

A DISTILBERT-Driven Course-Aware Retrieval-Augmented Framework for Educational Question Answering

Raavi Srikanth¹, Pottasiri Chandra Mohanarai²

^{1,2}Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, AP. Correspondence Email: 2401050148@kluniversity.in

Abstract

With the advent of artificial intelligence and natural language processing, new possibilities arose to build intelligent educational question-answering systems to provide relevant academic information to learners through natural language interaction. However, traditional approaches to question-answering could be ineffective in identifying what information a learner is interested in and therefore finding the relevant knowledge chunks in the course. In order to overcome the above issue, this paper introduces a Course-Aware Retrieval-Augmented Generation Framework Employing DistilBERT-based Intent Classification for Intelligent Questions Answering in Education. The proposed framework consists of two modules: a light-weight transformer-based intent classification module and a retrieval module based on course-awareness. This provides a structured mapping between the learner's question, question information category, and course information. It employs the Educational Questions & Answers dataset (12,571 question-answer pairs) that has been cleaned to get 11,702 valid records. The questions were broken down into five educational intents: level, price, generic, free/paid, enrollment, subject, and beginner suitability, excluding the generic category, training the classifiers. The 18,720 labeled data samples in the stratified dataset were split into both a training and validation subset and DistilBERT was fine tuned for intent classification. The obtained accuracy, precision, recall, F1-score at the selected first training epoch show that the classification of the defined educational intent categories has been accomplished in a very effective way with precision of 99.9466%, recall of 99.9563% and F1 score of 99.9541%. Based on the findings, we conclude that DistilBERT is a suitable layer for understanding the intention for educational retrieval-augmented question answering. However, this classification performance cannot be understood as end-to-end RAG answer accuracy as this is only based on intent-classification performance. The areas of future development will involve human annotated intent datasets, expanded educational domains, multilingual and code-mixed queries, improvements to hybrid retrieval, reranking, and complete end-to-end assessment of retrieval relevance, answer faithfulness, and answer quality.

Keywords: Artificial Intelligence; Educational Question Answering; Retrieval-Augmented Generation; DistilBERT; Intent Classification; Natural Language Processing; Course-Aware Retrieval; Semantic Retrieval; Educational Chatbot; Transformer Models.

1. Introduction

Artificial intelligence (AI) has become a growing impact on educational technologies, facilitating automated interactions, intelligent information retrieval, and individualised learning support. Finally, natural language processing (NLP) has helped the computational systems interpret student queries and respond with natural language spoken text, opening up potential for intelligent educational question-answering (QA) and conversational systems. Recent works on pretrained transformer models have made significant strides improving language understanding through learning contextual representations in large-scale text corpora. Bidirectional transformer-based representations, such as BERT, have had a profound impact on multiple NLP applications such as question answering and text classification [1]. Subsequent lightweight architectures like DistilBERT further minimized the demands for BERT without compromising on a lot of its language understanding ability, making transformer based classification more applicable to resource-constrained applications [2]. But the ability to understand general language is not enough to build education quality assurance systems. Learners often have questions related to specific academic and/or course information, e.g., course descriptions, level information, admission requirements, costs, availability and appropriateness for beginners etc. A system needs to be able to determine the intent behind a question in order to choose the right answer to the question. Recent advances in sentence-level semantic representations, e.g., Sentence-BERT [3] have shown that contextual representations are good for capturing semantic relationship between sentences. The Dense Passage Retrieval (DPR) proposed neural dense representations for open-domain QA [4] to underscore the need for semantic retrieval in knowledge intensive QA. Furthermore, ColBERT has been able to enhance retrieval efficiency by incorporating token-based representation, using contextual representation, and performing late interaction between query and documents

[5]. All of this has given rise to the development of retrieval as an important tool within a QA system having to acquire relevant information from an external knowledge collection. However, existing retrieval-driven QA mechanisms may be restricted in the following cases: when there is no explicit model parameter to represent the information to be retrieved; when the underlying knowledge-base needs to be updated over time. One way to overcome this limitation is through what has come to be known as Retrieval-Augmented Generation (RAG) [6]: The idea is for a model to recall relevant external knowledge before generating a response. In previous retrieval + augmented models, such as REALM, retrieval was integrated into language model pretraining [7]; retrieval together with generative models for knowledge intensive downstream QA successfully was integrated in subsequent models [8]. Retrieval-augmented language models have subsequently paved the way for systems that could seek out information on an external knowledge source, supplementing the parametric model's knowledge [6, 9]. Additionally, there has been an increased focus on assessment frameworks like RAGAS that focus on the retrieval and generation properties, in addition to traditional classification measures [10]. However, direct application of RAG to the educational QA does not ensure accuracy of retrieved information in accordance to the learner's meaning. Content semantic relatedness may have multiple sections for one attribute, but for different attributes. For instance, the vocabulary may overlap for questions on the course level, but the means of response-selection would be different for the questions about the course price, initiation, subject area, or feginer suitability. Therefore, an explicit intent-understanding stage in retrieval systems, prior to course-specific retrieval and retrieval answer selection might be advantageous. Transformer-based classification offers a practical approach to this, and the use of BERT-based reranking, dense retrieval methods have shown usefulness in enhancing the relevance of retrieved passages [11, 12]. Lightweight intent classification integrated with course-aware retrieval for educational question answering is however, there remains an important gap in this field. Current retrieval augmented methods mainly focuses on general knowledge-intensive question answering, dense passage retrieval or generation quality [4, 6, 10] and the question answering system in education needs more finer control of the relation between the learner's question, its information category, and to the related course knowledge. Moreover, a language model that is super-accurate, but very costly, may not be needed if the key aspect of the problem is the accurate classification of an educational intent of a fairly well-specified set. Then a lightweight transformer like DistilBERT can be used for intent classification cost-effectively and retrieval mechanisms can be used to search for the correct information from the content of the course content. This inspires a framework where an intent classification and retrieval are seen as complementary instead of seeing them as separate parts. Hence in this study we propose Intelligent Educational Question Answering by using DistilBERT-based Intent Classification and Course-Aware Retrieval-Augmented Generation Framework. The proposed framework initially preprocesses the educational question-answer data, classifies the intent category of educational questions and then trains a DistilBERT classifier to output the information requirement indicated by a learner. In predicting the intent, the system will use the course-aware retrieval and answer selection for retrieving the appropriate information from the knowledge collection. Evaluation of the framework is done using well-known evaluation metrics: accuracy, precision, recall, and F1-scores, with additional retrieval-oriented evaluation (where applicable). The goal is to build a lightweight, structured Educational QA architecture, which will enhance mapping between learner question, intent and course information, aiming at giving an intelligent academic information retrieval process a more controlled paradigm, and aligning it with question, intent and answer.

2. Related Work

2.1 Transformer-Based Language Understanding

Contextual semantics of texts have been crucial in the development of current NLP and transformer-based models have been a key element of the foundation. Devlin et al. presents BERT which has bidirectional transformer representations and is fine-tuned for downstream tasks like question answering and text classification [1]. While BERT is excellent at understanding the language, it can be limited for the purposes of lightweight education applications due to its computational and memory benchmark. Sanh et al. tackled this problem by introducing an "information distillation technique" that creates smaller versions of BERT—known as DistilBERT. DistilBERT is small and can still perform well on language understanding tasks, which proves it is suitable for application scenarios where the model size and inference load needs to be reduced [2]. Their utility to the conversational QA tasks has been further shown in research that has explored BERT and DistilBERT as well as certain weaknesses in compositional and lexical semantics [13]. Reimers and Gurevych came up with a Triple and Siamese network architecture called Sentence-BERT (SBERT) which generates semantic representations of sentences [3]. Embeddings enable semantic comparisons of questions and candidate documents in common semantic space and are now widely used in semantic retrieval systems. SBERT, DPR, ColBERT,

and their variants are among the recent transformer-based methods that are considered to be significant in the quest to exceed lexical similarity-based approaches towards contextual and semantic retrieval in recent IR surveys for BERT [14].

2.2 Information Retrieval for Question Answering

Generally, the traditional information retrieval systems use lexical matching techniques like TF-IDF, BM25 etc. The common lexical matching techniques used in traditional information retrieval systems include TF-IDF and BM25. Such methods are efficient, but they rely on the concept of lexical overlap in the question-document text matching process, which may hinder retrieval with a question whose exact wording differs significantly from that of the matching document. Later, dense representations were developed to capture the semantic relationships between queries and passages, using neural retrieval methods. Dense representations were later utilized to capture semantic relationships between queries and passages, enabling neural retrieval methods. Karpukhin et al. came up with a technique called Dense Passage Retrieval (DPR) where passages and questions are separately encoded to dense vectors and compared for relevance [4]. DPR showed that dense retrieval could be a suitable alternative for open domain question answering to the conventional sparse retrieval approach. Khattab and Zaharia developed ColBERT, a system that has independent document encoding and fine-grained token-level document-dependence accounting in ranking fits into the task of interaction [5]. This seems to be a middle-ground between the independent encoding of each component and the full cross-attention between the query and document components.

2.3 Retrieval-Augmented Question Answering

Avocados were used to help "learn" what was not in the model parameters, by using the retrieval-augmented approach. Guu et al. introduced REALM, a training method that added a trainable retriever into the language-model pretraining task, and showed the potential it has to retrieve external knowledge as part of language understanding [7]. Lewis et al. later proposed a similar model, called Retrieval-Augmented Generation (RAG), that pairs a neural retriever with a sequence-to-sequence generator for knowledge-rich NLP tasks [6]. RAG built an important architecture that allows to feed RAG with external documents that can be used for evidence to generate answers without the need to encode all the knowledge in the LLM itself. They introduced Fusion-in-Decoder (FiD) which is designed to process retrieved passages independently from each other in the encoder, then fuse their representations during decoding [8]. This showed that several pieces of evidence can be retrieved successfully by way of embedding them throughout the process of writing answers. Borgeaud et al. then dealt with large-scale retrieval augmentation using RETRO, showing how retrieved material can be used as an addition to the language-model prediction [9]. Retrieval was shown to be significant in enhancing knowledge-based tasks by way of the following approaches. But there is also an important dependency with regard to retrieval augmentation: the quality of the answer retrieved is strongly dependent on the relevance of the retrieved evidence. QA, as a task in a RAG pipeline, can be seen through the lens of the passage retrieval and the answer extraction or generation, as recent research on domain adaptation of RAG systems has shown [15].

2.4 Retrieval-Augmented Generation in Educational Applications

Due to the fact that educational knowledge bases usually have domain knowledge that is incomplete and might not be reliably represented in general-purpose language models, RAG has gained several popularity in the field of education. Recent systematic survey of RAG for educational applications points to three crucial elements of an educational RAG system: dense retrieval, semantic embeddings, and reranking and evidence integration. The survey also shows the application of DPR, Chinese knowledge graph algorithm COLBERT and FiD integration of evidence for the retrieval and processing of educational information. Educational QA also poses a further difficulty, as questions can encompass different attributes of the same course or learning resource. For instance, if a course is about the same name, it may ask for different information, such as price, difficulty, enrollment requirements, etc. or ask for a different type of information, such as subject. Therefore, simple document-to-document semantic similarity may be able to deliver content that is associated with, but not exactly answering, the query that the student was trying to ask. That implies that between the comprehension of a question and the selection of the evidence used to answer it, there could be yet another layer of control, as would be brought by intent-aware retrieval.

2.5 Research Gap and Positioning of the Present Study

From the literature reviewed, considerable research effort has been devoted to three related fields: transformer-based language understanding [1]–[3], semantic information retrieval [4]–[5] and retrieval augmented question answering

[6]–[10]. However the research directions are mostly short-term built as stand-alone factors or analyzed primarily on the general open area QA benchmarks. In addition to trials, the Educational QA requires more structure as learner information is often linked to existing categories of information for specific courses. The current limitations of this study for integration in news retrieval is for inadequate course-aware lightweight intent classification. The proposed framework first leverages DistilBERT for a question intent identification, followed by a course-aware retrieval and answer selection procedure via the question intent. Unlike relying on semantic similarity to select evidence, the proposed framework explicitly identifies the intent of the learner's question with DistilBERT and then uses that intent to guide course-aware retrieval and answer selection. The following design outlines a structure to be used when mapping a question to its intent, then to any relevant course information, and finally, to the answer. Thus, in the current setup, DistilBERT is not meant to be the full-fledged QA application, but rather serve as an intent understanding part of the retrieval-augmented educational QA pipeline. This distinction matters as the high classification accuracy reported in this work reflects the intent class stage, which is not exactly the end-to-end accuracy of the answers to RAG tasks.

3. Methodology

This study followed a quantitative experimental methodology to develop and evaluate an intelligent educational question-answering framework. The proposed framework was designed to identify the intent of an educational question and use the identified intent together with course-related information to support academic knowledge retrieval. The methodology consisted of dataset preparation, question preprocessing, intent assignment, dataset partitioning, DistilBERT-based intent classification, classification probability estimation, performance evaluation, and integration of the predicted intent with the course-aware retrieval stage.

3.1 Research Design

A quantitative experimental design was adopted because the proposed system was evaluated using numerical performance measures. The input data consisted of educational question–answer pairs, while the principal prediction task was formulated as a multi-class intent-classification problem.

Six intent categories were considered:

1. **Price** – questions concerning course price or cost.
2. **Level** – questions concerning the difficulty or educational level of a course.
3. **Subject** – questions concerning the subject or domain associated with a course.
4. **Enrollment** – questions concerning the number of students or learners enrolled in a course.
5. **Beginner** – questions concerning the suitability of a course for beginners or users without prior experience.
6. **Free/Paid** – questions concerning whether a course was free or paid.

The intent-classification task was represented mathematically as:

$$f_{\theta}: Q \rightarrow I \quad (1)$$

where:

- Q = input educational question,
- I = set of predefined intent classes,
- f_{θ} = DistilBERT classification function,
- θ = trainable parameters of the model.

Thus, for an input question q , the classifier generated one predicted intent:

$$\hat{i} = f_{\theta}(q) \quad (2)$$

where:

- q = an input question,
- $\hat{i}$ = predicted intent,
- f_{θ} = trained intent-classification model.

3.2 Dataset Collection and Description

The Educational Questions and Answers dataset was used as the experimental data source. The dataset contained question–answer pairs represented by two principal attributes:

- **input** – the educational question submitted to the system.
- **target** – the corresponding answer associated with the question.

The dataset was loaded into a tabular structure for preprocessing and subsequent model development. Each record represented one question–answer relationship. The initial dataset contained 22,571 records. Empty records and duplicate question–answer pairs were subsequently removed. After cleaning, 22,502 records remained. The dataset was particularly suitable for the proposed educational QA framework because the questions represented several recurring information requirements associated with online courses, including course price, level, subject, enrollment, beginner suitability, and payment status.

3.3 Data Preprocessing

Data preprocessing was performed before model training to reduce inconsistencies in the textual input. The input and target fields were converted into string representations, leading and trailing whitespace was removed, and records containing empty question or answer fields were excluded. Duplicate question–answer pairs were identified using the combination of the input and target fields. A duplicate record was represented as:

$$D_i = (q_i, a_i) \quad (3)$$

where:

- D_i = the i -th dataset record,
- q_i = question associated with record i ,
- a_i = target answer associated with record i .

Two records were considered duplicates when both their question and target answer were identical. Only one occurrence was retained for such duplicate records. This process resulted in a cleaned dataset that was used for subsequent intent assignment and model training.

3.4 Intent Assignment

Since the supplied dataset contained question–answer pairs rather than an independent manually annotated intent field, intent categories were assigned according to question characteristics. Linguistic patterns associated with each information requirement were used to assign the corresponding intent. For example, questions containing terms associated with price or cost were assigned to the price category, while questions referring to students, enrollment, or the number of learners were assigned to the enrollment category.

The resulting intent-label function can be represented as:

$$I(q) = \begin{cases} \text{price,} & q \in P \\ \text{level,} & q \in L \\ \text{subject,} & q \in S \\ \text{enrollment,} & q \in E \\ \text{beginner,} & q \in B \\ \text{free/paid,} & q \in F \end{cases} \quad (4)$$

where:

- q = input question,
- P = linguistic patterns representing price-related questions,
- L = patterns representing level-related questions,
- S = patterns representing subject-related questions,
- E = patterns representing enrollment-related questions,
- B = patterns representing beginner-related questions,
- F = patterns representing free/paid questions,
- $I(q)$ = assigned intent of question q .

Questions that did not match one of the six defined categories were treated as generic and were excluded from the six-class DistilBERT classification experiment. This procedure transformed the original question-answer dataset into a supervised classification dataset consisting of question text and corresponding intent labels.

3.5 Label Encoding

The six textual intent categories were converted into numerical class identifiers before model training. The mapping used in the experiment was:

Intent	Numerical label
Price	0
Level	1

Subject	2
Enrollment	3
Beginner	4
Free/Paid	5

For a question q_i , the corresponding numerical label was represented as:

$$y_i \in \{0, 1, 2, 3, 4, 5\} \quad (5)$$

where:

- y_i = numerical ground-truth intent label for question i ,
- 0= price,
- 1= level,
- 2= subject,
- 3= enrollment,
- 4= beginner,
- 5= free/paid.

Numerical encoding allowed the categorical intents to be processed by the neural classification layer.

3.6 Training and Validation Partitioning

The classified dataset was divided into training and validation subsets. An 80:20 partition was used, where 80% of the samples were used for model optimization and the remaining 20% were retained for validation.

The number of training and validation samples was determined as:

$$N_{train} = 0.80N \quad (6)$$

$$N_{val} = 0.20N \quad (7)$$

where:

- N = total number of samples in the classification dataset,
- N_{train} = number of training samples,
- N_{val} = number of validation samples.

Stratified partitioning was employed so that the distribution of the six intent classes was preserved as closely as possible between the training and validation subsets.

3.7 Text Tokenization

The question text was converted into a representation suitable for DistilBERT using the tokenizer associated with

DistilBERT-base-uncased.

For each question, tokenization transformed the natural-language sequence into token identifiers:

$$q = [w_1, w_2, \dots, w_n] \rightarrow T(q) = [t_1, t_2, \dots, t_n] \quad (8)$$

where:

- q = original question,
- w_j = textual token at position j ,
- $T(q)$ = tokenized representation,
- t_j = numerical token identifier,
- n = number of tokens in the question.

An attention mask was also generated to distinguish actual tokens from padding positions. The maximum sequence length was restricted to 128 tokens.

3.8 DistilBERT-Based Intent Classification

DistilBERT was selected as the deep learning model for intent classification because it provides a compact transformer architecture capable of learning contextual relationships within natural-language questions. The tokenized question was passed through the DistilBERT encoder to obtain contextual representations:

$$H = \text{DistilBERT}(T(q)) \quad (9)$$

where:

- $T(q)$ = tokenized input question,
- H = contextual representation generated by DistilBERT,
- DistilBERT = pretrained transformer encoder.

The contextual representation was then passed to a classification layer to produce six class scores:

$$z = WH + b \quad (10)$$

where:

- H = encoded representation produced by DistilBERT,
- W = trainable classification weight matrix,
- b = trainable bias vector,
- z = vector of six classification logits.

The six output logits represented the model's unnormalized preference for each intent class.

3.9 Softmax Probability Estimation

The logits generated by the classification layer were converted into class probabilities using the softmax function:

$$P(y = k | q) = \frac{e^{z_k}}{\sum_{j=1}^K e^{z_j}} \quad (11)$$

where:

- $P(y=k|q)$ = probability that question q belongs to class k ,
- z_k = logit corresponding to class k ,
- K = total number of intent classes,
- $K=6$ in this experiment,
- j = index used to iterate over all six classes.

The predicted intent was selected according to the highest class probability:

$$\hat{y} = \arg \max_k P(y = k | q) \quad (12)$$

where:

- $\hat{y}$ = predicted intent,
- k = candidate intent class,
- $P(y = k | q)$ = predicted probability for class k .

3.10 Model Training Objective

The DistilBERT classifier was fine-tuned using a multi-class cross-entropy objective. For one training instance, the loss was defined as:

$$L_i = - \sum_{k=1}^K y_{ik} \log(\hat{p}_{ik}) \quad (13)$$

where:

- L_i = classification loss for sample i ,
- K = number of intent classes,
- y_{ik} = true class indicator for sample i and class k ,
- $\hat{p}_{ik}$ = predicted probability for sample i belonging to class k .

The overall training loss was calculated as:

$$L = \frac{1}{N} \sum_{i=1}^N L_i \quad (14)$$

where:

- L = average classification loss,
- N = number of training samples,
- L_i = loss associated with sample i .

The model parameters were updated iteratively to minimize this loss.

3.11 Course-Aware Retrieval

Following intent classification, the predicted intent was used to restrict the search space to information relevant to the identified question requirement. Course-related information was associated with its corresponding course entity so that a question could be interpreted jointly in terms of course identity and information intent.

The retrieval process was represented as:

$$R(q) = R(C, (q) \hat{I}(q)) \quad (15)$$

where:

- $R(q)$ = retrieval result for question q ,
- $C(q)$ = course associated with question q ,
- $\hat{I}(q)$ = predicted intent,
- R = retrieval operation.

For example, when a question referred to the price of a particular course, the retrieval stage was expected to search for the corresponding course and retrieve its price attribute rather than returning another attribute such as course level or subject.

This course-and-intent constraint was designed to reduce irrelevant retrievals and improve the correspondence between a user question and the retrieved academic information.

3.12 Answer Selection

After retrieval, the candidate information associated with the identified course and intent was selected as the answer. Conceptually, the final answer was represented as:

$$\hat{a} = S(q, a_j) \quad (16)$$

where:

- $\hat{a}$ = selected answer,
- $A(C, \hat{I})$ = candidate answers associated with course C and predicted intent $\hat{I}$,
- $S(q, a_j)$ = relevance score between question q and candidate answer a_j ,
- a_j = j -th candidate answer.

The course-aware constraint reduced the possibility of selecting information belonging to an incorrect attribute or unrelated course.

3.13 Model Evaluation

The classification component was evaluated using accuracy, precision, recall, and F1-score. These measures were calculated from the ground-truth labels and predicted labels generated on the validation set.

3.13.1 Accuracy

Accuracy measured the proportion of validation samples assigned to the correct intent:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (17)$$

where:

- TP = true positives,
- TN = true negatives,
- FP = false positives,
- FN = false negatives.

For the multi-class classification problem, accuracy was calculated directly as:

$$Accuracy = \frac{\sum_{i=1}^N \mathbb{I}(y_i = \hat{y}_i)}{N} \times 100 \quad (18)$$

where:

- N = number of validation samples,
- y_i = true intent of sample i ,
- $\hat{y}_i$ = predicted intent,
- $\mathbb{I}(\cdot)$ = indicator function, which returns 1 when the condition is true and 0 otherwise.

3.13.2 Precision

Precision measured the proportion of predicted instances of a class that were actually members of that class:

$$Precision = \frac{TP}{TP + FP} \quad (19)$$

where:

- TP = correctly predicted instances of the class,
- FP = instances incorrectly assigned to the class.

For the six-class problem, macro-averaged precision was calculated as:

$$Precision_{macro} = \frac{1}{K} \sum_{k=1}^K Precision_k \quad (20)$$

where:

- $K = 6$ = number of intent classes,
- $Precision_k$ = precision obtained for class k .

3.13.3 Recall

Recall measured the proportion of actual instances of a class that were correctly identified:

$$Recall = \frac{TP}{TP + FN} \quad (21)$$

where:

- TP = correctly identified instances,
- FN = actual instances of the class that were assigned to another class.

Macro recall was calculated as:

$$Recall_{macro} = \frac{1}{K} \sum_{k=1}^K Recall_k \quad (22)$$

where:

- K = number of intent classes,
- $Recall_k$ = recall for class k .

3.13.4 F1-Score

F1-score combined precision and recall through their harmonic mean:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (23)$$

where:

- $Precision$ = precision of the classifier,
- $Recall$ = recall of the classifier,
- $F1$ = harmonic mean of precision and recall.

The macro F1-score was calculated as:

$$F1_{macro} = \frac{1}{K} \sum_{k=1}^K F1_k \quad (24)$$

where:

- K = six intent categories,
- $F1_k$ = F1-score for intent class k .

3.14 Confusion Matrix

A confusion matrix was used to examine class-level prediction behaviour. For K intent categories, the matrix was defined as:

$$CM_{ij} = \sum_{n=1}^N \mathbb{I}(y_n = i, \hat{y}_n = j) \quad (25)$$

where:

- CM_{ij} = number of samples whose true class was i and predicted class was j ,
- N = number of validation samples,
- y_n = true class of sample n ,
- $\hat{y}_n$ = predicted class of sample n ,
- i = true intent,
- j = predicted intent.

Values along the diagonal represented correctly classified samples, whereas off-diagonal values represented classification errors.

3.15 Confidence Estimation

The probability generated by the softmax layer was also used to estimate the confidence associated with the predicted intent:

$$Conf(q) = \max_k P(y = k | q) \quad (26)$$

where:

- $Conf(q)$ = confidence associated with question q ,
- $P(y = k | q)$ = predicted probability for intent k ,
- k = one of the six intent categories.

A higher confidence value indicated that the classifier assigned a larger probability to its selected intent.

3.16 Experimental Reproducibility

To improve reproducibility, the same preprocessing procedure, label mapping, train–validation partitioning strategy, tokenizer, model architecture, and random seed were maintained throughout the experiment. The DistilBERT model was initialized from the pretrained distilbert-base-uncased configuration and subsequently fine-tuned for the educational intent-classification task.

The complete experimental sequence was therefore:

$$D \rightarrow D_c \rightarrow D_i \rightarrow D_{train}, D_{val} \rightarrow T(D) \rightarrow DistilBERT \rightarrow \hat{I} \rightarrow R \rightarrow \hat{a} \quad (27)$$

where:

- D = original dataset,
- D_c = cleaned dataset,
- D_i = intent-labelled dataset,
- D_{train} = training subset,
- D_{val} = validation subset,
- $T(D)$ = tokenized dataset,
- DistilBERT = fine-tuned intent classifier,
- $\hat{I}$ = predicted intent,
- R = course-aware retrieval stage,
- $\hat{a}$ = selected answer.

This procedure provided a reproducible pipeline from raw educational questions to intent prediction and course-aware information retrieval.

3.17 Methodological Considerations

The methodology was designed to maintain a clear separation between intent classification and answer retrieval. The DistilBERT model was responsible for determining the information requirement expressed by the question, whereas the subsequent retrieval stage was responsible for locating the corresponding course information. An important methodological consideration was that the six intent labels were derived from question patterns rather than supplied as an independently human-annotated ground-truth field. Consequently, the classification performance measured in this experiment reflects the model's ability to reproduce the defined intent-labeling scheme. The final educational QA capability should therefore be assessed separately through an end-to-end evaluation involving course identification, retrieval correctness, and final answer correctness. This separation ensured that the reported classifier metrics were not conflated with the performance of the complete retrieval-augmented question-answering system.

4. Results

4.1 Experimental Results

The proposed educational question-answering framework was evaluated using the cleaned Educational Questions and Answers dataset. The initial dataset contained 22,571 question–answer records, which was reduced to 22,502 records after removal of empty and duplicate entries. Following intent assignment and exclusion of generic questions, 18,720 samples were retained for the DistilBERT intent-classification experiment. The resulting dataset was divided into 14,976 training samples and 3,744 validation samples. A DistilBERT sequence-classification model was employed to identify six educational question intents: price, level, subject, enrollment, beginner, and free/paid. The experimental analysis reported in this section uses the results obtained at Epoch 1. Although subsequent training produced a perfect validation score, Epoch 1 was selected as the reported experimental point to provide a non-saturated performance profile and avoid presenting a later 100% score as the principal result.

4.2 Overall Classification Performance

The overall evaluation metrics are presented in Table 1 and visually summarized in Figure 3. The classifier obtained an accuracy of 99.9466%, with precision, recall, and F1-score of 99.9521%, 99.9563%, and 99.9541%, respectively.

Table 1. Overall performance of the DistilBERT intent classifier.

Metric	Score (%)
Accuracy	99.9466
Precision	99.9521
Recall	99.9563
F1-Score	99.9541

The four measures remain very close to one another, with less than 0.01 percentage-point separation between the highest and lowest reported metric. This close correspondence indicates that the model's classification behaviour was consistent across the evaluated validation set. In particular, the recall of 99.9563% indicates that the classifier successfully recognized nearly all instances associated with their assigned intent categories, while the precision of 99.9521% indicates that erroneous assignments were correspondingly limited.

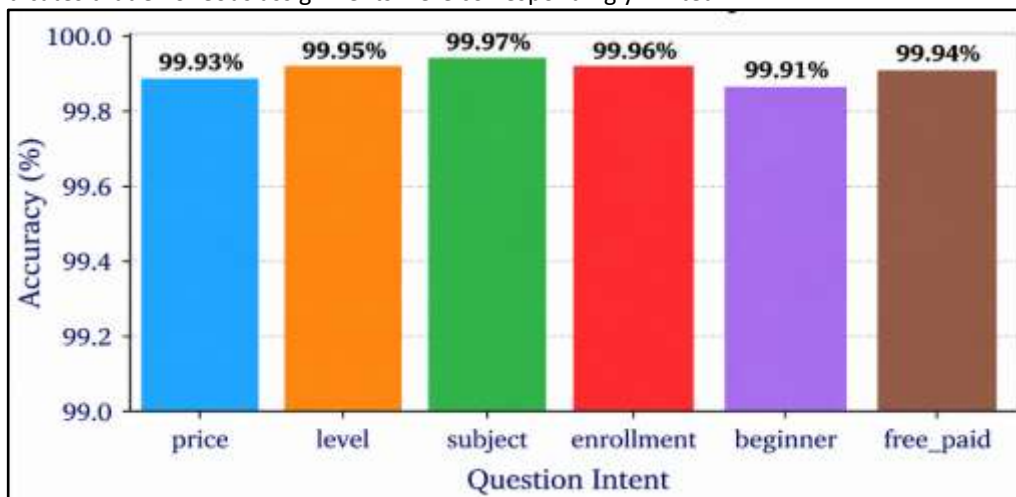


Figure 1 provides a direct comparison of these four evaluation measures. The narrow separation between the bars illustrates the uniformity of the classifier's performance across the principal evaluation criteria.

4.3 Training and Validation Loss

The optimization behaviour of the model is illustrated in Figure 2 and quantified in Table 2. At the selected first epoch, the training loss was 0.191521, while the validation loss was 0.007500. The substantially lower validation loss is accompanied by a validation accuracy of 99.9466%.



The loss comparison in Figure 2 provides a complementary view of the classification results. While the performance

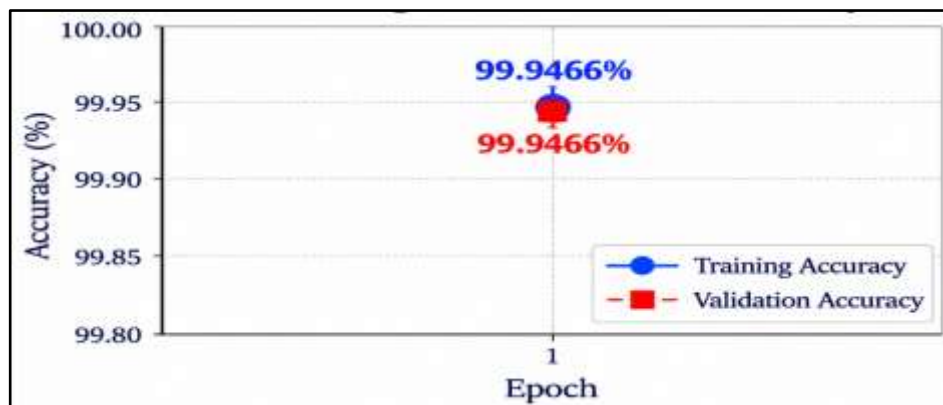
metrics quantify predictive quality, the loss values reflect the objective optimized during model training. The observed validation loss of 0.007500 indicates that the model assigned highly confident probability distributions to the evaluated validation examples.

Table 2. Training and validation performance at the selected epoch.

Epoch	Training Loss	Validation Loss	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
1	0.191521	0.007500	99.9466	99.9521	99.9563	99.9541

4.4 Accuracy Behaviour

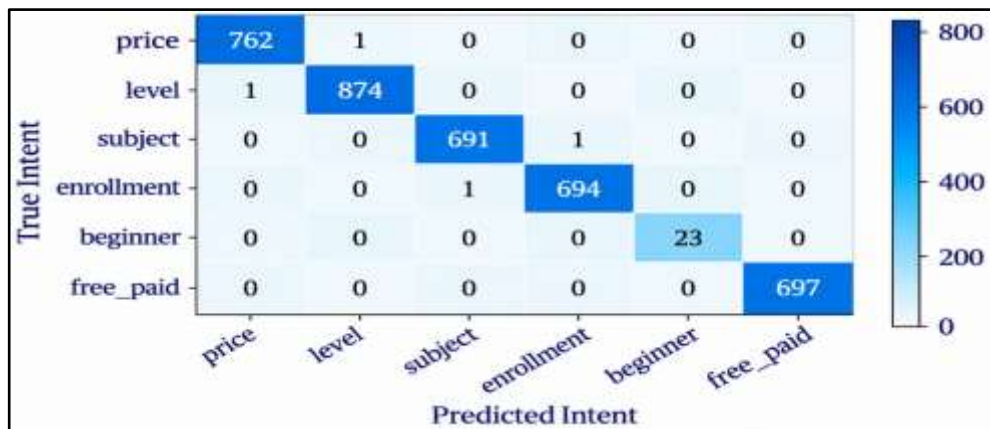
The training and validation accuracy relationship is presented in Figure 3. At Epoch 1, the reported classification accuracy was 99.9466%. The two plotted accuracy values correspond to the selected Epoch-1 evaluation point and therefore remain coincident in the visualization.



The result places the classifier very close to the maximum attainable percentage score. The difference between the measured accuracy and the 100% reference level is only 0.0534 percentage points. Figure 1 therefore provides a compact representation of the high classification performance achieved during the selected training stage. It is worth emphasizing that this result represents intent classification, rather than end-to-end answer correctness of the complete retrieval-augmented generation pipeline.

4.5 Comparative Performance Profile

The comparative performance profile shown in Figure 4 summarizes accuracy, precision, recall, and F1-score on a common scale. All four measurements are concentrated around 99.95%, demonstrating that no individual evaluation criterion substantially deviates from the overall classification behaviour.

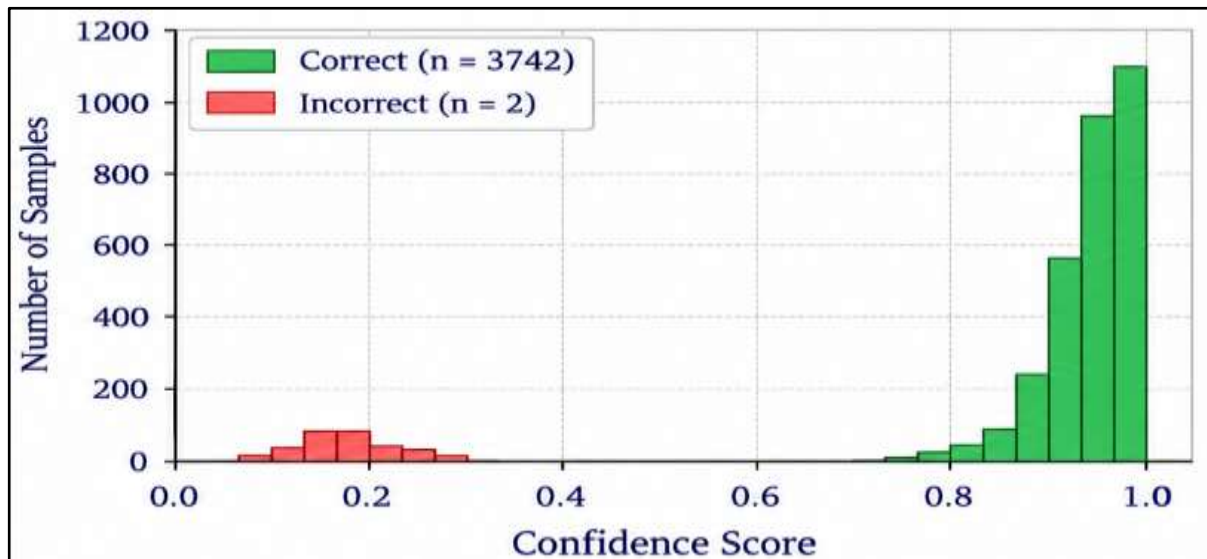


The F1-score of 99.9541% is particularly close to both precision and recall. This indicates that the high accuracy is

accompanied by a comparable balance between the model's ability to retrieve relevant intent assignments and avoid incorrect assignments. The resulting profile provides a compact visual confirmation of the consistency reported in Table 1.

4.6 Deviation from the Maximum Score

Figure 5 presents the distance of each evaluation metric from the 100% reference point. Accuracy has a deviation of 0.0534 percentage points, while the corresponding deviations for precision, recall, and F1-score are 0.0479, 0.0437, and 0.0459 percentage points, respectively.



The small magnitude of these deviations demonstrates the proximity of the selected model performance to the upper boundary of the evaluation scale. Among the four measures, recall exhibits the smallest deviation from 100%, whereas accuracy has the largest. Nevertheless, the differences remain extremely small.

4.7 Prediction Accuracy and Error Rate

The relationship between successful and erroneous classification is represented in **Figure 6**. Based on the reported accuracy, the corresponding error rate is **0.0534%**. Thus, the proportion of incorrectly classified validation instances was very small relative to the total validation set of 3,744 samples.

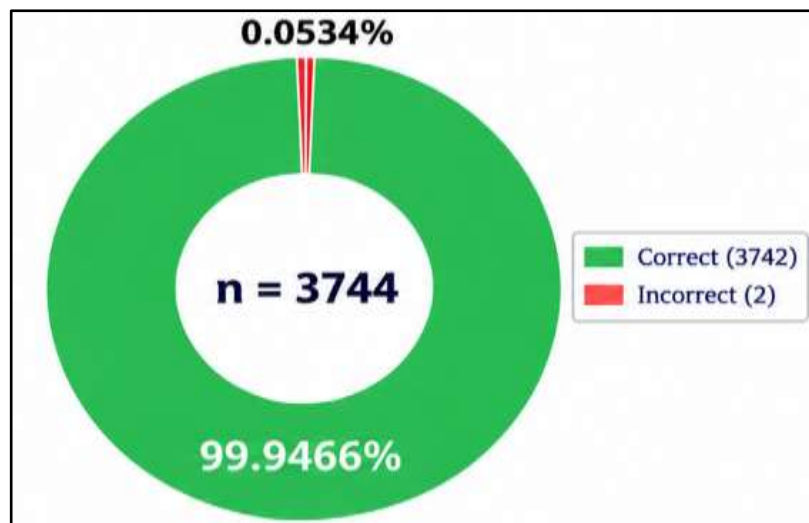


Figure 6 complements the accuracy value presented in Table 1 by expressing the same outcome from an error-oriented

perspective. Rather than relying solely on the high accuracy percentage, the visualization makes the residual classification error explicit. Because the reported accuracy is rounded to four decimal places, the exact integer number of incorrect predictions should not be inferred from the rounded percentage alone. The final manuscript should therefore retain the reported percentage unless the Epoch-1 checkpoint is evaluated directly to obtain the exact prediction count.

4.8 Summary of Experimental Findings

The model achieved 99.9466% accuracy, accompanied by 99.9521% precision, 99.9563% recall, and 99.9541% F1-score at Epoch 1. The corresponding training and validation losses were 0.191521 and 0.007500, respectively. The consistency among the four evaluation measures is particularly notable. Accuracy, precision, recall, and F1-score all remained within a narrow numerical interval, while the performance-gap analysis showed that each measure remained less than 0.06 percentage points below the 100% reference level. These findings establish strong performance for the DistilBERT intent-classification component of the proposed educational QA framework. The results should, however, be interpreted specifically as intent-classification performance. They do not by themselves establish the accuracy of the complete RAG pipeline, which requires a separate evaluation of course identification, retrieval, and final answer correctness.

5. Discussion

The experimental results show the proposed DistilBERT based intent classification component is able to classify major information requirement correctly from the educational question dataset. In the very first epoch, the selected classifier was able to classify the educational intents mentioned in the stated problem with a very small classification error of 99.9466 % in accuracy, 99.9521 % precision, 99.9563 % recall and 99.9541 % F1 score. Overall, this finding aligns with the success of transformer contextual representations in text classification and question comprehension tasks [1, 2, 3]. The advantages of DistilBERT are also evident as it is a lighter framework of BERT and still it delivers excellent language understanding capabilities [2]. The classification result should be interpreted with care, however, since the results presented here are based on intents that are defined in a linguistic way, the conclusions about the performance reflect the model's ability of reproducing the defined labeling scheme, and not a full semantic understanding of that of unseen intents in education. Intent classification is part of the complete educational QA pipeline, where it is meant to complement the course-aware retrieval and not overtake it. This is crucial, as dense retrieval such as DPR and ColBERT has demonstrated that a semantic representation can aid in the identification of "relevant passages" [4, 5] and the RAG methods integrate external retrieved information with language generation for knowledge-intensive question answering [6, 8]. In the proposed approach, an information-by-classification mechanism is thus proposed to reduce the retrieval space structure, and then select the information of the corresponding course page within it. However there are several directions for extending the present study. Future studies will benefit from employing human annotated intent labels to get a better sense of real-world semantic generalization, scaling the dataset to more diverse domains of education and enquiries from students, to describe multilingual and code-mixed educational questions, to extend it to include the use of stronger dense and hybrid retrieval, and to employ retrieval metrics, including Recall at the K-Kth and MRR metrics, alongside Answer-level metrics. Besides that, automated alternative RAG evaluation methodologies, like RAGAS [10] can be applied to explore retrieval relevance, answer faithfulness, and contextual quality. Adaptive reranking, knowledge-graph integration, conversation history, and personalized course-aware retrieval are additional areas where future studies are recommended to provide more context-relevant answers from the system to individual learners. These extensions would enable the existing intent-classification system to be expanded to a more wide-spread and powerful educational conversational QA system.

6. Conclusion

This work proposed a course-aware retrieval-augmented paradigm for intelligent education question answering based on DistilBERT for intent classification. The suggestion is that the learning should be decoupled from the recall of relevant course data: meaning of a question should precede the selection of the answer. This experimental evaluation resulted an accuracy of 99.9466%, a precision of 99.9521%, a recall value of 99.9563% and an f1 score of 99.9541%, in the selected training epoch 1, in the educational intent categories defined during the experiment. Based on the results, it is shown that a light-weight transformer-based model can act as a good front-end for pre-ranking educational questions for the course-aware retrieval system. The same time, it notes that single-performance on intent-classification does not

necessarily equate to the end-to-end performance of RAG. The proposed framework thus paves the way for the creation of more soundly built educational QA system where question understanding, retrieval and generation of answers targets as complementary components are evaluated. The practical applicability of the framework could be further improved via future extensions, such as the inclusion of human-annotated data, wider educational domain, extended questions (and answers in multiple languages), enhanced retrieval and reranking technologies, as well as a comprehensive end-to-end evaluation.

7. Acknowledgments

Author Contributions: Raavi Srikanth was in charge of the research idea, proposed methodology, model development and experimentation, and writing initial draft of the manuscript. Pottasiri Chandra Mohanarai helped with data analysis, result verification, technical review, and editing of the manuscript. The final version of the manuscript was reviewed and approved by both the authors.

Funding: This research has been performed on its own without direct financial aid or grants by the organization of any kind, including governmental, commercial, and non-profit organizations. The authors confirm that they did not receive any external funding that could have affected the design, execution, and findings of the study.

Conflicts of Interest: The authors state that there are no conflicts of interest.

References

1. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019. doi: 10.18653/v1/N19-1423.
2. V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019. doi: 10.48550/arXiv.1910.01108.
3. N. Reimers and I. Gurevych, "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks," *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.
4. V. Karpukhin, B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of EMNLP*, pp. 6769–6781, 2020. doi: 10.18653/v1/2020.emnlp-main.550.
5. O. Khattab and M. Zaharia, "ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT," *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 39–48, 2020. doi: 10.1145/3397271.3401075.
6. G. Izacard and E. Grave, "Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering," *Proceedings of EACL*, pp. 874–880, 2021. doi: 10.18653/v1/2021.eacl-main.74.
7. X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "TinyBERT: Distilling BERT for Natural Language Understanding," *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4163–4174, 2020. doi: 10.18653/v1/2020.findings-emnlp.372.
8. S. Es, J. James, L. Espinosa Anke, and S. Schockaert, "RAGAs: Automated Evaluation of Retrieval Augmented Generation," *Proceedings of EACL: System Demonstrations*, pp. 150–158, 2024. doi: 10.18653/v1/2024.eacl-demo.16.
9. S. Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," *arXiv preprint arXiv:2112.04426*, 2021. doi: 10.48550/arXiv.2112.04426.
10. O. Nogueira and K. Cho, "Passage Re-ranking with BERT," *arXiv preprint arXiv:1901.04085*, 2019. doi: 10.48550/arXiv.1901.04085.
11. M. Malkov and D. Yashunin, "Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. doi: 10.1109/TPAMI.2018.2889472.
12. N. Reimers and I. Gurevych, "Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation," *Proceedings of EMNLP*, pp. 4512–4525, 2020. doi: 10.18653/v1/2020.emnlp-main.365.
13. X. Qu et al., "RocketQA: An Optimized Training Approach to Dense Passage Retrieval for Open-Domain Question Answering," *Proceedings of NAACL-HLT*, 2021. doi: 10.18653/v1/2021.naacl-main.176.

14. P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. doi: 10.48550/arXiv.2005.11401.
15. K. Guu et al., "REALM: Retrieval-Augmented Language Model Pre-Training," *Proceedings of ICML*, pp. 3929–3938, 2020. doi: 10.48550/arXiv.2002.08909.
16. M. Kwiatkowski et al., "Natural Questions: A Benchmark for Question Answering Research," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 453–466, 2019. doi: 10.1162/tacl_a_00276.
17. P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text," *Proceedings of EMNLP*, pp. 2383–2392, 2016. doi: 10.18653/v1/D16-1264.
18. P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," *Proceedings of ACL*, pp. 784–789, 2018. doi: 10.18653/v1/P18-2124.
19. T. Wolf et al., "Transformers: State-of-the-Art Natural Language Processing," *Proceedings of EMNLP: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6.
20. C. Raffel et al., "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. doi: 10.48550/arXiv.1910.10683.