LLMs and multilingual historical corpora in a digital history project

Reflections from the Berlin workshop

Jeffrey C. Wolf

1. Introduction

My expectations for the Technical University of Berlin's LLM workshop were unclear beforehand. What would participants learn? How were others using LLMs in their HPSS research? Even the definition of Large Language Models remained confusing. Were they simply the models that powered the new generative AI tools, like ChatGPT, or something more? What was meant by "large" anyways? I was uncertain about the extent to which our ERC-funded VERITRACE project could use LLMs; there were no plans to offer chatbot-type access to the VERITRACE corpus, and even if there were, the volume of data made this approach impractical. Developments were also changing so rapidly in the field that the workshop's relevance to ongoing work in the coming weeks, let alone months, remained murky. Still, I hoped it would all become more clear at the Technical University in Berlin, and indeed, it did.

Before offering more specific reflections on the questions above, I will provide a summary of the VERITRACE project, and how we imagined various ways that LLMs might help us.

2. The VERITRACE project: context and challenges

The VERITRACE project[1] traces the influence of prominent ancient wisdom writings throughout the early modern period, focusing on their function in natural philosophical discourse (Schilt, 2022). It applies sophisticated digital analysis to a large corpus of early modern texts, tracing referenced and unreferenced uses of the *Corpus Hermeticum* (including the *Asclepius*), the *Chaldean Oracles* and *Sibylline Oracles*, and the *Orphic Hymns*. It analyses how these texts were used and with what sentiment they were discussed

1 https://veritrace.eu/

by proponents and antagonists, examining how key transmission episodes influenced these debates. VERITRACE will provide the first comprehensive analysis of ancient wisdom writings' influence on early modern natural philosophy using methodologies never employed at this scale in the early modern history of science.

VERITRACE draws on printed books, the period's most ubiquitous intellectual materials. While early modern debates involved oral discussion and manuscript circulation, these pertained to small circles. Books were everywhere, connecting authors and readers across Europe and beyond. Even focusing solely on a set of core languages—works in Latin, English, German, French, Dutch, and Italian—the number of surviving early modern books is staggering, presenting significant difficulties and opportunities.

Several challenges uniquely characterise VERITRACE as a digital humanities project:

1. **Multilingualism**: VERITRACE source material spans at least six languages, both modern and classical. Since many NLP techniques were developed for English users, this multilinguistic nature creates unique challenges, particularly with limited support for classical languages like Latin.
2. ***Longue durée***: Spanning almost two hundred years (1540–1728) our corpus adds diachronic complexity. Interpretations applicable to smaller data slices cannot be assumed to apply universally, given changing historical contexts and semantic meanings.
3. **Big Data**: With hundreds of thousands of texts, simple search processes and data management are inadequate. The project requires resource-intensive solutions given the collection sizes.
4. **Complex Integration**: Data from different institutional sources collected over long periods creates inherent integration and harmonisation challenges, requiring careful documentation of cleaning and transformation rules.

VERITRACE also confronts familiar distant reading challenges: OCR quality issues, parameter-dependent NLP techniques, and similar concerns.

3. Distant reading

Meaningfully tracing ancient wisdom writings' influence case-by-case would require massive research teams or drastically reduced scope. Digital techniques from distant reading (Moretti, 2005, 2013; Underwood, 2019), closely related to natural language processing, enable large corpus analysis, identifying patterns and uncovering both prominent and neglected works: Margaret Cohen's the 'great unread' (Cohen, 2009; Reid, 2019). Early modern writers rarely referenced source material explicitly, creating key project challenges.

Digitisation advancements have expanded distant reading applications. Improved OCR technology yields meaningful results despite suboptimal text recognition (Hill and Hengchen, 2019; Kurhekar et al., 2021). Online repositories like the Bibliothèque nationale de France provide standardised data, facilitating large-scale analysis.

VERITRACE's Distant Reading Corpus (DRC) consists of several hundred thousand works from major European collections in Latin, French, German, Dutch, English, and Italian:

- **EEBO-TCP** (Early English Books Online): ∼68,000 English and Latin texts (1540–1700)
- **Gallica** (Bibliothèque nationale de France): ∼20,000 books (1540–1728) in French and various languages
- **Digitale Sammlungen of the Bavarian State Library**: ∼340,000 books (1540–1728) in Latin, German, and other languages

These sources enable historical claims about the *prisca sapientia* tradition, its prevalence, and changing interest levels. By interrogating truly representative samples, reasonable claims about interest and prevalence become possible. A rigorous statistical approach underpins this methodology, with chosen sources forming the basis for representative sampling of European printed books (1540–1728).

4. Close reading too

VERITRACE combines state-of-the-art distant reading techniques on large corpora with close reading of carefully selected Renaissance and early modern texts.

VERITRACE's Close Reading Corpus (CRC) includes all relevant editions of the *Corpus Hermeticum*, *Chaldean Oracles*, *Sibylline Oracles*, and *Orphic Hymns* published during the Renaissance and early modern period, plus works drawing heavily on these writings and promoting *prisca sapientia* ideas, such as Ficino's *Theologia platonica* and Steuco's *De perenni philosophia*: approximately 80 works initially, which we have already expanded to c.150.

While a readership and geographical dissemination census would be valuable, as previous censuses of Copernicus's *De Revolutionibus*, Vesalius's *De Fabrica*, and Newton's *Principia* revealed a wealth of information (Palumbo, 2018; Margocsy et al., 2018; Feingold and Svorenčík, 2020), this remains impractical within our project's scope. Nonetheless, we will catalogue as many works in the tradition as we can.

This combination of distant and close reading methodology combines granular interpretation of select works with broader context. Previously, ancient wisdom texts' influence on natural philosophy was confined to case studies involving major figures—Copernicus, Kepler, Bacon, More, Newton. But early modern natural philosophy extended far beyond select individuals. While many never documented their ideas, thousands did. Ancient wisdom writings were widely read, with new Greek, Latin, and vernacular editions appearing regularly throughout the early modern period.

5. Large language models: do they have a place?

Where Large Language Models fit into our approach represents both an opportunity and a challenge unique to projects working with multilingual historical corpora. We did not immediately think of LLMs as a critical tool in our project; instead, our initial motivation

was more intellectual. We wanted to offer scholars the ability to uncover similar-meaning passages across texts on a wide scale.

We knew we could offer traditional keyword search tools that allow researchers to perform their own searches of the full VERITRACE corpus—a helpful tool for user-driven scholarship—but such searches are limited by requiring the user to know precisely what they are searching for, prior to use. And search queries are limited to keywords or short phrases.

We envisioned a valuable complement: searches where no such prior knowledge is necessary, where the tool could identify similar passages across the entire corpus. This would be based on textual and semantic similarity between query text(s) and comparison text(s). This led us to develop a Text Matching tool that functions as a kind of "early modern plagiarism detector"—an effective, surface-level tool focused on identifying matching passages and phrases through keyword and vocabulary analysis (Wolf et al., 2024).

However, our multilingual corpus presents challenges that surface-level matching cannot address. Traditional lexical matching excels at identifying shared vocabulary within languages but proves nearly useless across linguistic boundaries. Given that our corpus spans six core languages with Latin predominating, we also needed "Semantic Matching"—a tool capable of identifying semantic similarity between textual passages regardless of the language in which that meaning was encoded. This cross-linguistic semantic matching represents a valuable, complementary tool, enabling scholars to identify, for instance, passages written in Latin that are semantically similar to passages in Italian, thereby bridging the "multilingual gap."

This approach aligns with what recent digital humanities scholars have called "text similarity analysis" or "text reuse analysis," such as the work of the Helsinki Computational History Group[2]. This has proven particularly valuable for humanities research in identifying patterns of textual borrowing and influence across large corpora (Rosson et al., 2023; Ryan et al., 2023).

6. Embeddings models and the historical challenge

This semantic matching capability depends fundamentally on vector embeddings; specifically, Large Language Models of the full-context, bi-directional, encoder variety (Devlin et al., 2018) rather than the generative decoder models exemplified by GPT architectures (Vaswani et al., 2017). As recent scholarship has demonstrated, word embedding models can capture semantic information in large corpora by learning distributional properties of words—how often particular words appear in specific contexts—making them particularly valuable for historical research focused on tracing conceptual change over time, despite some drawbacks researchers must be aware of (Wevers and Koolen, 2020).

The workshop crystallized several critical insights about model selection that had not been top of mind before. Most significantly, I realised that our corpus, containing texts printed over 200 years (1540–1728), necessitated accounting explicitly for "semantic

2 https://www.helsinki.fi/en/researchgroups/computational-history

shift"—the fact that words change meaning over extended time periods. This challenge has been recognised in recent computational linguistics research, where scholars have proposed statistical laws governing semantic change: words used more frequently tend to change at slower rates (the law of conformity), while polysemous words change more rapidly regardless of frequency (the law of innovation) (Hamilton et al., 2016).

This temporal dimension creates a fundamental decision: should we employ a single embeddings model capable of capturing semantic change over our 200-year span, adding considerable complexity to the training process, or utilise different models for different time periods, which introduces critical issues in semantic alignment between models? Neither approach is clearly superior to the other and each one involves significant trade-offs that can affect the reliability of results. The importance of thinking rigorously about this choice only became clear to me at the LLM workshop in Berlin.

7. Four critical challenges for historical multilingual embeddings

Our workshop discussions helped clarify four interconnected challenges that any embeddings model must address when working with historical multilingual corpora of *longue durée*. First, **multilingual capability** requires robust support across our six core languages, with particular strength needed in Latin, which comprises the majority of our textual data. Recent research has shown that pre-trained multilingual models face significant challenges with historical content (due, in part, to OCR noise and outdated spellings), particularly when dealing with low-resource historical languages, though fine-tuning using contrastive learning can significantly increase accuracy (Michail et al., 2025).

Second, **OCR noise** presents ongoing difficulties. Studies have demonstrated that OCR noise may impact (though not catastrophically) the performance of downstream NLP tasks (Hill and Hengchen, 2019) with character-level mistakes proving problematic for embeddings-based topic models (Zosa et al., 2021). Our corpus contains substantial OCR noise, requiring models that can handle such imperfections or preprocessing pipelines that reduce OCR errors without introducing new distortions.

Third, the **out-of-domain** training problem reflects a broader issue in digital humanities applications of contemporary NLP tools. Recent studies demonstrate that most embedding models are trained on contemporary textual corpora, and this introduces a modern bias, pushing historical semantic representations closer to modern usages and obscuring historical contexts (Hamilton et al., 2016; Kutuzov et al., 2018). Indeed, that was one of the motivations for Qui and Xu to create HistBERT (Qui and Xu, 2022). The workshop's opening keynote "Large-scale text analysis for the study of cultural and societal change" by Pierluigi Cassotti and Nina Tahmasebi particularly stressed this limitation, noting how current LLMs, trained predominantly on modern data, can misrepresent historical language use.

What this suggests is that, without the right tools, we would struggle not just with representing historical semantics but with the entire **semantic shift** inherent in our corpus's 200-year time span (Kutuzov et al., 2018). Our early modern corpus requires

methodologies that can handle both the temporal scope and the linguistic complexity of pre-modern texts.

8. Methodological considerations and next steps

The workshop presentations provided valuable insights into how other HPSS researchers are addressing similar challenges. Projects using LLMs for computational epistemology, citation extraction from historical sources, and Retrieval Augmented Generation databases demonstrated various approaches to incorporating modern NLP tools into historical research workflows. These examples align with broader trends in digital humanities, where scholars are increasingly using vector embeddings not merely as technical tools but as means of exploring discursive spaces and semantic relationships within large textual collections.

Particularly relevant were discussions of metadata enrichment and diachronic analysis approaches that parallel our own VERITRACE objectives. Projects working with historical newspapers and other time-stamped corpora have developed sophisticated methodologies for tracking conceptual evolution, often employing temporal alignment strategies that train separate models for different periods while maintaining comparability across time.

Based on these workshop insights, our approach to model selection requires balancing multiple competing priorities. We are primarily interested in sentence-level embeddings models rather than word-level approaches, as our research focuses on identifying similar textual passages rather than individual word relationships.

Our current assessment suggests that no single existing model adequately addresses all four challenges outlined above. Models may excel in multilingual support but lack historical training data, or handle OCR noise sufficiently but prove inadequate for our temporal scope. Consequently, we anticipate the need to fine-tune a base model specifically for our corpus characteristics.

For base model selection, we are evaluating Microsoft's multilingual-e5-large-instruct (Wang et al., 2024), which offers strong multilingual support, including Latin. However, this choice depends on comprehensive evaluation across our specific requirements: Latin language capacity, OCR noise tolerance, and sensitivity to semantic shift. Our evaluation framework must assess whether the base model provides sufficient performance across these dimensions to justify subsequent fine-tuning to achieve reliable results.

The OCR quality issue may necessitate parallel development of correction models or preprocessing pipelines. Some recent studies have shown that while OCR post-correction can significantly improve downstream task performance, the effectiveness varies considerably depending on the specific characteristics of the historical corpus and the intended application (Todorov and Colavizza, 2020). For VERITRACE, this means carefully balancing preprocessing costs against potential improvements in semantic matching accuracy.

9. Conclusion

The Berlin LLM workshop provided a crucial perspective on the challenges and opportunities facing digital humanities projects seeking to incorporate advanced NLP methodologies. For VERITRACE, the key insight was recognising that our multilingual, temporally-extended corpus requires specialised solutions that address the interconnected challenges of multilingual support, OCR quality, domain adaptation, and semantic shift.

Rather than seeking to apply existing models wholesale, we must develop evaluation frameworks that can assess model performance against our specific research requirements and create fine-tuning approaches that address our corpus's unique characteristics.

In some ways, we were spared some of the hard decisions that await scholars who use Transformer-based LLMs more extensively than we do. Critical issues highlighted with these newer models include the substantial computational resources required for training and inference, the black-box nature of their decision-making processes that complicates interpretability for humanities research, and the ongoing concerns about hallucination and factual accuracy when working with historical texts. Additional challenges emerge around data provenance and the difficulty of ensuring that training data does not inadvertently influence results when analyzing historical corpora.

Our workhorse is not a chat-based model but an embeddings model on the encoder side that focuses specifically on semantic similarity tasks without the complexities of text generation. These encoder-based embedding models have been more thoroughly tested across diverse applications and have been around longer with established track records. It is hoped that they can form a solid basis for our Text Matching tool, even if many implementation details have not yet been finalised.

Finally, the workshop reinforced the value of interdisciplinary dialogue between digital humanities practitioners and domain-specific experts. Our specific use case—semantic matching across languages and centuries in early modern natural philosophy texts—represents the kind of specialised application that can push the boundaries of current methodologies, while addressing substantive historical research questions. As LLM capabilities continue to evolve rapidly, fostering collaboration between technical researchers and scholars focused on historical practices of science will be essential to fully realise their potential across historical research. The LLM Workshop in Berlin was a laudable example of that.

References

Cohen M (2009) Narratology in the Archive of Literature. *Representations* (108): 51–75.

Devlin J, Chang M-W, Lee K et al. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv.* Epub ahead of print. arXiv:1810.04805v2.

Feingold M and Svorenčík A (2020) A Preliminary Census of Copies of the First Edition of Newton's *Principia* (1687). *Annals of Science* 77(3): 253–348.

Hamilton WL, Leskovec J and Jurafsky D (2016) Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Berlin: Association for Computational Linguistics, pp. 1489–1501. Available at: https://aclanthology.org/P16-1141/ (accessed 18 November 2025).

Hill MJ and Hengchen S (2019) Quantifying the Impact of Dirty OCR on Historical Text Analysis: Eighteenth Century Collections Online as a Case Study. *Digital Scholarship in the Humanities* 34(4): 825–843.

Kurhekar P, Nigam S and Pillai S (2021) Automated Text and Tabular Data Extraction from Scanned Document Images. In: *Data Management, Analytics and Innovation. Proceedings of ICDMAI 2021* (eds N Sharma, N Chakrabarti, VE Balas, AM Bruckstein), pp.169–182. Singapore: Springer Singapore.

Kutuzov A, Øvrelid L, Szymanski T. et al. (2018) Diachronic Word Embeddings and Semantic Shifts: A Survey. In: *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp.1384–1397. Association for Computational Linguistics. Available at: https://aclanthology.org/C18-1117/ (accessed 18 November 2025).

Margocsy D, Somos M and Joffe SN (2018) *The Fabrica of Andreas Vesalius: A Worldwide Descriptive Census, Ownership, and Annotations*. Leiden: Brill.

Michail A, Raclé C, Opitz J. et al. (2025) Adapting Multilingual Embedding Models to Historical Luxembourgish. In: *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*. Albuquerque, New Mexico, USA, pp.291–298. Association for Computational Linguistics. Available at: https://aclanthology.org/2025.latechclfl-1.26/ (accessed 18 November 2025).

Moretti F (2005) *Graphs, Maps, Trees: Abstract Models for Literary History*. London: Verso.

Moretti F (2013) *Distant Reading*. London: Verso.

Palumbo M (2018) Books on the Run: The Case of Francesco Patrizi. In: Zwierlein C and Lavenia V (eds) *Fruits of Migration: Heterodox Italian Migrants and Central European Culture 1550–1620*. Leiden: Brill, pp.197–216.

Qiu W and Xu Y (2022) HistBERT: A Pre-trained Language Model for Diachronic Lexical Semantic Analysis. *arXiv.* Epub ahead of print. arXiv:2202.03829.

Reid D (2019) Distant Reading, 'the Great Unread', and 19th-Century British Conceptualizations of the Civilizing Mission: A Case Study. *Journal of Interdisciplinary History of Ideas* (15).

Rosson DE, Mäkelä E, Vaara V, et al. (2023) Reception Reader: Exploring Text Reuse in Early Modern British Publications. *Journal of Open Humanities Data* (9).

Ryan Y, Mahadevan A and Tolonen M (2023) A Comparative Text Similarity Analysis of the Works of Bernard Mandeville. *Digital Enlightenment Studies* (1): 28–58. DOI: 10.61147/des.6.

Schilt CJ (2022) Traces de la Verité: The Reappropriation of Ancient Wisdom in Early Modern Natural Philosophy. VERITRACE (ERC-2022-STG-101076836). Available at: https://veritrace.eu/wp-content/uploads/2023/04/Project-Traces-de-la-Verite-Condensed.pdf (accessed 18 November 2025).

Todorov K and Colavizza G (2020) Transfer Learning for Historical Corpora: An Assessment on Post-OCR Correction and Named Entity Recognition. In: *CEUR Workshop Proceedings* 2723, pp. 310–339.

Underwood T (2019) *Distant Horizons: Digital Evidence and Literary Change*. Chicago: University of Chicago Press.

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention Is All You Need. In: *Advances in Neural Information Processing Systems* 30.

Wang L, Yang N, Huang X, et al. (2024) Multilingual E5 Text Embeddings: A Technical Report. *arXiv*. Epub ahead of print. arXiv:2402.05672v1.

Wevers M and Koolen M (2020) Digital Begriffsgeschichte: Tracing Semantic Change Using Word Embeddings. In: *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 53(4): 226–243.

Wolf JC, Cantoni N, Kovács E, et al. (2024) The Challenges of Multilingualism in the Search for Ancient Wisdom: A Case Study of VERITRACE's Text Matching Tool. In: *Proceedings of the 4th Humanities-Centred AI (CHAI) Workshop at the 47th German Conference on Artificial Intelligence*, Würzburg, Germany, 23 September 2024. Available at: https://ceur-ws.org/Vol-3814/paper1.pdf (accessed 18 November 2025).

Zosa E, Mutuvi S, Granroth-Wilding M, et al. (2021) Evaluating the Robustness of Embedding-Based Topic Models to OCR Noise. In: *Towards Open and Trustworthy Digital Societies. 23rd International Conference on Asia-Pacific Digital Libraries, ICADL 2021, Proceedings*, 13133 (eds Ke H-R, Lee CS and Sugiyama K), Virtual Event, December 1–3, 2021, pp. 392–400. Lecture Notes in Computer Science. Singapore: Springer. Available at: https://hal.science/hal-03480518v1 (accessed 18 November 2025).