

Semantic Search for Ancient Inscriptions

Micah Tongen¹, Sara Sprenkle¹ , Rebecca Benefiel² , and Trevor Stalnaker³ 

¹ Computer Science Department, Washington and Lee University, Lexington, VA, U.S.A.

² Classics Department, Washington and Lee University, Lexington, VA, U.S.A.

³ Computer Science Department, William & Mary, Williamsburg, VA, U.S.A.

Abstract

Digital humanities research has revolutionized the study of ancient inscriptions by providing researchers with access to immense epigraphic corpora. However, traditional search methods for these databases rely primarily on exact or fuzzy keyword matching, limiting researchers' ability to find semantically related inscriptions. This paper presents a new approach to searching ancient inscriptions using vector embeddings and semantic similarity, implemented through a hybrid search system that combines semantic search with keyword matching and large language model re-ranking. Our system processes Greek and Latin inscriptions from the Ancient Graffiti Project database, embedding them in a high-dimensional vector space that captures semantic meaning beyond exact text matches. Our process is designed for reproducibility, using open data and code, and shows promise in preliminary evaluation. Our results demonstrate the system's capability to identify thematically related inscriptions that would be missed by traditional search methods, offering new possibilities for epigraphic research and discovery.

Keywords: semantic search, ancient graffiti, epigraphy, natural language processing, word embedding, classical languages

1 Introduction

The field of epigraphy (*i.e.*, the study of inscriptions) is an important subject for historians, offering direct insights into the thoughts, beliefs, and daily lives of ancient peoples. For example, the preserved cities of Pompeii and Herculaneum, with their thousands of inscriptions ranging from formal dedicatory texts to scrawled insults, provide an exceptionally rich corpus for understanding Roman society [3]. However, the sheer volume of epigraphic material presents significant challenges for researchers seeking to identify patterns, connections, and themes within these texts.

Current epigraphic databases mostly provide exact or fuzzy keyword matching on inscription text [11], an approach that limits researchers' ability to discover semantically related inscriptions. Consider a researcher searching for inscriptions about "victory": the returned results might miss relevant texts that discuss success, winning, and celebration, concepts that are thematically connected but don't use exactly the same term. Variations in spelling, word form (*e.g.*, declension, gender, tense, *etc.*), and abbreviation can also confuse these traditional keyword searches, forcing scholars to search the database manually or search for only portions of words, a technique that can yield false positives.

This paper begins to address these limitations by introducing a semantic search system for ancient inscriptions that uses modern natural language processing techniques. Our approach combines vector embeddings, which capture semantic meaning in a high-dimensional numerical representation, with traditional keyword matching. By using transformer-based models and large language model (LLM) re-ranking, the system provides thematically related inscriptions that share conceptual connections rather than just entries with words that match the query or are syntactically similar. We focus on ancient Roman graffiti, an especially challenging case due to the informality and orthographic inconsistency of the inscriptions [4], which seemed unlikely to yield robust embeddings. We argue that if semantic search

Micah Tongen, Sara Sprenkle, Rebecca Benefiel, and Trevor Stalnaker. "Semantic Search for Ancient Inscriptions." In: *Computational Humanities Research 2025*, ed. by Taylor Arnold, Margherita Fantoli, and Ruben Ros. Vol. 3. Anthology of Computers and the Humanities. 2025, 513–525. <https://doi.org/10.63744/xLUm5DRZmDDq>.

© 2025 by the authors. Licensed under Creative Commons Attribution 4.0 International (CC BY 4.0).

succeeds on a small dataset of linguistically challenging graffiti, it will generalize to larger and more standardized epigraphic corpora.

The contribution of this work is threefold: (1) we demonstrate how to apply semantic search techniques to ancient epigraphic corpora, (2) we present a hybrid search architecture that combines multiple retrieval strategies for improved precision, and (3) we offer a first step toward future developments in computational epigraphy that could transform how researchers discover and analyze ancient inscriptions.

2 Background

2.1 Digital Epigraphy

The field of epigraphy has been completely transformed by digital humanities researchers who have developed massive databases containing detailed information about hundreds of thousands of Inscriptions from ancient cities across Europe, north Africa, and the middle east [5; 9; 12]. Traditional epigraphic research required scholars to manually examine physical inscriptions and rely on printed corpora, limiting the scope and speed of their analysis. Digital epigraphy has provided unprecedented access to classical writings through user-friendly interfaces.

Digital encoding standards, particularly EpiDoc, have facilitated the creation of structured and interoperable databases [10]. There is an international community around EpiDoc that produces guidelines and tools for encoding scholarly and educational editions of documents, especially inscriptions and papyri, in TEI XML (Text Encoding Initiative Extensible Markup Language, a standard for representing texts in digital form). Projects like the ItAnt project [15] and CLaSSES [7] have used these standards to create searchable corpora that enable a systematic analysis of ancient texts.

Recent advances have also explored machine learning approaches for contextualizing ancient inscriptions. Assael *et al.* [1] have recently introduced Aeneas, a generative neural network that retrieves textual parallels and performs restoration and attribution tasks for Latin inscriptions, demonstrating how machine learning can assist historians in grounding their research by identifying epigraphic connections that might otherwise remain obscured. Such developments highlight the growing integration of computational methods in epigraphic scholarship.

Despite these advances, current search capabilities for epigraphic databases operate primarily with exact or partial text keyword matching. Researchers must know specific terms in the source language (*e.g.*, Latin or Greek) to find relevant inscriptions, and variations in spelling, abbreviations, and outdated words can prevent the discovery of related texts. This limitation becomes especially problematic when studying thematic concepts that might only be seen through various vocabulary choices or when analyzing vernacular graffiti that often deviates from traditional philology.

2.2 Vector Search and Semantic Similarity

Vector search represents a complete shift in information retrieval (IR), moving beyond keyword matching to find semantic relationships between texts [14]. Currently, vector search is used in search engines to retrieve documents, articles, web pages, or other textual content based on their similarity to the query [6], and serves as the foundation for Retrieval Augmented Generation (RAG) systems that combine retrieved information with language model outputs. Vector search enables users to find relevant information even if the exact terms used in the query are not present in the documents. This capability is especially important for historical texts, where vocabulary variation and linguistic evolution can distract from conceptual connections.

The foundation of vector search lies in embeddings that are created by sentence transformers, a type of neural network architecture designed to create high-dimensional vector representations of sentences or paragraphs [18]. These embeddings capture the semantic meaning of the text, enabling systems to understand relationships between concepts beyond lexical similarity.

For ancient languages, several specialized embedding models have been developed. Bamman and Burns [2] introduced Latin BERT, trained on 642.7 million words of Latin text, achieving state-of-the-art performance in part-of-speech tagging and enabling semantic search applications. Similarly, advances

in ancient Greek processing include AGILe, a lemmatizer that achieves above 80% accuracy [8] and sentence embedding models developed through multilingual knowledge distillation [13].

Recent work by Riemenschneider and Frank [19] has also demonstrated the potential of large language models (LLMs) for classical philology tasks, including text analysis and semantic understanding of Latin and Ancient Greek texts. Their exploration of cross-lingual capabilities in classical languages provides important context for multilingual approaches to ancient text processing, though their focus differs from the retrieval-oriented applications presented in this work.

Despite these developments, few models can effectively handle both Latin and Greek inscriptions within a single system. This limitation is particularly relevant for our epigraphic corpora, which contains texts in both languages. Furthermore, the informal nature of graffiti, with its non-standard spellings and abbreviations, presents additional challenges for traditional natural language processing approaches.

2.3 LLM-Based Re-ranking in Information Retrieval

Large Language Models (LLMs) have emerged as powerful tools for enhancing information retrieval (IR) through re-ranking techniques. Zhu *et al.* [20] survey the application of LLMs in IR, emphasizing their potential for re-ranking while noting challenges such as data scarcity, a particularly relevant concern for ancient language corpora. Research has shown that combining smaller embedding models with LLM re-ranking can outperform larger models in retrieval tasks [17], suggesting that hybrid approaches may be particularly effective for this project.

The application of LLM re-ranking to historical texts has shown promise in medieval Arabic texts [16], demonstrating the potential for these techniques to enhance retrieval precision in historical corpora. For epigraphic applications, LLM re-ranking offers the possibility of applying expert knowledge about thematic relationships and cultural contexts to refine search results beyond what standard computational measures might achieve.

3 Semantic Search Pipeline

Our semantic search system for ancient inscriptions employs a multi-stage approach that combines vector embeddings, hybrid search strategies, and LLM-based re-ranking. Figure 1 provides an overview of this process. The system architecture is designed to address the unique challenges of epigraphic texts while providing researchers with semantically relevant results.

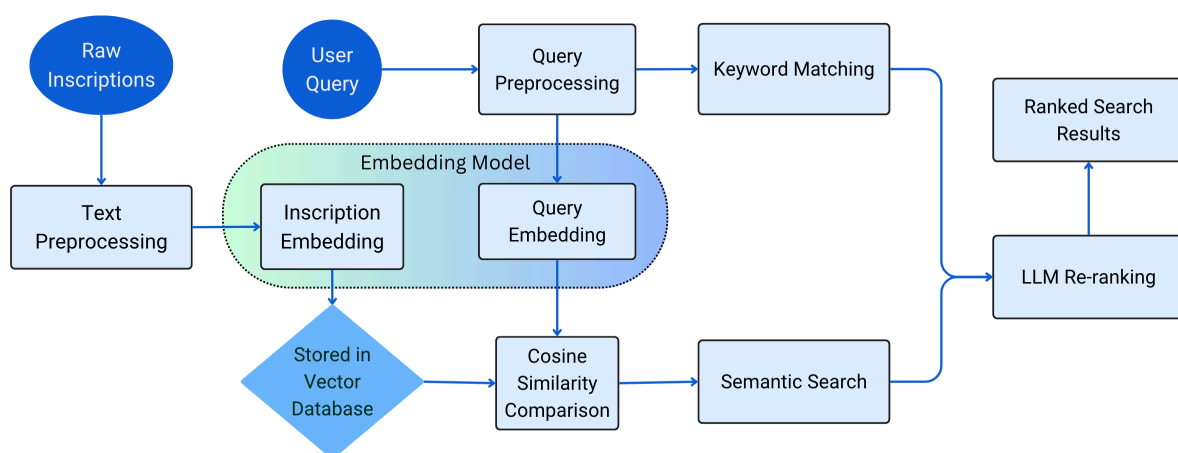


Figure 1: An Overview of our Semantic Search Pipeline.

3.1 Preparing the Vector Database

3.1.1 Text Preprocessing

Given the unique characteristics of ancient inscriptions, we developed a preprocessing pipeline for both Latin and Greek epigraphic texts. Our preprocessing addresses several challenges specific to ancient inscriptions while accounting for the differences between the two languages in our corpus.

The Ancient Graffiti Project (AGP) database provides inscriptions encoded in EpiDoc TEI XML format, containing rich structural and semantic annotations including abbreviation expansions, gap markers, and editorial interventions [10]. For our semantic search application, we retrieved the epigraphic markup from the AGP database and extracted only the textual content. Although some annotations might provide valuable information, we prioritized consistency for the embedding model by focusing on cleaned text.

For Latin inscriptions, the system applies several standardization steps. Text is converted to upper-case following standard epigraphic conventions, with Unicode normalization handling diacritical marks and character variations. Classical Latin character standardization converts ‘V’ to ‘U’ and ‘J’ to ‘I’ to reflect ancient orthographic practices. Common abbreviations are expanded using a custom dictionary of 33 common Latin abbreviations mapped to their expansions. This dictionary covers personal names, official titles, and religious formulae frequently encountered in Roman inscriptions. Compared to Latin texts, Greek inscriptions undergo minimal preprocessing, reflecting both their smaller representation in our dataset and different orthographic characteristics.

To improve embedding quality, the system adds semantic markers that identify thematic categories within inscriptions. These markers tag inscriptions with categories such as ‘death’, ‘family’, ‘religious’, or ‘official’. This tagging could potentially be expanded, but we decided to only include a small selection of categories as a general proof of concept.

3.1.2 Embedding the Inscriptions

The system uses a sentence embedding model to turn both the inscriptions and query into vectors. Inscriptions are then processed in batches of 32 to balance computational efficiency with memory usage, enabling the system to handle large corpora effectively. Then, to reduce bias toward longer texts, the system applies logarithmic length scaling that prevents longer inscriptions from dominating semantic relationships because of their size. All embeddings are normalized to unit length, ensuring that cosine similarity calculations focus on directional relationships rather than magnitude differences.

3.2 Conducting a Query

3.2.1 Query Preprocessing

User queries undergo the same preprocessing pipeline as the inscriptions, ensuring consistency in vector representation. The system implements limited query expansion for common concepts, adding semantically related terms from a manually curated set of associations. For example, queries about death-related concepts are expanded to include terms like “mortuus,” “obiit,” and “defunctus.” This expansion covers six primary thematic areas: death, dedication, family, religious, official, and memorial concepts. The expansion process relies on expert knowledge rather than automated techniques, ensuring precision over comprehensive coverage.

3.2.2 Hybrid Search Strategy

Rather than relying solely on semantic similarity, our system implements a hybrid approach that combines multiple search strategies:

Semantic Search: The processed query is embedded using the same model applied to inscriptions, and cosine similarity is computed between the query embedding and all inscription embeddings. Similarity scores range from 0 to 1, with all embeddings scoring above 0.5 advancing to the next step. This method identifies inscriptions that are semantically related to the query concept.

Keyword Search: A parallel keyword search identifies inscriptions containing exact matches or variations of query terms. This component uses enhanced tokenization that includes not only individual words but also bigrams and trigrams to capture multi-word expressions and proper names.

3.2.3 *Result Combination and Re-ranking*

The hybrid approach combines results from semantic and keyword searches using a weighted scoring system. Inscriptions that match semantic or keyword criteria receive corresponding boost scores, while the system prioritizes hybrid matches that contain the keywords as well as a similar embedding. The inscriptions are then ranked in two steps:

Initial Ranking: Results are initially ranked by a combined score that weights semantic similarity, measured by cosine distance, and keyword matching scores. The system retrieves a larger initial set of candidates to provide enough material for re-ranking. The combination weights are fine-tuned to balance semantic similarity scores with exact keyword match bonuses to optimize retrieval performance for epigraphic queries.

LLM Re-ranking: The LLM is prompted to evaluate each inscription’s relevance to the query concept, considering not just exact matches but also cultural and historical connections. For example, when searching for “love,” the LLM might recognize that an inscription mentioning Venus is thematically relevant due to the goddess’s association with love, while discarding proper names that happen to have similar embeddings. The LLM re-ranking process includes filtering mechanisms that remove completely irrelevant results while increasing the priority of inscriptions with subtle but meaningful connections to the query topic. This approach addresses the challenge of balancing precision and recall in semantic search systems.

4 **Proof of Concept: Searching the Ancient Graffiti Project**

The Ancient Graffiti Project (AGP) exemplifies the digital transformation of epigraphic studies, providing a database of more than 3000 handwritten inscriptions from Pompeii, Herculaneum, Stabiae, and Smyrna [3]. The AGP was selected as the dataset for this research due to its extensive coverage of handwritten inscriptions, which offer a unique glimpse into the daily lives and thoughts of ancient Romans. Additionally, the project is fully downloadable and encoded in EpiDoc, making it an open and accessible resource for researchers. The project integrates epigraphic and archaeological data, providing context about the urban settings in which the inscriptions were found. This contextualization is crucial for understanding the social and spatial dynamics of ancient graffiti [4].

Our system operates on inscription data from the AGP, specifically working with a dataset of 1,993 textual Latin and Greek inscriptions. This dataset represents a substantial portion of the handwritten inscriptions from Herculaneum and Pompeii, providing a diverse corpus. Each inscription includes AGP identification numbers and the original Latin or Greek text.

The dataset’s diversity is crucial for testing semantic search capabilities, as it encompasses various inscription types like dedicatory texts, memorials, poetry, personal messages, and casual graffiti. This variety ensures that the system can handle different linguistic registers and textual conventions present in ancient epigraphic corpora. Using a larger corpora, there would be more approximate inscriptions, but due to the size of our dataset, we chose a low semantic score threshold of 0.5 to maximize the number of conceptually related results while not including loosely related ideas.

For the embedding process, detailed in Section 3.1.2, we chose the paraphrase-multilingual-mpnet-base-v2 model from Sentence Transformers.¹ We selected this model for its speed, strong performance in semantic similarity tasks, and multilingual capabilities [18]. This model choice addresses the challenge of handling both Latin and Greek inscriptions within a single system while maintaining high-quality semantic representations.

¹ <https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2>

Following the embedding process, all embeddings are stored in ChromaDB,² chosen for its lightweight, file-based persistence, and simple Python API that supports efficient cosine similarity queries at scale. We then employ Google’s Gemini 2.0 Flash model,³ chosen for its fast inference on short contexts, simple Python API, and strong performance on semantic judgments, to re-rank the initial results based on thematic relevance and lexical similarity.

We ran our framework on a MacBook Pro with four Intel cores and 16 GB of memory. Our system including the code and dataset used can be accessed at <https://github.com/AncientGraffitiProject/semantic-search>.

5 Preliminary Results

In this section we present our preliminary results, including search case studies, performance tests, and our discussion and analysis of the results.

5.1 Search Case Studies

Quantitative evaluation of semantic search systems for ancient inscriptions presents unique methodological challenges that distinguish this domain from contemporary information retrieval tasks. Unlike many modern collections, ancient epigraphic corpora lack standardized benchmarks for thematic relatedness, making traditional metrics like precision and recall difficult to apply systematically. Additionally, the absence of comprehensive relevance judgments for ancient texts creates fundamental evaluation limitations. Determining whether an inscription mentioning “Venus” is relevant to a query about “love” requires cultural and historical expertise that extends beyond lexical similarity. This interpretive complexity makes automated evaluation metrics insufficient for capturing the system’s true utility to scholars.

Consequently, we use a series of case studies to evaluate our model. The queries were primarily developed by a subject specialist who was not affiliated with this research team. These examples highlight how our hybrid semantic search system surpasses traditional lexical methods, showing how vector embeddings and LLM reranking uncover thematically related inscriptions beyond exact word matches.

Each table lists the top inscriptions retrieved for the corresponding query. For each inscription, the table contains the inscription’s position after LLM reranking, the unique inscription identifier,⁴ the original Latin inscription, whether the match is “semantic,” “keyword,” or “hybrid,” cosine similarity, and a translation.

5.1.1 Search term: “*urbe*” (city)

Rank	agpID	Text	Match Type	Semantic Score	Translation
1	EDR128013	Urbanus	SEMANTIC	0.76	Urban (name)
2	EDR128014	Urbanus	SEMANTIC	0.76	Urban (name)

Table 1: Search results for “*urbe*”. LLM was given two inscriptions for reranking (no LLM reranking applied).

Although *urbe* and *urbanus* do not share identical characters, our embeddings correctly identify them as linguistic variants,⁵ underscoring semantic and lexical linking. This example demonstrates the model’s ability to connect grammatical forms of related concepts.

² <https://www.trychroma.com/>

³ <https://ai.google.dev/>

⁴ The full agpID from the Ancient Graffiti Project database is AGP-EDR followed by the identification number but were shortened for space.

⁵ Latin names were often derived from words.

5.1.2 Search term: “obiit” (died)

Rank	agpID	Text	Match Type	Semantic Score	Translation
1	EDR154190	Mortuus	SEMANTIC	0.8119	Dead
2	EDR157024	Mortuus Deius	SEMANTIC	0.7345	Dead (is) Deius
3	EDR161836	Θ mortuus Caerulus postridie Nonas	SEMANTIC	0.6118	(Theta) Caerulus (was) dead the day after the Nonas
4	EDR161837	mortuus Glorus postridie Nonas	SEMANTIC	0.5879	Glorus (was) dead the day after the Nonas
5	EDR156998	Mortuus Hercolanius	SEMANTIC	0.512	Dead (is) Hercolanius

Table 2: Search results for “obiit”. LLM was given five inscriptions for reranking.

While the term *obiit* (died) never appears verbatim in the database, each returned result is relevant to death, most prominently *mortuus* (dead). This confirms that the embeddings capture the conceptual relationship between “dying” and “being dead” rather than relying on exact string matches. This example demonstrates the system is able to understand some conceptual variations in Latin vocabulary.

5.1.3 Search term: “love”

Rank	agpID	Text	Match Type	Semantic Score	Translation
1	EDR167551	Ἔρω	SEMANTIC	0.7778	Eros
2	EDR167552	Ἔρω	SEMANTIC	0.7778	Eros
3	EDR150782	Venustus	SEMANTIC	0.8112	Charming
4	EDR177788	fellat	SEMANTIC	0.8042	Performs fellatio
5	EDR179282	Μνηστος	SEMANTIC	0.7903	Name meaning “wooed and wedded”

Table 3: Search results for “love”. LLM was given 25 inscriptions for reranking.

The English “love” query surfaces epigraphic gems: *Venustus*, a love-related “charming,” and “Ἔρω”, which keyword search alone would miss, showcase the system’s true discovery potential. However, most stunning is that this query returns the word “Μνηστος”, a Greek name with a strong connection to love that shares no lexical similarity with the English query, yet carries profound thematic relevance. The results of this search show incredible promise for epigraphers wanting to get a broad view of their database, exploring inscriptions related to a topic of their choosing.

5.1.4 Search term: “amat” (love)

Rank	agpID	Text	Match Type	Semantic Score	Translation
1	EDR168646	quisquis amat valeat pereat qui nescit amare bis tanto pereat quisquis amare vetat	KEYWORD	—	Whoever loves, may he be well. May he perish whoever does not know how to love. May he perish twice over whoever forbids love.
2	EDR187553	quisquis amat calidis non debet fontibus uti nam nemo flammis ustus amare potest	KEYWORD	—	Whoever loves ought not to use hot water for no one who has been burned is able to love flames.
3	EDR192096	Quisquis amat veniat Veneri lumbos volo	KEYWORD	—	Whoever loves let him come. I want to break Venus’ back
8	EDR167482	Quisquis amat	KEYWORD	—	Whoever loves
9	EDR149060	Unus amat unam Salutem Salutem	KEYWORD	—	One guy loves one girl greetings greetings
10	EDR124975	quos amat valeant	KEYWORD	—	Those whom he loves may they fare well
11	EDR140143	Amata	SEMANTIC	0.7339	Loved
12	EDR149179	Ἀφροδίτη	SEMANTIC	0.7224	Aphrodite

Table 4: Search results for “amat”. LLM was given 25 inscriptions for reranking. 13 results were reranked, but five were removed for space.

The results from searching “amat” showcase the hybrid approach’s effectiveness by combining extensive keyword matches with deeper semantic matches. The system finds multiple love poems containing the exact term, while also including deeper semantic matches, *Amata* and “Ἀφροδίτη”, which would have been missed by exact matching. This case demonstrates how hybrid search maintains precision while expanding discovery potential through conceptual links.

5.1.5 Search term: “donum” (gift)

Rank	agpID	Text	Match Type	Semantic Score	Translation
1	EDR189997	Vota Restitutus donum dedit	KEYWORD	—	Vows, Restitutus gave a gift

Table 5: Search results for “donum”. LLM was given one inscription for reranking.

When no semantically similar items exist, the model cleanly falls back on the single exact keyword match, preserving precision by avoiding irrelevant retrievals. Therefore, epigraphers who want to search

solely for exact matches will still be satisfied with the hybrid search which keeps relevant keyword matches. This search shows the system will not inflate results with bad semantic connections when genuine matches can't be found.

5.1.6 Search term: “*medicamentum*” (medicine)

Rank	agpID	Text	Match Type	Semantic Score	Translation
1	EDR167955	Augustalis Pierus Celadus Papiri Amandumum medicavit	SEMANTIC	0.6143	The Augustalis, Pier Celadus, cured Amandus the slave of Papirius

Table 6: Search results for “*medicamentum*”. LLM was given one inscription for reranking.

Despite no occurrence of *medicamentum* (medicine), the system finds a parallel medical inscription, demonstrating its ability to link related words and concepts. The moderate similarity score reflects that it is more of a conceptual connection than an exact match, with the high score threshold still excluding completely irrelevant results. This query shows the system’s ability to identify domain-specific semantic relationships.

5.1.7 Analysis of search case studies

These case studies collectively affirm that our semantic search approach and its simultaneous use of keyword matching provides a reliable, concept-driven alternative to exact-match queries. Lacking standardized epigraphic benchmarks, these qualitative examples serve as robust evidence that embedding-based retrieval can transform how scholars explore ancient inscriptions.

5.2 Performance

This system has performance metrics suitable for real-time research applications. System initialization (*i.e.*, initializing the embedding model, initializing the vector database, and linking the LLM API key) requires approximately 3.6 seconds on a MacBook Pro. This one-time overhead ensures optimal performance for iterative research workflows. The core search components exhibit excellent response times, with both semantic and keyword searches completing in under one second. However, the LLM re-ranking time is much more variable, ranging from 1 to 15 seconds, depending on the number of inscriptions needing evaluation.

Encompassing the complete pipeline from initial retrieval through final re-ranking, overall search performance averages 6-7 seconds per query. This response time is short enough for scholars to use efficiently, while maintaining the system’s sophisticated semantic understanding capabilities.

The system’s resource efficiency is another significant advantage for usability and accessibility. The entire search system requires less than 35 MB of storage space, enabling deployment on modest hardware configurations and facilitating integration into existing digital humanities infrastructure. Critically, the system relies exclusively on open-source libraries and free APIs, eliminating licensing costs and reducing barriers to adoption. This design philosophy makes sure the tool is accessible for researchers regardless of institutional resources, and it simplifies integration into existing websites and research platforms. The combination of minimal resource requirements and zero licensing costs makes the system particularly suitable for educational institutions and independent researchers working with limited budgets.

These performance characteristics demonstrate that sophisticated semantic search capabilities can be delivered efficiently, supporting the broader goal of making advanced natural language processing tools accessible to the epigraphic research community.

5.3 Discussion and Analysis

This semantic search approach proposes a paradigm shift in epigraphic methodology, offering researchers an exploratory alternative to traditional exact-match searching. While epigraphers typically prioritize exhaustive precision through systematic keyword searches, our system demonstrates the potential for discovery-oriented research that can find thematically relevant inscriptions that might otherwise remain hidden.

The system’s strength lies in its ability to find conceptual connections that transcend lexical boundaries. Traditional search methods, while precise, may overlook inscriptions that discuss related concepts using different vocabulary, variant spellings, or cultural references. Our hybrid approach preserves the reliability of keyword matching while expanding the scope of discovery, enabling researchers to explore themes and connections more intuitively.

Despite our successful results, we faced several significant challenges specific to ancient language processing during development:

- Experiments with various embedding models (*e.g.*, Latin BERT [2], PhilTa [19], bge-m3,⁶ and XLM RoBERTa⁷) revealed significant limitations: monolingual models failed on cross-lingual queries, while other multilingual models produced inconsistent results.
- Early system iterations disproportionately favored shorter inscriptions in similarity calculations, creating false matches based on text length rather than semantic content. This issue was resolved through logarithmic length normalization and LLM re-ranking, which evaluates thematic relevance independent of inscription length.
- The complexity of ancient Latin and Greek presented ongoing challenges in abbreviation expansion and translation accuracy. Expertise in ancient languages remains necessary for checking results. Translations garnered from the Google Translate API were not consistently accurate, but a Latin professor corrected translations and, most importantly, confirmed that the model returned relevant inscriptions.

Several aspects of the system’s behavior still require further investigation. The differential performance between semantically equivalent queries in different languages, for example “love” versus “amat,” suggests that multilingual embedding models may not capture cross-linguistic semantic relationships uniformly. Additionally, cases where semantically similar terms receive low similarity scores indicate potential limitations in the embedding model’s representation of ancient language concepts. The LLM re-ranking process, while generally effective, occasionally produces rankings that seem inconsistent with thematic relevance. All of these sources of variability reflect the challenges of applying modern language models to ancient texts, where cultural and historical contexts can be much different than contemporary training data.

Overall, the system’s computational efficiency and accessibility represent significant advantages for widespread adoption. Implementation requires minimal technical expertise; researchers only have to provide their inscription data in CSV format, adjust the semantic similarity threshold, and customize the LLM reranking prompt to match their research objectives.

Importantly, the system scales linearly with database size, making it suitable for larger epigraphic databases without significant increases in computational complexity. This scalability, combined with the lightweight architecture, positions the system as a practical tool for integration into existing digital humanities platforms and research workflows.

Our approach demonstrates that sophisticated natural language processing techniques can be made accessible to humanities researchers without requiring extensive technical infrastructure or expertise, potentially democratizing access to advanced computational tools in epigraphic research.

⁶ <https://huggingface.co/BAAI/bge-m3>

⁷ https://huggingface.co/docs/transformers/en/model_doc/xlm-roberta

6 Conclusions and Future Work

This paper presents a novel approach to searching ancient inscriptions that moves beyond traditional keyword matching to capture semantic relationships between texts. By combining vector embeddings, hybrid search strategies, and LLM re-ranking, our system demonstrates the potential for advanced natural language processing techniques to enhance epigraphic research. The semantic search system addresses fundamental limitations in current epigraphic query systems by identifying thematically related inscriptions that share conceptual connections rather than lexical overlap alone. This capability opens new possibilities for researchers to discover patterns, influences, and cultural connections within ancient corpora that might otherwise remain hidden.

While the current implementation only focuses on Latin and Greek inscriptions from Pompeii and Herculaneum, the underlying approach is broadly applicable to other ancient languages and epigraphic traditions. Additionally, several areas offer opportunities for improvement and expansion:

- **Multilingual Enhancement:** While the current system handles Latin and Greek inscriptions effectively, expanding support for other languages would benefit researchers working with diverse ancient corpora.
- **Quantitative Benchmarking:** Developing domain-specific evaluation metrics and annotated datasets for ancient inscriptions would enable more rigorous comparative analysis and performance measurement across different retrieval strategies.
- **Integration with Archaeological Data:** Incorporating spatial and archaeological metadata into the search and ranking process could enable location-aware searches that consider the physical and cultural contexts of inscriptions.
- **Real-time Collaboration:** Implementing collaborative search refinement features could enable researchers to collectively improve the system’s understanding of ancient texts and their relationships.

As digital humanities continues to evolve, the integration of modern AI technologies with traditional epigraphic scholarship will be a paramount step toward more sophisticated tools for historical research. By making thematic connections more discoverable and reducing the barriers to finding relevant inscriptions, semantic search systems have the potential to accelerate research and enable new forms of analysis that were previously impractical with manual methods.

Future developments in this area could transform how scholars approach ancient texts, moving from laborious keyword searches to intuitive, concept-based exploration of historical corpora. As these technologies mature, they promise to make the voices of the ancient world more accessible to researchers and, ultimately, to the broader public interested in understanding our shared human heritage.

Acknowledgements

Support for this project was provided by the Summer Research Scholars Program and the Lenfest Summer Fellowship Award at Washington and Lee University. This work also benefited from the thorough comments and suggestions by the anonymous reviewers.

References

- [1] Assael, Yannis et al. “Contextualizing ancient texts with generative neural networks”. In: *Nature* 645 (Sept. 2025), pp. 141–147. DOI: 10.1038/s41586-025-09292-5.
- [2] Bamman, David and Burns, Patrick J. “Latin BERT: A Contextual Language Model for Classical Philology”. arXiv:2009.10053. Sept. 2020. DOI: 10.48550/arXiv.2009.10053.
- [3] Benefiel, Rebecca, Sprenkle, Sara, Sypniewski, Holly M., and White, Jamie. “The Ancient Graffiti Project: Geo-Spatial Visualization and Search Tools for Ancient Handwritten Inscriptions”. en. In: *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*. Göttingen Germany: ACM, June 2017, pp. 163–168. ISBN: 978-1-4503-5265-9. DOI: 10.1145/3078081.3078104.

- [4] Benefiel, Rebecca R. and Sypniewski, Holly M. “Documenting Ancient Graffiti: Text, Image, Support and Access”. In: *Graffiti Scratched, Scrawled, Sprayed*, ed. by Ondřej Škrabal, Leah Mascia, Ann Lauren Osthof, and Malena Ratzke. De Gruyter, Nov. 2023, pp. 425–466. ISBN: 978-3-11-132630-6. DOI: 10.1515/9783111326306-014.
- [5] Cayless, Hugh, Roueche, Charlotte, Elliott, Tom, and Bodard, Gabriel. “Epigraphy in 2017”. In: *Digital Humanities Quarterly* 3, no. 1 (2009). ISSN: 1938-4122. URL: <https://dhq.digitalhumanities.org/vol/3/1/000030/000030.html>.
- [6] Clarke, Amelia, Hughes, Nathaniel, Banerjee, Priya, and Ogunrinde, Victor. “Semantic Search and Its Role in Knowledge Discovery”. May 2025. URL: https://www.researchgate.net/publication/392014461_Semantic_Search_and_Its_Role_in_Knowledge_Discovery.
- [7] De Felice, Irene, Marotta, Giovanna, and Donati, Margherita. “CLaSSES: a New Digital Resource for Latin Epigraphy”. In: *IJCoL. Italian Journal of Computational Linguistics* 1, no. 1 (Dec. 2015), pp. 125–136. ISSN: 2499-4553. DOI: 10.4000/ijcol.331.
- [8] de Graaf, Evelien, Stopponi, Silvia, Bos, Jasper K., Peels-Matthey, Saskia, and Nissim, Malvina. “AGILE: The First Lemmatizer for Ancient Greek Inscriptions”. In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, June 2022, pp. 5334–5344. URL: <https://aclanthology.org/2022.lrec-1.571>.
- [9] Elliott, Tom. “Epigraphy and Digital Resources”. In: *The Oxford Handbook of Roman Epigraphy*. Oxford: Oxford University Press, 2014, pp. 78–85.
- [10] Elliott, Tom, Bodard, Gabriel, Cayless, Hugh, and al., et. “EpiDoc: Epigraphic Documents in TEI XML”. 2006. URL: <https://epidoc.stoa.org/>.
- [11] Heřmánková, Petra, Ballsun-Stanton, Brian, and Laurence, Ray. “FAIR Turn in Epigraphy: Low Barrier Pathways to Quantitative and Reproducible Research in Latin Epigraphy”. In: *Computational Humanities Research*. 2024, pp. 649–661.
- [12] Heřmánková, Petra, Kaše, Vojtěch, and Sobotková, Adéla. “Inscriptions as Data: Digital Epigraphy in Macro-Historical Perspective”. In: *Journal of Digital History* 1, no. 1 (2021), pp. 99–141. DOI: 10.1515/jdh-2021-1004.
- [13] Krahm, Kevin, Tate, Derrick, and Lamicela, Andrew C. “Sentence Embedding Models for Ancient Greek Using Multilingual Knowledge Distillation”. In: *Proceedings of the Ancient Language Processing Workshop*, ed. by Adam Anderson, Shai Gordin, Bin Li, Yudong Liu, and Marco C. Passarotti. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria, Sept. 2023, pp. 13–22. URL: <https://aclanthology.org/2023.alp-1.2/>.
- [14] Monir, Solmaz Seyed, Lau, Irene, Yang, Shubing, and Zhao, Dongfang. “VectorSearch: Enhancing Document Retrieval with Semantic Embeddings and Optimized Search”. In: , no. arXiv:2409.17383 (Sept. 2024). arXiv:2409.17383 [cs]. DOI: 10.48550/arXiv.2409.17383.
- [15] Murano, Francesca, Quochi, Valeria, Del Grosso, Angelo Mario, Rigobianco, Luca, and Zinzi, Mariarosaria. “Describing Inscriptions of Ancient Italy. The ItAnt Project and Its Information Encoding Process”. In: *Journal on Computing and Cultural Heritage* 16, no. 3 (Aug. 2023), 53:1–53:14. ISSN: 1556-4673. DOI: 10.1145/3606703.
- [16] Muther, Ryan, Barber, Mathew, and Smith, David. “Querying the Past: Automatic Source Attribution with Language Models”. In: *Computational Humanities Research Conference*. 2023, pp. 344–355.
- [17] Rao, Arjun, Alipour, Hanieh, and Pendar, Nick. “Rethinking Hybrid Retrieval: When Small Embeddings and LLM Re-ranking Beat Bigger Models”. arXiv:2506.00049 [cs]. May 2025. DOI: 10.48550/arXiv.2506.00049.

- [18] Reimers, Nils and Gurevych, Iryna. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 3980–3990. DOI: 10.18653/v1/D19-1410.
- [19] Riemenschneider, Frederick and Frank, Anette. “Exploring Large Language Models for Classical Philology”. arXiv:2305.13698 [cs]. May 2023. DOI: 10.48550/arXiv.2305.13698. (Visited on 07/02/2025).
- [20] Zhu, Yutao, Yuan, Huaying, Wang, Shuting, Liu, Jiongnan, Liu, Wenhan, Deng, Chenlong, Chen, Haonan, Liu, Zheng, Dou, Zhicheng, and Wen, Ji-Rong. “Large Language Models for Information Retrieval: A Survey”. arXiv:2308.07107 [cs] version: 4. Sept. 2024. DOI: 10.48550/arXiv.2308.07107.