Gretino: a Greek and Latin Dataset to Benchmark Retrieval Systems in Classical Languages

Hawau Olamide Toyin[1], **Federico Iezzi**[2], **Elia Scapini**[2]
Giulio Federico[3], **Giovanni Puccetti**[3]
[1]MBZUAI - Abu Dhabi, UAE
[2]Dipartimento di Educazione e Scienze Umane - Universita di Modena e Reggio Emilia, Italy
[3]Institute of Science and Technologies of Information "A. Faedo" - CNR Pisa, Italy
[1]hawau.toyin@mbzuai.ac.ae, [2]name.surname@unimore.it, [3]name.surname@isti.cnr.it

Abstract

Semantic similarity search is a method for exploring large text corpora and retrieving conceptually related content. Although widely used in modern language applications, it remains underexplored in the context of classical literature, where it could provide scholars with tools to uncover meaningful connections across authors, genres, and languages, surpassing the limitations of rule-based or keyword search systems. To promote the adoption of semantic retrieval in classical languages, we introduce Gretino, the first benchmark dataset for evaluating semantic search systems in Latin, Ancient Greek, and cross-lingual settings. Gretino comprises 240 carefully designed queries, each paired with five semantically relevant passages in Latin and Greek. The dataset is divided into two subsets: Gretino Silver, consisting of 200 queries and 1,000 targets (evenly split between Latin and Greek), generated with the assistance of ChatGPT and subsequently reviewed; and Gretino Gold, a manually curated high-quality subset of 40 queries and 200 targets, fully based on authentic classical texts. We evaluate four pre-trained language models: GreBERTa, LaBERTa, PhilBERTA, and SPhilBERTa and demonstrate the potential of a contrastive learning approach based on SimCSE (Gao et al., 2021) for fine-tuning, showing that training on carefully curated bilingual corpora, with texts aligned in the two languages, can improve retrieval performance.

1. Introduction

The literature on transformer-based language models trained on ancient languages is growing over time and the development of AI-based tools for the study of Ancient Greek and Latin has accelerated dramatically with the release of several transformer-based language models (Riemenschneider and Frank, 2023a,b; Krahn et al., 2023). However, with the rise of these tools, there is a growing need for datasets and benchmarks to compare models in a fair way. This will allow scholars to make a more well-founded choice in preferring one over the other. To address this need, we propose a novel retrieval dataset and benchmark partially machine generated and partially collected from original sources to test the effectiveness of existing encoder models in retrieving semantically similar sentences.

More specifically, the dataset we propose, `Gretino`[1], can be used to evaluate how well models retrieve concepts similar to a query across a corpus allowing to compare them in vertical domains such as Latin and Ancient Greek and also in cross-lingual reference retrieval. In this work we describe how the dataset has been developed, we measure the performance of State of the Art language models trained on classical languages and finally we fine-tune the model we find to be the best among the existing one to improve its precision when searching for text passages in Ancient Greek, Latin and most notably in a cross-lingual scenario when one has a query in Latin and wants to search texts in Greek and vice-versa. The construction of Gretino is part of a broader effort in making sentence semantic retrieval available to scholars in historical languages that brought to the creation of the DaMSym retrieval suite[2].

2. Related Works

Since the transformer technology based on the self-attention mechanism (Vaswani et al., 2017) was integrated into the first BERT model (Devlin et al., 2019), new research possibilities have opened up in the field of natural language processing (NLP).

2.1. Ancient Greek models

We focus on Ancient Greek, broadly defined to include its Archaic, Classical, Hellenistic Koine, and Medieval Byzantine stages, an Indo-European language characterized by high morphological complexity and limited representation in contemporary digital resources. Early model attempts, such as Ancient Greek BERT (Singh et al., 2021) and AG-BERT (Yamshchikov et al., 2022), focused on morphosyntactic tasks and authorship attribution, relying on models for Modern Greek (Koutsikakis et al., 2020). In parallel, another RoBERTa model, Ancient Greek RoBERTa (Spanopoulos, 2022) came

[1]github.com/gpucce/index_encoders_eval_itserr

[2]https://damsym-itserr.d4science.org/

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 919–928
11–16 May 2026.

out, while with GRC-ALIGNMENT (Yousef et al., 2022), multilingual and text alignment was explored, exploiting XLM-RoBERTa. Models such as Ithaca (Assael et al., 2022) and Logion BERT (Cowen-Breen et al., 2023) have addressed specific application challenges: inscription restoration and scribal error correction. A significant breakthrough came with GreBERTa and GreTa (Riemenschneider and Frank, 2023a), the first models trained from scratch on Ancient Greek, with expanded corpus and preserved diacritics, surpassing predecessors in morphosyntactic tasks.

2.2. Latin models

The inaugural Latin BERT model was introduced in 2020 (Bamman and Burns, 2020). After Latin BERT, other solutions were tried to develop pretrained language models (PLMs) for Latin, sometimes, by the authors' admission, without good results (Mercelis and Keersmaekers, 2022; Mercelis, 2024). To track the exploitation of transformer technology for Latin NLP and, in doing so, to detect the development of these models until today, a good starting point is considering models and tools presented at EvaLatin (Sprugnoli et al., 2024), an evaluation campaign for NLP tools dedicated to Latin that occurs every two years (the last and third happened in 2024). To face changes in Latin over centuries and to create a general parser for Latin that works effectively regardless of the Latin linguistic period, at EvaLatin 2024 Rufus Behr introduced sentence embeddings (Behr, 2024). In this technique, the model encodes sentences of the same period into vectors with higher cosine similarity compared to sentences from another period. SBERT was produced by fine-tuning Latin BERT (Bamman and Burns, 2020) following Reimers and Gurevych (2019) (Reimers and Gurevych, 2019). According to the author, the inclusion of SBERT model in the pipeline did not significantly change the outcomes compared to previous state-of-the-art tools. In the EvaLatin 2024 campaign, the model gained second place for the Dependency Parsing task.

The aforementioned work of Riemenschneider and Frank (Riemenschneider and Frank, 2023a) led to the development of two monolingual Latin models: an encoder-only model called LaBERTa (based on RoBERTa), and LaTa, a T5-based encoder-decoder model[3].

[3]As has become customary, we refer here to their paper *Exploring Large Language Models for Classical Philology*; however, for the sake of precision, it should be noted that this article focuses exclusively on Ancient Greek and multilingual models (Greek, Latin, and English). For a brief description of the Latin models, one should instead consult the corresponding GitHub repository.

2.3. Multilingual models

To date, research institutes and universities have developed a relatively small number of multilingual models specifically designed for Latin and Ancient Greek. These initiatives address the need to enhance NLP pipelines for morphologically complex and low-resource languages, which are still underrepresented in digital corpora. A crucial turning point in this area occurred in 2023 with the introduction of PhilBERTA, an encoder-only model, and PhilTa, based on the T5 encoder-decoder architecture (Riemenschneider and Frank, 2023a). These were followed by SPhilBERTa (Riemenschneider and Frank, 2023b), a multilingual model derived via knowledge distillation (Reimers and Gurevych, 2020), using PhilBERTA as the student model. SPhilBERTa demonstrated particularly strong performance in identifying translations and semantic similarities between sentences in Latin, Ancient Greek, and English, showing remarkable efficacy in detecting intertextual allusions.

In the same year, knowledge distillation techniques were also applied to two fine-tuned models for Ancient Greek that are based on existing Multilingual models. These experiments led to the creation of the first models for classical languages evaluated on tasks such as Semantic Textual Similarity and Semantic Retrieval (Krahn et al., 2023). However, unlike PhilBERTA and SPhilBERTa, which jointly support Latin, Ancient Greek, and English, the models derived by Krahn et al. are limited to Ancient Greek and English.

A significant advancement came with the introduction of HLM-DeBERTa-V3 (Riemenschneider and Krahn, 2024). This model integrates advanced techniques of hierarchical tokenization and hybrid pretraining, combining Masked Language Modeling (MLM) and Replaced Token Detection (RTD), thereby substantially improving the handling of positional context and diacritic variation. This approach represents a major step forward both in adapting general-purpose architectures and in developing targeted solutions to overcome the limitations imposed by data scarcity and the morphological complexity of Ancient Greek. The tokenization strategy employed in HLM-DeBERTa-V3 is grounded in the Hierarchical Pretrained Language Models (HLM) framework (Sun et al., 2023), which adopts a dual-level representation, both intra-word and inter-word, particularly suited to capturing the morphological nuances of highly inflected languages.

Building on this architecture, Kevin Krahn subsequently developed shlm-grc-en (Riemenschneider and Krahn, 2024; Krahn et al., 2023), a multilingual model for Ancient Greek and English. This model is based on a modified version of the HLM architecture and was trained to produce sentence embeddings through multilingual knowledge distil-

lation. This model was used to embed English and Greek corpora for a semantic retrieval tool prototype currently operating[4].

In parallel with these academic efforts, a second strand of general-purpose multilingual models has emerged, primarily developed by major technology companies. Although not explicitly designed for classical languages, these models include Latin among the supported languages and provide useful benchmarks for evaluating NLP techniques in historical language settings. The pioneering model in this area was Multilingual BERT (mBERT) (Devlin et al., 2019), trained on over 100 languages, including Latin. It was followed by XLM-RoBERTa (Conneau et al., 2020), which was employed in the third edition of the EvaLatin campaign (2024) (Straka et al., 2024), and had previously been used in zero-shot emotion polarity detection tasks for Latin texts (Sprugnoli et al., 2023).

3. The Gretino Dataset

As previously mentioned, the only known study that has attempted to evaluate language models on a semantic NLP task for Ancient Greek is that conducted by Krahn et al. in 2023 (Krahn et al., 2023). In their work, the authors assessed model performance in semantic retrieval by compiling a dataset of 40,000 Ancient Greek passages, drawn from the Perseus and First1KGreek corpora. Approximately 50 queries in English, formulated as sentences or questions, were then associated with the relevant passages. The models' performance was evaluated using recall and mean average precision.

The present study introduces Gretino, a dataset specifically designed for Ancient Greek and Latin. This approach is in fact based on two distinct components: a synthetic dataset (Gretino Silver), consisting of data generated with the support of generative artificial intelligence and validated by domain experts, ensuring both control and replicability; and a second, more limited Gretino Gold dataset, composed of passages from Ancient Greek and Latin literature that serve as queries and their associated targets. Gretino represents the first dataset developed to evaluate cross-lingual retrieval performance between these two major languages of the ancient mediterranean world.

3.1. Gretino Silver

To construct Gretino Silver, a three-phase methodology was adopted, combining expert philological guidance with the support of generative artificial intelligence. Owing to its synthetic nature, we refer to it as the Silver dataset.

[4]Krahn makes this tool available at Semantic Search for Ancients Texts.

In the first phase, 100 query sentences were created in Italian (the native language of the authors) and in English. These sentences were deliberately designed by two specialists to incorporate concepts absent from the classical textual tradition, while carefully avoiding anachronisms (e.g., references to airplanes or trains). This methodological choice ensured full control over the expected retrieval outcomes during evaluation and facilitated the automatic computation of performance metrics. By "concepts absent from the tradition," we refer to propositions that an ancient or medieval author would not plausibly have expressed—for example, claims that women are more intelligent than men, that Jesus had been a butterfly in a previous life, or that elderly inhabitants of Antioch subsist on bark and worms. Such content was intentionally introduced to prevent the retrieval task from inadvertently matching pre-existing passages in extant corpora.

In the second phase, each of the 100 query sentences was processed individually, and the AI was prompted—through semi-structured and largely standardized instructions—to generate five target variations for each query. These variations introduced modifications in wording, syntax, length, and sentence structure, as well as controlled "distractor elements," while preserving the semantic core of the original sentence. In some cases, the variations involve only minor lexical or syntactic changes; in others, the correspondence with the query is limited to broader semantic equivalence.

In the third phase, all sentences—both queries and targets—were translated into Ancient Greek and Latin using AI systems (ChatGPT 3.5, 4, and 4o). As in the previous stages, expert supervision played a crucial role in reviewing and validating the outputs, thereby ensuring linguistic coherence and an acceptable level of translational quality.

The use of a synthetic silver standard offers several methodological advantages. Firstly, it enables strict control over the dataset's content: since the generated sentences are external to existing corpora, they can be interpolated into any textual collection, allowing for a-priori definition of the expected retrieval results. Secondly, this approach ensures efficiency by eliminating the need for extensive manual annotation, which would be inevitable if queries were based on real-world content. For instance, passages attesting that Aristotle was a student of Plato, that Prometheus gave fire to humans, or that the cosmos is composed of the four elements. Evaluating such queries would require the involvement of highly specialized experts for each individual case, with the near-impossible task of mastering the entire corpus content. Additionally, a substantial collective effort by expert annotators would be necessary to assess the accuracy of each

Gretino Silver: Examples of queries and targets

Queries:

Greek: λέγεται ὅτι αἱ γυναῖκες σοφώτεραι εἶναι τῶν ἀνδρῶν.
Latin: Dicitur mulieres sapientiores esse viris.
English: It is said that women are wiser than men.

Targets:

Greek: Περὶ τῶν γυναικῶν καὶ τῶν ἀνδρῶν ἐπυνθάνοντο οἱ φιλόσοφοι, οἳ ἐν ἰδιώμασι καὶ λόγοις διακρινομένων ἀναλύσεων ἀναθεωροῦντες, φασὶν ὅτι αἱ γυναῖκες, αἳ τῇ φρονήσει καὶ διανοίᾳ ἀνωτέραι εἰσὶν τῶν ἀνδρῶν.
Latin: De mulieribus et viris quaerebant philosophi, qui in proprietatibus et rationibus analysium distinctarum considerantes, dicunt mulieres prudentia et intellectu viris superiores esse.
English: The philosophers inquired about women and men and, upon examining distinguishing features and principles of differentiated analysis, asserted that women surpass men in wisdom and intelligence.

Figure 1: Example of synthetic queries and targets from the Gretino Silver Dataset

model's retrieval output. Thirdly, the silver standard provides replicability, offering a standardized benchmark for future model evaluations. In addition to the dataset itself, the approach used to create it, allow for a relatively fast development of new silver datasets built through the use of generative AI.

Finally, since each query and its corresponding target set are aligned in both Ancient Greek and Latin, the dataset not only supports monolingual evaluation in either language but also enables cross-lingual semantic retrieval experiments (e.g., Greek-to-Latin and Latin-to-Greek), thus opening avenues for cross-lingual benchmarking in classical languages.

3.2. Gretino Gold

The dataset, which we refer to as Gold due to its corpus-based nature, while not offering the same flexibility as the Gretino Silver dataset because of the greater effort required for manual curation, nonetheless constitutes a fundamental resource for evaluating the semantic retrieval capabilities of language models on authentic Ancient Greek and Latin texts. The construction of the Gold dataset

Gretino Gold: Examples of queries and targets

Queries:

Greek: Εὐρώπην τὴν Φοίνικος Ζεὺς θεασάμενος ἔν τινι λειμῶνι μετὰ Νυμφῶν ἄνθη ἀναλέγουσαν ἠράσθη, καὶ κατελθὼν ἤλλαξεν ἑαυτὸν εἰς ταῦρον καὶ ἀπὸ τοῦ στόματος κρόκον ἔπνει. (*Schol. Iliad.* M 307)
English: Zeus, having seen Europa, daughter of Phoenix, picking flowers in a meadow with the nymphs, fell in love with her. He descended [from the sky] and, transformed into a bull, emitted a scent of saffron from his snout

Latin: Secundum signum est taurus, ob id quod Jupiter in raptu Europae in taurum est versus, et inter sidera translatus. (Honorius Augustodunensis, *De imagine mundi*, I, 93)
English: The second sign is the bull, because Jupiter, during the abduction of Europa, turned into a bull and was transferred among the stars.

Targets:

Greek: ἐντεῦθεν ὡμολόγησαν οἱ ποιηταὶ ὡς ὁ Ζεὺς ταύρῳ ὁμοιωθεὶς Εὐρώπην ἥρπασεν (Georgius Cedrenus, *Compendium historiarum*).
English: From this, the poets agreed that Zeus, having taken the form of a bull, abducted Europa.

Latin: Praeterea scio hunc esse, in quem potissimum Iuppiter se convertit, cum exportavit per mare e Phoenice amans Europam. (Varro, *De agricultura*, II, 5)
English: Moreover, I know that this is the one into whom Jupiter most especially transformed himself when he carried off his beloved Europa across the sea from Phoenicia.

Figure 2: Example of queries and targets from the Gretino Gold Dataset.

was carried out through a manual and philologically guided process, selecting 20 literary *topoi* shared across the Greek and Latin traditions: 5 Christian (e.g., the *topos* of the persecution of Peter and Paul in Rome) and 15 non-Christian, drawn from mythology (such as Ariadne assisting Theseus in escaping the labyrinth of the Minotaur) or philosophy and cosmology (for example, the world composed of the four elements: water, air, earth, and fire).

For each *topos*, six sentences were selected in each language, one serving as the query and the remaining five as the target passages. Unlike the Silver dataset, the Greek and Latin sentences do not constitute literal translations of one another; nevertheless, owing to the cross-cultural and cross-linguistic presence of these *topoi* in ancient literature, this approach allows for cross-linguistic retrieval evaluations.

It is important to emphasize that these are long-standing *topoi*. For each *topos*, queries and targets were drawn from authors spanning different historical periods, thereby covering a broad diacronic range. Their continuous reworking within the Archaic, Classical, Late Antique, and Medieval traditions (both Greek and Latin) gives the dataset a distinctive value, allowing for the evaluation of models' ability to capture semantic correspondences that transcend stylistic variations, historical periods, and linguistic registers.

4. Experimental Setup

To understand if the Gretino dataset can be useful as a benchmark for retrieval systems in Latin and Ancient Greek, we test 4 pretrained models trained on either Ancient Greek: GreBERTa, Latin: LaBERTa or both: PhilBERTA and SPhilBERTa.

We test models in four language settings: (A) Greek: queries and documents are in Ancient Greek, (B) Latin: queries and documents are in Latin, (C) Greek to Latin: queries are in Greek and documents in Latin and (D) Latin to Greek: queries are in Latin and documents in Greek. It should be noted that cross-lingual settings (C) and (D) are possible because in Gretino Silver queries and targets are the translation from one language to the other while for Gretino Gold the queries and targets, even if not literal translations, are highly semantically related according to a domain expert evaluation.

We also devise two retrieval dataset settings where the lookup source contains: (α) `target`: here the retrieval source contains the target for the specific query and the irrelevant documents which are the targets for the other remaining queries; (β) `target+random`: here the irrelevant documents are a set of 1,000 random documents sampled from the Thesaurus Linguae Graecae (TLG) corpus -CD version- for Greek and the Perseus Digital Library (PDL) for Latin. In the case of the Silver dataset, it can be asserted with confidence that the random sentences are not semantically related to the Gretino Silver queries, for the reasons discussed in Section 3.1, as they contain concepts absent from the classical textual tradition. In contrast, for the Gretino Gold dataset, manual verification by experts ensured that the 20 Gold queries were not semantically correlated with either the 1,000 random Latin sentences or the 1,000 random Greek sentences.

The two dataset settings test two separate properties we wish Gretino to have: the `target` setting measures if the models are able to tell relevant documents from irrelevant ones when they are all created in the same way, this is particularly relevant for Gretino Silver dataset which is synthetically created and therefore could be out of distribution for models trained on ancient languages. The second, `target+random` measures if Gretino is useful as a benchmark to test models meant for research use cases for Humanities and Digital Humanities scholarship. By adding irrelevant documents from actual research corpora we want to test if models can identify the semantic difference between the selected targets and other documents in a larger search index.

To measure models performance, we report Mean Average Precision (MAP), Reciprocal Rank (recip-rank), precision at k (P@k) and normalized discounted cumulative gain (NDCG@k) for k = 5, 10. *MAP* measures the overall quality of the ranked results by averaging precision scores across all relevant items; *recip-rank* captures how highly the first correct prediction appears in the ranking; *P@k* computes the proportion of relevant items among the top k retrieved results, and *NDCG@k* assesses ranking quality by assigning higher weights to correctly ranked items that appear earlier in the list.

Greek Results: The left side of Table 1 shows the results achieved by the models on the Greek dataset. As expected Latin only models perform worst, i.e. LaBERTa shows the lowest value on all scores. PhilBERTA and SPhilBERTa perform generally worse than GreBERTa although comparably well. This is consistent with PhilBERTA being trained in both Greek and Latin and SPhilBERTa being trained specifically to align embeddings in Greek and Latin.

Interestingly, the performance on Gretino Silver is higher for all models that understand Greek, GreBERTa (P@5 goes from 0.89 to 0.81), PhilBERTA (P@5 goes from 0.96 to 0.71) and SPhilBERTa (P@5 goes from 0.91 to 0.79) showing that synthetic sentences are easy to retrieve also for models trained on ancient languages only. The similar performance drop observed in the target+random setting suggests that some texts in Ancient Greek may be ranked higher than their corresponding synthetic targets, despite the fact that, as previously noted, the queries and targets in Silver Gretino are conceptually distant from the cultural context of ancient and medieval Greek.

All these findings validate the relevance of Gretino as a benchmark for retrieval systems in

split	setting	model	MAP	recip-rank	P@5	P@10	NDCG@5	NDCG@10	MAP	recip-rank	P@5	P@10	NDCG@5	NDCG@10
			Greek						Latin					
silver	target	GreBERTa	0.92	0.99	0.89	0.48	0.91	0.95	0.69	0.97	0.63	0.36	0.71	0.76
		LaBERTa	0.67	0.93	0.62	0.35	0.69	0.74	0.89	0.97	0.86	0.46	0.89	0.92
		PhilBERTA	0.91	0.99	0.87	0.47	0.90	0.93	0.93	0.99	0.89	0.48	0.92	0.96
		SPhilBERTA	0.95	0.99	0.91	0.48	0.93	0.96	0.95	0.99	0.92	0.49	0.94	0.97
	target+random	GreBERTa	0.97	1.00	0.94	0.49	0.96	0.98	0.82	0.97	0.76	0.43	0.82	0.87
		LaBERTa	0.78	0.97	0.71	0.41	0.78	0.83	0.97	1.00	0.93	0.49	0.95	0.98
		PhilBERTA	0.98	1.00	0.96	0.49	0.97	0.99	0.99	1.00	0.98	0.50	0.99	0.99
		SPhilBERTA	0.95	1.00	0.92	0.48	0.94	0.97	0.96	0.99	0.93	0.48	0.95	0.97
gold	target	GreBERTa	0.85	0.95	0.81	0.47	0.83	0.90	0.38	0.66	0.33	0.23	0.38	0.45
		LaBERTa	0.52	0.77	0.47	0.32	0.51	0.61	0.64	0.90	0.58	0.34	0.64	0.70
		PhilBERTA	0.77	0.97	0.71	0.40	0.77	0.82	0.59	0.81	0.55	0.34	0.58	0.65
		SPhilBERTA	0.83	0.94	0.79	0.44	0.82	0.87	0.78	0.95	0.69	0.41	0.75	0.83
	target+random	GreBERTa	0.92	1.00	0.87	0.47	0.91	0.94	0.33	0.70	0.32	0.18	0.38	0.41
		LaBERTa	0.43	0.74	0.35	0.27	0.42	0.52	0.77	1.00	0.71	0.40	0.79	0.83
		PhilBERTA	0.82	0.97	0.76	0.45	0.81	0.88	0.75	0.97	0.72	0.38	0.78	0.80
		SPhilBERTA	0.83	0.97	0.76	0.45	0.81	0.89	0.72	0.93	0.63	0.38	0.71	0.77

Table 1: Performance scores on **Greek** and **Latin** retrieval task. The red and blue gradient shades indicate higher and lower performance respectively for each metric.

split	setting	model	MAP	recip-rank	P@5	P@10	NDCG@5	NDCG@10	MAP	recip-rank	P@5	P@10	NDCG@5	NDCG@10
			Latin+Greek						Greek+Latin					
silver	target	GreBERTa	0.08	0.12	0.07	0.06	0.06	0.09	0.05	0.08	0.03	0.03	0.03	0.04
		LaBERTa	0.08	0.13	0.07	0.06	0.06	0.09	0.25	0.42	0.23	0.15	0.24	0.28
		PhilBERTA	0.51	0.48	0.58	0.38	0.50	0.60	0.87	0.97	0.80	0.45	0.85	0.90
		SPhilBERTA	0.93	0.99	0.89	0.48	0.91	0.95	0.93	1.00	0.89	0.48	0.92	0.96
	target+random	GreBERTa	0.06	0.10	0.04	0.03	0.04	0.05	0.01	0.01	0.00	0.00	0.00	0.00
		LaBERTa	0.49	0.74	0.44	0.28	0.50	0.56	0.08	0.10	0.05	0.05	0.04	0.07
		PhilBERTA	0.96	0.99	0.94	0.48	0.95	0.97	0.40	0.42	0.33	0.33	0.30	0.48
		SPhilBERTA	0.90	0.99	0.87	0.47	0.90	0.93	0.93	0.99	0.89	0.48	0.92	0.96
gold	target	GreBERTa	0.24	0.44	0.18	0.13	0.21	0.25	0.15	0.21	0.09	0.12	0.09	0.17
		LaBERTa	0.15	0.21	0.11	0.10	0.09	0.14	0.27	0.38	0.23	0.15	0.24	0.28
		PhilBERTA	0.44	0.66	0.36	0.29	0.40	0.53	0.56	0.85	0.47	0.31	0.54	0.62
		SPhilBERTA	0.75	1.00	0.63	0.41	0.72	0.82	0.73	0.90	0.64	0.41	0.69	0.78
	target+random	GreBERTa	0.14	0.31	0.11	0.10	0.13	0.18	0.01	0.01	0.00	0.00	0.00	0.00
		LaBERTa	0.15	0.33	0.13	0.10	0.15	0.18	0.02	0.02	0.00	0.01	0.00	0.00
		PhilBERTA	0.61	0.85	0.54	0.34	0.60	0.68	0.15	0.20	0.10	0.12	0.10	0.17
		SPhilBERTA	0.71	0.94	0.65	0.38	0.72	0.77	0.58	0.81	0.50	0.33	0.56	0.65

Table 2: Performance scores on **cross-lingual** retrieval task. The red and blue gradient shades indicate higher and lower performance respectively for each metric.

Ancient Greek to test both model performance and their usability in real retrieval scenarios.

Latin Results: The right half of Table 1 shows the results when models are tested on the Latin dataset. The results mimic those for Greek. Specifically, Greek only models underperform on Latin, GreBERTa is the worst candidate. The performance gap with LaBERTa is comparable to the results for Greek. PhilBERTA and SPhilBERTa work mildly worse than LaBERTa but overall achieve comparable results.

As observed for Greek, Gretino Gold proves more challenging than Gretino Silver across all models, with a consistent performance gap in both the `target` and `target+random` settings. We interpret these results as further evidence that Gretino is a valuable benchmark for evaluating retrieval systems in Latin as well.

Cross Lingual Results: We test models also in a cross-lingual configuration. Table 2 reports the same metrics for the cross-lingual setting, where Latin queries are matched against Greek documents (Greek+Latin) and vice versa (Latin+Greek). The results are consistent with the previous ones, and the only model showing higher performance is SPhilBERTa. Despite being multilingual, in this context PhilBERTA shows a wider gap with SPhilBERTa than it does in the single language setting.

The Gold part of the dataset remains more challenging as it is for both languages when testing models in a single language. In contrast, the gap between `target` and `target+random` is less pronounced, showing that SPhilBERTa has a small performance difference when tested on synthetic texts. Notably, SPhilBERTa is partly trained on synthetic texts, although generated with a different

Language model, PhilTA instead of models from the GPT family.

Overall, these results confirm the value of the Gretino dataset as a meaningful benchmark for evaluating semantic retrieval systems in Ancient Greek, Latin, and across these two languages. Its design allows researchers to assess both fine-grained performance differences between language models and their applicability in broader and more realistic cross-lingual and domain-specific retrieval tasks. In particular, the combination of synthetically generated and manually curated sentences ensures that Gretino captures the complexity and nuances of classical texts, in both literal parallels and more nuanced semantic connections, making it a reliable tool for advancing retrieval research in the context of Digital Humanities.

5. Model Adaptation with SimCSE

With the goal to offer a semantic retrieval tool for Humanists and Digital Humanists scholar, we outline our preliminary efforts to fine-tune retrieval models with the aim of improving their performance in recognizing semantic features in sentences both in Ancient Greek, Latin and cross-lingual. In this framework it becomes clear that evaluation on the Gretino benchmark has been particularly valuable for assessing both the training dynamics and the quality of the training data. To explore this possibility, we collect two parallel dataset of Greek and Latin texts:

- SimCSE1: is tailored with a corpus of paired verses from the Greek and Latin Bible. This corpus is composed of two parts: (1) The first part contains around 6k pairs of Greek and Latin sentences collected from the Bible: we use the book numbering system to align a greater amount of couples, then we remove verses with less than 60 characters, we then calculate similarity with SPhilBERTa and we manually verify all the sentences with a similarity score minor than 0,65 keeping only good example couples according to domain-expert evaluation. We also eliminate pairs with similarity higher than 0.85 to avoid model collapse. (2) The second part contains around 1.5k pairs of sentences that are sensitively longer than those contained in the first part of this dataset. 700 are manually paired keeping only those with a lower similarity according to SPhilBERTa, while 800 are randomly selected by the higher similarity sentences.

- SimCSE2: This dataset is smaller, couples actual ancient texts with augmented data, does not contain texts from the Bible and is composed of two parts: (1) 300 pairs of Ancient Greek sentences, created by selecting genuine texts, translating and rephrasing them by a domain expert (modifying syntactic structures and lexical choices while preserving semantic meaning), and then back-translating them into Ancient Greek using DeepSeek and GPT-4. (2) 100 pairs of authentic Latin sentences paired with their Greek counterparts, generated through the same process.

After collecting the datasets, we fine-tune the SPhilBERTa model (*the best-performing model in all our experiments*) using the SimCSE loss (Gao et al., 2021), a contrastive learning technique that encourages the model to develop similar representations for sentences with the same meaning in different languages. We chose to perform SimCSE in a supervised way, feeding the model with the aforementioned domain-expert verified material and exploiting all the other sentences in batch as negative examples. Therefore, SPhilBERTa was further optimized using a SimCSE-style contrastive learning approach, applying CLS pooling, where the [CLS] token serves as the sentence-level representation.

Tokenization was performed with padding up to a maximum length of 128 tokens and automatic truncation of longer sequences. The model was trained for 10 epochs, with a batch size of 32, and an initial learning rate of 5e-5. Gradient clipping uses a threshold of 1.0. The loss function is the standard SimCSE contrastive loss, with a temperature parameter set to 0.05, encouraging the separation of semantically distinct sentence pairs. Evaluation and checkpoint saving were scheduled every 20 steps, with early stopping triggered after 2 evaluations without improvement in the evaluation loss. The best model was automatically saved at the end of training, with SimCSE1 stopping at 10 epochs and SimCSE2 at 6.

We remark that the authors of SPhilBERTa (Riemenschneider and Frank, 2023b) tested the model in a retrieval setting similar to ours, however they trained it using distillation from existing multilingual models, while we use a different approach.

Table 3 shows the results of the fine-tuned model on Gretino. We can see that the fine-tuned model performs better than the pretrained one. Specifically, the performance on single languages is mostly unchanged, while on the cross-lingual setting SimCSE2 systematically outperforms SPhilBERTa with P@5 differences of up to 6% in some settings.

6. Conclusions

In this work, we introduced Gretino, a novel benchmark dataset designed to evaluate the retrieval capabilities of language models in Ancient Greek and

split	setting	model	MAP	recip-rank	P@5	P@10	NDCG@5	NDCG@10	MAP	recip-rank	P@5	P@10	NDCG@5	NDCG@10
			Greek						Latin					
silver	target+random	SimCSE1	0.96	1.00	0.93	0.49	0.95	0.98	0.95	0.99	0.93	0.49	0.94	0.97
		SimCSE2	0.97	1.00	0.94	0.49	0.95	0.98	0.96	0.99	0.93	0.49	0.95	0.97
		SPhilBERTa	0.95	1.00	0.92	0.48	0.94	0.97	0.96	0.99	0.93	0.48	0.95	0.97
	target	SimCSE1	0.94	0.99	0.90	0.48	0.93	0.96	0.94	0.99	0.90	0.48	0.92	0.96
		SimCSE2	0.95	0.99	0.92	0.48	0.94	0.97	0.95	0.99	0.92	0.48	0.94	0.96
		SPhilBERTa	0.95	0.99	0.91	0.48	0.93	0.96	0.95	0.99	0.92	0.49	0.94	0.97
gold	target+random	SimCSE1	0.81	1.00	0.74	0.43	0.80	0.87	0.71	0.90	0.66	0.39	0.72	0.78
		SimCSE2	0.85	0.94	0.82	0.44	0.85	0.88	0.82	0.96	0.75	0.43	0.81	0.87
		SPhilBERTa	0.83	0.97	0.76	0.45	0.81	0.89	0.72	0.93	0.63	0.38	0.71	0.77
	target	SimCSE1	0.83	0.97	0.73	0.43	0.79	0.87	0.77	0.96	0.68	0.42	0.74	0.84
		SimCSE2	0.84	0.97	0.78	0.44	0.82	0.88	0.80	0.95	0.75	0.43	0.79	0.85
		SPhilBERTa	0.83	0.94	0.79	0.44	0.82	0.87	0.78	0.95	0.69	0.41	0.75	0.83
split	setting	model	Latin+Greek						Greek+Latin					
silver	target+random	SimCSE1	0.94	1.00	0.90	0.48	0.93	0.97	0.94	0.99	0.90	0.48	0.92	0.96
		SimCSE2	0.94	0.99	0.91	0.48	0.93	0.96	0.95	0.99	0.92	0.48	0.94	0.97
		SPhilBERTa	0.90	0.99	0.87	0.47	0.90	0.93	0.93	0.99	0.89	0.48	0.92	0.96
	target	SimCSE1	0.93	0.99	0.88	0.48	0.91	0.95	0.92	0.99	0.88	0.47	0.91	0.95
		SimCSE2	0.93	0.99	0.89	0.48	0.92	0.95	0.94	0.99	0.91	0.48	0.93	0.96
		SPhilBERTa	0.93	0.99	0.89	0.48	0.91	0.95	0.93	1.00	0.89	0.48	0.92	0.96
gold	target+random	SimCSE1	0.73	0.95	0.68	0.39	0.74	0.79	0.53	0.70	0.47	0.30	0.51	0.58
		SimCSE2	0.85	0.97	0.78	0.45	0.84	0.90	0.65	0.86	0.59	0.34	0.65	0.70
		SPhilBERTa	0.71	0.94	0.65	0.38	0.72	0.77	0.58	0.81	0.50	0.33	0.56	0.65
	target	SimCSE1	0.74	1.00	0.64	0.41	0.73	0.82	0.69	0.88	0.61	0.40	0.66	0.76
		SimCSE2	0.77	0.95	0.69	0.41	0.76	0.82	0.74	0.91	0.64	0.41	0.70	0.80
		SPhilBERTa	0.75	1.00	0.63	0.41	0.72	0.82	0.73	0.90	0.64	0.41	0.69	0.78

Table 3: Comparison of performance scores between SPhilBERTa-baseline and models fine-tuned using SimCSE. The red and blue gradient shades indicate higher and lower performance respectively for each metric.

Latin. The dataset comprises two components: Gretino Silver, a synthetic collection of query-target pairs, and Gretino Gold, a curated corpus drawn from authentic ancient texts. We evaluated four pre-trained state-of-the-art models and two SimCSE fine-tuned variants introduced in this paper, demonstrating that Gretino is well-suited for benchmarking retrieval systems in both monolingual and cross-lingual settings.

Our experiments with fine-tuning SPhilBERTa on a parallel corpus of Greek and Latin texts show performance gains, particularly in cross-lingual retrieval, confirming the effectiveness of domain-specific fine-tuning for semantic search tasks in classical languages.

Overall, Gretino represents a valuable resource for advancing research in computational philology and digital humanities. We believe it can serve as a robust benchmark for future research in retrieval systems for classical languages, and we hope it will foster the development of new models and tools to support the study of Ancient Greek and Latin.

In future work, we aim to collect larger and more diverse sets of unrelated documents, in order to make the benchmark more challenging and to encourage the development of more robust retrieval models for Ancient Greek, Latin, and classical languages more broadly. This will facilitate deeper integration of AI-based tools into existing Digital Humanities practices.

7. Acknowledgments

This paper was performed in the framework of the Italian Strengthening of the ESFRI RI RESILIENCE (ITSERR) project (cod. Progetto IR0000014 - CUP B53C22001770006) – Piano Nazionale di Ripresa e Resilienza (PNRR) – Missione 4, “Istruzione Ricerca”, Componente 2, “Dalla ricerca all’impresa”, Investimento 3.1, “Fondo per la realizzazione di un sistema integrato di infrastrutture di ricerca e innovazione”, finanziato dall’Unione Europea – NextGenerationEU – rif. Avviso MUR 3264/2021. ITSERR is an interdisciplinary and distributed Research Infrastructure for Religious Studies that aims to strengthen the RESILIENCE RI project in Italy. The necessity of a semantic retrieval tool was raised by the case studies that Elia Scapini and Federico Iezzi encountered during their work as Ph.D students for the fourth work package (WP4) of the ITSERR project. WP4 studies the semantics of the Nicene-Constantinopolitan Creed producing both humanities research and AI-based tools. In this framework they observed the lack of a tool capable of returning similar meanings, rather than similar words verbatim. Federico Iezzi and Elia Scapini are also members of the Fondazione per le Scienze Religiose (FSCIRE) in Bologna. Hawau Olamide Toyin was funded by the grant as a visiting scholar.

References

Yannis Assael, Thea Sommerschield, Brendan Shillingford, Mahyar Bordbar, John Pavlopoulos, Maria Chatzipanagiotou, Ion Androutsopoulos, Jonathan Prag, and Nando Freitas. 2022. Restoring and attributing ancient texts using deep neural networks. *Nature*, 603:280–283.

David Bamman and Patrick J. Burns. 2020. Latin BERT: A contextual language model for classical philology. *CoRR*, abs/2009.10053.

Rufus Behr. 2024. Behr at EvaLatin 2024: Latin dependency parsing using historical sentence embeddings. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 198–202, Torino, Italia. ELRA and ICCL.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034. Association for Computational Linguistics.

Charlie Cowen-Breen, Creston Brooks, Barbara Graziosi, and Johannes Haubold. 2023. Logion: Machine-learning based detection and correction of textual errors in greek philology. In *Proceedings of the Ancient Language Processing Workshop*, pages 170–178. INCOMA Ltd., Shoumen, Bulgaria.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Koutsikakis, Ilias Chalkidis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2020. GREEK-BERT: The greeks visiting sesame street. arXiv:2008.12014. arXiv.

Kevin Krahn, Derrick Tate, and Andrew C. Lamicela. 2023. Sentence embedding models for Ancient Greek using multilingual knowledge distillation. In *Proceedings of the Ancient Language Processing Workshop*, pages 13–22, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Wouter Mercelis. 2024. KU leuven / brepols-CTLO at EvaLatin 2024: Span extraction approaches for Latin dependency parsing. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 203–206, Torino, Italia. ELRA and ICCL.

Wouter Mercelis and Alek Keersmaekers. 2022. An ELECTRA model for Latin token tagging tasks. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 189–192, Marseille, France. European Language Resources Association.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023a. Exploring large language models for classical philology. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15181–15199, Toronto, Canada. Association for Computational Linguistics.

Frederick Riemenschneider and Anette Frank. 2023b. Graecia capta ferum victorem cepit. detecting latin allusions to ancient greek literature. In *Proceedings of the Ancient Language Processing Workshop*, pages 30–38. INCOMA Ltd., Shoumen, Bulgaria.

Frederick Riemenschneider and Kevin Krahn. 2024. Heidelberg-boston @ SIGTYP 2024 shared task: Enhancing low-resource language analysis with character-aware hierarchical transformers. In *Proceedings of the 6th Workshop on Research*

in Computational Linguistic Typology and Multilingual NLP, pages 131–141. Association for Computational Linguistics.

Pranaydeep Singh, Gorik Rutten, and Els Lefever. 2021. A pilot study for BERT language modelling and morphological analysis for ancient and medieval greek. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 128–137. Association for Computational Linguistics.

Andreas Spanopoulos. 2022. Language models for ancient greek.

Rachele Sprugnoli, Federica Iurescia, and Marco Passarotti. 2024. Overview of the EvaLatin 2024 evaluation campaign. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 190–197. ELRA and ICCL.

Rachele Sprugnoli, Francesco Mambrini, Marco Passarotti, and Giovanni Moretti. 2023. The sentiment of latin poetry. annotation and automatic analysis of the odes of horace. *IJCoL. Italian Journal of Computational Linguistics*, 9(1). Number: 1 Publisher: Accademia University Press.

Milan Straka, Jana Straková, and Federica Gamba. 2024. ÚFAL LatinPipe at EvaLatin 2024: Morphosyntactic analysis of Latin. In *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA) @ LREC-COLING-2024*, pages 207–214, Torino, Italia. ELRA and ICCL.

Li Sun, Florian Luisier, Kayhan Batmanghelich, Dinei Florencio, and Cha Zhang. 2023. From characters to words: Hierarchical pre-trained language model for open-vocabulary language understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3605–3620, Toronto, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ivan P. Yamshchikov, Alexey Tikhonov, Yorgos Pantis, Charlotte Schubert, and Jürgen Jost. 2022. BERT in plutarch's shadows. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6071–6080, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tariq Yousef, Chiara Palladino, Farnoosh Shamsian, Anise d'Orange Ferreira, and Michel Ferreira dos Reis. 2022. An automatic model and gold standard for translation alignment of ancient greek. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5894–5905. European Language Resources Association.