Ancient Words, New Demands: Lessons from Recent Work on Greek, Latin, and Biblical Reception*

Retrieving Biblical References in Greek and Latin Patristic Texts

Davide Dainese[1,*], Marco Zanella[1,2]

[1]Department of History and Cultures (DiSCi), University of Bologna
[2]Department of Mathematics, University of Padua

Abstract

Identifying biblical quotations and allusions in early Christian literature is hard for computational methods: the Bible was never a single fixed text, ancient authors reshaped what they reused, and Ancient Greek and Latin are low-resource languages. We report on recent work that frames the problem as ranked retrieval, with Clement of Alexandria, Nonnus of Panopolis, and Augustine as test cases, and we argue that the decisive factor is knowledge enhancement: making philological knowledge operational within the pipeline. This happens at three points. A graded annotation scheme replaces the quotation/allusion dichotomy with nested levels of similarity. A rule-based parser turns the critical apparatus of the Septuagint into a knowledge graph and into virtual versions of each verse, which are indexed as historically attested data augmentation. Parallel Latin versions of the Bible supervise contrastive fine-tuning with self-mined hard negatives. Signals grounded in the textual tradition improve retrieval, while generic ones, such as paraphrases generated by large language models or unconstrained synonym expansion, degrade it. An error analysis maps the remaining failures to their causes and shows that the hardest cases require explicit exegetical knowledge. Throughout, the system ranks candidates and the scholar judges their relevance.

Keywords

Information Retrieval, Sentence Embeddings, Knowledge Enhancement, Critical Apparatus, Biblical Intertextuality, Ancient Greek, Latin, Digital Humanities

CEUR-WS.org/Vol-4314/ReDHinvited.pdf

1. Enhancing Knowledge

At least three preliminary problems confront anyone seeking to identify biblical quotations and allusions in early Christian literature. The first is methodological and historiographical: whereas an explicit quotation is relatively straightforward to recognize, we have much less robust criteria for determining when a lexical similarity, an image, a line of argument, or a motif should properly be regarded as allusive. Traditional editorial practice generally distinguishes between secure references and looser parallels, often through the opposition *cf./non-cf.*; yet this convention merely records a scholarly decision without resolving the underlying theoretical problem [1].

The other two problems are properly historical. First, there has never been a single Bible, understood as a unitary and stably fixed text. To remain within the scope of the case study of our PRIN2022 project, the Septuagint (i.e. the main ancient Greek translation of the Hebrew Scriptures) was transmitted in a plurality of manuscript variants and, in antiquity, coexisted with other Greek versions [2]. Different ethnolinguistic and cultural groups could also rely on scriptural corpora and textual traditions that did not coincide at all. Identifying a biblical reference therefore means more than simply recognizing a verse: whenever possible, it also requires asking which form of that text was known to the author and within which textual tradition it should be situated.

ReDH 2026: Rethinking the Future of Digital Humanities through Ethics and Artificial Consciousness, co-located with AIxIA 2026, October 6–9, 2026, Perugia, Italy

*Invited talk.

*Corresponding author.
davide.dainese@unibo.it (D. Dainese); marco.zanella5@unibo.it,marco.zanella@unipd.it (M. Zanella)
0000-0002-9505-9739 (D. Dainese); 0009-0000-1208-6743 (M. Zanella)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-05

The second historical problem concerns the way in which ancient Christian authors (those whom we conventionally call the Church Fathers) used Scripture. Their familiarity with the biblical text was such that it is often difficult to separate what derives directly from the text from what belongs instead to the history of its interpretation. A given formulation may depend on the biblical text itself, on an earlier exegetical interpretation, on a tradition that had already become sedimented, or on the author's own elaboration. Moreover, the closer we move toward the earliest phases of Christianity, the more the scarcity of surviving sources makes it difficult to reconstruct these intermediate stages with precision.

Against this problematic background, we set out to develop a tool that could help scholars navigate biblical-patristic intertextuality. To this end, we selected two very different authors. The first is Clement of Alexandria, who lived between the second half of the second century and the first decades of the third. His chronological position is particularly useful because he stands relatively close to the period in which the main Greek biblical traditions were taking shape and circulating, and he is one of the earliest Christian authors whose exegetical practice can be followed in substantial detail.

In his case, however, the question of the intertextual relations between Scripture and his writings is intertwined with a broader problem concerning the very nature of his oeuvre. Part of it has come down to us in fragmentary or incomplete form, and one of the major issues in Clementine studies is to understand the relationship between these materials and Clement's works that has survived in complete form [3, 4, 5]. For this reason, reconstructing Clement's use of Scripture is by no means marginal: identifying which biblical passages he employs, in what textual form, and which recurring patterns emerge across different works can also help us understand the organization and continuity of his thought.

The second author is Nonnus of Panopolis, a fifth-century Egyptian Christian poet and the author of a hexameter paraphrase of the Gospel of John. His distance from Clement is both chronological and literary: philosophical-theological prose gives way to a poetic rewriting shaped by metrical constraints, epic diction, and extensive exegetical reworking. Biblical material can therefore be transformed through lexical variation, thematic evocation, typological reinterpretation, or exegetical mediation, reducing (and at times almost entirely eliminating) verbal overlap with the source text.

Clement and Nonnus, then, are not simply two independent case studies. They were selected precisely because they are sufficiently different to function as a historical-philological benchmark: proximity to and distance from the ancient textual traditions of the Greek Bible; prose and poetry; reuse that remains relatively close to the source text and reuse involving radical poetic transformation. Comparing them allows us to assess which criteria for identifying textual reuse can be generalized and which instead depend on genre, period, or the specific literary practice of the author [6].

At this point, however, a fourth and more theoretical problem emerges. Strictly speaking, we did not build a tool capable of mapping intertextuality itself. Rather, we developed a tool designed to identify possible relationships between textual segments on the basis of different types and degrees of similarity. Yet intertextuality, understood in hermeneutic terms, cannot be reduced to an isolated relation between a text A and a text B. It is, rather, the network within which already interpreted texts are brought into relation through successive practices of reading, reception, and transformation [7]. Our system, therefore, identifies possible connections; it does not yet reconstruct, in the proper sense, the network that gives those connections their meaning.

This distinction, however, forced us to separate two levels that we had initially tended to conflate, leading to an insight that we regard as a collateral result which ultimately became one of the project's main theoretical milestones: both levels can, at least in part, be formalized. The first is the level of the event; the second, that of its conditions of possibility. A quotation, a paraphrase, or a lexical reuse can be traced back to historically situated events: someone read, remembered, copied, transmitted, or reused a text, all operations that have a beginning and an end. A different level is that of historical structures, that is, the conditions that make such an event possible and intelligible: linguistic competences, exegetical traditions, conceptual schemes, and cultural expectations.

In our case, the Bible can therefore be understood, at least heuristically, not only as a source from which individual expressions are drawn, but as a kind of cultural filigree running through early Christian literature, and thus as something that performs a genuinely structural function. The most interesting

methodological consequence of our work is that this second level too can, at least in part, be described, graded, and formalized. The computational problem, then, concerns not only the identification of intertextual events, but also the possibility of representing some of the historical conditions that make those events possible.

There is, however, one respect in which our wager seems to have paid off particularly clearly: the role of interaction among humanists, digital humanists, and computer scientists. The most significant outcome of the experiment did not simply consist in adding artificial-intelligence techniques to existing algorithmic procedures. Rather, it became clear that part of the knowledge produced by philology can be formalized and made operational within the computational process.

In our case, this meant taking seriously a feature that is obvious to a philologist but normally disappears when a *critically established text* is converted into a *digital dataset*: a critical edition does not contain only the text established by its editors. In its critical apparatus, it also records variant readings attested by manuscripts and by other witnesses to the textual tradition. We therefore sought to transform this information from a reference tool intended solely for human readers into structured data, linking the edited text to the alternative forms documented throughout its transmission history.

It is in this sense that we would speak of a process of *knowledge enhancement*. Our aim was not simply to provide a model with more raw data, nor to ask it to generate possible equivalences autonomously (users will ask this and evaluate the corresponding results). Rather, our purpose was to construct a search space constrained by domain knowledge: interrelated, but above all historically attested and scientifically verifiable data. Interdisciplinary collaboration therefore does not come into play only downstream, when the humanist checks or validates the machine's output; it already operates upstream, in defining the very object that the machine must be able to process.

This point is also important because it sets the limits of what we ask the system to do. We do not want to delegate historical or exegetical judgement to it. We want to place it in a position to generate candidates that are *useful* for research, drawing both on textual similarity and on formalized philological knowledge, while leaving the assessment of their relevance to the scholar. The technical section will make this distinction explicit by separating the concept of reference from those of similarity, retrieval, and relevance: the system can measure correspondences and rank candidates, but the historical relevance of a relation remains a domain-specific scholarly judgement.

At this point, we can therefore move from the historical-philological formulation of the problem to its computational formalization: defining what we mean operationally by a reference, how different degrees of similarity are represented, how this measure is turned into a retrieval task, and how the plurality of the biblical text can be made computationally accessible without losing the provenance of the information. This is where the next section begins.

2. Task and Similarity Model

As we have seen, critical editions of patristic texts record biblical references through a binary convention: a reference introduced by *cf.* is an allusion, one without it a quotation [8, 9]. Text-reuse research has moved to the opposite extreme, with typologies of up to forty-five reuse types [10]; tools such as Tesserae [11] match vocabulary at scale, and BiblIndex [12] indexes references already identified by editors. None of these fits a plural source quoted from memory [13] and reshaped by syntax, metre, and exegesis. We therefore replace categories with a graded similarity measure for annotation and a ranking task for retrieval.

Terminology. A *reference* is a relation, asserted by an annotator or by prior scholarship, between a span of a patristic text (the *query*) and a biblical verse it reuses (the *target*). *Similarity* is a computed correspondence between two strings: the annotation score S below, assigned by experts, or the model score s used for ranking. *Retrieval* ranks all candidate verses for a query. *Relevance*, whether a candidate is a genuine reference, is a scholarly judgement the system does not compute; in evaluation, the annotated target is the only relevant candidate, a simplification discussed in Section 5.

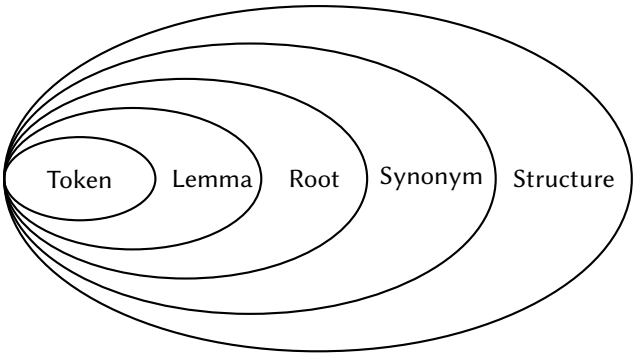


Figure 1: The five nested levels of the annotation scheme. Each level considers only the tokens left unmatched by the levels inside it; the first four contribute to the score S, the fifth flags purely conceptual parallels.

2.1. Graded annotation score

Annotation is performed on the query side, with the verse as the unit on the target side, over five nested levels of correspondence (Figure 1) [1, 14]. Let a query span consist of the tokens $t_1, \dots, t_n$ of the edited patristic text, and let $\mathscr{V}_i$ be the variant readings of t_i in the apparatus of that edition. For a candidate verse v, let $E(v)$ be the tokens of its edited text and $A(v)$ those of its apparatus. The *token-level* weight of position i is

$$w_i = \begin{cases} 1 & \text{if } t_i \in E(v), \\ 1/2 & \text{if } t_i \in A(v) \text{ or } \mathscr{V}_i \cap E(v) \neq \emptyset, \\ 1/4 & \text{if } \mathscr{V}_i \cap A(v) \neq \emptyset, \\ 0 & \text{otherwise,} \end{cases} \tag{1}$$

taking the first case that applies. Matches involving the apparatus are discounted because they depart from the text established by the editors. The token score is $S_1 = \frac{1}{n}\sum_i w_i$. Each outer level $\ell \in \{2, 3, 4\}$ (lemma, root, synonym) considers only the positions U_ℓ with no match at inner levels, and adds those that match at level ℓ:

$$S_\ell = S_{\ell-1} + \frac{1}{n}\left|\{\, i \in U_\ell \,:\, m_\ell(t_i, v)\,\}\right|, \tag{2}$$

where m_2 tests for a shared lemma, m_3 for a shared root, and m_4 for synonymy as judged by the annotator, including contextual synonymy. The fifth level, *structure*, flags thematic, syntactic, rhetorical, or conceptual parallels that leave no token-level trace, and adds nothing to the score.

Two examples from Augustine's *De Genesi ad litteram* [15] illustrate the computation. In *Terra autem erat inuisibilis et inconposita, et tenebrae erant super abyssum; et spiritus dei superferebatur super aquam* (I,13), 11 of the 17 tokens occur in the Vulgate text of Gen 1:2, *erant* in its apparatus, and *aquas*, a variant in Zycha's apparatus, in its text: $S_1 = (11 + 2 \cdot \frac{1}{2})/17 \approx 0.71$. In Augustine's Matt 13:52 (*scribam eruditum in regno Dei similem esse patrifamilias proferenti de thesauro suo nova et vetera*), the quotation becomes an infinitive clause and reaches only $S_1 = 0.625$; lemmas recover *scribam*, *similem*, *esse*, and *proferenti* ($S_2 = 0.8125$), and synonymy the remaining *eruditum* (*doctus*) and *Dei* (*caelorum*, contextually). References with $S_1 = 0$, such as *in aedificatione turris post diluvium* (I,2) for Gen 11:4, are kept to test non-literal retrieval.

2.2. Retrieval task and metrics

Let $\mathscr{C} = \{v_1, \dots, v_N\}$ be a corpus of verses and $\mathscr{Q}$ a set of queries, each with one annotated target $v_q^* \in \mathscr{C}$. A system scores every pair with a function $s(q, v)$ and ranks $\mathscr{C}$ by decreasing score; r_q denotes the rank of v_q^*. For a dense encoder f, $s(q, v) = \cos\big(f(q), f(v)\big)$ [16]. We report

$$\mathrm{R@}k = \frac{1}{|\mathscr{Q}|}\sum_{q \in \mathscr{Q}} \mathbf{1}[r_q \le k], \qquad \mathrm{MRR} = \frac{1}{|\mathscr{Q}|}\sum_{q \in \mathscr{Q}} \frac{1}{r_q}. \tag{3}$$

Because each query has a single target, R@k moves in steps of $1/|\mathscr{Q}|$.

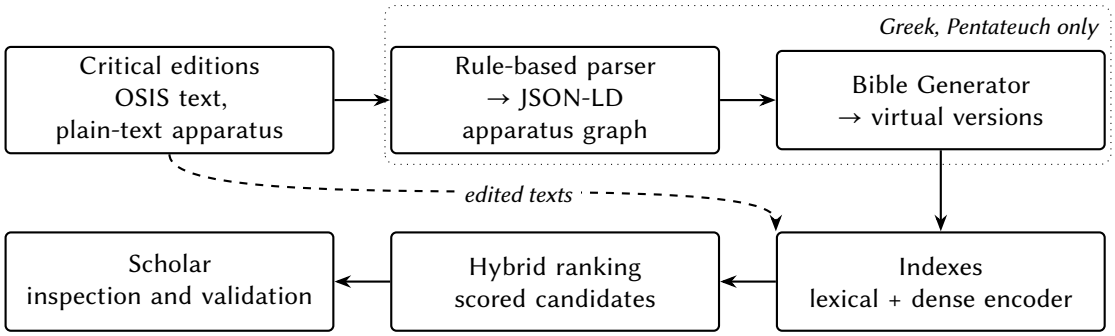


Figure 2: The retrieval pipeline. Edited texts (Greek Old Testament, Vulgate, Vetus Latina) are indexed directly; for the Greek Pentateuch, the apparatus is parsed into a graph and expanded into virtual versions, which are indexed alongside the edited text. The output is a ranked list of candidates for the scholar to validate.

3. Data Pipeline

Apparatus-aware retrieval requires the plurality of the biblical text in machine-actionable form, whereas existing resources offer a single edited text without apparatus (the Thesaurus Linguae Graecae) or an apparatus reproduced typographically [17]. Figure 2 shows our pipeline; this section covers its first three stages.

3.1. Parsing the apparatus

The Göttingen Septuagint provides its edited text in OSIS XML, with an explicit book–chapter–verse hierarchy, but its two apparatuses, textual and hexaplaric, only as plain text. An entry such as

1:1 ἐποίησεν] επλασεν 664; εκτισεν Ios. I 27

must be decomposed into a locus (Gen 1:1), a lemma (ἐποίησεν), and two readings with their witnesses: ἔπλασεν in manuscript 664, and ἔκτισεν in the indirect tradition, through Josephus, *Ant.* I,27 [18]. In the hexaplaric apparatus each reading must also be attributed to a translator (Aquila, Symmachus, Theodotion). The rule-based parser of the EnGraSept dataset [19] derives its rules from a "philological dictionary", maintained jointly by the developer and a digital philologist, which defines each element of the apparatus (lemma, reading, reading type such as omission or addition, witnesses, abbreviations) with its expected structured output: in domain-driven-design terms, the ubiquitous language of the two disciplines [20]. Development was "anytime correct": rules for the most frequent structures were made reliable first and coverage extended incrementally, so that every parsed entry is structurally correct even while the corpus is incomplete. The parser currently covers the Pentateuch.

3.2. Apparatus graph and virtual versions

The parser emits JSON-LD, which the search engine indexes directly and which, being RDF-compatible, makes the apparatus a FAIR knowledge graph [21]. The model reuses the Critical Apparatus Ontology [22] and the Critical Edition Ontology [23], adding classes specific to the Göttingen tradition such as witnesses and their carriers. Relations implicit on the printed page become queries: which witnesses share a reading, how close two witnesses are, which variants two patristic authors share.

The Bible Generator turns the graph into *virtual versions*: for each verse and witness, the edited text with that witness's readings applied (for the entry above, Gen 1:1 with επλασεν for witness 664). Witnesses not cited at a locus are assumed to agree with the edited text. Since this standard convention flattens editorial uncertainty, virtual versions represent only the variants recorded by the editors, and each reading keeps a link to its apparatus entry.

Table 1
Evaluation benchmarks. Each query is a patristic span with one annotated target verse; the corpus is the pool of candidates.

Language	Queries	References	Corpus (verses)
Greek	Clement of Alexandria (prose)	78	Old Testament (21,365)
	Nonnus of Panopolis (hexameter)	32	
Latin	Augustine, *De Genesi ad litteram*	192	Vulgate (35,057)
	Augustine, *De Genesi ad litteram*	170	Vetus Latina (20,791)

3.3. Corpora and benchmarks

The Greek retrieval corpus is the whole Greek Old Testament, 21,365 verses; the New Testament is not included, and virtual versions exist only for the Pentateuch. For Latin, the corpora are Jerome's Vulgate in the Weber–Gryson edition [24] and the Vetus Latina in Sabatier's reconstruction [25], the only complete one; the Beuron edition [26], incomplete but rich in variants, was used in annotation. Sabatier's third volume was transcribed with OCR and revised manually, since the available scans were too poor.

Domain experts annotated gold references on INCEpTION [27], resolving disagreements by discussion. Each annotation records the span, any introductory formula (*scriptum est*, *dicente Deo*), the target in each biblical version, and the scores of Section 2 (Table 1). The Vulgate set is spread almost evenly over four bands of S_1 (0–25%, 25–50%, 50–75%, 75–100%); the Vetus Latina set is skewed towards near-verbatim references (83 in the top band). Variants from editions under copyright are used only for indexing and training, and shown to users as metadata.

4. Retrieval Models

The dense encoder is one component of a hybrid engine (Section 4.3). Sections 4.1 and 4.2 evaluate it in isolation, varying only the encoder f, so that a better one can be swapped into the pipeline. Since corpora, benchmarks, and systems differ, figures are not comparable across subsections.

Protocol. All encoders produce 768-dimensional embeddings. Fine-tuning uses Adam with a learning rate of 10^{-6} and batches of 32 queries; for Latin, each query is paired with 7 sampled negatives, and a run takes about 6 hours on a single NVIDIA A40 GPU. Greek uses the same optimiser settings, with positives and negatives built according to the strategy. Training data comes from the biblical corpora, so target verses are seen as documents, as is standard in retrieval; query–target pairs are used only for evaluation.

4.1. Latin: contrastive fine-tuning with self-mined hard negatives

We compared three encoders pretrained on Latin: Latin RoBERTa [28], Latin BERT [29], and LaBERTa [30]. Without fine-tuning, mean pooling of the output token embeddings outperformed the `[CLS]` embedding for nearly all models, and Latin BERT and LaBERTa outperformed Latin RoBERTa [31]; the two stronger models were fine-tuned with mean pooling.

Commentary–verse pairs are too few for training, so fine-tuning uses a surrogate task: a Vulgate verse q and the corresponding Vetus Latina verse d^* (or vice versa) form a positive pair. With a set $\mathcal{N}_q$ of negatives, the model minimises the InfoNCE loss [32]

$$\mathcal{L} = -\sum_{(q,d^*)\in\mathcal{B}} \log \frac{\exp s(q,d^*)}{\exp s(q,d^*) + \sum_{d\in\mathcal{N}_q} \exp s(q,d)} \tag{4}$$

Table 2
Latin dense retrieval on *De Genesi ad litteram* (Recall@k, %): pretrained encoders and fine-tuning without and with self-mined hard negatives (HN). 192 queries for the Vulgate, 170 for the Vetus Latina.

		Vulgate		Vetus Latina	
Model	Fine-tuning	R@1	R@10	R@1	R@10
Latin BERT	none	33.3	47.9	35.3	48.8
	w/o HN	38.5	58.9	35.9	54.1
	w/ HN	**47.4**	59.9	38.8	55.3
LaBERTa	none	34.4	52.6	33.5	47.6
	w/o HN	41.1	**64.6**	37.1	**62.4**
	w/ HN	43.2	63.5	**41.8**	61.2

Table 3
Greek dense retrieval with sphilBERTa by fine-tuning strategy; best value per column in bold. With 78 and 32 queries, one query moves R@k by 0.013 on Clement and by 0.031 on Nonnus: many differences in the table amount to one or two queries.

	Clement (78)			Nonnus (32)		
Strategy	MRR	R@10	R@50	MRR	R@10	R@50
Pretrained baseline	.252	.321	.385	.023	**.094**	.125
Supervised pairs	.262	.346	.462	.019	.063	.156
Multiple negatives ranking loss	.254	.346	.410	**.046**	**.094**	.125
Apparatus-based augmentation	.271	.346	.397	.028	.063	.188
LLM-generated paraphrases	.210	.256	.372	.002	.000	.063
Segment-level embeddings	.252	.321	**.513**	.013	.031	.156
Domain adaptation (MLM)	.214	.295	.346	.011	.031	.063
Weak contrastive (TF-IDF)	**.290**	.372	.436	.030	.063	.188
Weak contrastive (English Bible)	.263	**.397**	.449	.037	**.094**	.125
Translation-based weak supervision	.236	.282	.410	.021	.063	**.250**
Cross-lingual distillation	.034	.038	.064	.001	.000	.031
Hard cases, continual fine-tuning	.259	.359	.449	.029	**.094**	.094
Hard cases, retraining from scratch	.057	.064	.077	.000	.000	.000

over a batch $\mathcal{B}$, where $\mathcal{N}_q$ holds the sampled negatives and the other queries' positives and negatives. *Self-mined hard negatives* are the top-k neighbours of each d^* in the Vulgate as encoded by the pretrained model; these near neighbours are incorrect matches, which makes the objective harder [14].

Fine-tuning improves every configuration (Table 2). Hard negatives raise R@1 in all four model–corpus pairs, most for Latin BERT on the Vulgate (38.5 to 47.4), while at R@10 their effect is mixed. The largest gains are on the hardest references: for $S_1 < 0.25$ on the Vulgate, fine-tuned LaBERTa goes from 0.0 to 15.7 in R@1 and from 9.8 to 41.2 in R@10. Synthetic positives and hard negatives generated by a large language model and validated by philologists also improve Latin retrieval [33].

4.2. Greek: a comparison of fine-tuning strategies

For Ancient Greek we fine-tuned sphilBERTa, a sentence encoder built on a RoBERTa model pretrained on Ancient Greek [34], with twelve families of strategies, all trained on the Greek Old Testament: supervised pairs; multiple negatives ranking loss; triplets by relative similarity; segment-level training; augmentation by apparatus rules or by paraphrases from large language models; masked-language-model adaptation [35]; weak contrastive learning, with pairs inferred from rare shared vocabulary (TF-IDF) or from aligned English verses (King James Version); translation-based weak supervision; cross-lingual distillation [36]; lemma overlap with hard negatives; same-book batching; and extra training on hard cases, continuing from the best model or from scratch. Table 3 reports a selection.

On Clement, weak contrastive learning gives the most consistent gains: the TF-IDF variant has the best MRR (0.290 against 0.252 for the baseline), the English-Bible variant the best R@10 (0.397 against 0.321, i.e. six more queries in the top ten). Segment-level training gives the best R@50 (0.513), which matters when the encoder feeds a reranker. Four strategies fall below the baseline for identifiable reasons. Cross-lingual distillation (MRR 0.034) forces the Greek space to imitate an independent English one through lossy translations, destroying the geometry cosine retrieval relies on. Masked-language-model adaptation (0.214) optimises token prediction, an objective misaligned with sentence similarity. Paraphrases from large language models (0.210) are unreliable Ancient Greek, so treating them as equivalent to their sources injects false positives. Retraining from scratch on hard cases alone (0.057) drops the easy pairs and collapses the model; continual fine-tuning on the same cases stays at 0.259. On Nonnus, almost every strategy stays near the baseline, with gains only at R@50. Given the sample sizes, these results indicate trends only.

4.3. The hybrid engine

The full engine indexes the edited texts and, for the Pentateuch, the virtual versions, combining deterministic structured retrieval with the dense component. Users select the breadth of similarity (exact match, inflections, roots, synonyms, or embedding-based structures), mirroring the annotation levels; segment-level retrieval splits verses and aggregates their matches. Indexing the virtual versions raised R@10 on Clement from about 0.4 to nearly 0.7 and moved relevant verses earlier; for Nonnus the gain was modest, as his reworkings exceed the variation recorded by the tradition [19]. Experimental query expansion with synonym groups from Ancient Greek WordNet [37] lowered both recall and MRR: the groups were over-inclusive, and synonymy depends on context (“father” and “God” are interchangeable only in some passages).

5. Error Analysis

Aggregate metrics measure *benchmark correctness*: whether the one annotated target is ranked high. *Scholarly usefulness*, whether the candidates a scholar inspects lead to valid insight, requires a separate assessment. The two diverge in both directions: other genuine references to the same passage count as errors, and irrelevant candidates ranked above the target affect the score only through the target’s rank. We therefore analysed recall failures of the hybrid engine on the Greek benchmarks: queries whose target was missed or ranked too low to be noticed, with the top fifteen of fifty candidates inspected. The targets considered are restricted to the Pentateuch, where virtual versions exist; candidates are drawn from the whole Old Testament. For Nonnus, whose hypotext is the Gospel of John (not in the corpus), the targets are the Old Testament passages the *Paraphrase* evokes beyond John. Rankings refer to a testing phase that ended in 2025. Table 4 summarises the failure types observed.

On Clement, 18 of the 78 queries (23%) returned no positive result. The misses are nonetheless informative, and some are useful. For *Paed.* 1.6.41.2, τὸ μάννα οὐρανόθεν ἐπερρέετο, the benchmark target Exod 16:4 (ἐγὼ ὕω ὑμῖν ἄρτους ἐκ τοῦ οὐρανοῦ) was missed, but Exod 16:31, where the manna is compared to white coriander seed (ὡς σπέρμα κορίου λευκόν), was ranked second. In *Paed.* 1.6 Clement describes the Logos as nourishing milk, explains milk as transformed blood, and compares it to seed: the non-target candidate fits the argument better than the target. This is a benchmark error and a scholarly success. Lower in the same list, Wis 14:21 (rank 11) is pertinent for a reader who knows the *Protrepticus*, while Tob 11:9 is plausible but marginal.

The Nonnus cases in Table 4 form a scale. Where query and target share vocabulary, the target is reached, if late. Where the allusion is spread over lines, query design compensates in part: combining *Par.* 3.71 and 3.72 into one query puts Deut 8:15, which has serpent, bite, and desert together, first, and brings the engine into the right pericope. Where the link is purely exegetical, as in the devil ἀνδροκτόνος, ἐξότε κόσμου of *Par.* 8.126, which paraphrases John 8:44 and is related by tradition to the serpent of Gen 3, no surface signal exists: the only candidate combining an evil being and killing, the demon of Tob 3:8, comes 30th. Across both authors, then, some failures have engineering remedies

Table 4
Representative failure types of the hybrid engine on Greek queries (ranks among 50 candidates). Remedies other than apparatus indexing are proposals that remain to be evaluated.

Failure type	Query → target	Observed behaviour and cause	Remedy
Unindexed variant	*Paed.* 1.5.23.1 → Gen 22:6	rank 23 on the edited text; the citing text follows another branch of the tradition	index virtual versions (rank 1)
Orthography	*Paed.* 1.6.41.2 → Exod 16:31	μάν vs μαν in two editions; matched only once the editions were linked	normalise orthography and diacritics
Morphology	*Par.* 3.72 → Num 21:6	ὄφεις in the target, ὄφιν in the query and in neighbouring verses; the target never surfaces	lemma-level indexing
Lexical distractor	*Par.* 4.102 → Exod 29:12	rank 33; bull and horns co-occur in the law on the goring ox (Exod 21:28–36)	rerank with pericope and context features
Distributed allusion	*Par.* 3.71–72 → Num 21:4–6	each line alone fails; the composite query reaches the pericope	multi-line and sliding-window queries
Poetic form	*Par.* 2.1–4 → Exod 19:11	metrical τριτάτη, unattested in the Bible, matches only at root level; rank 7	root-level matching
Conceptual link	*Par.* 8.126 → Gen 3:4	no shared vocabulary; the link exists only in exegetical tradition; not retrieved	explicit exegetical knowledge

(orthography, morphology, query granularity), while the conceptual cases require knowledge that lies outside the texts.

6. Lessons Learned

Taken together, the Greek and Latin work supports five lessons. Each concerns what ancient texts demand of AI methods.

Textual plurality is signal. The variants of the critical apparatus are historically attested and encode documented transformations (substitution, omission, addition). Indexing them raised recall for Clement substantially, and the apparatus-aware scores make it possible to record that a reference follows a branch of the tradition other than the one printed. The apparatus is a form of data augmentation controlled by philology. Its value is highest when reuse stays within the range of attested variation, as in Clement's prose, and falls when reuse is poetic rewriting, as in Nonnus. Not every philologically legitimate variant is equally informative computationally, and adding data indiscriminately adds noise.

Signal quality matters more than quantity. In both languages, the strategies that worked drew clean training signals from the structure of the biblical corpus itself: parallel Latin versions, rare shared Greek vocabulary, aligned English verses. The strategies that failed imported signals from outside the domain: unvalidated paraphrases from large language models, geometry distilled from an English encoder, a pretraining objective misaligned with retrieval. The contrast with the Latin branch, where synthetic data helped after philological validation, locates the problem in synthetic ancient-language text that nobody has checked.

Genre and unit of analysis matter. The same methods that give measurable gains on prose barely move the needle on hexameter paraphrase. The biblical verse, the natural unit of the source side, is

often the wrong unit on the target side: Nonnus distributes an allusion over several lines, and composite queries partly compensate for it, just as segment-level retrieval helps on Clement. Evaluation should report results separately by genre, as we did.

Similarity and relevance diverge. Embedding models favour local lexical convergence and statistically frequent configurations. They are weakest where ancient reuse is most characteristic: metaphorical condensation, allusions distributed over a passage, and connections mediated by an interpretive tradition. This follows from how distributed representations work in general. The appropriate role for the model is heuristic: to widen the set of candidates a scholar inspects, within a manageable range, and to suggest questions. Failures are diagnostic: they mark where expert interpretation remains indispensable.

Responsibility is designed in. The architecture makes each suggestion checkable: every generated reading can be traced to the apparatus entry it comes from, structured resources make results verifiable, and the default assumptions of the Bible Generator are documented. Every inclusion or exclusion of a reference is the scholar's decision, and historical or exegetical significance requires more than a rank. Copyright is respected by exposing protected material as metadata, and the apparatus is published openly as a FAIR knowledge graph. For a field in which authority, attribution, and interpretation are the object of study, these properties are requirements.

Outlook. Four directions follow. Qualitative analysis should extend from failures to the additional suggestions returned alongside correct hits. Benchmarks must grow beyond the Pentateuch and beyond the current small sets. Retrieval should operate on units larger than the verse, especially for poetry. Finally, the hybrid architecture should represent exegetical knowledge explicitly, including typological relations, traditional associations between passages, and documented patterns of interpretation, so that connections which leave no trace on the surface of the text become reachable. Interpretation would remain the scholar's task.

Acknowledgments

Section 1 was written by Davide Dainese, and Sections 2–6 by Marco Zanella.

This paper brings together work carried out with many colleagues, whom we thank: first Arianna Rotondo, deputy principal investigator of Resilient Septuagint; then Anna Mambelli, Laura Bigoni, Isabella Pignocco, Riccardo Stefani, and Fabio Tutrone; and Davide Caffagni, Federico Cocchi, Marcella Cornia, and Rita Cucchiara (University of Modena and Reggio Emilia), who led the work on the Latin models. We also thank the other members of the Resilient Septuagint team: Luca Arcari, Laura Carnevale, Giorgia Sampò, and Gianluca Scatigno.

This work was supported by the PRIN project *Resilient Septuagint. An Initial Exploration of the Semantics of Killing and Healing in the Septuagint and its Reception in Patristic and Late Antique Sources – 3rd cent. BCE–5th cent. CE* (Prot. 20229E83B3 – CUP J53D23013060001).

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] D. Dainese, A. Mambelli, Intertestualità tra Bibbie e antichi commentari cristiani: l'esempio di *simul* nel *De Genesi ad litteram* di Agostino, Lexicon Philosophicum: International Journal for the History of Texts and Ideas 11 (2023–2024) 39–65. doi:10.19283/lph-202411.872.

[2] E. Bons, D. Candido, D. Scialabba, La Settanta. Perché è attuale la Bibbia greca?, San Metodio, Siracusa, 2016.
[3] M. Havrda, V. Hušek, J. Plátová (Eds.), The Seventh Book of the Stromateis. Proceedings of the Colloquium on Clement of Alexandria (Olomouc, October 21–23, 2010), Brill, Leiden and Boston, 2012.
[4] M. Havrda, The So-Called Eighth Stromateus by Clement of Alexandria, Brill, Leiden and Boston, 2017.
[5] P. Ashwin-Siejkowski (Ed.), Clement of Alexandria, volume 5 of *Papers Presented at the Seventeenth International Conference on Patristic Studies Held in Oxford 2015, ed. by Markus Vinzent*, Peeters, Leuven, Paris, and Bristol (CT), 2017.
[6] D. Dainese, A. Rotondo, Pathways of texts: From ancient traditions to new computational frontiers, in: L. Arcari, L. Carnevale, D. Dainese, A. Rotondo (Eds.), Lost in Translation: Bibles, Libraries, Webs of Books, Marietti1820, Bologna, 2026, pp. 9–18.
[7] S. Alkier, Intertextual dispositions: The semiotic power of intertextuality as the decentralization of meaning, in: L. Arcari, L. Carnevale, D. Dainese, A. Rotondo (Eds.), Lost in Translation: Bibles, Libraries, Webs of Books, Marietti1820, Bologna, 2026, pp. 107–114.
[8] J. Allenbach, Étapes, moyens et méthode d'analyse pour la constitution du fichier microphotographique des citations de l'Écriture chez les Pères, Université de Strasbourg, Strasbourg, 1967.
[9] J. Allenbach, A. Benoît, D. A. Bertrand, et al. (Eds.), Biblia Patristica: index des citations et allusions bibliques dans la littérature patristique. Vol. 1: Des origines à Clément d'Alexandrie et Tertullien, CNRS, Paris, 1975.
[10] M. Büchler, Informationstechnische Aspekte des Historical Text Re-use, Ph.D. thesis, Universität Leipzig, 2013.
[11] N. Coffee, J.-P. Koenig, S. Poornima, C. W. Forstall, R. Ossewaarde, S. L. Jacobson, The Tesserae project: Intertextual analysis of Latin poetry, Literary and Linguistic Computing 28 (2013) 221–228. doi:10.1093/llc/fqs033.
[12] L. Mellerin, New ways of searching with BiblIndex, the online index of biblical quotations in early Christian literature, in: C. Clivaz, A. Gregory, D. Hamidović (Eds.), Digital Humanities in Biblical, Early Jewish and Early Christian Studies, Brill, Leiden, 2014, pp. 177–190.
[13] H. A. G. Houghton, The earliest Latin translations of the Bible, in: H. A. G. Houghton (Ed.), The Oxford Handbook of the Latin Bible, Oxford University Press, New York, 2023, pp. 1–18.
[14] A. Mambelli, L. Bigoni, D. Dainese, F. Tutrone, D. Caffagni, F. Cocchi, M. Zanella, M. Cornia, R. Cucchiara, The biblical heritage in ancient Latin Christian literature: Advancing intertextual mapping through sentence embeddings, Umanistica Digitale 22 (2026) 157–186. doi:10.60923/issn.2532-8816/22160.
[15] J. Zycha (Ed.), Sancti Aureli Augustini De Genesi ad litteram libri duodecim, volume 28/1 of *Corpus Scriptorum Ecclesiasticorum Latinorum*, Tempsky, Prague and Vienna, 1894.
[16] N. Reimers, I. Gurevych, 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), 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/.
[17] T. Kauhanen, H. Kalavainen, Automated semantic tagging of the Göttingen Septuagint apparatus, TC: A Journal of Biblical Textual Criticism 25 (2020) 145–147.
[18] J. W. Wevers (Ed.), Genesis, volume I of *Septuaginta: Vetus Testamentum Graecum auctoritate Academiae Scientiarum Gottingensis editum*, Vandenhoeck & Ruprecht, Göttingen, 1974.
[19] I. Pignocco, M. Zanella, From critical apparatus to knowledge graph: The ENRiched GRAnular SEPTuagint case study, in: L. Arcari, L. Carnevale, D. Dainese, A. Rotondo (Eds.), Lost in Translation: Bibles, Libraries, Webs of Books, Marietti1820, Bologna, 2026, pp. 57–67.
[20] E. Evans, Domain-Driven Design: Tackling Complexity in the Heart of Software, Addison-Wesley, Boston, 2004.
[21] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, et al., The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016) 160018. doi:10.1038/sdata.2016.18.

[22] F. Giovannetti, The critical apparatus ontology (CAO): Modelling the TEI critical apparatus as a knowledge graph, in: E. Spadini, F. Tomasi, G. Vogeler (Eds.), Graph Data-Models and Semantic Web Technologies in Scholarly Digital Editing, BoD, Norderstedt, 2021, pp. 125–139.
[23] C. Martignano, Critical edition ontology: A concept model for digital critical editions, Umanistica Digitale 17 (2024) 71–94.
[24] R. Weber, R. Gryson (Eds.), Biblia Sacra iuxta Vulgatam Versionem, 5th ed., Deutsche Bibelgesellschaft, Stuttgart, 2007.
[25] P. Sabatier (Ed.), Bibliorum Sacrorum Latinae versiones antiquae seu Vetus Italica, Reginaldus Florentain, Reims, 1743–1751. 3 vols.
[26] B. Fischer, R. Gryson, W. Thiele, et al. (Eds.), Vetus Latina: Die Reste der altlateinischen Bibel nach Petrus Sabatier neu gesammelt und herausgegeben von der Erzabtei Beuron, Herder, Freiburg im Breisgau, 1949–.
[27] J.-C. Klie, M. Bugert, B. Boullosa, R. E. de Castilho, I. Gurevych, The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation, in: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, 2018, pp. 5–9. URL: https://aclanthology.org/C18-2002/.
[28] P. B. Ströbel, RoBERTa base Latin cased v1, https://huggingface.co/pstroe/roberta-base-latin-cased, 2022.
[29] D. Bamman, P. J. Burns, Latin BERT: A contextual language model for classical philology, arXiv:2009.10053, 2020. arXiv:2009.10053.
[30] F. Riemenschneider, A. Frank, Exploring large language models for classical philology, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, 2023, pp. 15181–15199. URL: https://aclanthology.org/2023.acl-long.846/.
[31] D. Caffagni, F. Cocchi, A. Mambelli, F. Tutrone, M. Zanella, M. Cornia, R. Cucchiara, Benchmarking BERT-based models for Latin: A case study on biblical references in ancient Christian literature, in: Proceedings of the 21st Conference on Information and Research Science Connecting to Digital and Library Science (IRCDL 2025), volume 3937 of *CEUR Workshop Proceedings*, 2025. URL: https://ceur-ws.org/Vol-3937/short11.pdf.
[32] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, arXiv:1807.03748, 2018. arXiv:1807.03748.
[33] D. Caffagni, F. Cocchi, A. Mambelli, et al., Generating synthetic data with large language models for low-resource sentence retrieval, in: Linking Theory and Practice of Digital Libraries: 29th International Conference on Theory and Practice of Digital Libraries, TPDL 2025, Springer, Cham, 2026, pp. 36–52. doi:10.1007/978-3-032-05409-8_4.
[34] F. Riemenschneider, A. Frank, Graecia capta ferum victorem cepit. detecting Latin allusions to Ancient Greek literature, in: Proceedings of the Ancient Language Processing Workshop, 2023, pp. 30–38. URL: https://aclanthology.org/2023.alp-1.4/.
[35] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8342–8360. URL: https://aclanthology.org/2020.acl-main.740/.
[36] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, arXiv:1503.02531, 2015. arXiv:1503.02531.
[37] G. Scatigno, Analysis of anachronistic lemmas and semantic fields in Ancient Greek WordNet, in: Proceedings of the 13th Global Wordnet Conference, Global WordNet Association, Pavia, 2025, pp. 95–99. URL: https://aclanthology.org/2025.gwc-1.11/.