

Work-in-Progress: PiazzaPlus–Enhancing Piazza Search with Semantic Matching

Ibraheem El Sheikha, University of Toronto

Ibraheem El Sheikha is an undergraduate student at the University of Toronto pursuing a BAsC in Computer Engineering.

Salma Emara, University of Toronto

Salma Emara is an Assistant Professor, Teaching Stream in the Department of Electrical and Computer Engineering at the University of Toronto. She received her B.Sc. in Electronics and Communications Engineering from the American University in Cairo in 2018, and her Ph.D. in Computer Engineering from the University of Toronto in 2022. Her Ph.D. research focuses on improving reinforcement learning algorithms to solve problems in computer networking algorithms. Currently, she is interested in building software-tools for programming education and pedagogical practices that build testing and debugging skills for beginner programmers.

Work-in-Progress: PiazzaPlus—Enhancing Piazza Search with Semantic Matching

1. Abstract

Many courses use discussion forums like Piazza to support learning outside the classroom, where students rely on built-in search tools, including Piazza's native search to find relevant information. However, these systems typically use keyword-based matching and lack semantic understanding. As a result, when students phrase their queries differently from how related posts are written, irrelevant threads are often retrieved, causing students to miss existing answers and ask duplicate questions. To address this gap, we propose PiazzaPlus: a hybrid semantic-keyword retrieval engine to help students find relevant answers. Our pipeline starts with preprocessing posts to clean up text, auto-captioning images using a large multimodal model (LMM) and concatenating captions with the text body of the post. Preprocessed posts are chunked into documents, which are then vectorized using an embedding model, and the generated vectors are stored in a vector database. At query time, a keyword search retrieves the top 100 keyword-similar posts, which are then re-ranked by a semantic stage based on similarity in meaning to the user's query. Our hybrid search is accessible via a browser extension interface. To date, we have extensively tested PiazzaPlus quantitatively. We evaluated retrieval quality using mean average precision (MAP), which is the average of precision scores at the ranks where relevant items occur, averaged across queries. We also used mean reciprocal rank (MRR), the average of the reciprocal of the rank of the first relevant result. Furthermore, we used average Precision@3 (P@3), which is the fraction of the top three results that are relevant. A human expert evaluating the relevance of suggested posts indicates that our engine achieved a 102.5% improvement in MAP, a 143.5% increase in MRR, and a 72.4% improvement in P@3 over Piazza's native search. An independent evaluation using an LLM-as-a-judge approach demonstrated corresponding improvements of 138.4%, 165.5%, and 102.7%, respectively. These results support our hypothesis that semantic-keyword integration resolves context-mismatch issues. We discuss qualitative instructor feedback based on students' informal comments. The feedback informs our future work to scale deployment, conduct field trials, collect formal feedback from students and explore LLM-driven summarization to further facilitate learning.

2. Introduction

Question-and-answer forums, such as Piazza, are now central to communication in large university courses due to their positive impact on "students' (i) academic performance, (ii) perceived levels of learning, and (iii) sense of being part of a learning community" [1]. However, a major limitation to the utility of these platforms depends on how quickly students and instructors can surface prior answers that match their specific wording, context, or images.

Piazza's built-in search is keyword-based: users enter terms and the system returns posts that contain particular keywords, with optional filters (unread, instructor posts, unresolved) [2]. While effective for exact term overlap, such keyword searching can miss relevant posts that use different phrasing or synonyms [3], [4], and ignores data captured by previously posted images altogether. This tends to produce long, noisy result lists in high-volume courses—conditions that encourage posting near-duplicate questions. Bhardwaj[5] argues that the high prevalence of semantically identical questions fragments knowledge across multiple threads, which impedes

efficient information retrieval, thereby degrading user experience. This motivates other approaches beyond keyword matching.

Semantic searching addresses this gap by representing queries and documents in a vector space and retrieving by similarity in meaning (e.g., paraphrases), rather than only exact word matches. Prior work [6] shows that pairing keyword searching with embedding-based retrieval often yields improved performance than either alone. As noted in [7], coupling semantic search with keyword filters streamlines the retrieval process. This hybrid method boosts precision and aids users who are exploring related themes or are unsure of the precise terminology.

To mitigate the limitations of the existing keyword-only searching mechanism on Piazza, we present PiazzaPlus, a browser extension that uses semantic matching to interpret the meaning behind user queries [8]. By implementing a semantic searching layer on top of a strong keyword searching layer, users can quickly surface the most contextually relevant posts to them reliably without the need for extra filtering. At query time, PiazzaPlus first creates a top-100 list using an attested keyword searching algorithm, then re-evaluates the list through a semantic perspective, making a general keyword match list much more targeted to the user's intention behind the query.

3. Motivation

Large courses generate many Piazza posts each term. When students search, they often know their problem but not the exact terminology. Keyword-only search thus fails when paraphrases don't match the wording in prior posts. Information captured by screenshots in prior posts are not reflected at all in search results, making it challenging to find relevant posts in many engineering courses. Since the list of relevant posts from native Piazza keyword-search is long and noisy, it causes students to post near-duplicate questions, burdening both students and instructors.

Keyword matching rewards literal token overlap, but students frequently search in synonyms or approximations. Pure semantic search helps but can over-generalize and surface off-topic results. Our motivation is to combine the strengths of both keyword and semantic searching to create the ideal search tool for Piazza. BM25 provides high precision on exact matches and established ranking behavior [9]; an embedding layer restores recall for paraphrases and conceptually similar posts, including content recovered from image captions [7].

4. Relevant and related work

Discussion forums such as Piazza are critical for engineering education, providing an outlet for asynchronous peer learning that has been shown to improve student engagement and course performance [1]. However, as courses grow, they suffer from the "continuous accumulation" problem: valuable answers get buried under hundreds of posts, forcing students to rely on search tools that are often inadequate [10].

The primary limitation of native forum search is the "Vocabulary Problem," a fundamental human-computer interaction challenge where users (students looking for old posts) rarely use the exact keywords chosen by authors (authors of old posts) [11]. When a student's query does not lexically match existing posts, keyword search fails, leading to the creation of duplicate questions. Research from Xu and Yuan [10] highlights that this redundancy "inevitably wastes manpower and time," making it harder for future students to find a definitive answer, while

increasing burden on instructors because they spend time answering, or at least reading, every version of the same question [12].

To mitigate this, recent work in engineering education has shifted toward semantic analysis. While earlier studies focused on using NLP for sentiment tracking, recent approaches use convolutional neural networks (CNNs) to measure the semantic similarity between a new user query and archived posts, significantly outperforming keyword baselines [10]. PiazzaPlus adapts a similar architecture for the educational domain, using hybrid retrieval to retrieve posts with different terminology.

5. Methodology

The following sections detail the technical components of the system, ranging from the development environment and system architecture to the specific data preprocessing and hybrid retrieval algorithms employed.

5.1 Development environment

PiazzaPlus was implemented as a browser extension run on a lightweight server backend. The backend authenticates to an application programming interface (API) to ingest posts from registered courses. Authentication is performed using course-specific "bot" accounts provided by course instructors; coordinators add these accounts to the relevant Piazza course, thereby granting access to PiazzaPlus to read that course's posts. The system then normalizes the post content and embeds post data using an embedding model, storing the resulting vectors in a vector database. To ensure high relevance, we retrieve the top 10 results found at the intersection of BM25 keyword matching and semantic vector search, ensuring posts are both textually and contextually aligned with the query.

5.2 System architecture

The system has two main parts, shown in Figure 1: a browser extension that provides the query interface and a backend server that manages ingestion, preprocessing, indexing, and searching. Course-specific login credentials are stored in a private configuration file maintained by the system administrator or course instructor. For each participating course, a dedicated Piazza bot account, added to the course by its coordinator, is used for access. This account is accessed by the backend to retrieve post data.

The API ingests new or updated posts and writes normalized records to a database maintained for that course. A database builder then processes this content, enriching it with image captions generated by an LMM (currently gpt-4o-mini) before inserting it in a Chroma vector database. This extensible architecture allows for seamless upgrades to future models. Searching is handled by the backend's retrieval service and exposed through HTTP endpoints that the extension calls directly.

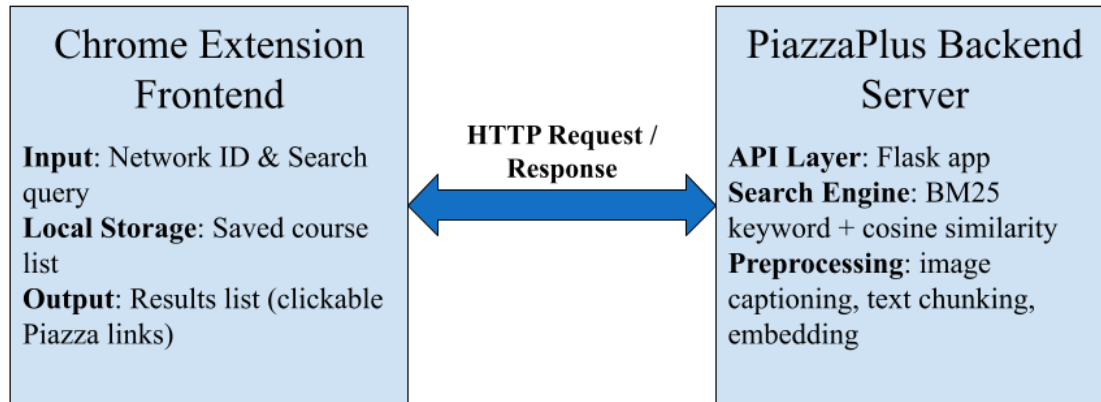


Figure 1: System architecture of PiazzaPlus

5.3 Data Preprocessing

In order to produce cleaner input for both keyword and semantic indexing, posts are first normalized by converting HTML to plain text and eliminating Markdown artifacts like inline images and link targets. The complete preprocessing workflow is illustrated in Figure 2. Redirects are used to first resolve URLs to stable CDN locations for postings with images. Then the following user prompt is used to fetch, base64-encode, and send each image to an image-aware multimodal model:

"Please describe this image to someone who is struggling in the course. Please describe all drawings and transcribe any text. Use up to 200 words."

In order to incorporate visual content (such as screenshots with equations) for both keyword search and semantic retrieval, created captions are appended to the post's text and stored as a separate field. A sentence tokenizer set up to produce one chunk per sentence (no overlap) is then used to chunk the caption-augmented text into sentences.

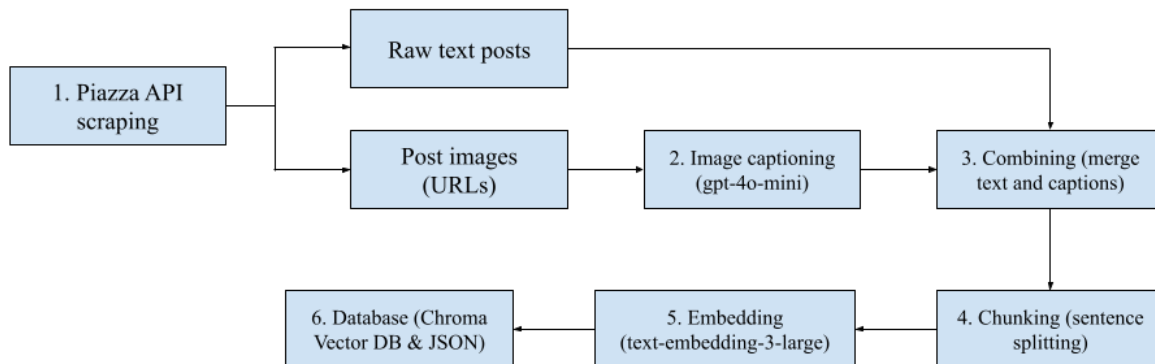


Figure 2: Data preprocessing pipeline for PiazzaPlus

5.4 Retrieval Pipeline

PiazzaPlus search works in two steps, as shown in Figure 3. First, a keyword stage (BM25) retrieves a candidate set by treating the query as a bag of terms and comparing those tokens to tokenized document text stored in the index; this emphasizes exact keyword overlap with length-normalization [9]. In our implementation, the query is tokenized and matched against the tokenized corpus to select the top-100 candidate post IDs.

Second, a semantic stage re-ranks these candidates by cosine similarity between the single query embedding and the sentence-level chunks of each candidate post stored in the vector index; for each post we keep the maximum chunk similarity as its score. This combined approach preserves high precision on direct keyword matches while recovering paraphrases and image-augmented text surfaced during preprocessing.

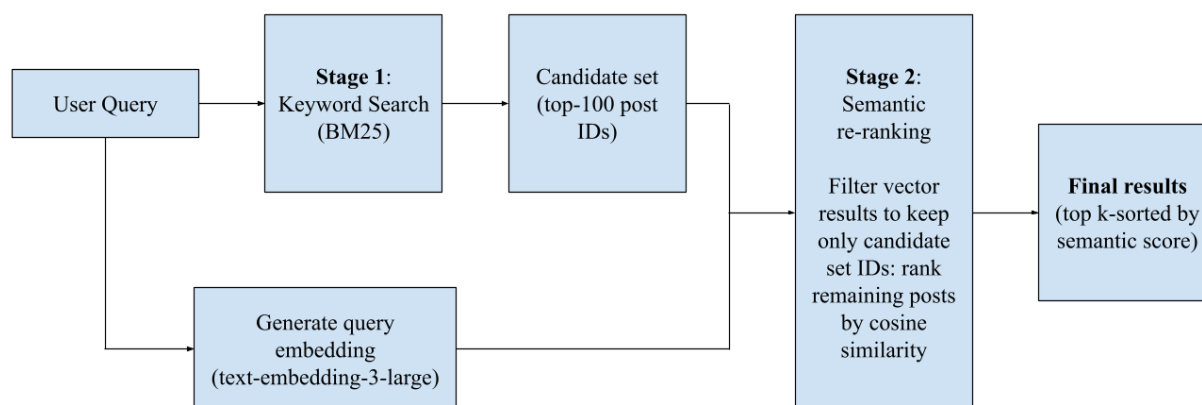


Figure 3: Query-time retrieval pipeline for PiazzaPlus

6. Chrome Extension Design

To ensure the tool fits seamlessly into a student's study workflow, the client interface was implemented as a lightweight Google Chrome extension. The user interaction flow consists of two stages: configuration and retrieval. Recent research by Shrivastava and Karmakar [13] shows that "users always prefer working with simple interface elements," noting that complex designs "can frustrate users in achieving their learning goals." Consequently, we prioritized a minimalist single-input interface to minimize cognitive load and maximize user acceptance.

6.1 Configuration and Persistence

When launched, the extension displays a configuration view where users input the target course's "Piazza Network ID." Supported courses are stored in the browser's local storage, creating a "Saved Courses" list, shown in Figure 4(a). This design choice is critical for usability; it ensures that returning students can access the search interface with a single click, without needing to re-enter course codes.

6.2 Search Interface

Once a course is selected, the view transitions to the query interface, shown in Figure 4(b). Unlike Piazza's native search, which offers filtering options (e.g., instructor posts, unread posts), the PiazzaPlus interface provides a single, minimalist text area. This constraint encourages students to input natural language questions rather than keyword fragments. When a search is executed, the extension sends the query and course ID to the backend via an HTTP POST request.

6.3 Result Presentation

The results are rendered as a list of hyperlinks containing the post ID and subject line. To facilitate rapid browsing, clicking a result opens the corresponding Piazza thread in a new background tab, allowing the student to maintain their search context in the popup while exploring multiple potential answers.

Overall, the decoupled architecture of PiazzaPlus ensures that the client remains lightweight, as all heavy computation, including embedding generation and vector similarity search, is offloaded to the backend server.

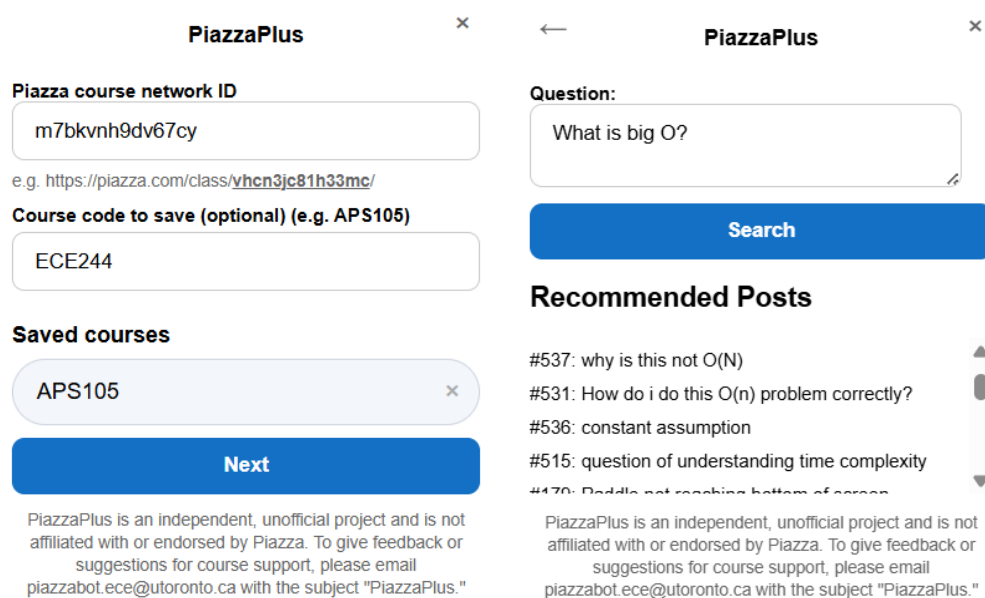


Figure 4(a) (left): Configuration screen of PiazzaPlus; Figure 4(b) (right): Query and results screen of PiazzaPlus

7. Evaluation

PiazzaPlus was evaluated using a benchmark of 40 text-only queries from APS105: Computer Fundamentals, a first-year programming course at the University of Toronto. Queries were paraphrased or created by the authors based on course experience, rather than copied directly from Piazza. To establish initial ground-truth performance, a human evaluator independently

judged the retrieved results, blind to the source system. A post was marked as relevant if it provided a complete or partial answer to the query, and irrelevant if it did not. Following this manual evaluation, we employed an LLM-as-a-judge (gpt-4o-mini) to serve as an independent second rater, a methodology shown to reliably approximate human preference and relevance judgments [14]. The model was prompted to act as an expert teaching assistant and was supplied with the query alongside the full thread context: the post title, original question body, and, if available, the instructor and endorsed student answers. It was instructed to strictly output a binary 'RELEVANT' or 'NOT_RELEVANT' label based on whether the thread provided a complete or partial resolution to the query. Due to the differing retrieval behaviors of the two search tools, the total number of evaluated trials varied for both raters. PiazzaPlus always retrieved 10 results per query, of which the top three were analyzed, yielding a full set of 120 trials. Conversely, native Piazza often returned fewer than three results, including instances that yielded 'NO_RESULT,' resulting in only 41 valid trials for evaluation. To measure the inter-rater agreement between the human and the LLM, we calculated raw agreement and Cohen's Kappa separately for each system. The analysis yielded a Cohen's Kappa of 0.65 (83% raw agreement) for the native Piazza baseline, and 0.79 (90% raw agreement) for PiazzaPlus. It should be noted that the smaller sample size for native Piazza (41 trials versus 120) may have an effect on the stability of its agreement scores. Nevertheless, these scores indicate high agreement [15], confirming that the relevance judgments are consistent between raters.

Both systems were tested on the same queries. PiazzaPlus returned the top 10 ranked results using its hybrid BM25 and embedding-based pipeline, while Piazza's native search returned all keyword matches in its "Relevance" priority mode (as opposed to its "Date" chronological mode). Retrieval effectiveness was measured with three metrics: mean average precision (MAP), mean reciprocal rank (MRR), and precision at rank three (P@3). These were selected to reflect realistic student use, where users typically review only the first few results. Efficiency and latency were not measured.

7.1 Results and Discussion

Figure 5 summarizes the retrieval performance of Piazza's native keyword search and PiazzaPlus on the APS105 benchmark. Across all three metrics, PiazzaPlus substantially outperformed the baseline.

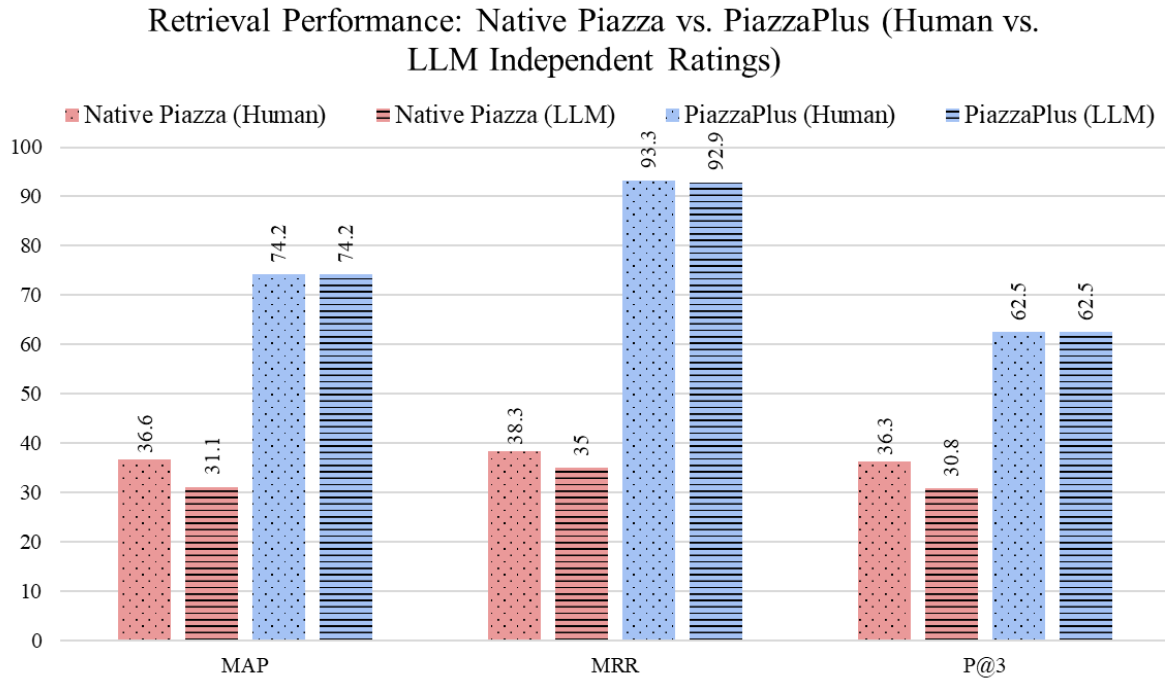


Figure 5: Retrieval performance comparison between native Piazza and the proposed PiazzaPlus system

PiazzaPlus more than doubled the MAP score, indicating that relevant posts were consistently retrieved in higher positions. Its MRR approached 1.0, which means the first relevant result almost always appeared at the top of the search results. P@3 also improved significantly. This suggests that students using PiazzaPlus often found a useful answer without having to look beyond the first few results.

Image retrieval was another area of improvement. Piazza's search does not support image content at all. In contrast, PiazzaPlus could index captioned images from past posts and return them accurately for relevant queries. While formal image retrieval metrics are not reported in this work-in-progress (WIP), preliminary qualitative testing confirms that the image captioning pipeline successfully retrieves posts containing only screenshots of code or diagrams—content that is invisible to the native keyword search.

Overall, these results suggest that PiazzaPlus greatly improves accuracy compared to Piazza's native search. This is especially true for paraphrased questions and posts with visual content. These improvements may lead to fewer duplicate questions, lighten the workload for instructors, and make it easier for students to find older discussions. We acknowledge specific limitations in this preliminary evaluation protocol. First, the study was conducted on a single introductory programming course (APS105), which may limit generalizability to non-technical or writing-intensive domains. Second, the use of 40 author-generated queries—while derived from genuine course concepts—may not fully capture the variability of real-time student search behavior. Future work will expand this evaluation to include diverse course datasets and live user query logs to further validate these findings.

7.2 Classroom Integration and Instructor Feedback

Following the offline evaluation of PiazzaPlus, we deployed the tool in a second-year programming course running live during the Fall 2025 term, with an enrolment of approximately 400 students. From the instructor’s perspective, PiazzaPlus was positioned as a supplementary search tool to encourage students to consult existing discussion threads before posting new questions. Informal instructor observations suggest that students used the tool primarily to search for previously asked questions relevant to their own issues prior to submitting a post on Piazza, as well as to relocate posts they had found earlier but did not bookmark. Several students reported that they were able to find answers to their questions through PiazzaPlus and, as a result, chose not to post a new question, reducing redundancy in the discussion forum. In addition, a small number of students provided informal feedback on the user interface, specifically suggesting that the course code be stored in the browser’s local storage so that it would persist across sessions and not need to be re-entered each time the extension is used. While this feedback was anecdotal and not collected through a formal user study, it provided early insight into both the perceived usefulness of the tool and opportunities for improving usability in future iterations.

8. Future Work

To validate the generalizability of PiazzaPlus, future evaluations will expand beyond programming-centric courses to include more academic subjects such as mathematical, writing-based, and non-technical courses. We plan to administer formal student surveys to quantitatively measure user satisfaction and the tool’s perceived impact on learning efficiency. Technically, we aim to evolve the retrieval pipeline to support multimodal inputs, allowing students to search using images directly to find relevant past posts. Finally, we will continue to iteratively refine the user interface by integrating this formal survey data with the informal verbal feedback received during initial deployments to ensure a smooth student experience.

9. Conclusion

In large educational courses, students frequently struggle to find existing answers on discussion forums because native search tools rely heavily on exact keyword matching, leading to missed information and duplicate questions when terminology differs. To address this problem, this WIP paper introduces PiazzaPlus, a hybrid retrieval engine that combines traditional keyword search (BM25) with semantic embeddings to understand the actual meaning behind student queries. Our preliminary evaluation demonstrates that this semantic-keyword integration successfully resolves the context-mismatch issues inherent in student searches. Specifically, PiazzaPlus significantly outperformed the native keyword search, achieving a 102.5% improvement in MAP as observed by a human evaluator. An independent LLM-as-a-judge reinforced these substantial gains with a 138.4% improvement in MAP, demonstrating high inter-rater agreement. Ultimately, these findings confirm that hybrid semantic retrieval is a highly viable and effective solution for improving information accessibility in educational forums.

Acknowledgements

This research was supported by the Edgar McAllister Foundation Undergraduate Summer Research Award within the The Edward S. Rogers Sr. Department of Electrical & Computer Engineering at the University of Toronto. The authors owe a very special thanks to Hassan El-Sheikha, whose valuable assistance with code development and paper proofreading were indispensable to the success of this work. The authors also gratefully acknowledge Roland Tan and Jaro Pristupa for providing the server infrastructure, system administration, and essential technical support that made the deployment of PiazzaPlus possible.

References

- [1] A. Squires and D. Rigby, "A mixed methods evaluation of online discussion tools in higher education," *International Review of Economics Education*, vol. 47, p. 100298, 2024.
- [2] "Search for posts," Piazza. [Online]. Available: <https://support.piazza.com/support/solutions/articles/48000616747-search-for-posts> [Accessed Oct. 1, 2025].
- [3] W. Chen et al., "Deep Learning for Instance Retrieval: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 6, pp. 7270–7292, 2023.
- [4] N. Nagpal, "Query Expansion for Information Retrieval using Word Embeddings: A Comparative Study," in *2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS)*, 2022, pp. 289–293.
- [5] S. Bhardwaj, "Semantic similarity in community forum questions: Case study on Quora dataset," *Journal of Umm Al-Qura University for Engineering and Architecture*, vol. 16, no. 4, pp. 1719–1728, 2025.
- [6] A. Ahluwalia, B. Sutradhar, K. Ghosh, I. Yadav, A. Sheetal, and P. Patil, "Hybrid Semantic Search: Unveiling User Intent Beyond Keywords," *arXiv [cs.IR]*, 2024.
- [7] R. Ishmam, S. Emara, and H. Timorabadi, "Work-in-Progress: TextMatch - A Semantically Enhanced Textbook Recommendation System," in *2025 ASEE Annual Conference & Exposition*, 2025.
- [8] B. Latard, J. Weber, G. Forestier, and M. Hassenforder, "Towards a Semantic Search Engine for Scientific Articles," in *Research and Advanced Technology for Digital Libraries*, 2017, pp. 608–611.
- [9] S. Robertson and H. Zaragoza, "The Probabilistic Relevance Framework: BM25 and Beyond," *Foundations and Trends in Information Retrieval*, vol. 3, pp. 333–389, Jan. 2009.
- [10] Z. Xu and H. Yuan, "Forum Duplicate Question Detection by Domain Adaptive Semantic Matching," *IEEE Access*, vol. 8, pp. 56029–56038, 2020.
- [11] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, Nov. 1987.
- [12] K. Iyer, G. Ewoorkar, A. Patil, H. Gummani, and A. Parkavi, "Efficient Model for Searching and Detecting Semantically Similar Question in Discussion Forums of e-learning Platforms," in *2019 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT)*, 2019, pp. 1113–1118.

- [13] S. Syamala, A. Shrivastava, and S. Karmakar, "Identifying interface design factors impacting user experience in digital learning platforms- A pilot study," *Social Sciences & Humanities Open*, vol. 11, p. 101662, 2025.
- [14] L. Zheng et al., "Judging LLM-as-a-judge with MT-bench and Chatbot Arena," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, New Orleans, LA, USA, 2023.
- [15] M. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica*, pp. 276–282, Jan. 2012.