Work in progress: TextMatch - a semantically enhanced textbook recommendation system

Department of Electrical and Computer Engineering, University of Toronto

Abstract

Browsing through a large assortment of books is often necessary for both instructors and students to find relevant resources. Manually searching through several volumes of textbook material can be time-consuming and unproductive. Online platforms typically rely on keyword searches, but these approaches may fail when users are unsure of the right terms. Semantic search offers a solution by interpreting the conceptual meaning behind queries, enabling more contextually relevant recommendations than keyword searches. In this work, we propose TextMatch, which implements semantic search atop an existing keyword-based system, improving the relevance of returned materials. After performing an initial keyword-based retrieval, TextMatch applies a multi-stage semantic reranking process to refine results. These stages leverage vector embeddings to reorder items by semantic similarity, ultimately yielding more meaningful recommendations. We extensively evaluated TextMatch both quantitatively and qualitatively. While results showed a 300% improvement in Mean Reciprocal Rank (MRR) and a 150% increase in Precision@10 compared to a purely keyword-based baseline—highlighting the substantial potential of semantic reranking—statistical significance tests indicated that the small sample size limited firm conclusions about performance. Nevertheless, user surveys revealed a strong preference for TextMatch's semantic search features, confirming the utility of incorporating contextual understanding. By blending keyword precision with semantic depth and offering flexible configuration options, TextMatch provides a more intuitive and effective resource discovery experience for students and instructors. As a work in progress, future iterations will focus on larger-scale evaluations, deeper faculty engagement, and integration with broader academic resources.

Introduction

Numerous sources have shown that students with access to high-quality educational resources demonstrate significantly improved learning outcomes, including higher grades and better conceptual understanding [1], [2], [3], [4]. While students often appreciate the familiarity of traditional textbooks, their high cost and limited availability create significant financial barriers [5]. These constraints can impede the learning process, leaving students to either forego essential materials or resort to lower-quality, yet more affordable alternatives. The rise of free online re-

sources has partially alleviated this burden, providing a broader array of educational content at reduced cost. The effectiveness of online learning tools, however, depends heavily on their quality and relevance [6]. In an increasingly digital landscape, learners must navigate a vast array of materials, often struggling to identify the most conceptually aligned resources for their needs. Finding the best match for a student's academic goals—especially when budgets are tight and precise keywords are unknown—is not trivial, due in large part to the sheer volume of available content.

Online search platforms, such as those used by libraries and educational repositories, offer a more efficient starting point, utilizing keyword search algorithms to rapidly filter massive databases [7]. However, traditional keyword searches heavily depend on exact or near-exact term matches. They struggle when users mistype terms, use uncommon synonyms, or cannot recall precise keywords [8]. Moreover, purely keyword-driven approaches fail to capture the deeper contextual underpinnings of a query, often returning results that appear relevant at first glance but are not contextually aligned with the user's true intent. While large language models (LLMs) have emerged as a flexible solution, they carry the risk of "hallucination"—fabricating or misrepresenting information when sufficient context is lacking [9], [10].

To mitigate the limitations of existing works, we propose TextMatch, that uses semantic search to interpret the meaning and context behind user queries [11]. When semantic techniques are applied on top of a verified, curated database of textbooks, users gain access to reliable, contextually rich recommendations without the misinformation risk associated with unconstrained LLM-driven searches. TextMatch exemplifies this approach by integrating semantic reranking into an existing keyword-based framework, blending the best of both worlds. Instead of relying solely on literal keyword matches, TextMatch reevaluates initial search results through a semantic lens, prioritizing materials that align more closely with the user's underlying intent. This layered method enhances precision, reduces the time spent filtering irrelevant results, and supports users who may not know the exact terms or who wish to explore related concepts.

In this paper, we focus on a curated set of engineering and technology textbooks from the O'Reilly platform to demonstrate TextMatch's capabilities. As a work in progress, we plan to expand TextMatch's scope to additional disciplines and data sources, ultimately making it more versatile for diverse educational settings. We highlight how TextMatch's flexible, modular architecture supports different semantic search models and user preferences, presenting both quantitative and qualitative evidence—including improved mean Precision@10 and MRR scores and favorable user survey responses—to demonstrate its value. Ultimately, TextMatch not only broadens access to relevant educational content but also streamlines the discovery process, showing how large repositories of textbooks and learning materials can become more accessible, cost-effective, and aligned with evolving student and instructor needs.

System design of TextMatch

TextMatch employed a multi-step process that combined a traditional keyword search with one or two layers of semantic reranking. The objective was to return results that not only matched the user's query terms but also aligned more closely with underlying context and intended topics of interest.

Existing TextCraft design

TextCraft [12], the foundational system for TextMatch, is a web-based textbook recommendation platform designed to streamline the discovery and organization of educational resources. Catering to both students and instructors, it provides recommendations tailored to user-provided course outlines or topics. The platform operates on a robust database of over three million book chapters sourced from O'Reilly and employs the BM25 search algorithm to rank results based on term frequency and document relevance.

This keyword-driven architecture forms the starting point for TextMatch, which enhances TextCraft's functionality by introducing semantic search capabilities. By building upon TextCraft's efficient and user-friendly design, TextMatch leverages a reliable, curated database while addressing the contextual limitations of traditional keyword-based methods.

When initiating a search, users provided a "book name" as their main query and one or more additional "topics" to refine the results. The "book name" served as the primary query (hereafter the main query), and the "topics" fields contained more specific concepts, subtopics, or chapter titles the user wished to explore. This input structure, inherited from TextCraft [12], allowed users to capture both a broad subject area (via the main query) and granular points of interest (via topics).

The search algorithm began by performing a keyword-based retrieval using BM25 parameters inherited from TextCraft [12]. When the user submitted their main query and topics, the system first extracted up to 200 candidate books from a large, curated database, relying solely on keyword matches. This initial pass produced a broad but potentially imprecise set of results.

Extending on TextCraft to create TextMatch

After identifying these candidate books, TextMatch applied semantic reranking at the book-title level. It utilized a transformer-based embedding model (all-MiniLM-L6-v2 from the sentence-transformers library) to generate vector representations of the user's main query and each retrieved book's title. In addition, it embedded the user-provided topics and compared them to each book title. Two sets of cosine similarities were computed:

- Query-to-Book Title Similarity: Measures how closely each book title, in the 200 candidate books, corresponds to the main query.
- Topics-to-Book Title Similarity: Measures how well each book title, in the 200 candidate books, aligns with the user-specified topics.

TextMatch then combined these two similarity values using a weighted sum. By default, it assigned a weight of 0.6 to the query/book-title similarity and 0.4 to the topics/book-title similarity. This weighting scheme is configurable, enabling users or administrators to emphasize either general relevance (main query) or more detailed conceptual alignment (topics). Notably, this adjustability supports different use cases and may affect performance metrics, warranting further empirical tuning in future work. After computing the composite scores, TextMatch selected the top 10 most relevant books. This process represents the first semantic reranking stage (SS1), focusing on refining results at the book-title level.

For users requiring greater precision, TextMatch offered an optional second semantic reranking step focused on chapter-level details. Starting from the top 50 books identified in the first stage, it embedded and compared their chapter titles against the user's topics. This finer-grained analysis employed the same transformer-based model and cosine similarity approach but focused on internal book structure. By examining how well specific chapter titles aligned with the user's topics, TextMatch produced an even more targeted list of recommended books. Since this second-stage refinement was more computationally intensive, limiting it to the top 50 candidates ensured that the response time remained acceptable. An acceptable response time in this study was considered under 10 seconds, balancing computational cost with user experience in a real-world academic setting. After this second step, the top 10 results were displayed to the user. This chapter-level refinement represents the second semantic reranking stage (SS2).

Figure 1 shows the flow chart of TextMatch's search algorithm. The algorithm was designed to support continual refinement and easy integration of new technologies. The semantic reranking components—embedding model selection, scoring parameters, and weighting—are configurable through external files. This modularity will enable users and developers to adjust weights, experiment with different models, or introduce new embedding methods. The system logs every query, result, and performance measure in a structured local database, facilitating error handling, diagnostic analysis, and future experimentation. By mining these logs, developers could identify trends, compare configurations, and steadily improve TextMatch as educational needs and search technologies evolve.

Evaluation pipeline

Establishing reliable ground truth data is a fundamental challenge in evaluating search systems. Different individuals may identify different sets of books as "ground truth" for the same query, introducing subjectivity and variability. One strategy involves recruiting domain experts to label relevant books in their fields, but this is time-consuming and difficult to scale.

Instead, we adopted an alternative approach that considered practical user workflows. We assumed that most users, when facing an information need, would consult widely used search engines—primarily Google—or increasingly, LLM-based services like ChatGPT. While Google provides reliable but fragmented information, ChatGPT can produce more coherent, user-friendly summaries but may occasionally "hallucinate" or fabricate content. Users relying on ChatGPT must manually verify any suggested sources.

Despite these limitations, ChatGPT still offers a valuable starting point for discovering potentially relevant books, especially for non-expert users. We leveraged ChatGPT-4o to generate candidate ground truth data, then applied rigorous human verification to mitigate misinformation. Our procedure was as follows:

- We issued arbitrary queries (main query plus one or more topics) to ChatGPT, asking it to identify all relevant books within the O'Reilly database.
- For each recommended book, we manually verified its existence in the O'Reilly database and in our curated dataset. Only after this verification step did we include the book as part of our ground truth set.

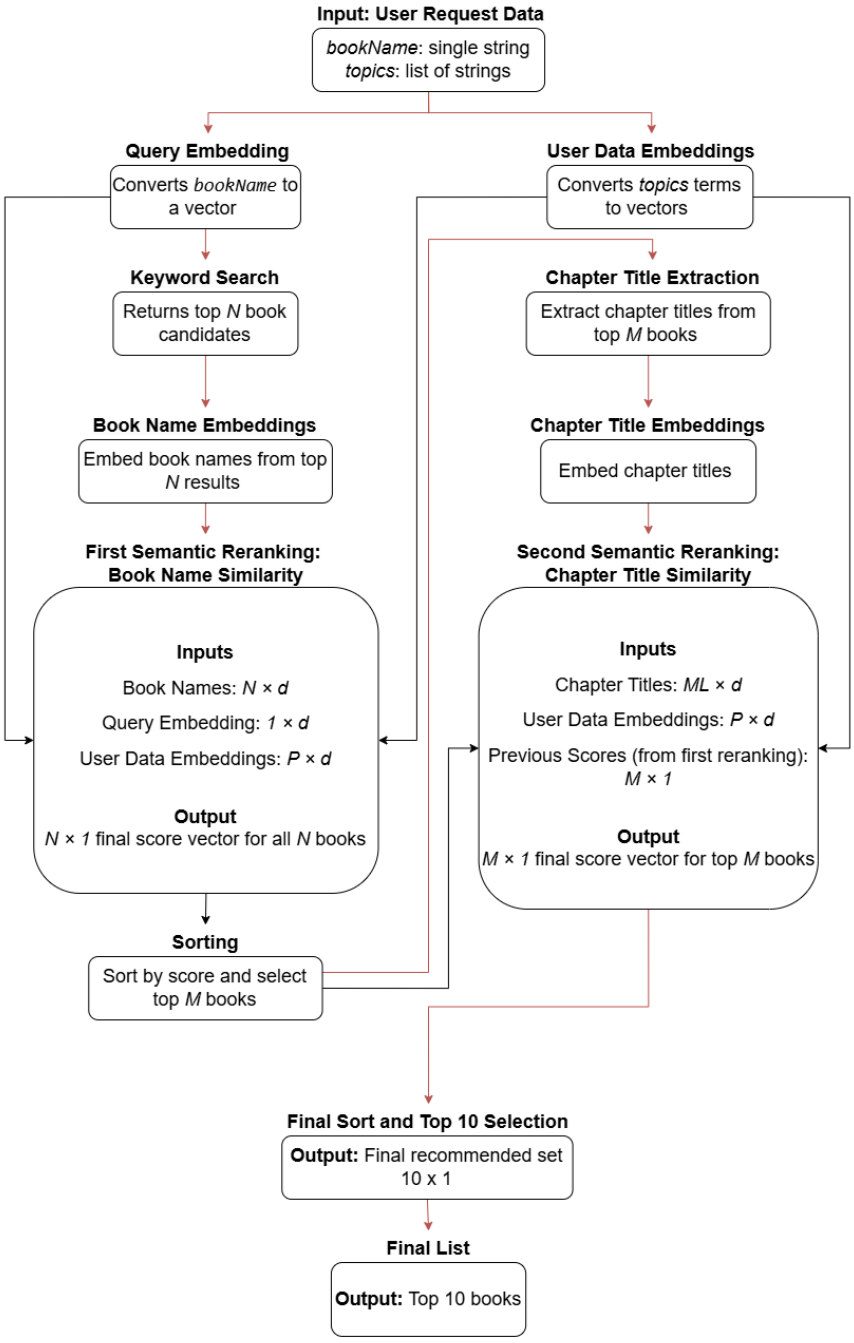

Figure 1: The flow chart of TextMatch's search algorithm, from user input (main query and topics) through keyword search, semantic reranking stages, and finally to the top 10 most relevant results.

This method combined ChatGPT's broad knowledge and user-friendly presentation with human verification, producing a set of ground truth data that was practical and reflective of typical user search strategies.

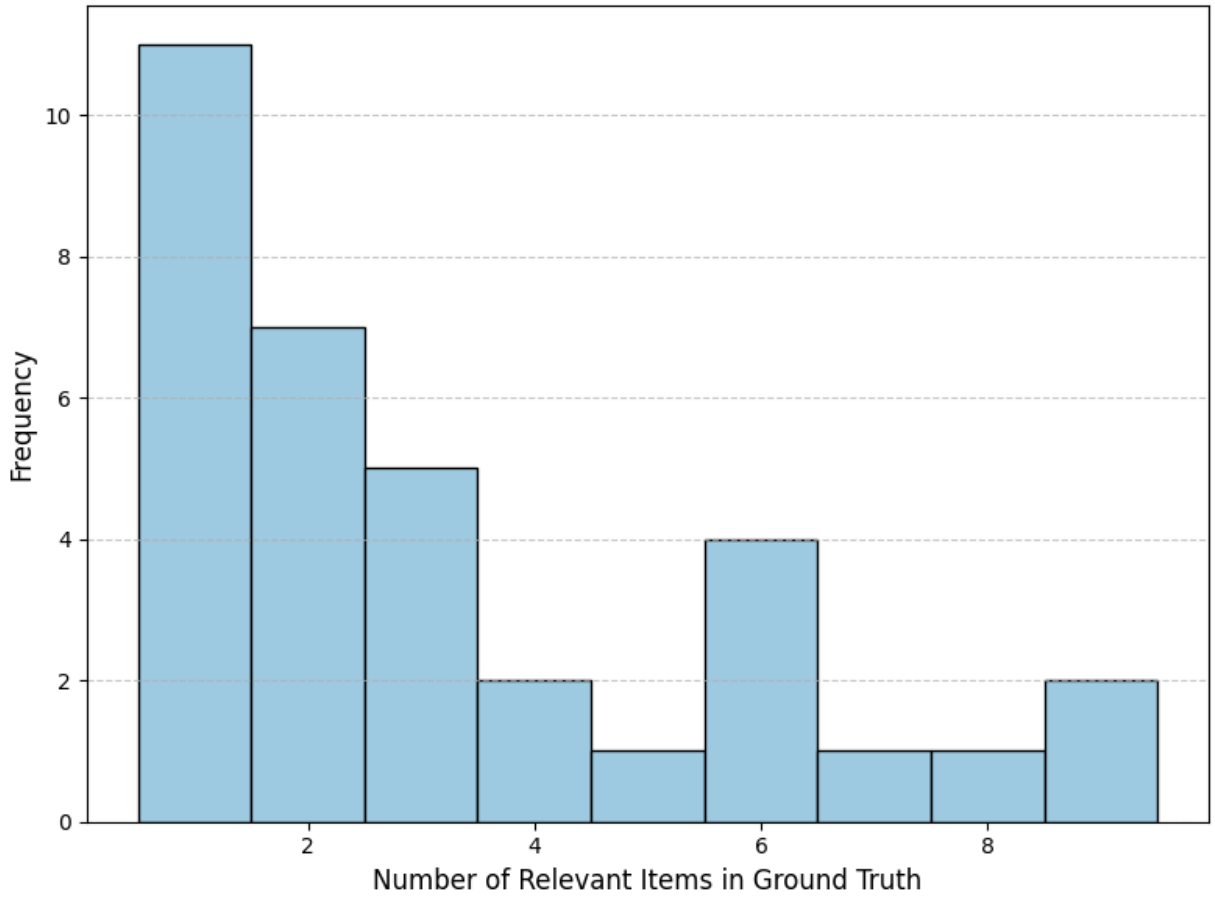

Figure 2: A histogram of the frequency distribution of relevant items in ground truth per query.

As shown in Figure 2, our initial ground truth dataset encompassed a reasonable variety of queries and relevant books. The histogram reveals a secondary peak around n=6, suggesting that some user queries tend to retrieve a similar number of relevant books. This may indicate clustering around specific topical requirements, where certain subjects naturally yield a consistent set of relevant materials. Future work may explore expanding the dataset and query pool to further analyze these distribution patterns. The manual verification process was very time-intensive, so we deemed the current distribution adequate for preliminary evaluations of TextMatch's semantic reranking performance.

After establishing the ground truth, we conducted a series of quantitative evaluations to compare TextMatch's performance against both the original TextCraft[12] system and established relevance metrics. Our primary objective was to ensure that TextMatch preserved TextCraft's keyword-based retrieval accuracy while meaningfully enhancing it through semantic analysis.

To gauge TextMatch's effectiveness, we focused on two primary metrics: Precision@10 and

Mean Reciprocal Rank (MRR), both established in the original TextCraft[12] study. These metrics evaluate different aspects of retrieval performance: Precision@10 measures the proportion of relevant books retrieved in the top 10 results, while MRR assesses how quickly a relevant book surfaces in the ranked list.

Precision@10 indicates the relevance of the top 10 results returned by the system. It is calculated as:

$$\text{Precision@10} = \frac{\text{Number of Relevant Items in Top 10}}{10} \tag{1}$$

This metric provides insight into the accuracy of the system within the top-ranked results. A higher Precision@10 value signifies that more relevant materials are present in the top 10 recommendations, enhancing user efficiency in finding suitable content.

Mean Reciprocal Rank (MRR) evaluates the rank position of the first relevant item for each query. It is defined as:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i} \tag{2}$$

Here, $|Q|$ represents the total number of queries, and rank_i is the rank position of the first relevant item for the i-th query. MRR provides a measure of how quickly users can locate the first relevant material in the results. A higher MRR score indicates that relevant books are ranked closer to the top, improving user satisfaction and reducing search effort.

In addition to these core metrics, we calculated Precision@3 and Precision@5 to evaluate performance at more stringent cutoffs. We also introduced Average Change in Rank (ACR) to quantify how much semantic reranking reshuffled the initial keyword-based results. A higher ACR suggested that the semantic layer actively modified the ranking beyond the baseline.

To further assess robustness and significance, we performed Wilcoxon signed-rank tests and computed Cohen's d effect sizes. The Wilcoxon test, a non-parametric statistical test, determined whether observed metric improvements were statistically significant [13]. Cohen's d quantified the magnitude of observed effects [14]. We applied these analyses to both the metrics and rank change data. Additionally, a user survey involving students and instructors familiar with both systems provided qualitative feedback on search accuracy, perceived speed, and overall satisfaction.

Results

Quantitative analysis

As shown in Figure 3, the full semantic reranking method (SS2) consistently outperformed keyword search (KS) and single-stage semantic reranking (SS1) across all evaluated metrics—MRR, Precision@10, Precision@3, and Precision@5. The Mean Reciprocal Rank (MRR) saw a fourfold increase, rising from 0.05 in KS to 0.20 in SS2, indicating a significantly higher likelihood of

retrieving relevant items in top ranks. Precision@10 more than doubled, increasing from 0.06 in KS to 0.15 in SS2, reflecting a marked improvement in the proportion of relevant items retrieved within the top 10 results. Similarly, Precision@3 and Precision@5 showed substantial gains, with SS2 achieving nearly double the values compared to KS—0.31 vs. 0.17 for Precision@3 and 0.38 vs. 0.23 for Precision@5.

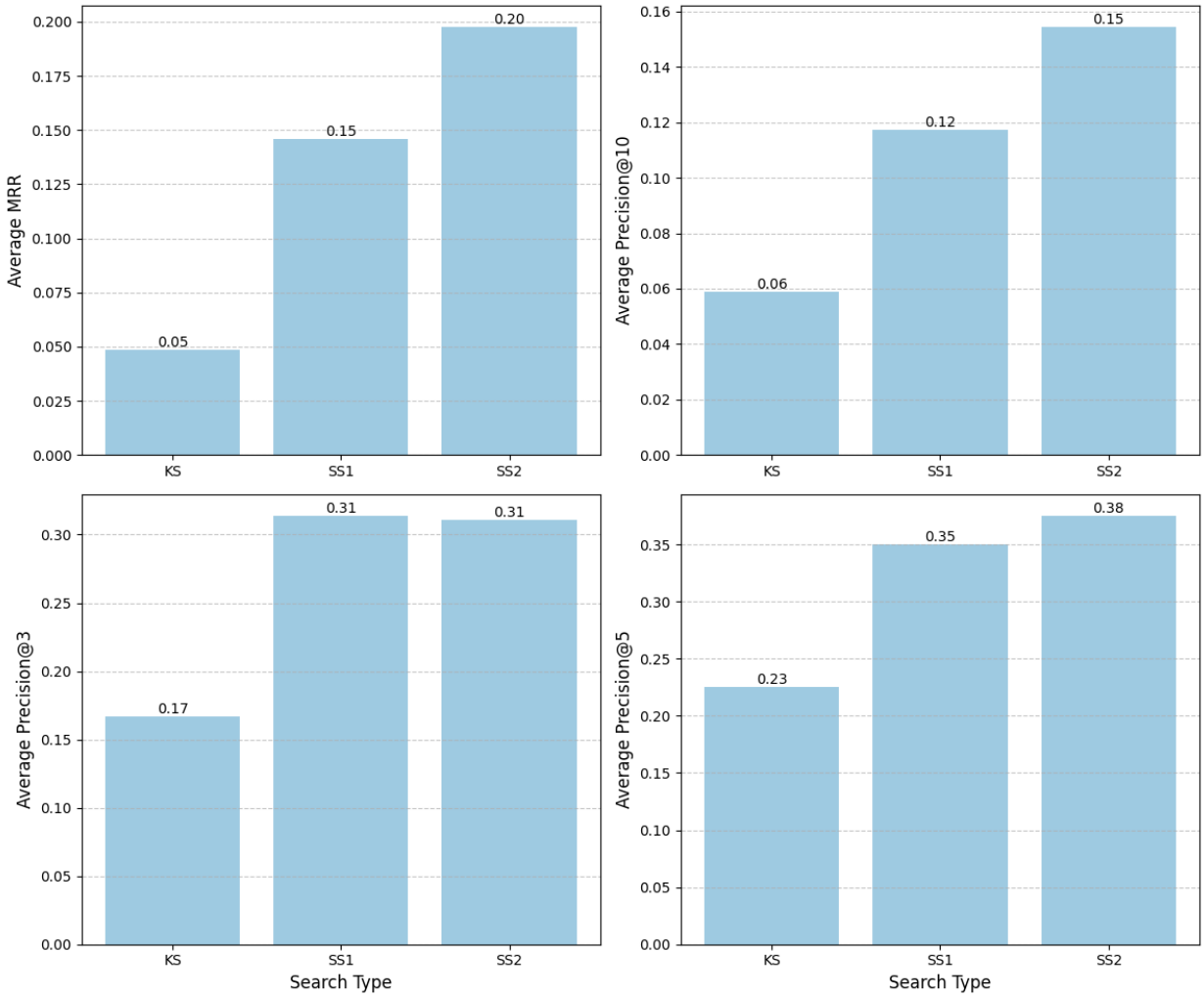

Figure 3: Bar plots showing the average Precision@3, Precision@5, Precision@10, and Mean Reciprocal Rank (MRR) for the three search methods. **KS** represents keyword search, **SS1** indicates semantic reranking using only book names, and **SS2** involves semantic reranking with both book names and chapter titles.

Figure 4 shows that while KS achieved the fastest average search time at 4.46 seconds, SS1 introduced a slight delay (5.37 seconds), and SS2 required 8.63 seconds. This additional latency reflects the computational cost of the second semantic reranking stage, which is why this stage is optional.

Figure 5 illustrates the stability of ACR across varying query lengths, with values ranging from 3.49 to 3.67. This suggests that semantic reranking impacts rankings consistently, regardless of query complexity.

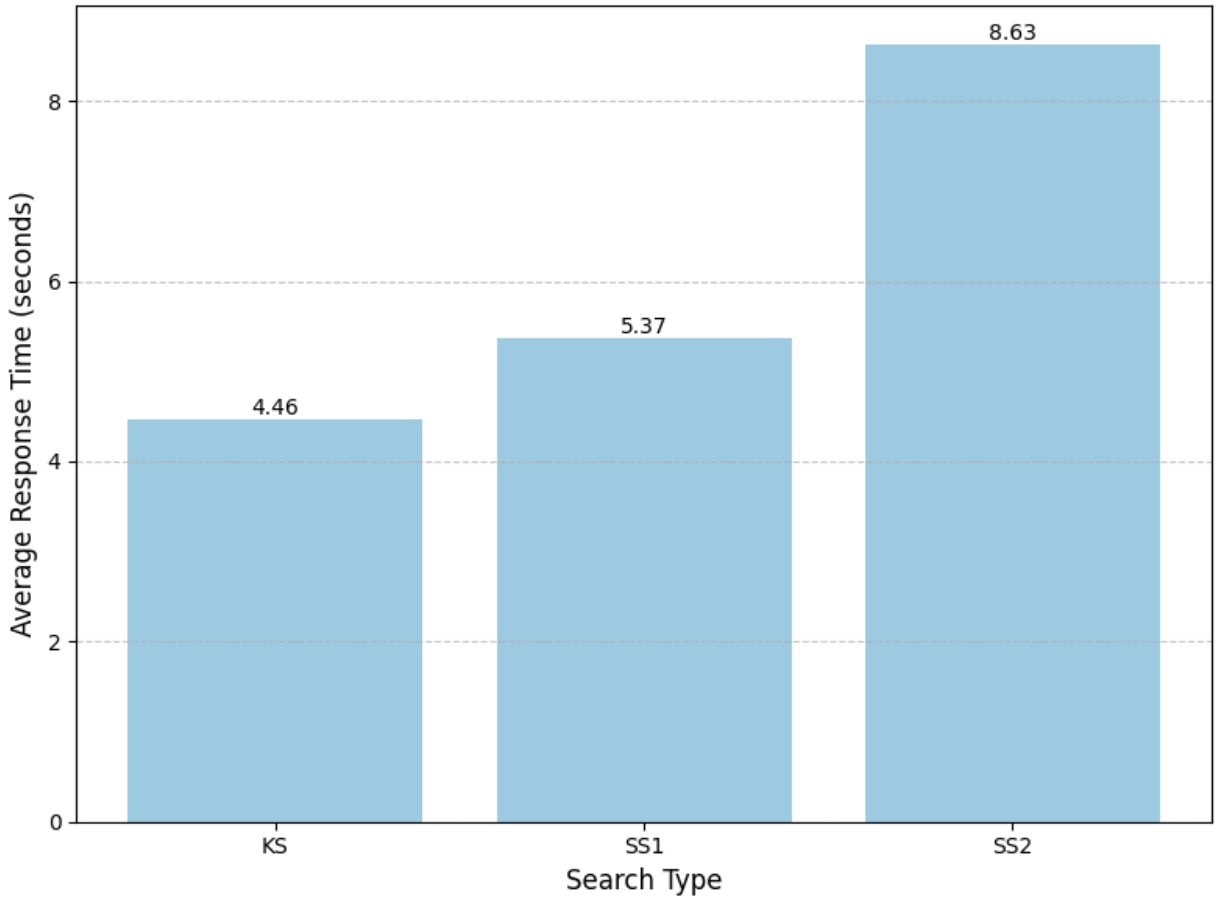

Figure 4: Bar plot showing the average search time in seconds for the three search methods. **KS** represents keyword search, **SS1** indicates semantic reranking using book names, and **SS2** involves semantic reranking with both book names and chapter titles.

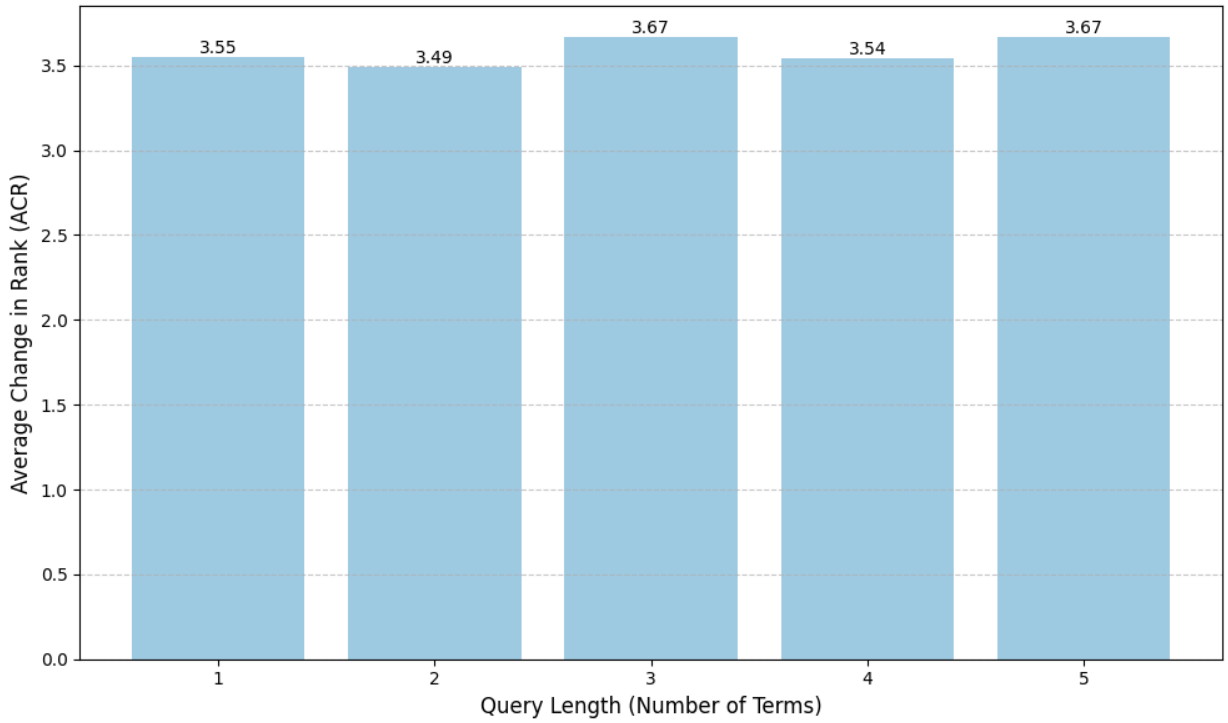

Figure 5: Average Change in Rank (ACR) as a function of query length

Table 1: Statistical Analysis Results

Metric	Wilcoxon Statistic	P-Value	Effect Size (Cohen's d)	Interpretation
MRR	11.00	0.38	0.32	Promising small effect
Precision@10	7.00	0.24	0.23	Promising small effect
Precision@3	9.00	0.39	0.28	Promising small effect
Precision@5	10.00	0.31	0.28	Promising small effect
Search Time Taken	0.00	0.01	2.87	Significant effect
Rank Changes	N/A	N/A	0.77	Moderate effect

Table 1 presents the statistical analysis comparing KS and SS2. For MRR and precision metrics at various cutoffs, Wilcoxon tests yielded p-values above 0.05, indicating no statistically significant differences. However, small-to-moderate effect sizes (Cohen's d) suggest meaningful practical improvements. Search time showed a statistically significant increase ($p = 0.008$) with a large effect size (Cohen's $d = 2.64$), and rank changes exhibited a moderate effect size (Cohen's $d = 0.77$), indicating noticeable reordering of results.

Qualitative analysis

The user survey results in Figure 6 demonstrate a strong preference for TextMatch. Among the 15 participants, 93.3% indicated that TextMatch provided more relevant recommendations, while only 6.7% chose TextCraft. Similarly, TextMatch was favored by 93.3% of respondents for helping find relevant books faster. However, when asked about response times, 100% of participants

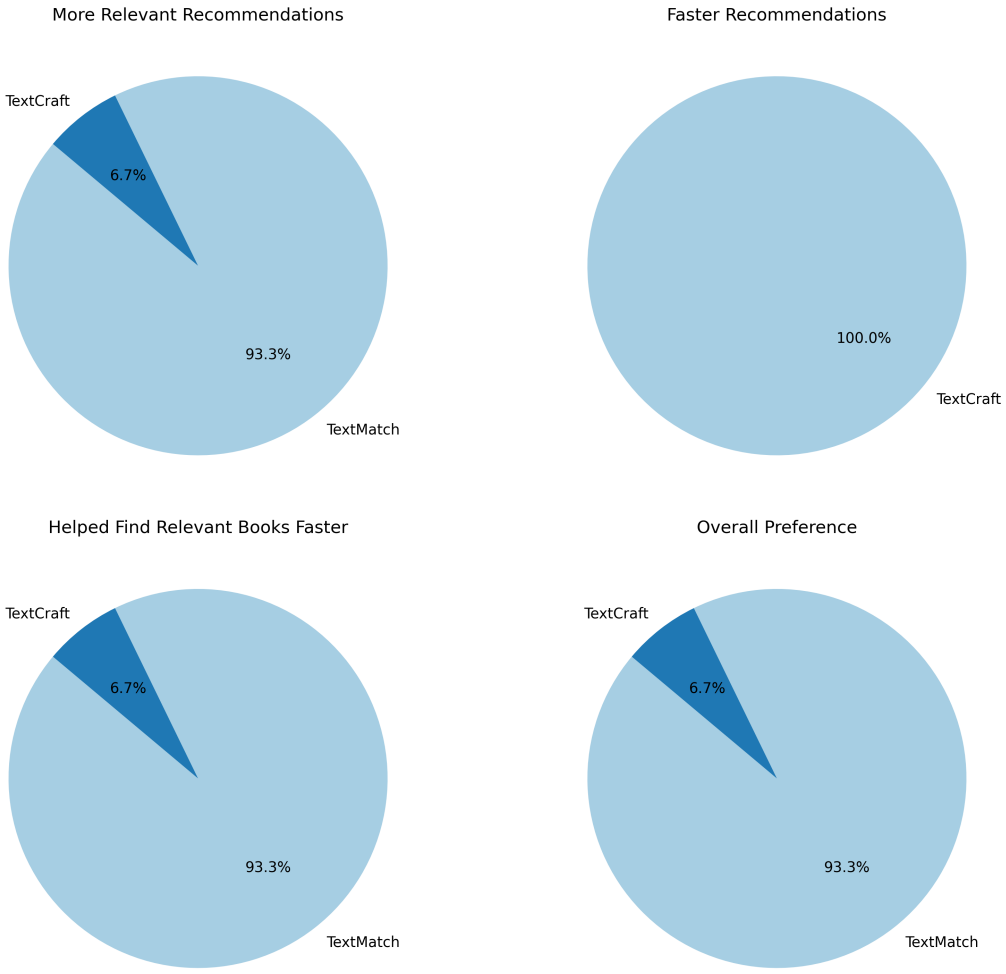

Figure 6: Survey results showing participant feedback across four key aspects of the two systems, TextCraft and TextMatch. The pie charts illustrate user preferences for more relevant recommendations (top-left), faster recommendations (top-right), help in finding relevant books faster (bottom-left), and overall system preference (bottom-right).

noted that TextCraft was faster. For overall system preference, 93.3% of participants expressed a clear preference for TextMatch.

Participants' additional comments highlighted TextMatch's ability to provide contextually relevant recommendations that closely matched their queries. Many users appreciated the semantic reranking for significantly reducing the effort required to locate the desired materials. While some acknowledged TextCraft's faster raw response times, they valued the higher quality and depth of results offered by TextMatch, deeming the slight delay a worthwhile trade-off. These insights demonstrate the importance of semantic enhancements in creating a more effective and user-friendly search experience.

Discussion

The results highlight the potential and value of semantic reranking in a keyword-based retrieval framework. While Wilcoxon tests did not reveal statistically significant improvements for most metrics, this likely reflects the small sample size rather than an absence of actual benefits. The observed increases in MRR and precision metrics, along with moderate effect sizes, indicate that semantic reranking provides tangible, practical advantages. SS2's higher MRR suggests that users encounter relevant items sooner, and better precision metrics imply that more useful results appear within the top ranks.

The ACR results suggest that semantic reranking consistently reshapes the initial keyword-based rankings, likely bringing more contextually aligned materials to the top. Although this reordering was not statistically significant by conventional standards, the moderate effect size indicates meaningful shifts toward greater contextual relevance.

Response time was the one metric with a statistically significant difference, reflecting the added computational overhead. However, this study was conducted on hardware with 16GB RAM, GTX 1660Ti GPU, and 512GB SSD. On modern, more powerful hardware, we would expect considerably faster response times, making the semantic enhancements even more practical in real-world deployments.

Crucially, the user survey validated the usefulness of TextMatch. Participants overwhelmingly favored TextMatch, citing its ability to produce more conceptually relevant recommendations and reduce the time needed to find suitable materials, despite the slower raw response. This suggests that the improvements introduced by semantic reranking are not merely theoretical but translate into a more positive, user-friendly search experience. Importantly, users appeared willing to tolerate additional latency in exchange for better-quality results.

From a broader perspective, these findings highlight the trade-offs involved in integrating advanced semantic techniques. While semantic reranking can improve relevance and user satisfaction, it may require additional computational resources and careful optimization. Future research could explore more efficient embedding models, caching strategies, or on-demand semantic reranking activated only when necessary. Additionally, expanding both the dataset and the pool of participants for evaluation could help clarify the trends observed here and provide more definitive statistical significance regarding the system's overall impact. At present, TextMatch remains a private prototype and is not openly accessible for direct testing. However, future itera-

tions may include an open-source or limited-access version to facilitate collaborative research and further validation of its semantic reranking capabilities.

Conclusion

TextMatch demonstrates the potential of semantic search to transform textbook recommendations by integrating conceptual understanding into a traditional keyword-based framework. By combining the precision of keyword search with semantic depth, TextMatch offers a richer and more intuitive search experience that aligns with user needs. The cumulative evidence—improved MRR and precision scores, along with overwhelmingly positive user feedback—strongly supports the efficacy of TextMatch's semantic reranking capabilities. Although our sample size limited the statistical significance of certain improvements, the observed trends highlight the system's practical advantages. User feedback further reinforces the value of semantic search, with participants overwhelmingly preferring TextMatch for its ability to provide relevant results efficiently, even at the cost of slightly longer response times. On modern hardware, this computational overhead would likely diminish, enhancing the practicality of TextMatch. As a prototype focused on O'Reilly engineering textbooks, TextMatch also lays the groundwork for broader adoption. Ongoing efforts will include incorporating faculty perspectives, integrating the system into institutional Learning Management Systems (LMS) or Open Educational Resources (OER) platforms, and expanding beyond engineering texts to other academic domains. By continually refining its modular design and semantic reranking framework, TextMatch stands poised to meet evolving user expectations, ultimately offering a path toward more accessible, efficient, and user-centered learning tools.

References

[1] R. Greenwald, L. V. Hedges, and R. D. Laine, "The effect of school resources on student achievement," *Review of educational research*, vol. 66, no. 3, pp. 361–396, 1996.

[2] D. Steiner *et al.*, "What we teach matters: How quality curriculum improves student outcomes," Tech. Rep., 2018.

[3] N. B. Colvard, C. E. Watson, and H. Park, "The impact of open educational resources on various student success metrics," *International Journal of Teaching and Learning in Higher Education*, vol. 30, no. 2, pp. 262–276, 2018.

[4] A. Kutu, N. Nzimande, and N. Grace, "Availability of educational resources and student academic performances in south africa," *Universal Journal of Educational Research*, vol. 8, pp. 3768–3781, 2020.

[5] M. Lebens, "Impact of textbook costs on student success: An opportunity to increase equity in mis courses by removing the textbook cost barrier," Tech. Rep., 2021.

[6] C. N. Akpen *et al.*, "Impact of online learning on student's performance and engagement: A systematic review," *Discover Education*, vol. 3, no. 1, pp. 1–15, 2024.

[7] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends® in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[8] A. Ahluwalia *et al.*, "Hybrid semantic search: Unveiling user intent beyond keywords," *arXiv preprint arXiv:2408.09236*, 2024.

[9] Z. Xu, S. Jain, and M. Kankanhalli, "Hallucination is inevitable: An innate limitation of large language models," *arXiv preprint arXiv:2401.11817*, 2024.
[10] L. Huang *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, 2023.
[11] B. Latard *et al.*, "Towards a semantic search engine for scientific articles," in *Research and Advanced Technology for Digital Libraries: 21st International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017, Proceedings*, vol. 21, Springer International Publishing, 2017.
[12] X. Fan, H. S. Timorabadi, and S. Emara, "Board 97: Work-in-progress: Textcraft: Automated resource recommendation for custom textbook creation," in *2024 ASEE Annual Conference & Exposition*, 2024.
[13] R. F. Woolson, "Wilcoxon signed-rank test," *Encyclopedia of Biostatistics*, vol. 8, 2005.
[14] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.