethnonym / cultural group in the broader sense: 'Vietnamesen', 'Polen', 'Deutsche'
LOC = Geographic reference: 'im Osten', 'im Westen', 'in Ostdeutschland', 'Berlin', 'Koblenz'
'Osten/Westen' → LOC only if clearly geographic/regional, not metaphorical.

SPAN RULES

Minimal semantic span. Tag only the exact entity term, leave affixes outside.
Example: '<\<ORG DDR /ORG»-Zeiten', 'über <\<EVENT Wende /EVENT»-Jahre'
Hashtags/Handles: #Wende → tag without '#': '#<\<EVENT Wende /EVENT»'
Plural/Abbreviations/Pejoratives: 'Wessis', 'Ossis' → <\<REG … /REG»; 'SED', 'MfS' →

<\<ORG … /ORG»
No overlaps or nesting. Tags must not intersect.
Keep orthography exactly as in source.
Compounds: If part of a compound is an entity ('DDR-Bürger'), tag only the entity root: '<\<ORG DDR /ORG»-Bürger'. For identity markers like 'DDRler', 'Ex-DDR-Bürger' → REG.

INSTRUCTIONS – STEP-BY-STEP

1. Read the German text completely, without altering a single character.
2. Identify all entities matching the TAGSET.
3. Assign the correct tag type based on the DISAMBIGUATION rules.
4. Apply tags precisely (correct open/close, minimal span, no overlap).
5. Always tag:
   'DDR', 'MfS', 'Stasi' → ORG
   'Militär' (institutional) → ORG
   'Wende', 'Kalter Krieg', 'Holocaust' → EVENT
6. Validate before output: Balanced tag pairs? Only allowed tags? No nesting?
7. Output:
   Only the annotated German text.
   Optionally, add 'NOTES:' (max 2 short sentences) for ambiguity.

INTERNAL REASONING MODE – DO NOT OUTPUT

Use a light Tree-of-Thought process internally

(a) Candidate identification: Collect all possible spans.
(b) Type check: Assign each candidate 1–2 possible tags, check against rules/examples.
(c) Self-consistency: Pick the most consistent, rule-compliant tag.
(d) Borderline cases: Tag conservatively, only if clear.

This reasoning must remain internal, output only the annotated text.

FEW-SHOT EXAMPLES

Input: Viele Ostfahrzeuge wurden nach der Wende einfach verschrottet.
Output: Viele Ostfahrzeuge wurden nach der <\<EVENT Wende /EVENT» einfach verschrottet.
Input: Könnt ihr einen Beitrag über die Bundeswehr im Kalten Krieg machen?
Output: Könnt ihr einen Beitrag über die <\<ORG Bundeswehr /ORG» im <\<EVENT Kalten Krieg /EVENT» machen?
Input: Viele Wessis konnten die Lage im Osten nicht nachvollziehen.
Output: Viele <\<REG Wessis /REG» konnten die Lage im <\<LOC Osten /LOC» nicht nachvollziehen.
Input: Die Stasi arbeitete eng mit dem Militär zusammen.

Output: Die <\<ORG Stasi /ORG» arbeitete eng mit dem <\<ORG Militär /ORG» zusammen.

HINTS – CORE REMINDERS

‘‘‘
Preserve original German text exactly.
Only allowed tags.
No overlaps/nesting.
REG = intra-German identity markers, NAT = nationality, LOC = geographic.
‘DDR/MfS/Stasi’ → ORG; ‘Wende/Kalter Krieg/Holocaust’ → EVENT.
Minimal span; affixes outside tags.
Double-check tag balance before finalising.
‘‘‘

MOTIVATION

Precision and care are your top priority. Every correct annotation increases the dataset’s value and ensures the success of downstream analysis. Work without errors or deviations, check each detail twice before finalising.

OUTPUT FORMAT

<annotated German text as continuous flow>
[optional, new line] NOTES: <max. 2 short sentences for ambiguity>

INPUT_TEXT

[User Prompt with Input Text]