Embedded in the Labyrinth: Investigating Latin Word Senses through Transformer-Based Contextual Embeddings and Attention

Vojtěch Kaše[1], Sarah Lang[2], and Petr Pavlas[1]

[1] Institute of Philosophy, Czech Academy of Sciences, Prague, Czech Republic
[2] Max Planck Institute for the History of Science, Berlin, Germany

Abstract

This paper explores how transformer-based models can enhance historical keyword-in-context studies through automatic word sense disambiguation (WSD). Using the Latin term *labyrinthus* as a case study, we analyze its contextual meanings across time and genre within the GreLa corpus. A Large language model provides preliminary sense labels, which we use to evaluate 64 embedding variants—contextual, attention-based, and co-occurrence-based—derived from XLM-R and Latin BERT. Our results show that combining embedding types yields the best performance. We also illustrate how attention-based embeddings capture meaningful diachronic patterns, offering promising directions for future research on semantic change and metaphor in historical texts.

Keywords: labyrinth, keyword-in-context, computational Latin philology, contextual word embeddings, automatic word sense disambiguation, word sense induction, semantic change detection, metaphor detection

1 Introduction

Keyword-in-context search has long been a foundational technique in distant reading, particularly within the digital humanities, where it supports exploratory analysis and facilitates close reading. Tools such as Voyant have made this method widely accessible, especially for scholars interested in tracing terms across large corpora. However, the analytical reach of keyword-in-context remains limited, particularly when dealing with polysemous terms whose meanings shift across genres, periods, and discursive domains.

It is within this methodological space that our study intervenes. We focus on the term *labyrinthus*, tracing its semantic evolution across the longue durée of Latin literature. Initially rooted in classical mythology, the word later developed into a metaphorical device used in theological, philosophical, and scientific contexts. While prior scholarship has examined its poetic and religious resonances [6], its appropriation within early scientific discourse remains largely unexplored, although some intriguing examples are known in alchemical texts [17; 20; 21].

Our goal is to go beyond surface-level co-occurrence analysis and explore whether recent advances in computational linguistics—particularly transformer-based models—can help disambiguate the contextual meanings of historically layered terms. We propose that *labyrinthus* serves as a valuable case study for assessing the utility of various embedding strategies in tasks such as word sense disambiguation (WSD) and diachronic semantic analysis.

Vojtěch Kaše, Sarah Lang, and Petr Pavlas. "Embedded in the Labyrinth: Investigating Latin Word Senses through Transformer-Based Contextual Embeddings and Attention." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 498–512. https://doi.org/10.63744/FuaAvdPMdtwW.

To this end, our methodology consists of five main steps. First, we extract all occurrences of the target term from the GreLa corpus [12], a morphologically annotated and lemmatized collection of Latin and Greek texts (see Appendix C). Second, we use large language models (LLMs) to generate preliminary sense labels for each instance. Third, we generate multiple variants of contextual and attention-based embeddings using two transformer models. Fourth, we train four different classifiers on these embeddings to predict the LLM-generated sense labels. Finally, we evaluate classification performance and interpret results with an eye to both methodological implications and the semantic contours of the term itself.

Rather than aiming for a definitive taxonomy of senses, our goal is exploratory: to assess the comparative performance of embedding strategies, uncover semantic patterns in the data, and reflect on the strengths and limitations of these methods for historical-linguistic inquiry. In doing so, we position this study as a bridge between traditional philological concerns and recent innovations in computational language modeling.

2 Literature Review

Latin word sense disambiguation (WSD) is a well-established, though still relatively underexplored, task in digital humanities and historical linguistics. Early work by Bamman and Crane [2] introduced bilingual WSD using parallel corpora, while more recent efforts have aimed to formalize annotation frameworks for diachronic semantics [22]. Broader multilingual approaches have also been developed: [19] demonstrated scalable WSD across 158 languages using word embeddings, and [18] showed the effectiveness of contextualized transformer models such as LatinBERT over static embeddings. Ghinassi et al. [9] extended the field further by constructing the largest annotated Latin WSD dataset to date using a multilingual pivoting strategy.

While our study is related to WSD, it introduces additional complexity. We focus not only on sense disambiguation but also on metaphorical extension and semantic change across historical contexts. These themes are increasingly central in computational semantics. The emergence of transformer-based contextual embeddings has significantly advanced both WSD and metaphor detection. Their advantages over static, type-based embeddings such as word2vec [24; 26] are well documented [3; 26]. Some studies have identified layers within the transformer architecture—particularly deeper layers—as especially rich in semantic information [31; 32], while attention distributions have been shown to carry useful disambiguation signals as well [11]. At the same time, some researchers argue for the continued relevance of count-based contextual representations, especially in historical-linguistic applications [8].

These methodological discussions form the backdrop to our analysis. We position *labyrinthus* as a historically rich and semantically complex term – one whose interpretive range spans myth, poetry, theology, and early scientific discourse. By testing multiple embedding types and classification strategies against automatically generated sense labels, we aim to contribute to an ongoing conversation about the utility of modern NLP tools for historical semantic inquiry.

3 Materials

Our analysis begins with the extraction of all occurrences of the term ‘labyrinthus’ from the GreLa corpus, along with essential metadata, including the name of the author, the title of the work, and the date of composition or publication. Given that GreLa provides morphologically annotated and lemmatized data, it enables the precise retrieval of all instances of the term without recourse to wildcard characters or fuzzy search techniques.

For each identified occurrence, we extracted four types of contextual information. First, we retrieved the raw text of the sentence in which the term appears. Second, we collected the full set of morphological annotations – comprising the lemma and part-of-speech tag – for each token within

the sentence. Third, we extracted a broader context window consisting of the sentence containing the target term, preceded and followed by one adjacent sentence each. Finally, we constructed a concordance of morphologically annotated data comprising the ten tokens preceding the target term, the term itself, and the ten tokens following it. In our experimental setup, we assess the utility of these different types of contextual data as inputs for computational models.

4 Methods

4.1 Automatic sense labeling

To obtain classified sense labels for the usage of the term 'labyrinthus' across the target sentences, the domain experts in our team inspected a random sample of sentences and designed a prompt for LLM to classify the sense into 6 different categories:

1 – **Mythological:** References to the Cretan myth or its main figures and settings (Daedalus, Minotaur, Ariadne, Theseus, Crete). Restricted to explicit mythic allusions.

2 – **Technical literal:** Descriptions of actual or imagined physical labyrinths such as buildings, caves, mines, or mechanical structures, used non-figuratively.

3 – **Confusion metaphorics:** Figurative uses expressing moral, spiritual, or intellectual confusion, entrapment, or the search for guidance and clarity.

4 – **Scientific complexity metaphorics:** Uses in scientific or natural-philosophical discourse where "labyrinthus" conveys systemic or structural intricacy (as in nature, geometry, alchemy, or method).

5 – **Medical anatomical:** Anatomical or physiological contexts, especially references to the inner ear and related bodily structures. Applies even within scientific treatises.

6 – **Ambiguous or indeterminate:** Bibliographic entries, cross-references, or fragmentary cases where the intended sense cannot be determined.

The prompt was then executed on our local server machine using Lamma 4 Scout (109B) LLM [23] as available through the ollama library [25] (The full prompt is available as Appendix A). Subsequently, we repeatedly inspected a random sample of 100 sentences to evaluate the performance of the LLM and the embedding-based models (see Appendix B).

4.2 Embeddings

Using these annotated instances, we proceeded to employ the XLM-RoBERTa-BASE (hence XLM-R) [5] and Latin BERT (hence LaBERT) [1] models to obtain contextualized word embeddings of the target token. To investigate the semantic signal distributed across the encoder stack, we extracted target-token embeddings from the last 5 hidden layers of the model (i.e., layers 8–12 in 1-based indexing) for both sentence and concordance contexts.

In addition to the out-of-the-box models, we further fine-tuned Latin BERT specifically for Word Sense Disambiguation (WSD) using a combination of manually and automatically sense-annotated resources. Training pairs were constructed from three complementary datasets: (1) the Latin subset of the SemEval lexical semantic change task [30], (2) the dictionary-based sense usage dataset accompanying the publication of Latin BERT [1], and (3) the silver sense-labeled corpus produced by [9]. Each pair consisted of two contextualized occurrences of the same lemma labeled as either expressing the same or different senses. The model was trained in a contrastive WiC-style setup, starting from a frozen encoder and subsequently unfreezing the top three transformer layers for controlled fine-tuning. This procedure yielded a WSD-optimized Latin BERT variant fine-tuned for word-sense discrimination in historical Latin.

To extract attention-based features, we computed layer-specific attention distributions centered on a target lemma. For each sentence or concordance window, we identified the subword tokens corresponding to the target lemma and retrieved the attention scores from the rest of the input to the target. We averaged the attention scores across the subword span and selected the top-k attention heads with the highest total attention mass. The final attention vector was obtained by averaging the attention distributions of these top heads. This vector was then aggregated by lemma: for each non-target lemma in the input, we summed the attention weights assigned to its subword tokens. This yielded a dictionary mapping each lemma to its total attention score and providing the individual subword-level contributions.

These token-weight dictionaries were then used to construct attention embeddings. For each context, we created a high-dimensional sparse vector representing attention weights over a shared vocabulary of lemmata, restricted to tokens morphologically tagged as nouns, verbs, adjectives, or proper names, and occurring in at least two contexts. These serve as our raw attention embeddings. To further reduce dimensionality and mitigate sparsity, we applied Truncated Singular Value Decomposition (SVD), projecting the raw attention vectors into a 400-dimensional dense space that retained over 95% of the variance in the original matrix.

This procedure yielded 30 plus 30 plus 18 embedding variants per instance of *labyrinthus*: 5 target-lemma contextual embeddings, 5 attention-based raw embeddings, and 5 attention-based SVD-reduced embeddings, for each of the two context types (sentence and concordance) and base model (XLM-R and LaBERT), together with 3 target-lemma contextual embeddings, 3 attention-based raw embeddings, and 3 attention-based SVD-reduced embeddings, for each of the two context types (sentence and concordance) and the WSD-optimzed LaBERT model (as only the last 3 layers of the encoder were finetuned).

As a non-transformer baseline, we also constructed word-document co-occurrence matrices from both sentence and concordance contexts, using the same lemmatized vocabulary as in the attention embeddings. Again, we retained both their sparse and SVD-reduced forms. Altogether, this yields a total of 82 individual embedding variants.

4.3 Word Sense Disambiguation Task

To evaluate the effectiveness of the embedding variants described above, we framed the task as a multiclass word sense disambiguation (WSD) problem. Each instance of the target lemma *labyrinthus* was assigned a sense label using our large language model (LLM)-based classification pipeline (described in Appendix A). We treated these labels as ground-truth classes and tested how well different embedding representations could predict them.

We experimented with four supervised classification algorithms, using the embedding vectors as input features:

- **Multinomial Logistic Regression (LR)** [33]: A linear classifier that models class probabilities via a softmax function over linear combinations of input features.
- **Random Forest (RF)** [4]: An ensemble of decision trees trained via bootstrap aggregation (bagging), with random feature selection at each split to reduce overfitting and increase robustness.
- **Histogram-based Gradient Boosting (HGB)** [14]: A boosting method that discretizes continuous input features into histograms, enabling fast training and efficient memory usage while capturing non-linear patterns.
- **Multilayer Perceptron (MLP)** [28]: A feed-forward neural network with multiple hidden layers trained via backpropagation, capable of learning complex non-linear representations.

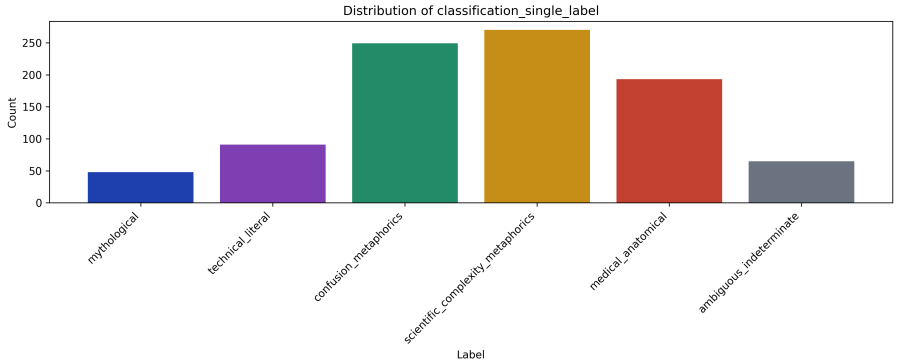

Figure 1: Distribution of *labyrinthus* sense labels across the dataset

These classifiers were selected to capture a range of inductive biases and strengths across different types of input representations. Logistic Regression offers a fast, interpretable baseline and performs well on dense, low-dimensional data such as contextual embeddings. However, it struggles with sparse or non-linearly separable inputs. Random Forests are better suited for sparse and categorical features and can model non-linear interactions, though they may be less effective on high-dimensional dense inputs. Histogram-based Gradient Boosting provides a balance between flexibility and efficiency, handling both sparse and dense features while modeling complex interactions. Finally, the Multilayer Perceptron is well-suited for heterogeneous high-dimensional data, especially when combining sparse attention vectors with dense embeddings. Its strong non-linear modeling capacity makes it effective in capturing distributed semantic signal across diverse embedding variants.

All the below introduced analyses and visualizations were implemented using the Python 3 programming language [29]. We make the code available for reuse via Github: https://github.com/CCS-ZCU/labyrinth.

5 Results

5.1 Automatic Sense Labels and Embedding Projections

The automatic sense labeling task yielded a distribution of *labyrinthus* instances across the six categories described in the previous section as visualized in Figure 1. This distribution provides a first overview of the dominant and rare senses identified by the LLM classifier.

As an initial sanity check and exploratory step, we examined 2-dimensional projections of the embedding variants using three dimensionality reduction techniques—UMAP, t-SNE, and PCA—each with three different parameterizations. These visualizations helped assess whether different senses exhibit distinguishable structure in embedding space.

Figure 2 shows the projections of the sentence-based co-occurrence baseline embeddings (reduced to 400 dimensions using SVD). Despite their simplicity, these embeddings reveal some meaningful clustering—most notably for the *medical_anatomical* sense (highlighted in red), which appears to be clearly separated. For comparison, Figure 3 displays projections based on concatenated contextual and attention embeddings from layer 10 (1-based index) of WSD-optimized Latin BERT. These patterns appear more diffuse, suggesting a richer and less easily interpretable feature space.

Full projections for all 82 embedding variants and their combinations are available in the project repository under the path `./figures/labyrinthus_projections_*.png`.

Figure 2: 2D projections of the co-occurrence baseline embeddings (400D SVD of sentence-level co-occurrence matrix) using UMAP, t-SNE, and PCA.

5.2 Sense Classification Evaluation

To quantitatively evaluate the semantic signal captured by different embedding variants, we trained supervised classification models with embeddings as predictors and LLM-assigned labels as classes. In addition to models based on individual embedding types, we evaluated combinations of contextual and attention embeddings, both raw and SVD-projected, as well as combinations with the baseline co-occurrence vectors. In total, this resulted in 186 input configurations. For each of the input configurations, we trained classification models using four different classification algorithms (LR, RF, HGB, MLP). To account for random variation and to allow a more rigid performance evaluation, each input configuration and classification model pair were trained ten times to account for random variation, and macro-averaged F1 scores were computed across runs. This enabled us to identify the best performing classification algorithm for the different embedding types and their combinations (For full overview of the results, with mean F1 scores and standard deviation across the 10 training iterations for each classification model and each input configuration, see the CSV file `./data/labyrinthus_classification_results.csv` in the project repository).

Regarding individual classification algorithms, we observe that

- **Logistic Regression (LR)** performed reasonably well in dense contextual embeddings, but failed to detect signals in inputs based on sparse raw attention embeddings.

Figure 3: 2D projections of concatenated contextual and attention embeddings from layer 10 of WSD-optimized Latin BERT using UMAP, t-SNE, and PCA.

- **Random Forest (RF)** showed strong results with sparse attention embeddings, even after SVD projection, but underperformed on dense inputs.
- **Histogram-based Gradient Boosting (HGB)** offered balanced performance across input types but was consistently outperformed by MLP.
- **Multilayer Perceptron (MLP)** achieved the best performance across the majority of embedding variants, particularly with combined dense and sparse features.

The best-performing configuration was an MLP model trained on a concatenation of:

- contextual embeddings (XLM-R, concordance context, layer 8),
- raw attention embeddings (XLM-R, concordance context, layer 8), and
- the concordance-level co-occurrence baseline (raw).

This model achieved an average F1 score of 0.678365, with some runs reaching up to 0.689177.

Interestingly, the co-occurrence baseline alone yielded relatively strong performance, surpassing most attention-only and contextual-only models. This suggests that the distributional information captured by count-based methods provides complementary signal that is not fully recoverable from transformer-based embeddings.

We also find that across all model types, XLM-R embeddings generally outperformed those from Latin BERT, though the margin was often small. No consistent pattern emerged regarding layer-specific performance; while deeper layers (e.g., layer 12) sometimes provided the clearest signal for classification, lower layers (especially 8 and 9) occasionally yielded more visually coherent clusters in projection plots.

We observed a slight advantage of concordance-level context across the board, indicating that the fluctuating length of input sentences can negatively affect the performance of the models.

While using an LLM for both annotation and evaluation ensured internal consistency, it also risked a degree of circular validation, since the same model family defined and then tested the category boundaries. To address this limitation and independently assess reliability, we manually annotated a sample of 100 instances of labyrinthus and re-evaluated the best-performing classifiers against this human ground truth. The resulting confusion matrices (Figure 4) illustrate that the medical_anatomical sense is consistently the most distinct and reliably identified category. In contrast, the mythological and technical_literal senses appear more easily conflated with the two metaphorical classes, suggesting that their boundaries are semantically or contextually less sharp in the data.

For the more abstract or metaphorical usages, both attention-based embeddings and the co-occurrence baseline contributed essential discriminative cues, highlighting that distributional and contextual information capture complementary aspects of meaning. Notably, the WSD-optimized Latin BERT exhibited a relatively strong ability to distinguish the mythological sense, performing comparably to the best XLM-R variants despite the limited sample size.

Overall, this diagnostic test should be viewed as an initial validation rather than a definitive benchmark, given the small and manually curated evaluation set. Nevertheless, the results indicate that models trained on LLM-labeled data retain meaningful signal when applied to human-verified sense distinctions.

Finally, to explore temporal trends, we extended our 2D projection setup by adding a third dimension: the date of composition of the containing text. Figure 5 shows an example 3D projection using raw attention embeddings from layer 12 of WSD-optimized Latin BERT. Here we clearly observe the rise of the *medical_anatomical* sense and *scientific_complexity_metaphorics* in Early Modern texts, in contrast to the *mythological* and *technical_literal* sense, which dominates earlier periods. This illustrates the potential of transformer-based attention embeddings for studying semantic change over time.

6 Discussion

In this study, we explored how historically motivated research on keywords-in-context can be enriched by transformer-based approaches to word sense disambiguation (WSD). Using the Latin term labyrinthus as a case study, we designed a series of experiments to assess the effectiveness of various embedding strategies and their combinations in predicting the correct sense of the target word in context.

Several limitations of our approach warrant discussion.

First, the sense labels used as ground truth were generated through iterative prompting of a large language model (LLM). While the resulting categories were refined through close reading and expert review, we acknowledge that the boundaries between senses are fluid and subject to interpretative disagreement. In this light, the LLM-based labels function more as an operational taxonomy than as a fixed ground truth. Nevertheless, the observed stability of the classifications—both across runs and in comparison with human annotation—suggests that such LLM-guided annotation can serve as a viable proxy for large-scale exploratory work, especially when combined with targeted expert validation.

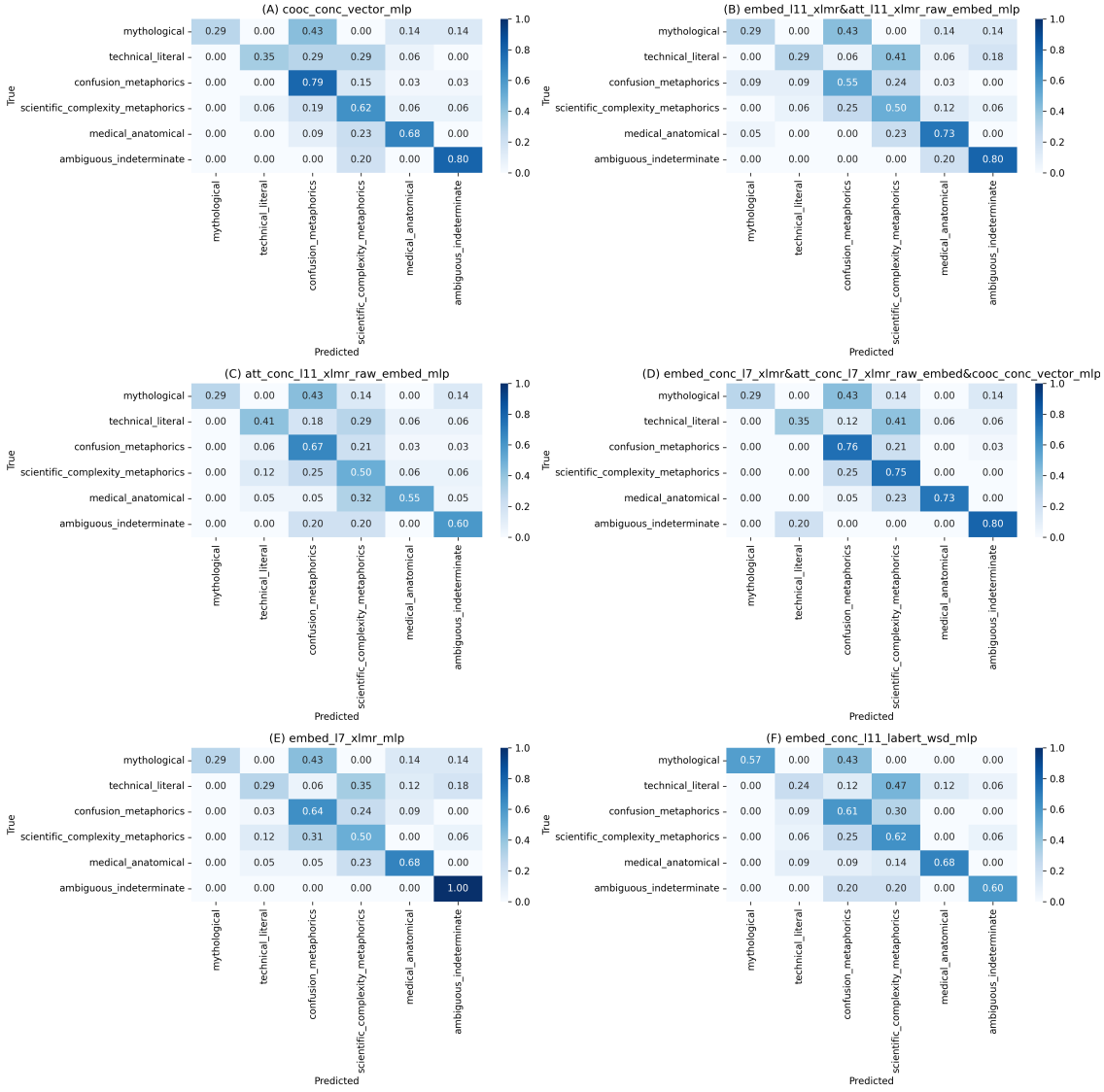

Figure 4: Confusion matrices of the six best-performing models evaluated on a manually annotated sample of 100 labyrinthus instances, showing F1 results per class and misclassification patterns.

Second, the conclusions we draw are necessarily limited by the scope of this single-term case study. Although labyrinthus provides a particularly rich and polysemous example—ranging from mythological and anatomical to philosophical and metaphorical uses—further work is needed to generalize these findings. In future research, we plan to extend our methodology to other semantically complex Latin terms with strong metaphorical potential, such as mercurius. The bilingual structure of the GreLa corpus also enables comparative analysis across Latin and Greek, offering new ways to model conceptual transfer and semantic interaction between the two languages.

Finally, our experiments only lightly addressed the diachronic dimension. Building on the initial 3D projections presented here, future work will explicitly model semantic change over time, tracing the emergence, dissemination, and transformation of senses across historical periods and genres.

Overall, this study demonstrates that the integration of LLM-based annotation with contextual, attention-based, and classical distributional embeddings can advance the precision and interpretive reach of computational philology. By linking quantitative modeling with philological

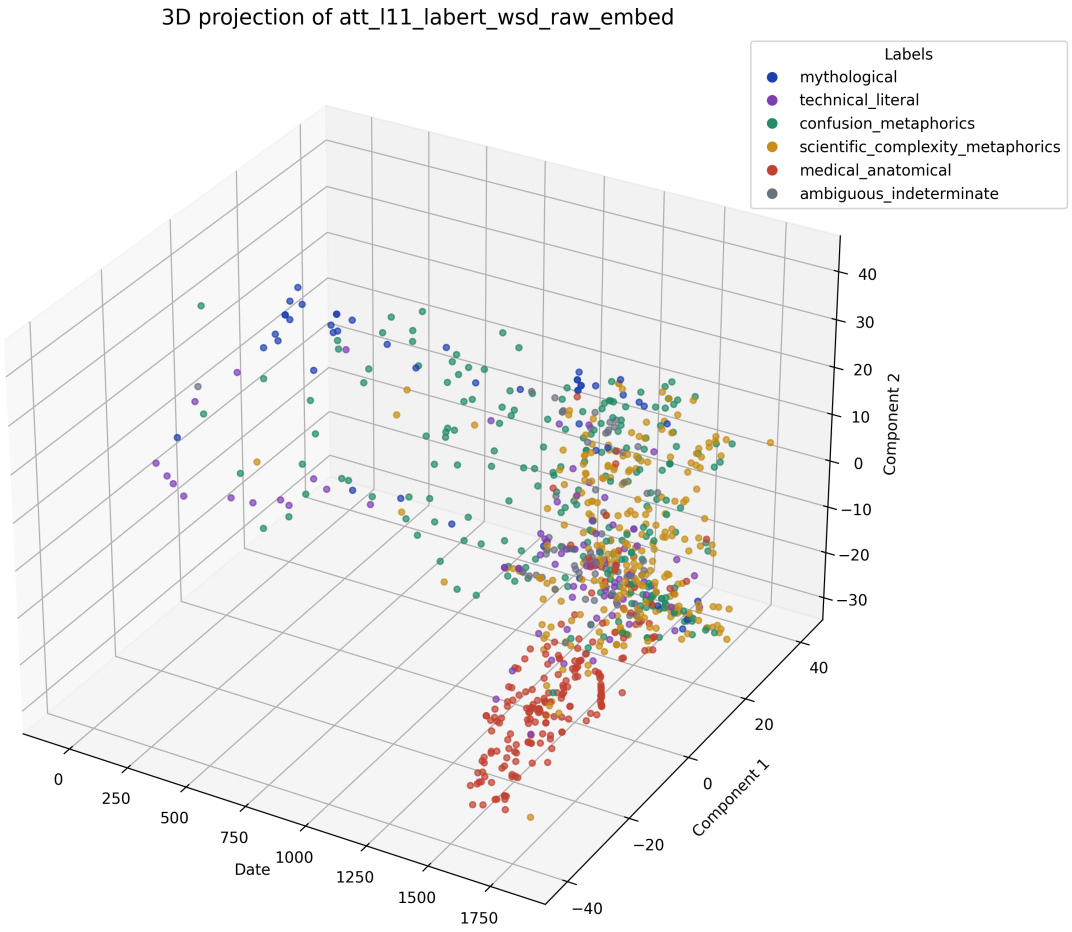

Figure 5: 3D projection of attention embeddings from layer 12 of WSD-optimized Latin BERT, with color-coded sense labels and date on the z-axis.

judgment, such hybrid methods can help bridge the gap between statistical regularity and historical meaning—a step toward more systematic, yet still interpretively sensitive, digital scholarship in the humanities.

Acknowledgements

The work of VK and PP was supported by the TOME project (Ministry of Education, Youth and Sports of the Czech Republic, ERC CZ, project no. LL 2320). Special thanks go to Farzad Mahootian (NYU New York), whose invitation of VK, SL, and PP to the workshop "Alchemy of Global Partnerships" (NYU Abu Dhabi, May 14–17, 2025) brought the three of us together and inspired the inception of this project.

References

[1] Bamman, David and Burns, Patrick J. "Latin BERT: A Contextual Language Model for Classical Philology". In: *arXiv preprint arXiv:2009.10053* (2020).

[2] Bamman, David and Crane, Gregory. "Measuring Historical Word Sense Variation". In: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. New York, NY, USA: Association for Computing Machinery, 2011, pp. 1–10. DOI: 10.1145/1998076.1998078.

[3] Bevilacqua, Michele, Pasini, Tommaso, Raganato, Alessandro, and Navigli, Roberto. "Recent Trends in Word Sense Disambiguation: A Survey". In: *International Joint Conference on Artificial Intelligence*. International Joint Conference on Artificial Intelligence, Inc., 2021, pp. 4330–4338.

[4] Breiman, Leo. "Random Forests". In: *Machine Learning 45*, no. 1 (2001), pp. 5–32.

[5] Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, and Stoyanov, Veselin. "Unsupervised Cross-Lingual Representation Learning at Scale". In: *arXiv preprint arXiv:1911.02116* (2019).

[6] Doob, Penelope Reed. *The Idea of the Labyrinth from Classical Antiquity through the Middle Ages*. Cornell University Press, 1992.

[7] Fröstl, Michael, Zathammer, Stefan, and Lang, Sarah. "Zur Transkription von Alchemica mithilfe der Transkribus-Software: Zu Handschriften, Drucken und dem NOSCEMUS GM 6 Modell". In: *Alchemistische Labore: Praktiken, Texte und materielle Hinterlassenschaften / Alchemical Laboratories: Practices, Texts, Material Relics*, ed. by Sarah Lang. Graz University Library Publishing, 2023, pp. 363–378.

[8] Geeraerts, Dirk, Speelman, Dirk, Heylen, Kris, Montes, Mariana, Pascale, Stefano de, Franco, Karlien, and Lang, Michael. *Lexical Variation and Change: A Distributional Semantic Approach*. Oxford: Oxford University Press, 2024.

[9] Ghinassi, Iacopo, Tedeschi, Simone, Marongiu, Paola, Navigli, Roberto, and McGillivray, Barbara. "Language Pivoting from Parallel Corpora for Word Sense Disambiguation of Historical Languages: A Case Study on Latin". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, 2024, pp. 10073–10084.

[10] Hedesan, Georgiana, Huber, Alexander, Kodetová, Jana, Kříž, Ondřej, Kubíčková, Jindra, Kaše, Vojtěch, and Pavlas, Petr. "Early Modern Latin Alchemical Prints (EMLAP) Corpus". 2025. DOI: 10.5281/zenodo.14765294.

[11] Ion, Radu, Păiş, Vasile, Mititelu, Verginica Barbu, Irimia, Elena, Mitrofan, Maria, Badea, Valentin, and Tufiş, Dan. "Unsupervised Word Sense Disambiguation Using Transformer's Attention Mechanism". In: *Machine Learning and Knowledge Extraction 7*, no. 1 (Jan. 2025), p. 10. DOI: 10.3390/make7010010.

[12] Kaše, Vojtěch. “GreLa (GitHub Repository)”. https://github.com/CCS-ZCU/GreLa. 2025.

[13] Kaše, Vojtěch, Söderholm, Harri, Vesala, Jimi, and Nikki, Nina. “Lemmatized Ancient Greek Texts Dataset (LAGT)”. 2024. DOI: 10.5281/zenodo.13889714.

[14] Ke, Guolin, Meng, Qi, Finley, Thomas, Wang, Taifeng, Chen, Wei, Ma, Weidong, Ye, Qiwei, and Liu, Tie-Yan. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”. In: *Advances in Neural Information Processing Systems 30* (2017).

[15] Korenjak, Martin. “Nova Scientia: Early Modern Scientific Literature and Latin (NOSCEMUS)”. Leopold-Franzens-Universität Innsbruck; ERC Advanced Grant (Nr. 741374, 2017–2023). 2023.

[16] Lang, Sarah. “A Machine Reasoning Algorithm for the Digital Analysis of Alchemical Language and Its Decknamen”. In: *Ambix 69*, no. 1 (2022), pp. 65–83. DOI: 10.1080/00026980.2022.2038428.

[17] Lang, Sarah. “Maier, Viatorium, 1618”. In: *Matthäus Merian d.Ä. und die Bebilderung der Alchemie um 1600*, ed. by Berit Wagner. Frankfurt: Virtuelle Ausstellung & Dynamische Wissensplattform, 2021.

[18] Lendvai, Piroska and Wick, Claudia. “Finetuning Latin BERT for Word Sense Disambiguation on the Thesaurus Linguae Latinae”. In: *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*. Taipei, Taiwan: Association for Computational Linguistics, 2022, pp. 37–41. DOI: 10.18653/v1/2022.cogalex-1.5.

[19] Logacheva, Varvara, Teslenko, Denis, Shelmanov, Artem, Remus, Steffen, Ustalov, Dmitry, Kutuzov, Andrey, Artemova, Ekaterina, Biemann, Chris, Ponzetto, Simone Paolo, and Panchenko, Alexander. “Word Sense Disambiguation for 158 Languages Using Word Embeddings Only”. arXiv preprint arXiv:2003.06651. 2020.

[20] Maier, Michael. *Viatorium, hoc est, de monitibus planetarum septem seu metallorum*. Oppenheim: Johann Theodor de Bry, 1618.

[21] Matton, Sylvain. *Le filet d’Ariane: Pour entrer avec sécurité dans le labyrinthe de la philosophie hermétique, précédé de variations alchimiques sur le symbole et le mythe du labyrinthe*. Paris: Gutenberg Reprints, 2006.

[22] McGillivray, Barbara, Kondakova, Daria, Burman, Annie, Dell’Oro, Francesca, Sabel, Helena Bermúdez, Marongiu, Paola, and Cruz, Manuel Márquez. “A New Corpus Annotation Framework for Latin Diachronic Lexical Semantics”. In: *Journal of Latin Linguistics 21*, no. 1 (2022), pp. 47–105. DOI: 10.1515/joll-2022-2007.

[23] Meta AI. “Introducing LLaMA 4: Advancing Multimodal Intelligence”. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. 2024.

[24] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S., and Dean, Jeff. “Distributed Representations of Words and Phrases and Their Compositionality”. In: *Advances in Neural Information Processing Systems*, ed. by C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger. Vol. 26. Curran Associates, Inc., 2013, pp. 3111–3119.

[25] Ollama. “Ollama: Run Large Language Models Locally”. https://ollama.com. 2023.

[26] Periti, Francesco and Tahmasebi, Nina. “A Systematic Comparison of Contextualized Word Embeddings for Lexical Semantic Change”. Mar. 2024. URL: http://arxiv.org/abs/2402.12011.

[27] Roelli, Philipp. “The Corpus Corporum: A New Open Latin Text Repository and Tool”. In: *Archivum Latinitatis Medii Aevi 72* (2014), pp. 289–304.

[28] Rosenblatt, Frank. “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain”. In: *Psychological Review 65*, no. 6 (1958), p. 386.

[29] Rossum, Guido Van and Drake, Fred L. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009. ISBN: 1441412697.

[30] Schlechtweg, Dominik, McGillivray, Barbara, Hengchen, Simon, Dubossarsky, Haim, and Tahmasebi, Nina. “SemEval-2020 Task 1: Unsupervised Lexical Semantic Change Detection”. In: *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. Barcelona (online): International Committee for Computational Linguistics, 2020, pp. 1–23. DOI: 10.18653/v1/2020.semeval-1.1.

[31] Teglia, Simone, Tedeschi, Simone, and Navigli, Roberto. “How Much Do Encoder Models Know About Word Senses?” In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. Vienna, Austria: Association for Computational Linguistics, 2025, pp. 2266–2277. DOI: 10.18653/v1/2025.acl-long.113.

[32] Tenney, Ian, Das, Dipanjan, and Pavlick, Ellie. “BERT Rediscovers the Classical NLP Pipeline”. In: *arXiv preprint arXiv:1905.05950* (2019).

[33] Theil, Henri. “A Multinomial Extension of the Linear Logit Model”. In: *International Economic Review 10*, no. 3 (1969), pp. 251–259.

[34] Zathammer, Stefan. “Noscemus Digital Sourcebook”. 2025. DOI: 10.5281/zenodo.15040256.

A LLM classification prompt

To support the annotation of word sense labels, we implemented a lightweight prompt-based classification system using a local LLM API (based on `llama4:17b-scout-16e-instruct-q4_K_M`). The model was prompted with a few-shot instruction template and returned a numerical label representing one of ten predefined semantic categories. Each occurrence of the word *labyrinthus* was classified into exactly one of the six categories.

Prompt The prompt included the classification instruction and a few-shot setup with three labeled examples. Each prediction was obtained by inserting the test passage as the final example:

```
You are a Latin philologist and semantic analyst.
Classify how "labyrinthus" is used in a Latin passage.

Return exactly ONE digit (0-5). No words, no punctuation.

Categories:
0 - mythological
Explicit reference to the Cretan myth (Daedalus, Minotaur, Ariadne, Theseus, Crete, Ovid). Not general archit

1 - technical_literal
Literal descriptions of built or natural labyrinths (buildings, caves, gardens, fishing traps, mines). No met

2 - confusion_metaphorics
Moral, spiritual, or epistemic confusion, entrapment, or rhetorical intricacy.
Common cues: error, confusio, filum (gratiae), via/exitus, anima/animus, peccatum, sophistae.

3 - scientific_complexity_metaphorics
Scientific or natural-philosophical contexts where "labyrinthus" symbolizes complexity, method, or structure
Common cues: methodus, hypotheseis, experimentum, calculi, astronomia, geometria, elementa, natura, systema.

4 - medical_anatomical
```

```
Anatomical/physiological context (esp. inner ear): auris, cochlea, vestibulum, tympanum, nervus, membrana, me

5 - ambiguous_indeterminate
Bibliographic/index mentions, titles, or genuinely unclear/fragmentary contexts.

Decision rules:
• Body parts/functions → 4
• Concrete real-world structures (literal) → 1
• Scientific/systemic metaphor → 3
• Moral/spiritual/epistemic confusion → 2
• Mythic narrative → 0
• Indeterminate or purely bibliographic → 5

Clarifications:
- Use 4 only for real anatomy (esp. inner ear). If the body is used metaphorically/structurally (e.g., cerebr
- Prefer 3 for scientific or methodical reasoning about complex systems, even when Ariadne/filum imagery appe
- Use 2 for human moral or intellectual struggle, not for technical/analytic complexity.
```

Prediction Protocol Each passage was submitted as a JSON payload via a REST API call. The model's prediction was parsed as an integer and matched against the predefined label map. This setup allowed us to rapidly assign semantic class labels to all instances of *labyrinthus* in the corpus for supervised evaluation of embedding-based classifiers.

B Discussion of results of the LLM labelling task

In the classification task, human annotations were compared with outputs from the LLM classification and the predictions of the best performing models. Overall performance was strong across methods, with high agreement between human and machine annotations. Discrepancies typically arose from differences in granularity or interpretive framing rather than from clear misclassifications.

The category system employed had a substantial impact on outcomes. Categories were selected to remain manageable, based initially on LLM-generated suggestions and refined to reduce ambiguity. However, the chosen set sometimes failed to reflect the nuances of metaphorical usage. Overlapping or closely related domains frequently produced borderline cases where both human and machine classifications varied.

However, misclassifications did occur. These were most serious when the LLM or the embedding based models failed to recognise clear disciplinary contexts, such as alchemical passages referencing Paracelsus, or astronomical texts dense with technical vocabulary. Such errors suggest limitations in the model, particularly in distinguishing overlapping terminology common in Early Modern texts. In these cases, domain-specific indicators that are readily identified by human reviewers were not consistently picked up by the automatic classifications.

The classification of metaphor relied on contextual cues rather than on function or type, echoing simple KWIC-based distributional semantics techniques previously suggested for the disambiguation of alchemical nomenclature through surrounding keywords [16]. While effective in many cases, this approach would benefit from refinement. Incorporating bibliographic metadata, field-specific specialist vocabulary and clearly defined category boundaries could improve accuracy, especially in ambiguous cases. For instance, texts from the *EMLAP* or *NOSCEMUS* corpora often signalled alchemical or scholarly contexts that the LLM failed to fully exploit.

Ultimately, model outputs largely aligned with human reasoning, especially where domain-specific vocabulary or concrete cues were present. However, future iterations would benefit from a finer-grained category system and explicit prompts to capture domain knowledge.

C Corpus Selection and Limitations

Motivation Given the size and heterogeneity of available datasets, manual inspection of all occurrences of the term *labyrinthus* proved impractical. Most instances are likely non-metaphorical and thus fall outside the scope of our investigation. Preliminary classification using LLMs and embedding models provided an effective means of narrowing the dataset to relevant cases.

The aggregate corpus and its components GreLa [12] constitutes an extensive corpus encompassing Greek and Latin literature from the eighth century BCE to the seventeenth century CE.

It comprises over 11,000 individual works, amounting to approximately 26 million sentences and 380 million tokens. The corpus is the result of an integration of several pre-existing collections.

Besides a Greek component derived from the LAGT (Lemmatized Ancient Greek Texts) Corpus [13], which consolidates ancient Greek texts from multiple sources, including the Perseus Digital Library, *The First Thousand Years of Greek*, Glaux, and the OGA corpus but is not directly relevant to our current case study, there is an extensive Latin component. It incorporates material from *Corpus Corporum* [27], a broad-based Latin literature repository, as well as two databases focused on early modern scientific and alchemical texts: NOSCEMUS [7; 15; 34], which covers early modern scientific literature, and *Early Modern Latin Alchemical Prints* (EMLAP, [10]).

Our corpora vary significantly in composition and purpose. The *NOSCEMUS* corpus, designed to trace the development of early modern science from circa 1450 to 1850, offers relatively balanced coverage across centuries. The *EMLAP* corpus focuses on sixteenth-century Latin alchemical texts, though some texts may predate this period as the inclusion in the corpus is based only on appearance date of the specific edition included. These two corpora represent the quite systematically assembled components for their specific aims. In contrast, the *Corpus Corporum* sources are more eclectic in their coverage. The corpus is shaped by availability of digitized sources, rather than a design aimed at thematic or chronological balance.

Representativeness This unevenness in the GreLa corpus introduces bias. Scientific and alchemical texts are likely overrepresented, while other genres remain under-sampled. The dataset is therefore not suitable for answering general questions about metaphor usage in Latin literature. However, it is well suited to tracing metaphorical patterns in early modern scientific and alchemical writing, which we are particularly interested in. We expect these to be particularly present in paratextual or introductory sections of alchemical works.

We also had to contend with the diachronic range of the corpus, which complicates metaphor identification due to shifting language use. Nonetheless, plotting metaphorical occurrences over time and by context remains potentially informative, provided these limitations are kept in view.

Although the corpus is imbalanced and incomplete, these features do not invalidate our method. These computational tools are used to guide close reading rather than replace it. In this sense, the visualisations and outputs are exploratory, pointing toward promising passages for detailed analysis in the form of close readings.

Indeed, the very metaphor that sparked our interest in this study – an intriguing alchemical use of *labyrinthus* as a metaphor – lies outside the scope of this corpus, underscoring that the methodological value of our study is not necessarily tied to the exhaustiveness of our corpus.

Critiques of digital humanities often target the representativeness of corpora. While such concerns are valid, they apply equally to traditional scholarship, which typically relies on a narrow set of texts due to practical constraints. Our approach seeks to expand the range of material available for interpretation while remaining conscious of its limitations. The digital method serves as a tool for identifying sources, not a substitute for historical analysis. As such, corpus imbalance, though significant, is not a decisive flaw for the kind of exploratory work undertaken here.