

Graphilosophy: Graph-Based Digital Humanities Computing with The Four Books

Minh-Thu Do^{1,2}, Quynh-Chau Le-Tran^{1,2},
Duc-Duy Nguyen-Mai^{1,2}, Thien-Trang Nguyen^{1,2},
Khanh-Duy Le^{1,2}, Minh-Triet Tran^{1,2}, Tam V. Nguyen³,
Trung-Nghia Le^{1,2*}

¹University of Science, VNU-HCM, Ho Chi Minh City, Vietnam.

²Vietnam National University - Ho Chi Minh, Ho Chi Minh City, Vietnam.

³University of Dayton, Dayton, Ohio, United States.

*Corresponding author(s). E-mail(s): ltnghia@fit.hcmus.edu.vn;

Contributing authors: 24C02018@student.hcmus.edu.vn;
24C02003@student.hcmus.edu.vn; 24C02006@student.hcmus.edu.vn;
24C02021@student.hcmus.edu.vn; lkdud@fit.hcmus.edu.vn;
tmtriet@fit.hcmus.edu.vn; tamnguyen@udayton.edu;

Abstract

The Four Books have shaped East Asian intellectual traditions, yet their multi-layered interpretive complexity limits their accessibility in the digital age. While traditional bilingual commentaries provide a vital pedagogical bridge, computational frameworks are needed to preserve and explore this wisdom. This paper bridges AI and classical philosophy by introducing Graphilosophy, an ontology-guided, multi-layered knowledge graph framework for modeling and interpreting The Four Books. Integrating natural language processing, multilingual semantic embeddings, and humanistic analysis, the framework transforms a bilingual Chinese-Vietnamese corpus into an interpretively grounded resource. Graphilosophy encodes linguistic, conceptual, and interpretive relationships across interconnected layers, enabling cross-lingual retrieval and AI-assisted reasoning while explicitly preserving scholarly nuance and interpretive plurality. The system also enables non-expert users to trace the evolution of ethical concepts across borders and languages, ensuring that ancient wisdom remains a living resource for modern moral discourse rather than a static relic of the past. Through an interactive interface, users can trace the evolution of ethical concepts across languages, ensuring ancient wisdom remains relevant for modern

discourse. A preliminary user study suggests the systems capacity to enhance conceptual understanding and cross-cultural learning. By linking algorithmic representation with ethical inquiry, this research exemplifies how AI can serve as a methodological bridge, accommodating the ambiguity of cultural heritage rather than reducing it to static data. The Source code and data are released at <https://github.com/ThuDoMinh1102/confucian-texts-knowledge-graph>.

Keywords: Digital humanities, Natural language processing, Knowledge graph, Confucian philosophy, AI interpretability, Cultural heritage

1 Introduction

The Four Books (¹), including The Great Learning (), The Doctrine of the Mean (), The Analects of Confucius (), and The Works of Mencius (), occupy a central place in East Asian intellectual and moral history. As the foundation of Confucian philosophy, these texts have shaped education, politics, and ethics across China, Vietnam, Korea, and Japan for over two millennia, while embodying enduring ideals of virtue and social harmony. Among many commentarial traditions surrounding The Four Books, Chinese-Vietnamese *Commentaries on The Four Books* () (Tuan 2017), a widely recognized luminary in East Asian philosophy, is notable for its pedagogical clarity and bilingual structure. A scholar of Eastern philosophy, Ly Minh Tuan structured this work to integrate Classical Chinese text, transliteration, Vietnamese translation, and commentary, making the sages thought accessible to modern readers. The work presents Classical Chinese texts with modern Vietnamese translations and notes, and its introduction underscores the continued relevance of Confucian virtues such as benevolence and altruism in modern life (Tuan 2017). This commentary thus helps bridge ancient Confucian ethics and contemporary moral concerns.

Despite their enduring influence, computational research on The Four Books and related commentaries remains scarce. Traditional digitization projects focus on text preservation and retrieval, rarely modeling the dynamic interpretive layers found in annotated works where commentary, translation, and source text interrelate. These gaps are exacerbated by broader issues in digital humanities and cultural heritage preservation. AI-assisted translation introduces conceptual asymmetries; mapping terms like *ren* (, benevolence) or *li* (, ritual propriety) into modern languages risks diluting universal ethical ideals into culturally specific, localized interpretations. Privileging a single translation or commentary tradition in AI models can inadvertently amplify specific localized perspectives while marginalizing others, raising critical questions of interpretive authority and representational bias (Zhu et al. 2024).

Recent advances in AI-driven text analysis and knowledge graphs (KGs) have expanded how large cultural corpora can be organized and explored, supporting access and relational interpretation in digital humanities research (de Jong 2009; Ferro et al. 2025; Haslhofer et al. 2019). While general-purpose infrastructures offer broad coverage, domain-specific cultural heritage graphs better capture historical and conceptual

¹https://en.wikipedia.org/wiki/Four_Books_and_Five_Classics

complexity, yet scholarship emphasizes that such systems are sociotechnical constructs whose representational choices shape interpretive authority (Suchanek et al. 2024; Vrandečić and Krötzsch 2014; Barzaghi et al. 2025; Bai and Hou 2023; Drucker 2020; D’Ignazio and Klein 2020; Liu 2012). Recent work highlights pluralistic graph-based models as a response to these concerns, but Confucian classics such as The Four Books remain challenging due to linguistic concision, polysemy, and dense commentary traditions that resist stable or discrete computational representation (Yuan et al. 2025; Foka et al. 2025).

To address this, we propose Graphilosophy, an ontology-guided, multi-layered KG framework for modeling The Four Books and their commentaries. Graphilosophy functions as an interpretive infrastructure that makes relationships among texts, translations, commentaries, speakers, and concepts explicit and navigable. Our system transforms *Commentaries on The Four Books* (Tuan 2017), which integrates the original Classical Chinese with a modern Vietnamese translation and pedagogically oriented commentar, into a structured, machine-readable dataset that supports semantic search, philosophical reasoning, and educational applications. We construct a multi-layered KG representation to model the intertextual relationships between doctrine and interpretation, an essential foundation for semantic understanding of Eastern philosophy. By externalizing interpretive structures instead of concealing them within opaque models, the system supports plural readings and reflexive engagement.

Our system addresses representational bias through a scalable and explicitly pluralistic design that treats knowledge modeling as an evolving, interpretive process rather than a fixed technical structure. Central to Graphilosophy is a modular KG that supports expansion across linguistic, interpretive, and philosophical dimensions, allowing multiple translations and expert commentaries to coexist and reducing linguistic bias and singular interpretive authority. Its layered and extensible ontology separates textual, linguistic, conceptual, and commentary dimensions so each can evolve independently while remaining connected, accommodating ambiguity and multiplicity as core features of Confucian philosophy. This design addresses concerns that reductive representational models flatten philosophical nuance and reproduce power imbalances (Drucker 2017; D’Ignazio and Klein 2020; Drucker 2020), aligning technical scalability with interpretive values central to digital humanities practice.

This paper addresses two interrelated questions concerning the design and implications of AI-mediated knowledge representations for classical philosophical texts. First, how can an ontology-guided, multi-layered knowledge graph framework support the representation of the semantic, interpretive, and translational plurality inherent in The Four Books, while making visible the asymmetries of language, conceptual framing, and interpretive authority embedded in their transmission? Second, to what extent can such a system meaningfully support philosophical learning and cultural preservation without reducing the openness, ambiguity, and historical situatedness of the original texts, and what challenges arise when attempting to align scalability and bias mitigation with the ethical and interpretive demands of this domain?

Our contributions are as follows:

- We assemble a digitally annotated corpus of The Four Books that integrates Classical Chinese texts, Vietnamese translations, and pedagogical commentaries,

making visible the layered and mediated nature of meaning across languages and interpretations.

- We propose a domain-specific KG that models textual, conceptual, and commentary relationships as interconnected layers, supporting interpretive plurality rather than fixed semantic closure.
- We develop exploratory and visual tools that enable navigation across these layers, supporting semantic exploration, teaching, and interpretive inquiry in digital humanities contexts.
- Through experiments and a user study, we show how AI-based representations can assist learning and research while preserving the central role of human interpretation in engaging with Confucian philosophical texts.
- The Source code and data are released at <https://github.com/ThuDoMinh1102/confucian-texts-knowledge-graph>.

2 Related Work

Digital humanities initiatives increasingly rely on computational approaches to preserve cultural heritage, utilizing large-scale platforms (Europeana², the Perseus Digital Library³, the Chinese Text Project⁴) that provide structured, multilingual access to historical materials. While these efforts effectively treat humanities texts as data to broaden access, they primarily emphasize digitization, search, and metadata organization. Consequently, they offer limited means to engage with the profound semantic, philosophical, and linguistic complexity that shapes classical works and their traditions of interpretation.

Similarly, Natural Language Processing (NLP) has become essential for engaging with premodern corpora (Haslhofer et al. 2019), adapting techniques to handle the interpretive challenges of archaic texts (Johnson et al. 2021) and utilizing multilingual models to enable large-scale cross-lingual alignment (Zhang et al. 2023). However, these computational developments remain heavily focused on surface-level linguistic processing and information retrieval. They provide limited support for modeling the rich intertextual and interpretive relationships through which historical and philosophical meaning is actually produced.

To capture this relational meaning, Knowledge Graphs (KGs) are increasingly adopted to structure cultural knowledge (Zheng et al. 2024; Cui et al. 2024). Yet, general-purpose KGs often privilege bibliographic structure over interpretive depth (Kokash et al. 2024). From a digital humanities and sociotechnical perspective, this emphasis risks treating cultural texts as stable, objective data points rather than dynamic sites of ongoing interpretation. For Confucian classics, linguistic concision and dense commentarial traditions expose fundamental tensions between current AI representations and the interpretive complexity of the domain.

In response, our study advances a domain-specific, ontology-guided KG that reflects the layered, dialogical nature of Confucian philosophy. By situating multilingual NLP within this framework, we treat language technologies not as neutral ends

²<https://www.europeana.eu/>

³<https://www.perseus.tufts.edu/hopper/>

⁴<https://ctext.org/>

in themselves, but as mediating infrastructures. Utilizing recent advances in graph analysis and multilingual semantic representation, this approach foregrounds relational meaning, interpretive plurality, and the ethical responsibilities of AI-mediated knowledge representation.

3 Proposed Dataset

To enable computational analysis, we construct The Four Books as a Classical ChineseVietnamese resource organized at three levels: (1) tri-parallel alignment incorporating phonetic bridging between the two languages, (2) dictionary-level lexical mapping for precise semantic correspondence, and (3) chapter-based exegesis capturing interpretive commentary and contextual meaning. Together, these components form an extended corpus that encodes the semantic, phonetic, and interpretive dimensions of Classical Chinese and Vietnamese, reflecting the linguistic depth and contextual sophistication of each chapter.

3.1 Dataset Construction

3.1.1 Data Source

The dataset consists of authoritative digital editions of The Four Books, including the original Classical Chinese texts and standard commentaries. We use Commentaries on The Four Books () (Tuan 2017) as the primary source for ChineseVietnamese translations and exegetical notes, chosen for its comprehensive coverage, reliable bilingual annotation, and pedagogical depth. All materials were drawn from open-access repositories and cross-checked against printed editions to ensure textual accuracy.

The dataset is structured around three interrelated components, Main Text, Dictionary, and Expert Analysis, which together reflect the layered and mediated nature of meaning in Confucian traditions. The *Main Text* consists of 2,222 sentences from The Four Books, organized according to their original textual hierarchy and presented in Classical Chinese alongside Sino-Vietnamese readings and modern Vietnamese translations, thereby preserving both linguistic structure and translational mediation. This component serves as the interpretive foundation of the dataset, enabling close engagement with source passages while situating them within broader pedagogical and semantic contexts. The *Dictionary* includes 5,344 entries of Classical Chinese characters that were consolidated into 2,788 unique entries to account for polysemy and contextual variation, treating lexical ambiguity as an interpretive condition rather than a defect to be eliminated. The *Expert Analysis* comprises 80 commentary entries authored by Ly Minh Tuan, providing pedagogical and interpretive perspectives that situate the texts within established traditions of explanation and learning. The separation of these components is an intentional design choice that supports transparency and interpretive plurality, allowing textual content, lexical interpretation, and commentary to evolve independently while remaining connected to the source texts, and framing computational structure as a mediating practice through which cultural knowledge is represented and transmitted.

3.1.2 Construction Pipeline

The dataset construction followed a staged workflow that treats corpus preparation as an interpretive practice. It involved Preprocessing, Alignment, Refinement, and Structuring, with attention to preserving textual integrity and contextual meaning. During *Preprocessing*, scanned sources were transformed into structured digital text, corrected, and normalized to remove layout artifacts using PaddleOCR⁵, segmented according to the original textual logic, and enriched with contextual information such as provenance and authorship. Lexical materials were extracted and contextualized using metadata (book and chapter) to reflect the role of vocabulary and polysemy in shaping meaning.

The *Strategy Alignment* explicitly differentiated among three corpus types: the Tri-Parallel Corpus, which aligns Classical Chinese passages with Sino-Vietnamese readings and modern Vietnamese translations; the Lexical Corpus, which connects dictionary entries to their textual contexts to address polysemy and usage; and the Exegesis Corpus, which links expert commentaries to corresponding canonical passages. Alignment combined automated procedures with expert review, underscoring that relational mapping across languages and interpretations is an inherently interpretive process.

Following automated processing, the *Refinement* stage treats correction and verification as acts of interpretive responsibility rather than purely technical cleanup. First, heuristic recovery was used to identify and correct digitization artifacts that could distort meaning or disrupt textual continuity. This was followed by a close manual audit carried out by domain experts, who reviewed all aligned textual units to ensure coherence and fidelity across the Classical Chinese, Sino-Vietnamese, and modern Vietnamese layers. Rather than prioritizing mechanical alignment alone, this process foregrounds accountability, transparency, and human judgment in the mediation of cultural texts, establishing a reliable and ethically grounded foundation for the subsequent multi-layered knowledge representation.

In the *Structuring* stage, all materials were organized into interoperable formats with consistent identifiers, enabling the corpus to function as an integrated whole. This structure supports analysis, teaching, and exploration while maintaining transparency in how texts, translations, and interpretations are connected.

3.2 Dataset Description

The finalized dataset comprises 2,222 segmented sentences, 80 scholarly commentary annotations, and a refined lexical dictionary of 2,788 entries, including 2,562 unique Classical Chinese characters and 23 core Confucian concepts. These materials are instantiated in an ontology-driven KG containing 16,468 nodes and 71,249 edges spanning textual, linguistic, and interpretive layers.

Rather than serving as a purely technical resource, this structure is designed to externalize how meaning is mediated through translation, commentary, and conceptual categorization in classical philosophy. The dataset is organized into three complementary components, including the *Tri-Parallel Corpus*, the *Lexical Dictionary*,

⁵<https://www.paddleocr.ai/>

and the *Exegesis Corpus*, each preserving a distinct dimension of interpretive plurality: translational mediation, lexical polysemy, and scholarly commentary, respectively. Together, they form an interoperable cultural dataset that links text, translation, lexicon, and expert interpretation through a unified ontology, enabling AI systems to *mediate*, rather than resolve, the complexity of classical philosophical knowledge.

4 Methodology

4.1 Overview

Our proposed Graphilosophy is a comprehensive, multi-stage framework that integrates advanced text processing, layered KG construction, and an interactive web interface powered by Google Gemini (Google 2023) for natural language querying. Our KG was built using the NetworkX library (Hagberg et al. 2008) in a modular manner. Rather than generating a monolithic graph, our system constructs six distinct yet interlinked layers, such as Textual, Linguistic, Conceptual, Commentary, Speaker, and Semantic (Figure 1). Each layer is processed independently to preserve its unique analytical function and data integrity before being interconnected through ontology-guided relationships. This modular design maintains interpretive clarity within each dimension while enabling a unified, multi-perspective understanding of The Four Books and their accompanying commentaries.

4.2 Ontology Design and Construction

The Graphilosophy framework employs a custom multi-layered ontology specifically designed to model the bilingual Classical Chinese–Vietnamese corpus of *The Four Books*. The ontology systematically structures textual, linguistic, conceptual, and interpretive dimensions across six interconnected layers (Meta, Textual, Linguistic, Conceptual, Commentary & Speaker, and Semantic), comprising 20 entity classes and 18 directed relationship types. This design supports multi-hop reasoning, cross-lingual retrieval, and interpretive plurality while maintaining modularity and scalability.

Figure 1 illustrates the ontology architecture, depicting the sequential flow from the Meta Layer through the Textual, Linguistic, Conceptual, and Commentary & Speaker Layers to the Semantic Layer, together with explicit cross-layer connections that unify the entire knowledge graph.

Table 1 provides a comprehensive summary of all entity classes and relationship types, grouped by layer.

Relations are generated through three distinct methods (fully automatic rule-based, semi-automatic with human validation, and fully manual expert-defined) and unified into a single directed graph via explicit cross-layer links (e.g., `HAS_HAN_FORM`, `CONTEXTUALIZES`, `EXPRESSES_CONCEPT`). This modular, ontology-guided structure externalizes interpretive processes, preserves semantic ambiguity inherent in Confucian philosophy, and provides a robust foundation for semantic search, philosophical reasoning, and educational exploration.

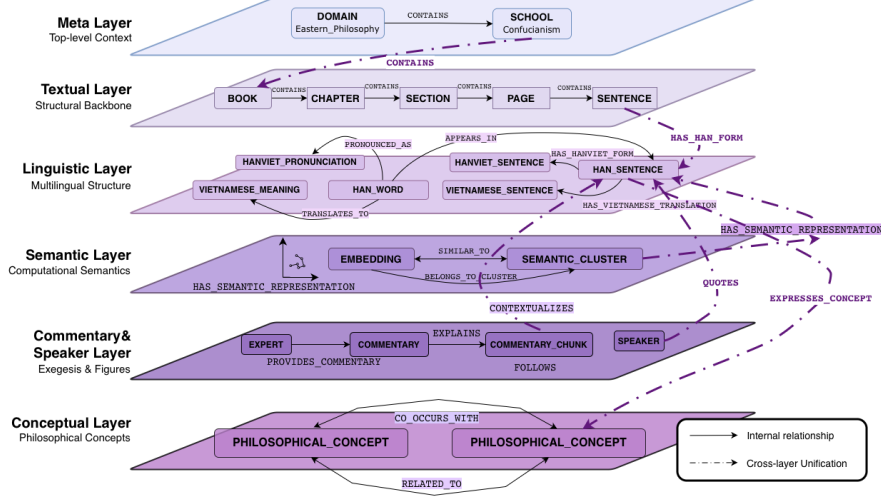


Fig. 1: The multi-layered ontology architecture of the Graphilosophy knowledge graph. The schema models the corpus across six distinct but interconnected layers to preserve structural, linguistic, and interpretive dimensions. Solid lines indicate intra-layer relationships, while dashed lines represent cross-layer unifications that enable complex, multi-hop reasoning.

4.3 Linguistic Processing

Classical Chinese presents unique computational challenges due to its brevity, polysemy, and context-dependence.

Polysemy Resolution via Contextual Embedding

When a Classical Chinese character has multiple possible meanings, the system adopts a context-sensitive matching approach, using Multilingual-e5-large embeddings and cosine similarity, rather than enforcing a single dictionary definition. Meanings are evaluated in relation to the surrounding passage to support interpretive coherence, without treating the result as definitive. This design reflects digital humanities commitments to preserving semantic ambiguity and aligns with concerns about limiting algorithmic authority in the interpretation of philosophical texts. Example ((o)): This character has three primary meanings in our dictionary: path/road, to speak/say, and doctrine/The Way. In the passage :! (Analects 4.15), the system interprets in relation to the surrounding philosophical discourse, foregrounding its doctrinal sense based on the philosophical context of Confucius addressing his disciple about his unifying principle.

Segmentation in Context

While Classical Chinese lacks punctuation, the system uses Ly Minh Tuans scholarly edition (Tuan 2017) as an interpretive reference rather than a neutral ground

Table 1: Entity classes and relationship types in the Graphilosophy KG ontology

Layer	Type	Name	Method
Meta	Entity	DOMAIN, SCHOOL	Predefined
Textual	Entity	BOOK, CHAPTER, SECTION, PAGE, SENTENCE	Auto
Textual	Relation	CONTAINS, FOLLOWS, APPEARS_IN	Auto
Linguistic	Entity	HAN_SENTENCE, HANVIET_SENTENCE, VIETNAMESE_SENTENCE	Auto
Linguistic	Entity	HAN_WORD, HANVIET_PRONUNCIATION, VIETNAMESE_MEANING	Auto
Linguistic	Relation	HAS_HAN_FORM, HAS_HANVIET_FORM, HAS_VIETNAMESE_TRANSLATION, TRANSLATES_TO, PRONOUNCED_AS	Auto
Conceptual	Entity	PHILOSOPHICAL_CONCEPT (Pattern matching)	Auto
Conceptual	Relation	EXPRESSES_CONCEPT, RELATED_TO, CO_OCCURS_WITH	Auto + semi-manual
Commentary & Speaker	Entity	EXPERT, COMMENTARY, COMMENTARY_CHUNK, SPEAKER	Manual / Pattern detection
Commentary & Speaker	Relation	PROVIDES_COMMENTARY, EXPLAINS, CONTEXTUALIZES, QUOTES	Manual + similarity > 0.75
Semantic	Entity	EMBEDDING, SEMANTIC_CLUSTER	Auto
Semantic	Relation	SIMILAR_TO, BELONGS_TO_CLUSTER, HAS_SEMANTIC_REP	Auto

Note: Total of 20 entity classes and 18 relationship types.

Method: Auto = fully automatic (rule-based); Semi = algorithm + human verification; Manual = expert-defined.

truth, and validates segment boundaries through cross-layer consistency with Sino-Vietnamese and modern Vietnamese translations. Ambiguous cases are resolved by favoring readings supported by both translation alignment and established commentary, while explicitly acknowledging the role of scholarly judgment. This design aligns

with digital humanities principles of interpretive transparency and with concerns about making human assumptions visible in computational text processing.

Handling Untranslatable Concepts

For philosophical concepts that cannot be fully captured in modern Vietnamese, the system preserves the original Classical Chinese term, links it to a broader conceptual category, and supplements it with expert commentary. For example, (*Nhón/Ren*) is represented through multiple Vietnamese approximations and contextualized as a core virtue through scholarly explanation, rather than reduced to a single translation. This layered representation treats untranslatability as an interpretive feature, aligning with digital humanities emphases on semantic plurality and concerns about avoiding reductive representations of ethical concepts.

Success and Failure Cases

Success Cases:

- *vs. Disambiguation*: Despite identical Sino-Vietnamese pronunciation “*Nhón*”, the system correctly distinguished (human/person) from (benevolence/virtue) in 98% of cases (2,178/2,222 sentences). The E5-Large embeddings position these characters in distinct semantic regions: clusters with pronouns and social roles, while clusters with virtue terms.
- *Cross-lingual Concept Retrieval*: A Vietnamese query for “*o hiu*” (filial piety) successfully retrieved 47 Classical Chinese sentences containing (*hiu*), even when the word “*o*” was absent from the original text, demonstrating effective semantic bridging.
- *Speaker Attribution*: The pattern-based detection correctly attributed 89% of quotations to speakers (Confucius: “”; Mencius: “”; disciples: “,” “”).

Failure Cases:

- *Phonetic Ambiguity*: The character can be read as *Lc* (joy/pleasure) or *Nhc* (music). In sentences like “,:,” (Analects 6.20), the system occasionally confuses these readings when sentence structure is sparse. This phonetic ambiguity remains a challenging case where embedding-based semantic disambiguation alone may be insufficient without additional phonetic or syntactic cues.
- *Implicit Subject Recovery*: Classical Chinese frequently omits subjects. In “,” (Analects 1.1), the implicit subject “one who learns” is not explicitly represented, limiting certain speaker-attribution queries.

These failure cases are not isolated edge cases but structurally predictable outcomes of embedding-based disambiguation applied to a language where phonetic and semantic information are not always recoverable from context alone; their systematic treatment is discussed in Section 5.4.

4.4 Semantic Chunking

This workflow segments classical texts and commentaries based on philosophical continuity rather than fixed boundaries, treating segmentation as an interpretive

intervention to preserve contextual integrity and navigate Classical Chinese polysemy. Linguistically, it integrates Classical Chinese, Sino-Vietnamese, and modern Vietnamese via a consolidated lexical resource that accommodates character-level ambiguity. By identifying core Confucian concepts through established taxonomies and contextual cues, the system explicitly supports multiple interpretations. Prioritizing this interpretive plurality over strict technical optimization limits algorithmic authority, ensuring responsible computational mediation of historical texts.

4.5 Interactive System Interface

The system provides an interactive platform for exploring, visualizing, and querying the KG via Google Gemini (Figure 2), which interprets natural-language queries and orchestrates a hybrid search mechanism combining structured graph traversal with exact textual matching. Rather than generating responses in isolation, Gemini operates over explicitly defined KG context: natural-language queries are resolved into localized graph neighborhoods to maintain interpretive clarity, while verbatim inputs trigger direct text matching to ensure philological accuracy. This design promotes transparent and verifiable interaction with classical texts, supporting close reading, comparative analysis, and pedagogical use.

5 Structural Evaluation

5.1 Core Graph Metrics and Sparsity

This unified KG is a highly complex and structured data system, comprising 16,468 nodes, which are interconnected by 71,249 edges representing meaningful relationships between them. Despite the large number of nodes and edges, the network exhibits an extremely low density of 0.000263, classifying it as a sparse network. This sparsity demonstrates that connections are selectively generated to represent meaningful, non-trivial semantic relationships, making the graph efficient for targeted queries.

5.2 Knowledge Graph Structural Validation

The internal consistency of the KG serves as a secondary validation of the segmentation quality. The graph currently hosts 16,468 nodes and 71,249 edges. A key metric is the `APPEARS_IN` relationship, which constitutes 41.3% of all edges (29,417 instances). This high density of connections between the 1,723 unique Classical Chinese words and the 2,222 segmented sentences confirms that the segmentation logic accurately mapped fine-grained linguistic units into their correct structural contexts.

5.3 Distribution and Layer Analysis

Initial accuracy tests showed strong linguistic results, with segmentation achieving approximately 90% accuracy in annotated samples. Entity recognition was strong for names and places but weaker for abstract terms. Structural analysis confirms the research’s focus on deep linguistic and semantic modeling (Figures 3, 4, and 5).

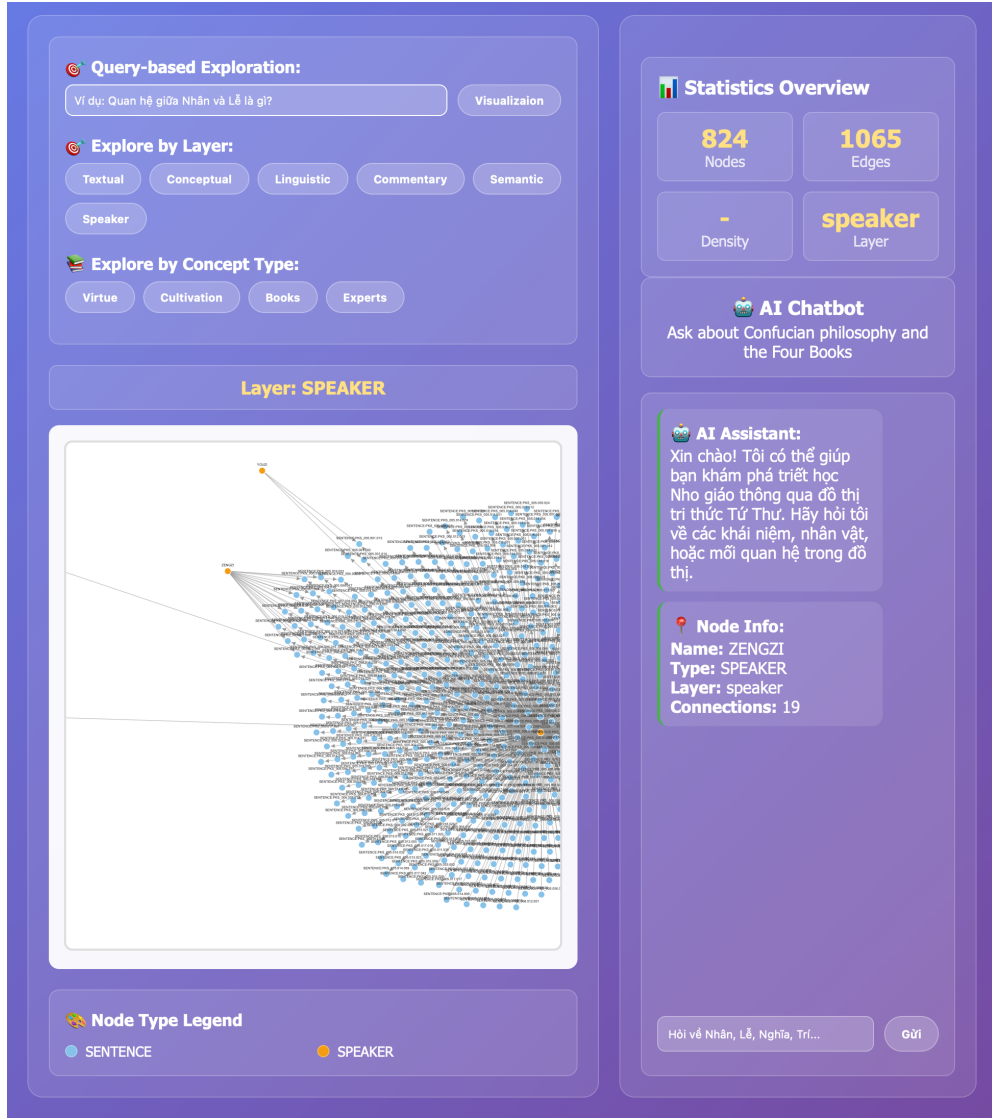


Fig. 2: User interface of the proposed system integrating layered KG visualization, semantic search, and commentary exploration. The interface enables cross-lingual retrieval and interpretive analysis of The Four Books through Gemini-powered natural language querying. Some instructions in the interface are displayed in Vietnamese to enhance usability for local users.

The KG exhibits a clear emphasis on linguistic and textual information, with nodes related to these forms constituting nearly 70% of all nodes, and the significant presence of EMBEDDING nodes (13.9%) further highlighting its reliance on a semantic layer for advanced information retrieval. This linguistic foundation is reinforced by

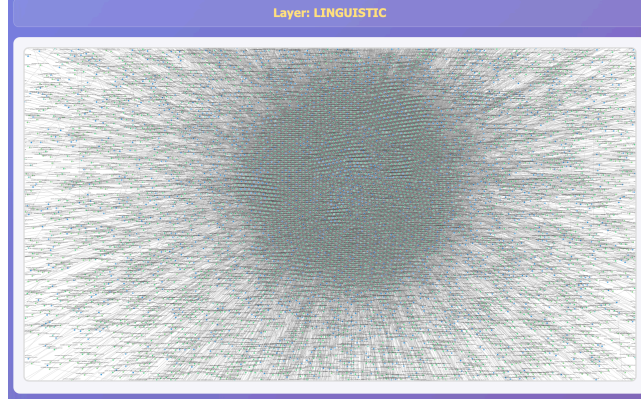


Fig. 3: The high density of the Linguistic Layer reflects the substantial volume of nodes and edges within this layer, establishing a robust foundation for subsequent semantic analysis.

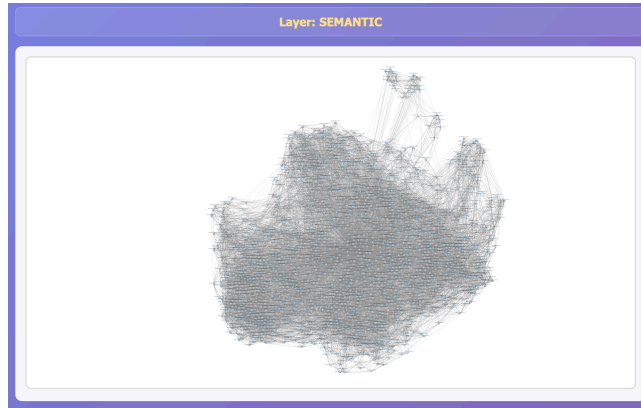
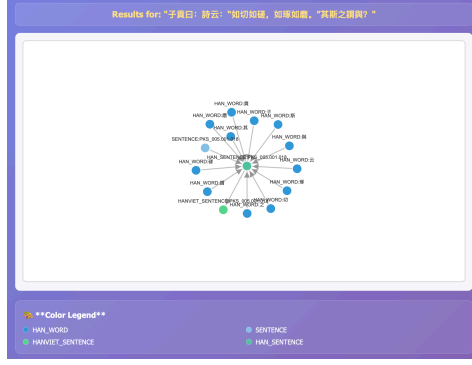


Fig. 4: Semantic Layer structure. Visualization of the dense network formed by EMBEDDING nodes and SEMANTIC_CLUSTERS, confirming the reliance on the semantic layer for similarity-based retrieval.

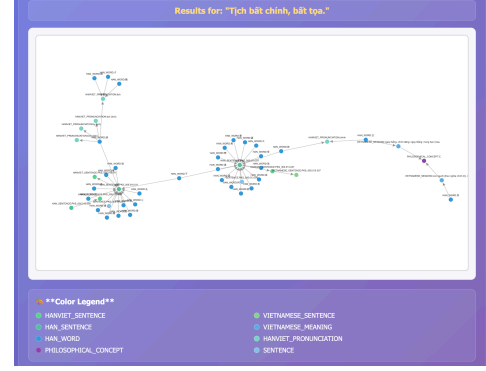
the edge distribution, where the APPEARS_IN relationship dominates (41.3%), confirming robust connections within the textual structure, while the crucial inclusion of HAS_SEMANTIC_REPRESENTATION and BELONGS_TO_CLUSTER relationships among the top relations validates the strong connectivity between textual units and their abstract semantic meanings, enabling sophisticated understanding and organization of information.

5.4 Error Analysis

From the validation in Section 5.3, remaining 10% of errors typically stem from:



(a) Focused subgraph for exact Classical Chinese query: “...” (Analects 1.4). The central HAN_SENTENCE node connects to constituent HAN_WORD nodes, with trilingual alignment to HANVIET_SENTENCE and SENTENCE nodes.



(b) Multi-cluster subgraph for Vietnamese semantic query: “Tch bt chờnh, bt ta.” Query retrieves multiple related passages, revealing cross-layer connections including PHILOSOPHICAL_CONCEPT nodes (magenta) and VIETNAMESE_MEANING nodes (light blue).

Fig. 5: Query-based focused visualization illustrating the BFS-based (depth = 1) search mechanism through Gemini model. Unlike full-layer visualizations (Figures 3, 4) which display the entire graph structure, these focused subgraphs present only the immediate neighborhood of query-matched nodes, reducing visual complexity while preserving interpretive context. (a) An exact single-sentence query produces a star-shaped subgraph centered on the matched passage. (b) A semantic query over Vietnamese text retrieves multiple thematically related passages, forming distinct clusters connected through shared linguistic and conceptual nodes.

- Complex intertextuality: Passages where the expert’s commentary and ancient quotes are deeply interwoven without explicit markers, occasionally causing the semantic coherence score to trigger a false boundary.
- Phonetic ambiguity (e.g., ; see Section 4.3) cannot be resolved by expanding the contextual window, implying that error reduction in this category requires architectural rather than parametric changes.

These cases are addressed through cross-layer connections: the Commentary Layer provides supplementary context that can disambiguate the Linguistic Layer when embedding-based methods fail, effectively distributing the interpretive burden across the graph rather than concentrating it in a single processing step.

5.5 Density Interpretation and Cross-Layer Connectivity

The comparison of individual layer densities offers crucial insights into the graph’s architecture: the significantly higher density observed in the Commentary (density: 0.004149) and Conceptual (density: 0.001368) layers (Figures 6 and 7) is structurally sound, indicating that scholarly notes and philosophical concepts form tighter, more

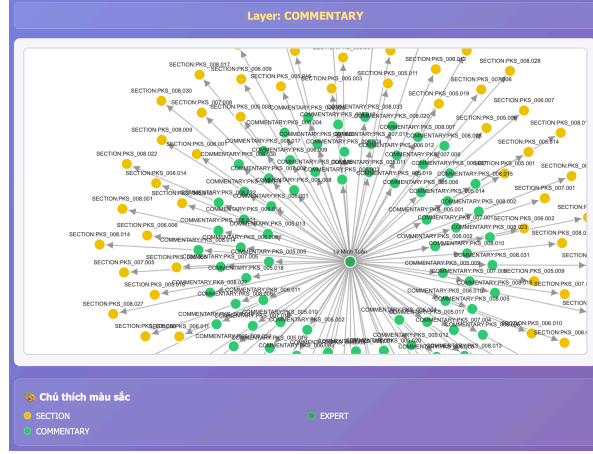


Fig. 6: Commentary Layer structure. The central EXPERT node (Ly Minh Tuan) connects directly to COMMENTARY nodes, illustrating the highly centralized and dense structure of the interpretive layer (Density: 0.004149).

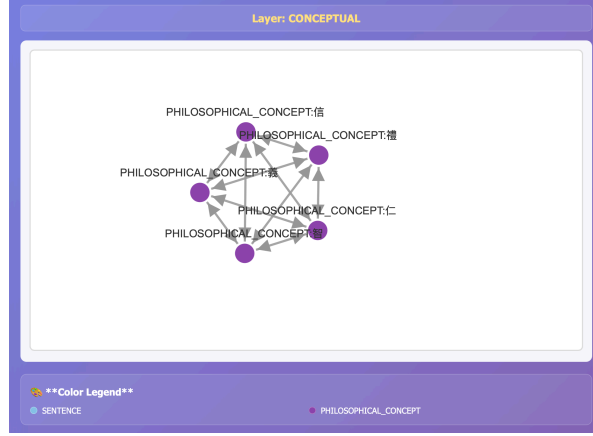


Fig. 7: Structural analysis of Conceptual Layer (Density: 0.001368). The visualization demonstrates that the Philosophical Concept nodes, representing core virtues (信, 禮, 義, 仁, 智), form tightly integrated clusters. This high degree of co-occurrence and relational density within the Conceptual Layer is structurally sound, indicating that these philosophical concepts form integrated networks essential for effective multi-hop reasoning and comparative analysis.

integrated clusters essential for effective multi-hop reasoning. Furthermore, the 12.6% of total edges dedicated to cross-layer connections serve as a critical bridge, facilitating sophisticated, multi-hop queries that traverse distinct information domains, such as linking a specific sentence to a broader philosophical concept or an identified speaker, thereby validating the complex structural design outlined in the methodology.

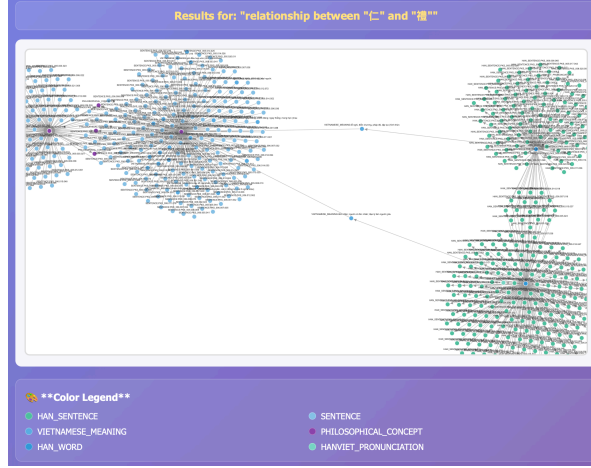


Fig. 8: Complex concept query result. The result for "relationship between 仁 and 禮" illustrates the initial subgraph expansion, showing two large clusters connected by Linguistic and Semantic nodes.

5.6 Conceptual Tracing and Intertextuality

Within this KG, entities such as Confucius, Mencius, and their disciples were meticulously extracted, alongside key concepts like 仁 (benevolence) and 禮 (ritual propriety) (Figure 8). The application of embedding-based similarity proved instrumental in tracing complex overarching themes, including governance, moral cultivation, and education, thereby moving retrieval capabilities beyond mere keyword matching towards genuine thematic discovery. For instance, these embeddings revealed clear and significant parallels between passages found in *The Mencius* and *The Analects* concerning the principles of benevolent leadership. Furthermore, the constructed networks effectively unveiled intricate relationships such as teacher-disciple links, nuanced commentary interpretations, and direct quotations that span across different books within the collection.

6 Preliminary User Evaluation

We conducted a pilot study to explore the pedagogical potential and scholarly utility of the system.

6.1 Study Protocol and Participant Recruitment

We recruited six participants (three males, three females, aged from 22 to 30) from a local university via emails and snowball sampling. The participants were graduate students, specialized in Educational Science and had prior academic exposure to Philosophy, allowing them to provide informed and critical feedback on both the instructional design and the system’s scholarly depth. At their arrival at the experiment environment, the participants were asked to fill a consent form. Then each

participant used the system to complete a set of targeted analytical tasks under ethical guidelines of the hosting institution. The set of tasks the participants had to complete includes:

- **Concept Tracing:** Tracking the evolution of Ren (- Benevolence) from The Analects to The Works of Mencius.
- **Comparative Analysis:** Comparing how different speakers discuss the virtue of Li (- Ritual/Propriety) using the Speaker and Semantic layers.
- **Intertextual Exploration:** Utilizing "Exact Text Search" to locate specific commentary nodes bridging multiple textual segments.

After completing the tasks, we conducted semi-structured post-study interviews to elicit participants qualitative perceptions, experiences, and feedback regarding the system. The interviews primarily focused on three aspects: learning benefits (Q1), the helpfulness of the AI components (Q2), and overall system usability (Q3). All interviews were audio-recorded to support subsequent transcription and analysis.

6.2 Knowledge Integrity and Interpretive Multiplicity

A core challenge in modeling classical texts is the ambiguity of commentaries that often apply to multiple sentences. To ensure scholarly integrity, Graphilosophy avoids arbitrary selection; instead, it implements multi-directional linking by creating concurrent edges for a single node. This decision ensures interpretive multiplicity, allowing users to cross-examine different meanings across layers. While this increases informational density and may cause "visual clutter," it is a calculated trade-off to prioritize data transparency and preserve the original text's complexity over oversimplification.

Based on this design decision, concept maps, faceted search, and passage-to-concept explanations were successfully deployed in classroom settings for the study, enabling learners to trace philosophical concepts and explore intertextual relationships more effectively (Figure 8). The case study confirmed that computational tools can clarify conceptual structures and highlight connections across The Four Books.

6.3 Results and Feedback

On average, each participant spent approximately 60 minutes in the user study, including 45 minutes completing the tasks and 15 minutes participating in the post-study interview. The audio recordings of the interviews were transcribed verbatim and analyzed using an inductive thematic analysis approach to identify major themes related to system utility and usability. Initially, one researcher developed the thematic codes through iterative, data-driven coding of the transcripts. These codes and the resulting themes were subsequently reviewed and discussed with two additional researchers to resolve discrepancies and reach consensus.

User Response Summary

Participant responses were characterized by the frequency of positive sentiment versus reported concerns (Table 2). The qualitative data gathered from the user study was analyzed and categorized into three primary themes: Scholarly Utility, Visual Complexity and AI Integration Pedagogical Support as below.

Table 2: Summary of participant response trends ($n = 6$).

Evaluation Dimension	Positive	Primary Concerns / Suggestions
Learning Benefit (Q1)	4/6	Visual clutter in complex subgraphs
AI Helpfulness (Q2)	5/6	Text formatting and lack of visual links
System Usability (Q3)	3/6	Language mix and navigation flow

- **Scholarly Utility:** A majority of participants noted that the graph-based approach made intertextual relationships visible and manageable. One user observed: "The graph-based approach makes intertextual relationships visible and manageable". The participants generally confirmed the value of concept tracing and relation visualization, noting that the system improved comprehension and supported course assignments.
- **Visual Complexity and Interaction Flow of Concept Graph:** Users reported that the interface became "messy" or "overwhelming" when dense layers were active (i.e., when many nodes and edges were displayed at once). This confirms the trade-off; while multi-directional linking preserves interpretive richness, it requires better filtering for non-experts. While many praised the aesthetics and innovative design of the interface, others recommended a more consistent flow for exploration of the interface. Some suggested simplifying the visualization by showing only the most relevant nodes, enlarging or reformatting chatbot text for readability, and integrating in-app guidance to help new users navigate the system.
- **AI Integration for Pedagogical Support:** The Gemini integration was lauded for accuracy and speed. Students highlighted that AI responses from the integrated Gemini chatbot were often clear and useful. However, it was criticized for a lack of "animation or clear visual cues" linking chatbot explanations to graph nodes, and for verbose, unformatted text blocks.

Study Conclusion

The preliminary study suggests that Graphilosophy improves accessibility to The Four Books while strengthening scholarly rigor through transparent representation and clear provenance of interpretations. The multi-layer graph structure and AI-assisted interaction show strong potential for both research and pedagogical use, particularly in clarifying conceptual organization and intertextual relationships.

At the same time, the study highlights areas for improvement, including clearer labeling, reduced visual density, and refinement of chatbot output. Limitations remain in segmentation accuracy, the handling of philosophically ambiguous terminology, and the capacity of semantic embeddings to capture nuanced meaning.

Furthermore, as a small exploratory pilot with a small sample size ($n = 6$), the findings also call for larger and more diverse studies to assess robustness, generalizability, and long-term educational impact.

7 Discussion

7.1 Societal and Ethical Dimensions

7.1.1 Interpretive Authority and Source Selection

The selection of Ly Minh Tuan’s commentary as the primary interpretive lens reflects deliberate pedagogical priorities: bilingual accessibility for Vietnamese learners and educational orientation over purely academic discourse. However, this choice inherently privileges a Vietnamese interpretive tradition, potentially marginalizing Korean scholarly interpretations and Japanese readings. Importantly, Ly Minh Tuan’s work already incorporates Zhu Xi’s *Sishu Jizhu* as a foundational reference, translating and annotating it alongside his own commentary; the system therefore mediates this neo-Confucian Chinese tradition indirectly rather than excluding it entirely. The framework’s horizontal scalability directly addresses the remaining gap: the ontology supports multiple expert nodes, allowing future work to incorporate Korean Samaejip commentaries without restructuring the core architecture.

7.1.2 Translation as Cultural Politics

Translations are never neutral (D’Ignazio and Klein 2020). Classical concepts resist perfect mapping to modern Vietnamese:

- (Ren): Translated as “*Nhón*” (benevolence), but the Vietnamese term carries Buddhist connotations absent in the original Confucian usage. Ly Minh Tuan addresses this by expanding through commentary nodes that explain it as “the totality of all virtues” rather than a single moral quality.
- (Li): Rendered as “*L*” (ritual/propriety), but the Vietnamese term emphasizes ceremonial aspects while the Classical Chinese encompasses broader social norms. The system represents this semantic gap through translation relations that preserve multiple meaning nodes rather than collapsing to a single translation.

7.1.3 Algorithmic Mediation and Possible Distortion

Graph structures prefer discrete, named relations and may struggle with ambiguity, indirect allusions, or deliberate contradictions characteristic of philosophical texts (Drucker 2017). For example, the *Analects* presents seemingly contradictory statements about that resist single-relation encoding. Our system addresses this through:

1. **Multi-hop queries:** Allowing users to traverse multiple relation paths rather than expecting single-edge answers.
2. **Commentary integration:** Expert annotations provide interpretive context that disambiguates apparent contradictions.
3. **Semantic clustering:** The similarity relation groups thematically related passages regardless of surface-level contradiction, enabling users to explore conceptual tensions.



Fig. 9: AI-generated narrative visualization from The Analects.

7.2 Pedagogical Applications

Building on the pilot study, we extend the multi-layer KG into generative storytelling by transforming philosophical passages into sequential visual narratives. Figures 9 and 10 illustrate how classical commentary bridges interpretive scholarship and AI-driven creativity. This prototype functions as a collaborative co-creation platform, assisting users with narrative composition, panel organization, and visual coherence. Ultimately, this framework lays the groundwork for AI-assisted cultural heritage storytelling, merging KGs, generative models, and interactive design to reimagine and preserve Confucian philosophy in accessible, multimodal formats.



(a) Liang’s cluttered study. Scrolls lie open everywhere, ink stains the desk, an unstrung bow rests among account books. (b) A tranquil garden. Master Chen sits beneath willow, sipping tea in stillness. (c) A small stone table holding a nearly dead bonsai with dry soil and withered branches. (d) Liang works anxiously in the garden. Soil splashes, uneven cuts, water floods the pot.



(e) Evening in the garden. The bonsai looks worse; soil soggy, leaves shriveled further. (f) Morning light. Liang gently loosens roots, replaces soil, waters lightly, and dew glimmers on bow and account places the bonsai in young leaves. (g) New shoots sprout from the bonsai’s branches. Morning is open, brush ready, books neatly stored away. (h) Liang’s study is tidy. A single scroll

Fig. 10: Qualitative results of the Philosophy-Unfolded The Great Learning (/ i Hc) system. Eight visual narratives depict sequential moral progressions derived from the canonical text and commentary through multimodal generative interpretation.

7.3 Limitations

While Graphilosophy demonstrates the potential of integrating NLP and knowledge graphs for classical text analysis, several limitations warrant acknowledgment.

Sample Size and Generalizability

The preliminary user evaluation involved only six participants from a single institution, limiting the generalizability of user feedback. All participants were graduate students in Educational Science with prior exposure to philosophy, which may not reflect the broader target audience of non-expert learners. Future work should include larger, more diverse cohorts across multiple educational contexts and cultural backgrounds to validate the system’s pedagogical effectiveness.

Table 3: Retrieval performance comparison between BM25, Semantic, and Hybrid approach.

Metric	BM25	Semantic (E5)	Hybrid	Best Performer
P@1	0.773	1.000	1.000	Hybrid/Semantic
P@3	0.652	1.000	1.000	Hybrid/Semantic
P@5	0.564	1.000	1.000	Hybrid/Semantic
P@10	0.505	1.000	0.995	Semantic
MRR	0.773	1.000	1.000	Hybrid/Semantic
NDCG@5	0.770	1.000	1.000	Hybrid/Semantic
NDCG@10	0.766	1.000	1.000	Semantic

The scope of interpretive authority is further constrained by the dataset’s reliance on a single scholarly edition; the implications of this choice, and the architectural provisions for expanding it, are discussed in Section 7.1.1.

Disambiguation Accuracy

The system achieves varying accuracy across linguistic challenges, including homophone disambiguation (vs.), phonetic ambiguity (as Lc/Nhc), and speaker attribution. Characters with sparse contextual cues or multiple valid readings in philosophical contexts remain challenging for embedding-based disambiguation.

This limitation is most acute for characters whose ambiguity is phonetic rather than semantic, a distinction elaborated in Sections 4.3 and 5.4.

Implicit Linguistic Features

Beyond phonetic ambiguity, the system does not currently attempt ellipsis recovery or anaphora resolution, a structural limitation of Classical Chinese that affects speaker-attribution queries more broadly (see Section 5.4 for concrete instances). Addressing this would require syntactic augmentation beyond the embedding-based approach adopted here.

Evaluation Methodology

The retrieval evaluation (Table 3) used a synthetic test corpus containing Confucian concepts and unrelated modern topics as true negatives. While this demonstrates discriminative power, it does not reflect the nuanced relevance judgments required for philosophical inquiry. More rigorous evaluation with expert-annotated relevance judgments from domain scholars would strengthen validity claims. Additionally, the perfect precision scores ($P@1 = 1.0$) may reflect the controlled nature of the test set rather than real-world retrieval performance.

Scalability Validation

Although the framework is designed for horizontal and vertical scalability, we have not yet validated performance with substantially larger corpora. The current graph of 16,468 nodes and 71,249 edges represents a single commentary tradition on four

books. Extending to multiple commentary traditions or additional philosophical texts may introduce computational and ontological challenges not yet encountered.

Visual Complexity Trade-off

User feedback consistently identified visual clutter as a usability concern, particularly when dense layers are active. The current implementation prioritizes interpretive multiplicity over visual simplicity a deliberate design choice that may limit accessibility for non-expert users. Future iterations should explore progressive disclosure mechanisms or adaptive filtering to balance scholarly completeness with user-friendly visualization.

8 Conclusion

This paper demonstrates how NLP and KG construction can deepen engagement with classical philosophical corpora in *The Four Books*, a domain where linguistic concision, interpretive plurality, and centuries of commentary tradition have resisted stable computational representation. By curating a trilingual corpus, designing a six-layer ontology, and anchoring retrieval in semantic embeddings, Graphilosophy advances concept tracing, intertextual analysis, and AI-assisted pedagogy while keeping interpretive authority visible and contestable.

The two research questions framing this study asked whether a multi-layered KG can represent interpretive plurality without reducing it, and whether such representation can support learning without foreclosing the openness of the original texts. The evidence presented here suggests that both questions admit qualified affirmative answers, and that the qualification in each case points toward the same underlying tension. Graphilosophy’s architecture preserves translational asymmetry, distributes commentary authority, and encodes polysemy as a navigable feature rather than an error to be corrected; yet the very density that enables this fidelity is precisely what makes the system cognitively demanding for non-expert users. This friction is not an implementation failure but a structural one: fidelity to philosophical complexity and accessibility for broad audiences are not straightforwardly compatible design goals.

Our preliminary evaluation indicates that although maintaining interpretive multiplicity leads to greater visual complexity, it substantially enhances transparency and reinforces the academic rigor of digital hermeneutic analysis. Future work will focus on expanding evaluation metrics, and extending the framework to additional Confucian and East Asian philosophical texts. Resolving the tension between interpretive completeness and pedagogical accessibility, through adaptive filtering, progressive disclosure, or differentiated user pathways, remains the central design challenge ahead, and perhaps the central methodological question for AI-mediated engagement with classical philosophical heritage more broadly.

Acknowledgments

This research is funded by Vietnam National University - Ho Chi Minh City (VNU-HCM) under Grant Number B2026-18-17.

References

- Bai B, Hou W (2023) The application of knowledge graphs in the chinese cultural field: the ancient capital culture of beijing. *Heritage Science* 11(1):77. <https://doi.org/10.1186/s40494-023-00922-7>
- Barzaghi S, Moretti A, Heibi I, et al (2025) Chad-kg: A knowledge graph for representing cultural heritage objects and digitisation paradata. *arXiv preprint arXiv:250513276*
- Cui Y, Yao S, Wu J, et al (2024) Linking past insights with contemporary understanding: an ontological and knowledge graph approach to the transmission of ancient chinese classics. *Heritage Science* 12(1):382. <https://doi.org/10.1186/s40494-024-01504-x>
- D’Ignazio C, Klein LF (2020) *Data Feminism*. MIT Press, <https://doi.org/10.7551/mitpress/11805.001.0001>
- Drucker J (2017) Non-representational approaches to modeling interpretation in a graphical environment. *Digital Scholarship in the Humanities* 33(2):248–263. <https://doi.org/10.1093/lc/fqx034>
- Drucker J (2020) *Visualization and Interpretation: Humanistic Approaches to Display*. MIT Press, <https://doi.org/10.7551/mitpress/12523.001.0001>
- Ferro S, Giovanelli R, Leeson M, et al (2025) A novel nlp-driven approach for enriching artefact descriptions, provenance, and entities in cultural heritage. *Neural Computing and Applications* pp 1–22. <https://doi.org/10.1007/s00521-025-11449-2>
- Foka A, Griffin G, Badri S, et al (2025) Tracing the bias loop: AI, cultural heritage, and bias-mitigating in practice. *AI & Society* 40(8):5823–5834. <https://doi.org/10.1007/s00146-025-02349-z>
- Google (2023) Gemini: A family of highly capable multimodal models. Tech. rep., Google DeepMind, technical Report
- Hagberg AA, Schult DA, Swart PJ (2008) Exploring network structure, dynamics, and function using networkx. In: the 7th Python in Science Conference (SciPy), Pasadena, CA, USA, pp 11–15, <https://doi.org/10.25080/TCWV9851>
- Haslhofer B, Isaac A, Simon R (2019) Knowledge graphs in the libraries and digital humanities domain. In: *Encyclopedia of big data technologies*. Springer, p 1080–1087, https://doi.org/10.1007/978-3-319-63962-8_291-1
- Johnson KP, Burns PJ, Stewart J, et al (2021) The classical language toolkit: An nlp framework for pre-modern languages. In: the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural

- language processing: System demonstrations, pp 20–29, <https://doi.org/10.18653/v1/2021.acl-demo.3>
- de Jong FM (2009) Nlp and the humanities: the revival of an old liaison. In: 12th Conference of the European Chapter of the ACL (EACL 2009), Association for Computational Linguistics (ACL), pp 10–15
- Kokash N, Romanello M, Suyver E, et al (2024) The brill knowledge graph: A database of bibliographic references and index terms extracted from books in humanities and social sciences. *Research Data Journal for the Humanities and Social Sciences* 9(1):1–21. <https://doi.org/10.1163/24523666-bja10036>
- Liu A (2012) Where is cultural criticism in the digital humanities? In: Gold MK (ed) *Debates in the Digital Humanities*. University of Minnesota Press, p 490–509, <https://doi.org/10.5749/minnesota/9780816677948.003.0049>
- Suchanek FM, Alam M, Bonald T, et al (2024) Yago 4.5: A large and clean knowledge base with a rich taxonomy. In: the 47th international ACM SIGIR conference on research and development in information retrieval, pp 131–140, <https://doi.org/10.1145/3626772.3657876>
- Tuan LM (2017) *T Th Bõnh Gii [Chinese-Vietnamese Commentaries on the Four Books]*. Religious Publishing House
- Vrandečić D, Krötzsch M (2014) Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57(10):78–85. <https://doi.org/10.1145/2629489>
- Yuan H, Li Y, Wang B, et al (2025) Knowledge graph-based intelligent question answering system for ancient chinese costume heritage. *npj Herit Sci* 13(1):198. <https://doi.org/10.1038/s40494-025-01776-x>, received: 04 December 2024; Accepted: 07 May 2025; Published: 21 May 2025; Version of record: 21 May 2025
- Zhang X, Thakur N, Ogundepo O, et al (2023) MIRACL: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics* 11:1114–1131. https://doi.org/10.1162/tacl_a_00595
- Zheng X, Li M, Wan Z, et al (2024) Knowledge mining and graph visualization of ancient chinese scientific and technological documents bibliographic summaries based on digital humanities. *Library Hi Tech* 42(6):1693–1721. <https://doi.org/10.1108/LHT-11-2022-0538>
- Zhu L, Mou W, Lai Y, et al (2024) Language and cultural bias in AI: Comparing the performance of large language models developed in different countries on traditional chinese medicine highlights the need for localized models. *Journal of Translational Medicine* 22:319. <https://doi.org/10.1186/s12967-024-05128-4>

Appendix A Dataset

A.1 Data Source

The dataset is organized into three primary components:

- **Main Text:** This component includes 2,222 sentences representing the fundamental structural and linguistic units of The Four Books. Key indexing fields include `file_id` (book identifier), `sect_id` (chapter and section identifier), `page_id` (page number), and `sent_id` (sentence identifier), which together define the hierarchical structure of the Textual Structure Layer. The linguistic data encompass the original Classical Chinese text (C), the Sino-Vietnamese (Han-Viet) phonetic (V), and the modern Vietnamese translation (M), serving as the foundation for the Linguistic and Semantic Layers.
- **Dictionary:** This component contains 5,344 entries of Classical Chinese characters, including the Chinese form, Sino-Vietnamese (Han-Viet) phonetic, and Vietnamese meanings. A key challenge lies in the polysemy of many characters, each may have multiple entries with distinct meanings or pronunciations. To address this, the system incorporates a semantic consolidation mechanism to merge related entries and minimize redundancy. After semantic consolidation to merge related entries and remove duplicates, the refined dictionary comprises 2,788 unique entries (see Section 3.2).
- **Expert Analysis:** This section includes 80 entries of expert-level commentary authored by Ly Minh Tuan, offering contextual and interpretive insights that enrich the Commentary Layer. Each commentary entry is algorithmically linked to its corresponding section (`sect_id`) in the main text, allowing cross-referencing between doctrinal passages and their scholarly interpretation.

A.2 Construction Pipeline

The dataset construction process was organized into three major stages: preprocessing, alignment, and structuring the corpus.

Preprocessing. All source documents were converted from PDF into structured digital text through semi-automatic extraction. This process involved several key steps:

- Digitization and normalization: An optical character recognition (OCR) correction and character standardization library (i.e., PaddleOCR⁶) was applied to ensure textual consistency, including the removal of headers, footers, and page artifacts.
- Segmentation: The base text and commentary were separated yet aligned to preserve interpretive relationships. Texts were divided into sentences or discourse units that matched the logical structure of the original works.
- Structural encoding: Classical Chinese characters in traditional script were standardized and represented in Unicode. Hierarchical structures, including book, chapter, section, and sentence, were encoded in XML/JSON format.

⁶<https://www.paddleocr.ai/>

- **Metadata curation:** Provenance information such as edition, commentary author, and variant readings was attached at the segment level.
- **Lexical processing:** Glossary sections were automatically detected and parsed using regular expressions (regex). Contextual metadata (book and chapter) was assigned to each lexical entry, ensuring accurate semantic anchoring.

Alignment Strategies. Distinct alignment strategies were employed for different corpus components:

- **Tri-Parallel Corpus:** A rule-based heuristic approach was developed to align the Classical Chinese text with its Sino-Vietnamese (Han-Viet) phonetic and modern Vietnamese translation. Explicit textual cues, such as "Translation:" markers and structural delimiters, were utilized to achieve consistent three-way alignment.
- **Lexical Corpus:** Lexical items were extracted from dictionary sections through regex-based parsing and forward-fill propagation of contextual metadata (e.g., book and chapter). Post-processing included deduplication, normalization, and removal of incomplete entries to produce a clean, queryable dictionary resource.
- **Exegesis Corpus:** Commentary sections were identified through rule-based parsing and text cleaning. Each commentary entry was linked to its corresponding canonical text and translation via structured metadata. This corpus captures interpretive layers that complement the bilingual and lexical resources.

Manual Refinement. We employed a two-stage refinement process consisting of heuristic recovery, in which a simple module was implemented to resolve OCR-induced artifacts (e.g., character substitutions, diacritic distortions, and mid-sentence line breaks), followed by an expert manual audit, during which all textual triplets were manually verified by domain researchers to ensure exact synchronization across linguistic layers. This deterministic procedure guarantees a strict 1:1:1 mapping between Classical Chinese, Sino-Vietnamese, and modern Vietnamese sentence nodes, reflected in identical sentence indices across layers, providing a robust foundation for the multi-layered graph.

Structuring and Export. All corpora were encoded into standardized formats (CSV, XML, and Excel) with unique structured identifiers (**File.Sect.Page.STC**). This design ensures interoperability among the tri-parallel, lexical, and exegesis corpora, forming a coherent foundation for downstream applications such as ontology-guided knowledge graph construction, semantic retrieval, and cross-lingual analysis.

A.3 Dataset Description

The finalized dataset consists of 2,222 segmented sentences, 80 commentary annotations, and a refined lexical dictionary comprising 2,788 entries, including 2,562 unique Classical Chinese characters and 23 domain-specific Confucian concept terms. Ontology instantiation yielded a fully structured knowledge graph containing 16,468 nodes and 71,249 edges, systematically representing entities, semantic relations, and hierarchical interconnections across textual, linguistic, and conceptual layers.

Tri-Parallel Corpus. The tri-parallel corpus extends the alignment into a three-layer structure comprising: (1) the original Classical Chinese text, (2) the

Sino-Vietnamese (Han-Viet) phonetic, and (3) the modern Vietnamese semantic translation. This corpus includes 2,222 aligned triplets derived from *Commentaries on The Four Books*. The tri-parallel design offers several advantages: it minimizes the gap between orthography and semantics, supports simultaneous phonetic and semantic learning, and enables multi-view NLP tasks such as phonetic-aware translation and multistage alignment. Although the rule-based approach ensures high precision in structurally consistent passages, it remains sensitive to irregular formatting and cases where a single Classical Chinese sentence corresponds to multiple Vietnamese interpretations.

Lexical Dictionary. The lexical dictionary is a dictionary-level resource extracted from the glossary sections of *Commentaries on The Four Books*. It contains 5,344 entries, each consisting of a unique identifier (ID), Classical Chinese character, Sino-Vietnamese (Han-Viet) phonetic, modern Vietnamese meaning, and source context (book and chapter of occurrence). This lexical corpus serves as a critical bridge between traditional vocabulary and modern computational linguistics. It underpins knowledge graph construction, facilitates semantic disambiguation, and supports the development of educational and language-learning applications focused on classical texts.

Exegesis Corpus. The exegesis corpus encompasses 80 commentary annotations of interpretive commentary derived from *Commentaries on The Four Books*. Each commentary passage is linked to its corresponding canonical text and translation through structured metadata, including book, chapter, and section identifiers. This corpus adds interpretive depth to the overall dataset, enriching the tri-parallel and lexical corpora with contextual explanations, philosophical insights, and scholarly interpretation.

Appendix B Ontology Relation Generation

Relations are generated through three distinct methods, each with specific validation requirements.

Fully Automatic (Rule-based)

These relations are generated through deterministic algorithms without human intervention:

- **Structural:** CONTAINS, APPEARS_IN, FOLLOWS, HAS_HAN_FORM, generated through hierarchical parsing of document structure and tri-parallel alignment.
- **Linguistic:** TRANSLATES_TO, PRONOUNCED_AS, dictionary lookup with contextual embedding disambiguation.
- **Speaker:** QUOTES, pattern-based regex detection for attribution markers (e.g., , ,).
- **Semantic:** BELONGS_TO_CLUSTER, SIMILAR_TO, HAS_SEMANTIC_REPRESENTATION, computed from Multilingual-E5-Large embeddings with fixed thresholds.

Semi-automatic (Embedding + Verification)

These relations combine algorithmic generation with sampling-based verification:

- **EXPRESSES_CONCEPT:** Character pattern matching against predefined taxonomy, validated through embedding clustering to confirm semantic coherence.
- **CONTEXTUALIZES:** Initial candidates generated via semantic similarity (cosine > 0.75), followed by 10% sampling-based manual verification.
- **RELATED_TO:** Co-occurrence frequency analysis combined with manual expert review for philosophical validity.

Manual

These relations are defined based on scholarly classification:

- **PROVIDES_COMMENTARY:** Expert attribution (current implementation: single expert Ly Minh Tuan).
- **Taxonomic relations:** BELONGS_TO_SCHOOL, PART_OF_DOMAIN, domain hierarchy predefined.
- **Concept taxonomy:** 23 core Confucian concepts manually categorized by thematic function.

Appendix C Core Confucian Concepts

Conceptual Layer is anchored by 23 fundamental Confucian concepts, categorized by thematic function to allow for nuanced tracing of philosophical development. Table C1 presents the complete taxonomy. This taxonomy enables the system to automatically map granular linguistic units to abstract philosophical entities through character pattern matching, preserving interpretive multiplicity while maintaining structural integrity across the graph. For example, when the character 仁 appears in a sentence, the system creates an EXPRESSES_CONCEPT edge linking that sentence to the PHILOSOPHICAL_CONCEPT: node, which is further connected to its categorical grouping (Virtue) via RELATED_TO edges.

Appendix D Semantic Chunking

The first stage of the pipeline focuses on transforming raw classical and modern texts into structured units while preserving their semantic meaning, which is critical given the complexity and polysemy of Classical Chinese. To effectively process long commentary passages, a semantic-aware adaptive chunking module dynamically segments text based on semantic coherence rather than fixed character limits. Each document is first tokenized and then divided into segments with a maximum length of $L = 512$ tokens and a look-ahead overlap of $O = 100$. Text encoding is performed using the Multilingual-E5-large model, ensuring that each chunk captures a coherent philosophical argument suitable for subsequent Retrieval-Augmented Generation (RAG) tasks. An automated validation check further improves reliability by switching to a simpler fallback strategy when initial chunk quality is low.

Table C1: Core Confucian concepts taxonomy.

Character	English	Vietnamese	Category
<i>Cardinal Virtues ()</i>			
	Benevolence	Nhón	Virtue
	Righteousness	Ngha	Virtue
	Ritual propriety	L	Virtue
	Wisdom	Trờ	Virtue
	Trustworthiness	Tồn	Virtue
<i>Self-Cultivation ()</i>			
	Virtue/Power	c	Cultivation
	Sincerity	Chón	Cultivation
	Correctness	Chờnh	Cultivation
<i>Cosmological Foundations</i>			
	The Way	o	Foundation
	Heaven	Thiỏn	Foundation
	The Mean	Trung	Harmony
	Harmony	Hòa	Harmony
<i>Social Relations ()</i>			
	Filial piety	Hu	Relation
	Fraternal respect		Relation
	Loyalty	Trung	Relation
	Forgiveness/Reciprocity	Th	Relation
<i>Learning and Education</i>			
	Learning	Hc	Learning
	Teaching	Giỏo dc	Learning
	Knowledge	Tri	Learning
<i>Political Order</i>			
	Ruler	Quón	Social
	Minister	Thn	Social
	People	Dón	Social
	Government	Chờnh	Social

The adaptive chunking module employs cosine similarity of Multilingual-E5-Large embeddings to preserve semantic integrity when segmenting long commentary passages. Unlike fixed-length chunking that may split coherent arguments mid-sentence, our approach detects natural topic boundaries through embedding-based coherence analysis. For a sequence of extracted sentences $S = \{s_1, s_2, \dots, s_n\}$, each sentence is encoded using the Multilingual-E5-Large model with the instruction prefix “passage:” to obtain embedding vectors. The semantic coherence score for sentence s_i is computed as the average cosine similarity with preceding sentences within a sliding window of size $w = 3$:

$$\text{coherence}(s_i) = \frac{1}{\min(i, w)} \sum_{j=\max(0, i-w)}^{i-1} \cos(\mathbf{e}_{s_i}, \mathbf{e}_{s_j}). \quad (\text{D1})$$

A topic boundary is identified when $\text{coherence}(s_i) < \theta$, where the threshold $\theta = 0.3$ was empirically determined to balance granularity with coherence preservation. When a boundary is detected, the current chunk is finalized and a new chunk begins, ensuring that semantically related content remains together. The chunking process also enforces size constraints: maximum chunk length $L = 512$ tokens (compatible with downstream RAG tasks) and minimum chunk size $M = 256$ characters (preventing overly fragmented segments). A post-hoc quality validation verifies content coverage $\geq 95\%$; if validation fails, the system falls back to simpler fixed-length chunking to ensure robustness. This mechanism is particularly important for Ly Minh Tuan’s commentary, which often develops philosophical arguments across multiple sentences. By respecting semantic boundaries rather than arbitrary character limits, the adaptive chunker preserves the interpretive integrity essential for downstream retrieval and AI-grounded question answering.

For linguistic processing, a custom dictionary processor is employed to load, validate, and consolidate lexical resources across Classical Chinese, Sino-Vietnamese (Han-Viet) phonetic, and modern Vietnamese layers. This module resolves character-level polysemy and harmonizes multiple interpretations to prepare data for the Linguistic Layer. Philosophical concepts are pre-tagged according to a defined taxonomy (e.g., virtue, cultivation), while sentence embedding and semantic similarity computation are performed using the Sentence Transformers framework. Entity and concept extraction builds upon predefined taxonomies (e.g., ,) in conjunction with dictionary-based lookups and pattern-matching rules, ensuring high coverage and precision in identifying core Confucian concepts.

Appendix E KG Layer-Specific Construction

The **Textual Layer** encodes the canonical hierarchy from BOOK to SENTENCE through CONTAINS relations, forming approximately 2,400 nodes and serving as the structural backbone of the graph.

The **Linguistic Layer** enriches this structure by linking Classical Chinese sentences to their Sino-Vietnamese (Han-Viet) phonetic and modern Vietnamese translations, and by connecting Classical Chinese words to dictionary entries via TRANSLATES_TO and PRONOUNCED_AS relations. This layer contains about 11,600 nodes and 37,700 edges, reflecting the complexity of cross-linguistic alignment.

The **Conceptual Layer** isolates philosophical notions through character-level pattern matching and associates them with sentences using EXPRESSES_CONCEPT and RELATED_TO relations, thereby capturing the underlying philosophical ideas conveyed in the text.

The **Commentary Layer** introduces expert annotations as EXPERT and COMMENTARY_CHUNK nodes, sequentially ordered through FOLLOWS relations and connected to the base text via EXPLAINS and CONTEXTUALIZES relations, linking interpretive insights to the canonical material.

Table F2: Retrieval effectiveness between Adaptive Semantic and Fixed Chunking.

Method	Recall@5	NDCG@5	Mean Similarity
Fixed Chunking (Baseline)	0.315	0.281	0.861
Adaptive Semantic Chunking (Ours)	0.380	0.333	0.863

The **Speaker Layer** identifies quotation attributions through rule-based detection (e.g., , ,), assigning statements to **SPEAKER** nodes connected by **QUOTES** relations.

Finally, the **Semantic Layer** introduces an additional dimension of connectivity by leveraging multilingual-e5-large embeddings to link each node with semantically related neighbors, organizing them into **SEMANTIC_CLUSTERS** based on cosine similarity thresholds.

Appendix F NLP Component Evaluation

F.1 Semantic Alignment and Retrieval Performance

To validate the Semantic Layer, we conducted a comparative test between our Multilingual-e5-large embedding model and the traditional keyword-based BM25 baseline. We used a synthetic test corpus containing both Confucian concepts (Relevant) and unrelated modern topics (True Negatives) to measure the model’s discriminative power. The quantitative results across multiple standard retrieval metrics are presented in Table 3. The empirical data demonstrates that the semantic approach achieves perfect precision and ranking scores ($P@1 = 1.000$, $MRR = 1.000$). While the BM25 baseline performs reasonably well, its accuracy significantly degrades as the retrieval depth (K) increases, dropping to a $P@10$ of 0.505. Our Hybrid method successfully maintains the high precision of the semantic model while utilizing RRF for robust ranking. These results demonstrate strong discriminative power under controlled conditions; their interpretation in open-domain settings is discussed in Section 7.

F.2 Adaptive Semantic Chunking Performance

For long commentary passages where fixed boundaries are absent, we evaluated our Adaptive Semantic Chunking against a standard Fixed-Length baseline ($L = 512$, $O = 0$). We used a test corpus of 200 annotated queries to measure retrieval effectiveness, as shown in Table F2. The 20.6% improvement in Recall@5 and 18.5% increase in NDCG@5 validate that our semantic-aware segmentation successfully preserves the integrity of philosophical arguments, making them more discoverable than arbitrary text splits.