A Retrieval-Augmented Question Answering System Using BERT for Stream-Specific Educational Content

Submitted in partial fulfillment of the
requirements of the Degree: M.Tech in Data science and engineering

By

Jaykumar Chaudhary

2022DC04341

Under the supervision of

Mr. Lalkar Eknath Chhadawelkar

Technical Evangelist (Cybage Software, Pune)

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE

Pilani (Rajasthan) INDIA

MAY, 2025

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

SECOND SEMESTER 2024-25

DSECLZG628T / AIMLCZG628T DISSERTATION

Name of Supervisor: Mr. Lalkar Eknath Chhadawelkar

Name of Student: Jaykumar Chaudhary

ID No. of Student: 2022DC04341

Courses Relevant for the Project & Corresponding Semester:

1. **Deep Leaning (Semester 3)**
2. **Natural language Processing (Semester 3)**

Abstract

In academic environments, students often struggle to quickly locate specific information across large, text-heavy educational documents such as course handouts. As elective subjects vary across semesters, navigating this material can be overwhelming, especially when trying to understand topics or explore subjects before making elective decisions. Traditional keyword-based search methods lack contextual understanding, leading to imprecise and inefficient information retrieval.

This dissertation proposes the design and development of a **Retrieval-Augmented Question Answering (QA) system** that leverages a **pretrained BERT-based model** to enable students to query academic content in natural language and receive accurate, contextually relevant answers. The system is primarily designed to process structured and semi-structured educational content — including course handouts — from the **Data Science** and **AI/ML streams**.

The proposed solution follows a **Retrieval-Augmented Generation (RAG) architecture** that separates the QA process into two components: **retrieval** and **reading**. First, a **retriever** uses **Sentence-BERT** to generate **embeddings** and identify relevant content, which is then indexed using **FAISS** for efficient **semantic search**. Next, a **reader model (BERT fine-tuned on SQuAD)** extracts the most likely answer span from the retrieved text.

Unlike conventional QA systems that require domain-specific fine-tuning or extensive labeled data, this system uses pretrained components and unsupervised document chunking. This makes it scalable, adaptable, and ideal for academic use. The system is designed to help students retrieve reliable answers to subject-related queries, assist in exam preparation, and support elective planning by giving clarity on subject depth and focus.

The expected result is a functional prototype capable of answering factual, definition-based, and conceptual questions using curriculum-aligned educational content. The system will be evaluated through QA metrics like **Exact Match** and **F1 Score**, along with qualitative feedback from users. This dissertation also discusses limitations such as the lack of support for **multi-hop reasoning** or generative answers. Nonetheless, it showcases a practical application of **NLP** in the academic domain, offering an intelligent way to support self-directed learning.

Key Words

BERT, Question Answering, Retrieval-Augmented Generation, Semantic Search, Educational NLP, Sentence-BERT, FAISS, Extractive QA

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI

SECOND SEMESTER 2024-25

DSECLZG628T / AIMLCZG628T DISSERTATION

Dissertation Outline (Abstract)

BITS ID No. 2022DC04341 **Name of Student:** Jaykumar Chaudhary

Name of Supervisor: Mr. Lalkar Eknath Chhadawelkar

Designation of Supervisor: Technical Evangelist

Qualification and Experience: M.Tech in Data Science and Engineering, 28 Years

Official E- mail ID of Supervisor: lalkar@cybage.com

Topic of Dissertation: A Retrieval-Augmented Question Answering System Using BERT for Stream-Specific Educational Content

J. C. Chaudhary

Signature of Student

Signature of Supervisor

Date: 23-MAY-2025

Date: 23-MAY-2025

Dissertation Outline Report

1. Project Work Title

A Retrieval-Augmented Question Answering System Using BERT for Stream-Specific Educational Content

2. Discussion on the chosen topic

A. Purpose of the Work and Expected Outcome

The purpose of this project is to develop a Retrieval-Augmented Question Answering (QA) system tailored to curriculum-aligned academic content from the Data Science and AI/ML streams. The system is designed to assist students in understanding key concepts, clarifying doubts, and exploring subject matter through natural language queries. The expected outcome is a functional prototype that enhances academic engagement and supports students in their learning journey, as well as in making informed decisions during elective selection by providing precise, contextually relevant answers from structured educational material.

B. Literature Review

A preliminary review of current literature in question answering systems, BERT-based models, retrieval mechanisms, and educational NLP has been conducted. Key foundational works include:

- **BERT** (Devlin et al., 2019), which introduced deep bidirectional transformers for language understanding.
- **Retrieval-Augmented Generation (RAG)** (Lewis et al., 2020), combining retrieval with reading/generation for knowledge-intensive tasks.
- **Dense Passage Retrieval** (Karpukhin et al., 2020), leveraging dense embeddings and FAISS for efficient semantic search.
- **Sentence-BERT** (Reimers & Gurevych, 2019), enabling fast semantic sentence embeddings useful for retrieval.
- Recent advances in **Generative AI (GenAI)** and large language models such as **GPT** have revolutionized natural language processing by enabling more accurate, context-aware language understanding and generation, which enhances retrieval-augmented question answering systems.

These works inform the architecture and methodology of the system, providing a foundation for building QA systems over academic and educational content.

C. Existing Process and Its Limitations

Currently, students rely on manual reading, Ctrl+F searches, or peer discussions to extract information from course handouts or other educational material. This is inefficient, especially when documents are lengthy or complex. Traditional keyword-based search methods lack semantic understanding and cannot handle varied question phrasing, making them unsuitable for precise academic inquiry.

D. Justification for the Selected Methodology

Using a Retrieval-Augmented Generation (RAG) architecture allows for the combination of semantic retrieval and deep language understanding. This eliminates the need for task-specific QA datasets while still providing accurate extractive answers. Leveraging pretrained models like BERT and Sentence-BERT reduces development time and improves performance without requiring additional domain-specific training.

E. Brief Discussion on the Project Work Methodology

The system will process curriculum-aligned academic content, such as course handouts and topic summaries, by:

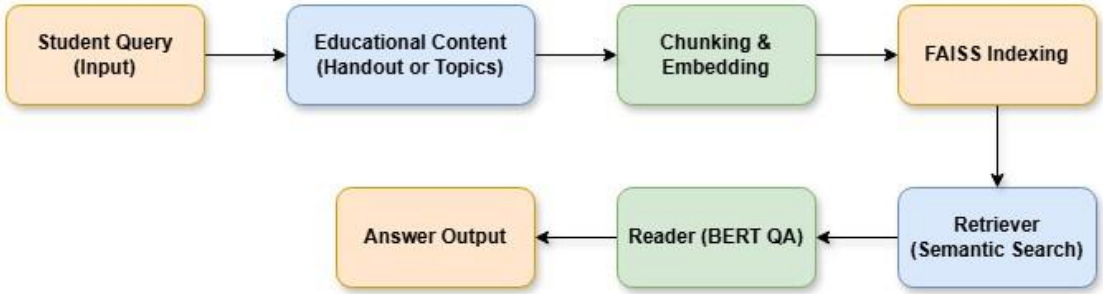
- Splitting text into semantic chunks
- Generating embeddings using Sentence-BERT
- Indexing chunks using FAISS for fast retrieval
- Accepting a student's natural language question and retrieving top-k relevant chunks
- Using a pretrained BERT QA model (fine-tuned on SQuAD) to extract the answer

Evaluation will be done using standard QA metrics like Exact Match (EM), F1 Score, and qualitative user feedback.

F. System Architecture and Workflow

The proposed Retrieval-Augmented QA system begins by splitting educational content (handouts or topics) into semantic chunks and encoding each chunk with Sentence-BERT. These chunk embeddings are indexed using FAISS to enable rapid, approximate nearest-neighbor search. When a student submits a question, it is also embedded via Sentence-BERT and used to query the FAISS index for the most relevant chunks. The retrieved content is then passed to a pretrained BERT QA model, which extracts the precise answer span. Finally, the system presents this answer—along with optional source context—back to the student, delivering fast, accurate, and context-aware responses.

Diagram illustrating the Retrieval-Augmented QA pipeline—student question embedding, FAISS-based semantic retrieval of content chunks, BERT QA answer extraction, and final response delivery.



F. Benefits Derivable from Work

- Helps students quickly access subject-specific academic content
- Supports elective subject exploration by offering insights into course material
- Reduces reliance on manual search or peer help
- Demonstrates a scalable application of NLP in the education domain

G. Other Supporting Details

The project also aligns with current educational technology trends and can be extended to include other academic documents in the future. It can be deployed as a web app or integrated into an academic portal to assist students more broadly.

3. Broad Area of Work

Natural Language Processing (NLP), Information Retrieval, Educational Technology, Applied Deep Learning

4. Objectives

The objectives of my project are as follows:

- To design and develop a **question answering system** capable of understanding and responding to student queries based on **academic content**.

- To implement a **retrieval-augmented architecture** that combines **semantic document retrieval** with a **BERT-based extractive reader**.

- To focus the **knowledge base** primarily on **course handouts** and other relevant **educational material** from the **Data Science** and **AI/ML** streams, ensuring relevant and useful responses.

- To evaluate the system's performance using standard **QA metrics** such as **Exact Match (EM)** and **F1 Score**, and to assess its usefulness through **student feedback**.

- To deliver a **usable** and **scalable prototype** that enhances access to **academic content** and supports students in both **learning** and **elective subject selection**.

5. Scope of Work

The scope of this dissertation is to design and develop a Retrieval-Augmented Question Answering (RAG) system that uses a pretrained BERT model to extract answers from semantically retrieved segments of curriculum-aligned academic content. While the system will initially be tested on course handouts from subjects within the Data Science and AI/ML curriculum, its design allows flexibility for integration with other structured educational resources. These documents will be processed, segmented, and embedded using Sentence-BERT, and indexed using FAISS for semantic search. The system will accept natural language questions and extract answers using a BERT QA model fine-tuned on the SQuAD dataset. This approach keeps the system efficient, focused, and ready for real-world academic use without needing domain-specific training.

6. Literature References

1. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.** NAACL-HLT. https://arxiv.org/abs/1810.04805

2. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). **Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.** NeurIPS. https://arxiv.org/abs/2005.11401

3. Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., & Yih, W.-t. (2020). **Dense Passage Retrieval for Open-Domain Question Answering.** EMNLP. https://arxiv.org/abs/2004.04906

4. Reimers, N., & Gurevych, I. (2019). **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.** EMNLP-IJCNLP. https://arxiv.org/abs/1908.10084

5. Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). **Deep Learning based Text Classification: A Comprehensive Review.** ACM Computing Surveys, 54(3), 1-40. https://arxiv.org/abs/2004.03705

7. Detailed Plan of Work

S.No.	Tasks or Subtasks	Start–End Date	Duration (weeks)	Specific Deliverable
1	Problem definition and finalize scope	Week 1	1	Project scope and architecture sketch
2	Literature review (RAG, BERT QA, Sentence-BERT, FAISS)	Week 2	1	Summary notes, list of tools & models, citations
3	Collection and preprocessing of course handouts	Week 3	1	Cleaned, structured handout corpus
4	Chunk documents & generate Sentence-BERT embeddings	Week 4	1	Chunked text + embeddings
5	Index content using FAISS	Week 5	1	FAISS vector store built
6	Develop retriever module & basic semantic search	Week 6	1	Retriever working
7	Integrate BERT QA model (Reader) & pipeline testing	Week 7-8	2	Basic QA pipeline working with test queries
	Mid-Semester - End of Week 8	**Demo: Basic working QA system**	**—**	**Minimal UI or CLI demo with question → answer workflow**
8	Develop simple user interface (Angular/React)	Week 9	1	Basic UI to query system
9	Expand handout coverage or improve retrieval	Week 10-11	2	Support for more subjects or improved chunking
10	Evaluation: EM, F1 Score, user feedback	Week 12	1	Evaluation metrics, feedback collection
11	Prepare draft report + visualizations/screenshots	Week 13-14	2	Draft dissertation with figures
12	Final polish, feedback incorporation, viva prep	Week 15-16	2	Final report, slides, working demo
	End-Semester - Final Viva & Submission	**End of Week 16**	**—**	**Complete system + final presentation**

Supervisor’s Rating of the Technical Quality of this Dissertation Outline

EXCELLENT / GOOD / FAIR/ POOR (Please specify): **EXCELLENT**

Supervisor’s suggestions and remarks about the outline (if applicable).

The outline demonstrates a clear structure and relevant scope for the dissertation. Looking forward to seeing how the research develops further.

Date: 23-MAY-2025

(Signature of Supervisor)

Name of the supervisor: Mr. Lalkar Eknath Chhadawelkar

Email Id of Supervisor: lalkar@cybage.com

Mob # of supervisor: +91 7774008916