Real-Time Query Management for Large Scale Online Classes

Kamal Bijlani , Sreya Sunil * , Navabhaarathi Asokan , Saran Dharshan SP , Vigneshwar E , Aadharsh Aadhithya A

Posted Date: 28 November 2024

doi: 10.20944/preprints202411.2232.v1

Keywords: Scalable Online Learning; Query Similarity Analysis; Automated Query Handling

 doi:10.20944/preprints202411.2232.v1

Article

Real-Time Query Management with FAQ Generation for Large Scale Online Classes

Kamal Bijlani [1], Sreya Sunil [1], Navabhaarathi [2], Saran Dharshan S.P [2], Vigneshwar E [2] and Aadharsh Aadhithya A [2]

[1] School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Faridabad, India
[2] School of Artificial Intelligence, Amrita Vishwa Vidyapeetham, Coimbatore, India
* Correspondence: sreya@amritaai.org

Abstract. This study introduces an AI-driven approach for real-time query management and FAQ generation, addressing the challenges of scalability and immediate responsiveness in large-scale online learning environments. We introduce a sophisticated student query-handling pipeline that augments the platform's responsiveness and scalability by embed- ding course material into a high-dimensional vector space, organizing it through topic clustering, and constructing an FAQ database pre-validated by instructors. During live sessions, student queries are embedded and evaluated against the FAQ (Frequently Asked Question) knowledge base using a dissimilarity metric. Queries exceeding a predefined dissimilarity threshold are escalated to instructors, while lower-dissimilarity queries are addressed by a Retrieval-Augmented Generation (RAG) model, op- timized to ensure minimal latency and high accuracy in real-time re- sponses.

Keywords: scalable online learning; query similarity analysis; au- tomated query handling

1. Introduction

In recent years, online learning has firmly established itself as a central mode of educational delivery, offering scalable and adaptable solutions that transcend traditional geographical and infrastructural limitations. With advancements in technology, digital classrooms have become essential in democratizing access to education, facilitating seamless participation of hundreds or even thousands of students concurrently. However, as observed in large-scale deployments such as the T10KT project initiated by IIT Bombay through the A-View platform [1], managing real-time interactivity and ensuring sustained engagement within such expansive settings pose significant challenges. These issues are exacerbated by connectivity constraints and high student-to-teacher ratios, which hinder personalized interactions and timely feedback.

This challenge in large-scale online classes is compounded by the overwhelming influx of questions during live sessions, often outstripping instructors' capacity for immediate responses. As Wang and Woo (2007) [2] explored, synchronous and asynchronous communication modalities in online learning settings each present distinct benefits and limitations in terms of interaction and engagement. To address these challenges, recent studies advocate for the integration of artificial intelligence (AI) in education, particularly in areas such as natural language processing (NLP) and machine learning (ML), which enable intelligent automation in managing student interactions. Notably, Heilman and Smith (2010) [3] introduced a statistical approach to automatic question generation, which informs the basis for frequently asked questions (FAQ) systems that dynamically adapt to students' inquiries. In response to these needs, our research proposes a system for real-time query management with FAQ generation that leverages AI to manage and prioritize student questions efficiently during live sessions. By combining NLP techniques for question classification with ML algorithms for prioritizing queries, our approach ensures that repetitive questions are managed through automated responses, while unique or contextually complex queries are escalated to instructors for direct

 doi:10.20944/preprints202411.2232.v1

engagement. Through this integration, the proposed system aims to streamline the interactive capacity of online platforms, enhancing the quality of interaction without burdening instructors, thus addressing scalability challenges in online education and promoting educational equity and access.

2. Related Work

The foundation of this study is built upon several influential works in the domain of AI-enhanced educational technology. Research on large-scale online learning environments, such as those highlighted by Luckin et al. (2016) [4], underscores the potential of AI to support personalized and adaptive learning. Their study emphasizes AI's capacity to cater to diverse learning needs, a premise that informs the development of AI-driven solutions for query management in our system. The integration of real-time query handling mechanisms in educational settings has gained traction with the rise of MOOCs and other digital learning platforms. Drachsler and Kalz (2016) [5] introduced the MOOC and Learning Analytics Cycle (MOLAC), identifying challenges in data integration and student privacy, which are critical considerations in designing our query-handling system. Similarly, McInnes and Pedersen (2013) [6] explored semantic similarity and relatedness metrics to enhance word sense disambiguation, contributing foundational techniques for improving automated FAQ systems in educational contexts. Building on the advancements in automated question-answer generation, Aithal, Rao, and Singh (2021) developed mechanisms for generating question-answer pairs and assessing question similarity in educational platforms. These methodologies, particularly the use of cosine similarity for filtering questions, have been instrumental in our design, helping to ensure that only contextually relevant questions are prioritized for instructor review [7]. Furthermore, in large-scale blended classrooms, hybrid architectures such as those explored in anonymous case studies demonstrate the importance of scalable and flexible learning designs. This case study highlights the practical considerations in balancing real-time feedback and engagement in mixed learning environments, offering valuable insights for the development of our own system. In addition, Al-Zahrani and Alasmari (2024) [8] examined the operational and ethical implications of AI in education, raising critical considerations for data privacy and responsible AI use in learning settings. These ethical frameworks are essential to our approach, ensuring that the deployment of AI in query management adheres to privacy standards and fosters an inclusive learning environment. In summary, this study aims to contribute to the growing field of AI in education by presenting a sophisticated AI-driven query management pipeline, designed to enhance large-scale online class environments. Through automated FAQ generation and similarity-based filtering, our system effectively addresses the core challenges of online learning, facilitating high-quality, real-time student engagement and personalized instruction. The proposed methodology exemplifies a sustainable approach to managing large volumes of student inquiries, supporting educational scalability, and promoting equitable access across digital platforms

3. Methodology

In our system(Fig. 1), instructors can upload textbooks or course material directly into the platform as a primary knowledge source. For initial development and testing, we selected Data Communications and Networking by Behrouz Forouzan [9], a widely taught textbook in undergraduate programs at Amrita University. This textbook offers a structured, modular approach to understanding core networking concepts, which aligns well with our objectives for building topic-specific knowledge clusters. The selected chapters serve as the foundation for the FAQ generation process, and embeddings are created for each section to facilitate topic clustering and subsequent retrieval.

3.1. Semantic Embedding

 doi:10.20944/preprints202411.2232.v1

To ensure optimal content representation, the textbook content is processed chapter-by-chapter, allowing for a logical segmentation of ideas. Key concepts, phrases, and sentences are identified and extracted from each chapter, thus creating a focused dataset that retains the contextual essence of each topic. This approach is vital for ensuring the integrity of subsequent embeddings, as each concept is aligned with its original chapter context. The next step is to transform these extracted concepts into high-dimensional semantic vectors using the paraphrase-mpnet-base-v2 model [10], a transformer-based model optimized for capturing subtle semantic relationships between phrases and sentences. The choice of this model is based on its proven effectiveness in generating dense, context-rich embeddings that reflect the deeper connections among similar concepts. In this configuration, each concept is encoded as a 768-dimensional vector, where semantic similarity is preserved through proximity in the vector space. These embeddings form the foundation for effective clustering and similarity comparisons, as they allow each question-answer pair to be mapped within the same space. By structuring the embeddings on a chapter basis, the system ensures that retrieval and clustering are efficient and that chapter-specific alignment is readily achievable.

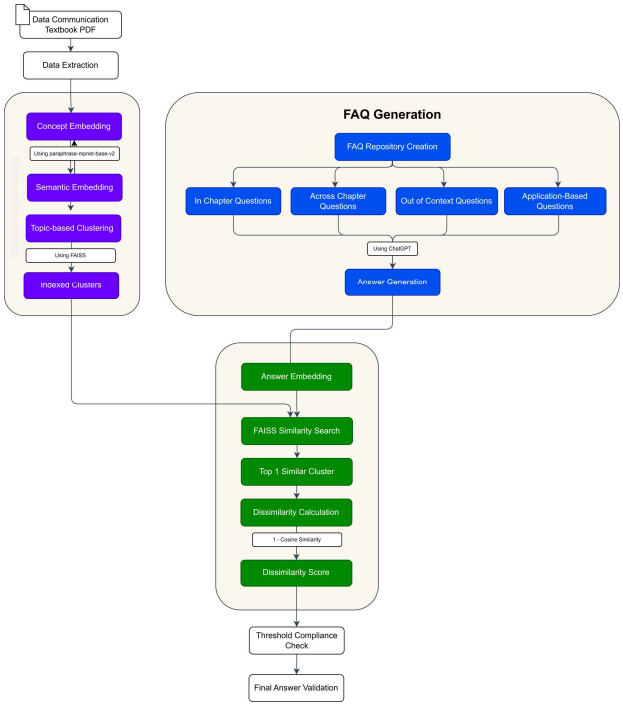

Figure 1. Proposed Pipeline.

3.2. Organizing Content through Clustering and FAISS Indexing

With embedded concepts prepared, the methodology proceeds to organize these vectors into clusters, utilizing the FAISS (Facebook AI Similarity Search) [11] library for scalable and efficient similarity-based retrieval. The primary goal here is to structure the high-dimensional space into conceptually cohesive groups, or clusters, that reflect the textbook's thematic organization. To accomplish this, we employ the k-means clustering algorithm within FAISS. K-means was selected for its suitability in handling large datasets and compatibility with vector-based representations, making it an ideal choice for organizing embedded concepts. Each cluster in this system represents semantically related concepts, and each cluster has a centroid vector that acts as a central reference point. This centroid vector plays a crucial role in subsequent similarity searches, where it is used to locate the most contextually appropriate cluster for a given query. The clustering process is further optimized by

 doi:10.20944/preprints202411.2232.v1

adjusting the number of clusters, or k, on a per-chapter basis. Chapters with dense or diverse content are segmented into a higher number of clusters to capture the full range of ideas, while simpler chapters are divided into fewer clusters to prevent excessive fragmentation. This configuration not only preserves chapter-specific context but also enhances the FAQ system's ability to handle detailed queries within specific chapters. The distance metric used for clustering is L2 (Euclidean distance), which complements the cosine similarity metric employed in dissimilarity measurements, creating a cohesive system that integrates distance and similarity for accuracy and efficiency. After clustering, the FAISS index is fine-tuned for fast retrieval. One critical parameter in FAISS, nprobe, which controls the number of clusters scanned during similarity searches, is adjusted to achieve an optimal balance between retrieval speed and accuracy. Higher values of nprobe increase accuracy by scanning more clusters but can lengthen processing time. In this system, the nprobe setting is customized to meet the project's real-time response requirements, with additional adjustments based on user query characteristics. [12]

3.3. FAQ Repository Compilation and Response Generation

The system's FAQ repository is meticulously curated to cover a range of question types, each associated with different expectations of alignment and contextual relevance. Four primary categories are defined: in-chapter, across-chapter, out-of-context, and application-based questions. Each category is crafted with unique alignment expectations, thus allowing the system to tailor its response generation approach. In-chapter questions are limited to single chapters, aligning closely with specific content and expected to have low dissimilarity scores. Across-chapter questions require cross-referencing multiple sections, resulting in slightly higher dissimilarity ranges that allow for some deviation in context. Outof-context questions necessitate knowledge beyond the textbook, producing high dissimilarity scores due to their broader knowledge requirements. Applicationbased questions extend textbook concepts to real-world scenarios, similarly showing higher dissimilarity due to the broader application scope. For answer generation, OpenAI's ChatGPT [13] is employed to ensure that responses are contextually relevant. ChatGPT accesses chapter-specific embeddings to provide content alignment for in-chapter and across-chapter questions, ensuring that the system can generate responses that reflect the original textbook material. To maintain high alignment, each generated answer is anchored within the context of related chapter embeddings, a process that constrains ChatGPT's output, reducing the likelihood of off-topic information and ensuring consistency with the original material. Additionally, for critical questions, particularly in out-of-context or application-based categories, generated answers undergo post-processing and are subject to human review to verify relevance and accuracy.

3.4. Dissimilarity Measurement and Threshold Establishment

The measurement of dissimilarity is essential for evaluating the alignment of generated answers with textbook content, providing a quantitative basis for threshold compliance. For each generated answer, an embedding vector is created, and the dissimilarity score [14] is calculated by comparing it with the most similar cluster centroid retrieved from the FAISS index. This comparison utilizes the 1 cosine similarity formula, where lower dissimilarity scores indicate a stronger alignment with the textbook content.

The system defines a dissimilarity threshold for each question category to assess compliance. In-chapter questions are expected to have a dissimilarity range between 0.2 and 0.4, reflecting close alignment. Across-chapter questions allow for a moderate range of 0.4 to 0.5, accommodating cross-referenced answers. For out-of-context and application-based questions, dissimilarity scores above 0.5 are acceptable due to the broader knowledge required. Threshold compliance is measured as the percentage of answers in each category that meet the expected dissimilarity range, providing a clear indication of the system's accuracy in generating contextually relevant answers.

 doi:10.20944/preprints202411.2232.v1

To quantify the alignment between a generated answer and the textbook content, we define the Dissimilarity Score based on cosine similarity.[15] Let:

- $\mathbf{v}_A$ be the embedding vector of the generated answer.
- $\mathbf{v}_C$ be the embedding vector of the closest cluster centroid (retrieved from
- the FAISS index).

The **Cosine Similarity** between these vectors is calculated as:

$$\text{Cosine Similarity}(\mathbf{v}_A, \mathbf{v}_C) = \frac{\mathbf{v}_A \cdot \mathbf{v}_C}{\|\mathbf{v}_A\| \cdot \|\mathbf{v}_C\|}$$

where:

- $\mathbf{v}_A \cdot \mathbf{v}_C$ is the dot product of vectors $\mathbf{v}_A$ and $\mathbf{v}_C$,
- $\|\mathbf{v}_A\|$ and $\|\mathbf{v}_C\|$ are the magnitudes (Euclidean norms) of the vectors $\mathbf{v}_A$ and
- $\mathbf{v}_C$, respectively.
- The cosine similarity yields a value between -1 and 1, where:
- A value of 1 indicates maximum similarity,
- A value of 0 indicates no similarity,
- A value of -1 indicates maximum dissimilarity.

To obtain a measure of dissimilarity, we subtract the cosine similarity from 1. Thus, the **Dissimilarity Score** is defined as:

$$\text{Dissimilarity Score} = 1 - \frac{\mathbf{v}_A \cdot \mathbf{v}_C}{\|\mathbf{v}_A\| \cdot \|\mathbf{v}_C\|}$$

Expanding this, we have:

Thresholds for Dissimilarity Scores by Question Category

Each question type has a specific range for acceptable dissimilarity scores:

- **In-Chapter Questions**: $0.2 \leq$ Dissimilarity Score ≤ 0.4
- **Across-Chapter Questions**: $0.4 \leq$ Dissimilarity Score ≤ 0.5
- **Out-of-Context and Application-Based Questions**: Dissimilarity Score > 0.5

A compliance check is performed by evaluating whether the dissimilarity score of each answer falls within the specified range for its question category. The compliance rate (as a success metric) is defined as:

$$\text{Compliance Rate} = \frac{\text{Number of Answers Within Threshold}}{\text{Total Number of Answers in Category}} \times 100$$

Table 1 provides an overview of dissimilarity ranges, mean, median, standard deviation, and the number of questions for each question category. These ranges define the expected alignment levels of each question type with the textbook content, aiding in threshold compliance assessment. The dissimilarity scoring framework not only establishes alignment thresholds but also enables efficient query handling by organizing incoming questions in descending order of dissimilarity.

Table 1. Summary of dissimilarity ranges, mean, median, standard deviation (StD), and question count for FAQ categories.

Category	Range	Mean	Median	StD	No. of Questions
Out of context	0.893 0.935	0.919	0.922	0.011	30
Application Based	0.881 0.946	0.912	0.909	0.016	31
Cross-Chapter	0.337 0.530	0.469	0.515	0.078	9
In-Chapter	0.318 0.540	0.475	0.510	0.066	42

Table 2. Summary of Expert Score Dissimilarity Statistics for Question Categories.

 doi:10.20944/preprints202411.2232.v1

Label	Mean	Median	Mode
Out of Context	0.740000	0.7	0.7
Application Based Questions	0.758065	0.8	0.7
Cross-Chapter	0.255556	0.2	0.2
In-Chapter	0.240476	0.2	0.2

The comparison between GPT-derived Dissimilarity Scores(Table 1) and expert-assigned scores(Table 2) shows a general alignment in relative values, with GPT scores consistently trending higher across categories. This indicates that GPT's model assesses questions as slightly more dissimilar compared to expert evaluations. For the Application-Based and Out-of-Context categories, both GPT and expert scores reflect higher mean values (0.9117 vs. 0.7581 for Application-Based, and 0.9194 vs. 0.7400 for Out of Context). This alignment suggests these types are perceived as more contextually detached from the core content, implying that they draw on broader or external knowledge applications. In contrast, Cross-Chapter and In-Chapter questions show lower mean dissimilarity scores in both methods. Notably, the expert-assigned scores are significantly lower (0.4691 vs. 0.2556 for Cross-Chapter and 0.4752 vs. 0.2405 for In-Chapter), reflecting a stronger contextual alignment. This supports the expectation that questions within or closely related to a single chapter naturally exhibit higher similarity to the core material. This overall pattern reinforces the validity of both scoring methods while highlighting GPT's tendency to slightly amplify dissimilarity, especially for questions extending beyond direct content boundaries.

3.5. Real-Time Query Handling and Response Flow

During live online sessions, incoming queries from students are processed in real time by the question management system. Each query is first embedded and evaluated for dissimilarity against the FAQ knowledge base using the FAISS index. This dissimilarity scoring mechanism enables the system to rank each query in descending order by relevance, as outlined in Table 2. Queries that exhibit high dissimilarity scores, meaning they fall outside the FAQ content, are automatically flagged and rerouted to the professor for direct handling. These high-dissimilarity queries often indicate unique or complex questions that require specialized instructor input, beyond the scope of the automated system. Conversely, questions with low dissimilarity scores, which closely match the FAQ knowledge base, are managed autonomously by the Retrieval-Augmented Generation (RAG) model. The RAG model leverages pre-existing FAQ clusters and dynamically synthesizes contextually relevant answers for low-dissimilarity queries. By efficiently responding to common questions in real-time, the system reduces the workload on instructors, allowing them to focus on more unique or challenging inquiries. The arrangement of questions by dissimilarity score (see Table 2) ensures that routine queries are handled by the automated system, while complex questions receive prompt instructor attention.

RAG serves as a post-retrieval layer that enriches initial answer suggestions pulled from the clusters. This is crucial, as RAG can synthesize responses that are not just semantically similar but also contextually nuanced, leading to a more refined answer that fits both in-chapter and across-chapter contexts. Upon retrieval, RAG applies quality control metrics to refine responses, ensuring that generated answers meet critical relevance and precision standards. The Answer Relevancy Score (88.65%) prioritizes responses that directly align with the query, confirming that the answer is both accurate and useful. The Contextual Relevancy Score (42.5%) assesses the depth of alignment with broader textbook themes, ensuring that responses contribute meaningfully within a wider conceptual scope. Additionally, the Contextual Precision Score (88.25%) verifies that responses are finely tuned to chapter-specific content, which is essential for questions requiring tight content alignment. Finally, the Hallucination Score (74.6%) helps limit the inclusion of non-relevant or speculative information, maintaining strict adherence to the textbook's factual content.

 doi:10.20944/preprints202411.2232.v1

Table 3. Sample questions organized by dissimilarity scores. Each question is labeled based on its alignment category, with dissimilarity scores indicating the degree of contextual relevance to the course material.

Chapter Name	Question	Label	DisScore
Chapter 4: Digital Transmission	If designing a video streaming platform, how would you approach bandwidth allocation?	Application Based questions	0.946
Chapter 5: Analog Transmission	How would you implement QAM in a high-speed internet environment?	Application Based questions	0.941
Chapter 5: Analog Transmission	How could AM be leveraged in emergency broadcasting scenarios?	Out of context	0.935
Chapter 6: Bandwidth Utilization	For a telemedicine application with varying bandwidth requirements, how would you manage data?	Out of context	0.934
Digital Transmission	How does Manchester encoding provide synchronization in data communication?	In-Chapter	0.540
Digital Transmission	What are the two primary transmission modes in digital communication?	In-Chapter	0.538
Digital Transmission	What are the challenges of long-distance digital communication?	Cross-Chapter	0.530
Bandwidth Utilization: Multiplexing and Spreading	What are the advantages of Code Division Multiplexing?	Cross-Chapter	0.529

4. Conclusion

This paper introduces an initial framework for automated query management within large-scale virtual classrooms, utilizing semantic embeddings, topic clustering, and dissimilarity scoring to improve the efficiency and accuracy of query responses. By categorizing questions based on their alignment with a curated FAQ database, the system enables the RAG model to handle frequent, routine questions autonomously, while higher-dissimilarity queries are escalated to instructors. Future developments will include replacing traditional models with LLMs to automate response generation fully, further reducing response times and enhancing contextual accuracy. Additionally, more sophisticated topic clustering algorithms will be developed to refine query categorization, and adaptive threshold tuning will be explored to dynamically adjust dissimilarity thresholds based on the content and complexity of incoming queries. Real-time content updates and continuous learning mechanisms will also be integrated, allowing the system to evolve with new material and improve its response capabilities over time. Ultimately, this progression will enable the platform to support complex, cross-disciplinary queries, fostering a more comprehensive and adaptive virtual learning environment.

References

1. Amrita e-Learning Research Lab, "A-VIEW: Real-time Collaborative Multimedia E-Learning," in *Proceedings of the 2011 ACM Multimedia Conference and CoLocated Workshops ACM International Workshop on Multimedia Technologies for Distance Learning*, MTDL'11, 2011.
2. Q. Wang and H. L. Woo, "Comparing asynchronous online discussions and face-toface discussions in a classroom setting," *British Journal of Educational Technology*, 2007.
3. M. Heilman and N. A. Smith, "Question generation via overgenerating transformations and ranking," tech. rep., 2009.
4. R. Luckin and W. Holmes, "Intelligence unleashed: An argument for ai in education," 02 2016.
5. H. Drachsler and M. Kalz, "The mooc and learning analytics innovation cycle molac: a reflective summary of ongoing research and its challenges," *J. Comp. Assist. Learn.*, 2016.
6. B. T. McInnes and T. Pedersen, "Evaluating measures of semantic similarity and relatedness to disambiguate terms in biomedical text," *Journal of Biomedical Informatics*, 2013.
7. S. Aithal, A. Rao, and S. Singh, "Automatic question-answer pairs generation and question similarity mechanism in question answering system," *Applied Intelligence*, 2021.
8. A. Al-Zahrani and T. Alasmari, "Exploring the impact of artificial intelligence on higher education: The dynamics of ethical, social, and educational implications," *Humanities and Social Sciences Communications*, 2024.
9. B. A. Forouzan, *Data Communications and Networking (McGraw-Hill Forouzan Networking)*. McGraw-Hill Higher Education, 2007.

 doi:10.20944/preprints202411.2232.v1

10. N. Reimers and I. Gurevych, “paraphrase-mpnet-base-v2,” 2020.
11. M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli,
12. L. Hosseini, and H. Jégou, “The faiss library,” 2024.
13. S. Kale, G. Khaire, and J. Patankar, “Faq-gen: An automated system to generate domain-specific faqs to aid content comprehension,” 2024.
14. OpenAI, J. Achiam, S. Adler, and S. Agarwal, “Gpt-4 technical report,” 2024.
15. F. Lan, “Research on text similarity measurement hybrid algorithm with term semantic information and tf-idf method,” *Hindawi*, 2022.
16. D. Gunawan *et al.*, “The implementation of cosine similarity to calculate text relevance between two documents,” 2018.