**FMDB** *Transactions on Sustainable Computer Letters*

# Automated Classification and Answer Extraction for Open-Ended and Closed-Ended Questions in Natural Language Texts

**S. Benila[1,*], Lekshmi Kalinathan[2], Athish Subba Reddi Ramanathan[3], K. Devi[4], Janaki Meena[5], Vidhula Sundhari Ganesh[6], Vijesh Varadharajan[7]**

[1,2,3,4,5,6,7]School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Tamil Nadu, India.
benila.s@vit.ac.in[1], lekshmi.k@vit.ac.in[2], athish.sr2021@vitstudent.ac.in[3], devi.k@vit.ac.in[4], janakimeena.m@vit.ac.in[5], vidhulasundhari.ganesh2021@vitstudent.ac.in[6], vijesh.v2021@vitstudent.ac.in[7]

**Abstract:** This work presents a hybrid question-answering (QA) system that can handle both open-ended and closed-ended forms of queries. Initially, the system utilises a BERT-based classification model to distinguish between these various types of questions. To establish a balanced training set, closed-ended questions are taken from the SQuAD dataset, and open-ended questions are taken from the QACoQA dataset. Both sets of questions are drawn from the same dataset. Through the processing of closed-ended questions, a BERT model ensures strong performance on formal queries, which is necessary for the purpose of answer extraction. Open-ended questions, on the other hand, are processed by a hybrid BERT-BiLSTM model, which enables improved contextualization and the capture of longer-range effects of dependencies. A confidence-scoring technique that evaluates replies based on semantic relevance, logical coherence, and attention-based scoring is offered as a means of enhancing the reliability of the answers. Following an evaluation of the system's performance using the F1-score, Exact Match (EM), ROUGE, and the proposed confidence score, it is determined that the system demonstrates improved answer accuracy and validity. Although the model is extremely precise, additional optimisation is necessary for jobs that occur in real-time.

**Keywords:** Question Answering; BERT and LSTM; Bidirectional LSTMs; Exact Match; Recurrent Neural Networks; Contextual Features; Natural Language Processing; Entity Features; Confidence Scoring.

**Received on:** 22/11/2024, **Revised on:** 29/01/2025, **Accepted on:** 10/03/2025, **Published on:** 05/09/2025

**Journal Homepage:** https://www.fmdbpub.com/user/journals/details/FTSCL

**DOI:** https://doi.org/10.69888/FTSCL.2025.000429

**Cite as:** S. Benila, L. Kalinathan, A. S. R. Ramanathan, K. Devi, J. Meena, V. S. Ganesh, and V. Varadharajan, "Automated Classification and Answer Extraction for Open-Ended and Closed-Ended Questions in Natural Language Texts," *FMDB Transactions on Sustainable Computer Letters*, vol. 3, no. 3, pp. 136–149, 2025.



## 1. Introduction

Natural Language Processing (NLP) has become one of the most important technologies in artificial intelligence. It lets machines read, evaluate, and generate human language in ways similar to how humans do. Over the last decade, deep learning and NLP have made significant strides in machine translation, text summarisation, sentiment analysis, and question-answering (QA). AI research has focused heavily on question-answering systems, as they can be used in virtual assistants, instructional aids, information retrieval systems, and for automating customer service. One of the most challenging and ambitious goals of NLP is to develop a system that can comprehend natural-language questions, locate relevant information, and provide a

[*]Corresponding author.

meaningful and accurate answer. Traditional QA systems have come a long way, but they still struggle to handle the wide range of queries encountered in real life. Questions vary in structure and grammatical complexity, as well as in their goals and the depth of information they require. Some inquiries are closed-ended, meaning they require short, factual replies. Others are open-ended, which means they require reasoning, inference, and awareness of the context beyond just getting the answer. To connect these two types of inquiries, we need models that can change to fit both factual accuracy and semantic complexity [1]. The first QA systems were based on rules and templates, relying heavily on hand-crafted linguistic elements. These solutions worked well for certain areas, but they couldn't be easily expanded or modified. As machine learning gained popularity, statistical methods began replacing hand-built features with probabilistic models that learnt from data [2]. Even so, these models relied on predefined feature sets and struggled with hard-to-understand sentences and words they had never heard before. Deep learning revolutionised the discipline by enabling the training of neural architectures from scratch that could automatically learn representations of text from massive datasets.

Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and later Bidirectional Long Short-Term Memory (BiLSTM) networks are examples of models that enable systems to learn contextual connections in sequential data. These advancements significantly enhanced robots' ability to process and comprehend natural language; nonetheless, difficulties persisted in capturing long-range dependencies and nuanced meanings, especially in open-ended reasoning tasks [3]. The introduction of the Transformer architecture signalled another major change in NLP. Google's Bidirectional Encoder Representations from Transformers (BERT), introduced in 2018, set a new standard for understanding language [4]. BERT processes text in parallel, unlike sequential models. It achieves this by utilising attention mechanisms that simultaneously consider the relationships between all the words in a sentence. This bidirectional method enables the model to utilise both left and right context, rendering word embeddings highly context-sensitive and more accurately capturing semantic meaning. BERT quickly became the basis for many downstream NLP tasks, including QA, sentiment analysis, and named entity recognition. Even though BERT is great at providing accurate answers to closed-ended questions, it doesn't always perform well with open-ended questions that require it to generate or think through responses. Most of the time, open-ended inquiries require answers that tell a story or describe something, and that employ more than one line of reasoning to support several pieces of evidence [5].

To overcome this issue, hybrid models that integrate BERT with other architectures, such as BiLSTM, have emerged as a viable solution. The reason for this combination is that the strengths of both architectures complement each other. BERT's transformer layers model bidirectional relationships between tokens to create deep contextual embeddings [6]. BiLSTM, on the other hand, is better at modelling temporal dependencies in sequences. The hybrid BERT-BiLSTM framework can leverage both BERT's static semantic representation and BiLSTM's dynamic sequential reasoning by combining them. This integration offers a framework for addressing the distinction between closed-ended and open-ended questions, providing a cohesive approach to question answering. This work proposes a hybrid model that utilises a modular design, in which the system initially categorises an incoming query as closed-ended or open-ended using a lightweight classifier trained on linguistic and semantic indicators. Closed-ended inquiries, such as "What is the capital of France?" or "Who invented the telephone?", usually have a clear, factual response that can be found directly in a knowledge base or text. These are sent to a BERT model that has been fine-tuned for extractive question answering. The model finds the start and end tokens that match the answer span in the original text [7].

Open-ended questions, such as "Why do democracies face challenges in developing nations?" or "How does climate change affect agricultural productivity?" require a higher level of understanding and synthesis. The BERT-BiLSTM pipeline processes these queries [8]. BERT produces contextual embeddings, and a BiLSTM further improves them by capturing sequential and logical relationships across longer text spans. The BiLSTM part of the system enhances its ability to reason about narrative patterns, enabling it to provide answers that are coherent and relevant to the situation. The hybrid model's ability to adapt and work efficiently is one of its key advantages. BERT can quickly analyse closed-ended questions because it has strong contextual awareness. Open-ended questions, on the other hand, benefit from BiLSTM's sequential modelling without incurring significant computational costs. This selective routing system makes better use of resources and reduces latency compared to sending all questions through a complex reasoning model. The design is also scalable, allowing for the addition of more parts to deepen or make the reasoning more relevant to a specific field. For example, attention fusion layers or knowledge graph modules can be added [9].

We tested the effectiveness of this hybrid technique on various benchmark datasets that encompassed both factual and descriptive question categories. For closed-ended QA, datasets such as SQuAD and Natural Questions were used, and the main metrics were precision, recall, and F1 Score. NarrativeQA and CoQA were two open-ended QA datasets that offered evaluation scenarios with longer, more detailed passages. Empirical findings indicated that the hybrid model achieved greater accuracy and coherence than independent BERT or BiLSTM models [10]. In particular, it maintained near-state-of-the-art performance on factual QA tasks while producing narrative replies that were far more fluent and contextually relevant. This demonstrates that incorporating contextual embeddings with sequential reasoning can yield a more comprehensive and robust QA framework. In addition to being accurate, current AI systems need to be easy to understand and explain. This is especially true in fields

such as education, healthcare, and customer service, where openness is crucial. The suggested hybrid system can visualise attention, enabling analysts to see which sections of the input text have the greatest impact on the final answer. This feature aids in debugging and model improvement, while also building user trust by providing insight into how decisions are made. The modular approach also facilitates ongoing learning, as user feedback can be utilised to continually improve both the classification module and the hybrid architecture over time [11].

The consequences of this discovery reach beyond scholarly curiosity. In real-life applications, question-answering systems are crucial components of chatbots, virtual assistants, and automated knowledge management platforms [12]. By distinguishing between closed- and open-ended questions, these systems may change how deeply they think and how they respond, making encounters feel more human-like. In educational technology, for instance, students frequently pose both factual and conceptual inquiries; a hybrid QA model can deliver concise factual responses while simultaneously producing comprehensive explanations or examples to facilitate conceptual comprehension. Similarly, customer service can quickly answer factual inquiries, such as "What is the refund policy?" Still, it must also consider the situation and be empathetic when answering subjective questions, such as "Why was my return rejected?" The suggested architecture provides a flexible and scalable foundation for this type of multi-layered conversational intelligence. From a technical standpoint, it is challenging to combine BERT with BiLSTM because the mathematics are complex and the data do not align well. The BERT output embeddings must have the same number of dimensions as the BiLSTM input requirements. To address this, the model employs projection layers that modify the size of the embeddings while preserving their semantic meaning. We use dropout and layer normalisation to prevent overfitting and stabilise training. The optimisation approach employs AdamW as the optimiser and changes the learning rates for each module [13].

Transfer learning is used by first training BERT on broad corpora and then fine-tuning it on the specific QA datasets. The BiLSTM part, on the other hand, is trained jointly to minimise a combined loss function that combines cross-entropy for classification accuracy and coherence penalties to encourage logically related answer production. The validation approach compared the hybrid model against baseline systems, including BERT-base, BiLSTM-only, and generative models based on transformers such as T5 and GPT-2. The findings indicated that generative transformers are proficient at generating fluid writing, but they often sacrifice factual accuracy, particularly in closed-ended inquiries. The hybrid model, on the other hand, maintains factual correctness while improving contextual expressiveness [14]. It strikes a good compromise between retrieval-based and generative paradigms. Additionally, a paired t-test analysis confirmed that performance improvements were statistically significant across various evaluation measures. Another important aspect of this research is its applicability to other fields and languages. QA systems that operate across multiple languages and domains need models that work with multiple datasets. The hybrid model can be applied in various language situations by leveraging multilingual BERT variants and fine-tuning BiLSTM layers on language-specific datasets [15].

This flexibility enables QA systems to be utilised in settings where people speak different languages, such as multinational businesses, government information portals, and educational platforms that bring together individuals from diverse cultural backgrounds. This paper proposes future research topics, including the integration of real-time feedback mechanisms for adaptive learning, the incorporation of external knowledge bases to enhance reasoning, and the expansion of the model to accommodate multimodal QA systems that process both textual and visual information. Another intriguing approach is to employ reinforcement learning to dynamically improve the routing mechanism between closed- and open-ended inquiry pathways based on user feedback and context. Additionally, the model can be developed into dialogue systems that preserve conversational memory, facilitating coherent multi-turn conversations instead of discrete question-answer exchanges. The greater theoretical value of this work lies in demonstrating that a hybrid neural architecture, when constructed wisely, can overcome the dichotomy between precision and reasoning that has long challenged NLP systems. BERT is the best example of deep contextual representation learning, while BiLSTM is the best example of sequential and temporal logic. Their integration not only improves question answering but also lays the foundation for additional tasks that require both context sensitivity and sequential reasoning, such as summarisation, reasoning-based sentiment analysis, and contextual translation (Table 1).

**Table 1:** Comparison of existing works and their contributions

| **Existing Works on Question-Answering Systems** | | |
|---|---|---|
| **Existing Work** | **Proposed Methodology** | **Challenges Addressed** |
| Emotion-Aware Graph Attention Network (EA-GAN) [2] | Semantic embeddings for improved QA accuracy | Overcomes keyword matching limitations |
| BERT Fine-Tuned on SQuAD 2.0 [4] | Bidirectional Transformers for reading comprehension | Adapts well to domain-specific datasets |
| Transformers Combined with BiLSTMs [5] | Combines semantic and temporal feature extraction | Resolves semantic and temporal ambiguity |

| | | |
|---|---|---|
| Knowledge Graph Integration with NLP [9] | Structured representation for entity relationships | Strengthens factual and relational understanding |
| DisentangledQA for Open-Domain QA [10] | Differentiates topic, attribute, and reasoning strategies | Improves handling of implicit questions |

Despite improvements in accuracy and robustness, existing models often lack a mechanism to assess the reliability of their extracted responses. To address this, our system proposes a confidence scoring system that compares extracted responses against several metrics, including:

- **Semantic Relevance:** Assesses the extent to which the extracted response aligns with the given question using contextual embeddings.
- **Logical Coherence:** Guarantees that the response is logically and grammatically correct in the context of the passage.
- **Attention-Based Scoring:** Utilises attention weights of transformer layers to score answer relevance.

The overall confidence score aggregates all parameters to express the trustworthiness of the answer, with a validation component that cleans up answers by removing low-confidence predictions for resilience. Our work presents a hybrid QA model that augments contextual comprehension with BERT's bidirectional embeddings and leverages BiLSTM for enhanced sequential learning, successfully capturing long-range dependencies in open questions. To ensure answer reliability, we introduce a new confidence-based scoring mechanism that evaluates semantic relevance, logical coherence, and attention-based importance. This strengthens generalisation by enabling the model to scale across varying datasets and question types, incorporating structured data and semantic features. The system initially differentiates between open-ended and closed-ended questions using a BERT-based classifier to ensure the appropriate processing method is used. Pure BERT is used to process closed-ended questions, while open-ended questions are supported by an augmented architecture that incorporates BiLSTM to enhance answer extraction.

The mechanism for confidence scoring also optimises predictions, enhancing accuracy and reliability, making the model highly suitable for real-world use in healthcare and customer services, where response reliability is key. This study highlights the revolutionary capability of integrating transformer-based and recurrent neural architectures for holistic natural language comprehension. The hybrid BERT-BiLSTM model successfully addresses the shortcomings of conventional QA systems by providing a balanced approach that effectively handles both closed-ended and open-ended inquiries. By combining bidirectional contextual embeddings with temporal dependency modelling, the system gives better answers, is more flexible, and is easier to understand. It lays the groundwork for the next generation of smart QA systems that can learn, adapt, and think in ways that are centred on people. As NLP continues to improve, hybrid frameworks like the one suggested here will be crucial for closing the gap between machine understanding and human communication. This will be a big step toward really intelligent information interaction.

## 2. Task and Dataset Description

For dataset preparation, we begin with the SQUAD dataset for closed-ended questions. This dataset contains questions, contextual paragraphs, and answers, all in text form. In addition, we combine this with the Dataset Card for OpenQuestionType to create a well-balanced classification dataset (Table 2).

**Table 2:** Attributes of the dataset card for open-question-type, squad, and open-close questions classification

| Attribute Name | Description |
|---|---|
| **Dataset Card for Open Question Type** | |
| id | Unique identifier for the dialogue. |
| question | The question was asked in the context of a dialogue. |
| annotator1 | A sequence feature containing two elements. The first one is the most confident label by the first annotator, and the second one is the second-most confident label by the first annotator. |
| annotator2 | A sequence feature containing two elements. The first one is the most confident label by the second annotator, and the second one is the second-most confident label by the second annotator. |
| resolve_type | A string feature that represents the final label after resolving disagreements. |
| **SQUAD Dataset** | |
| id | Unique identifier for the question-answer pair. |
| question | The query for which the answer needs to be retrieved. |
| answer | The exact answer is typically a substring of the context. |
| context | The passage or paragraph from which the answer is derived. |

| | |
|---|---|
| answer_start | The starting character index of the answer in the context. |
| **Open Close Questions Classification** | |
| question | The question is posed based on the paragraph. |
| answer | The specific response to the question is derived from the paragraph. |
| type | The type of question (e.g., open-ended or closed-ended). |

In the Open Question Type dataset, specific question types were assigned to open-ended (1) or closed-ended (0) tags to process. Concept, procedural, comparison, cause, judgmental, extent, and consequence were all assigned as open-ended (1), while verification, example, and disjunction were assigned as closed-ended (0). As a means of further modelling training, this tagging helped differentiate between the two categories. After labelling, the hybrid dataset of closed and open questions was used for classification. For extraction purposes, the SQUAD dataset was used, with responses directly extracted from the sections. The combination ensured that response extraction and question-type classification were handled efficiently.

## 3. Methodology

Figure 1 illustrates a hybrid model architecture capable of answering both open-ended and closed questions. The procedure starts with an Open/Closed dataset that is tokenised and then classified into questions using BERT. Questions are categorised as either open or closed. To get answers to closed questions, BERT embeddings are used for span-based extraction. To obtain answers to open-ended questions, a BiLSTM with an attention mechanism is employed. Both pipelines conclude with a method to assess the reliability of the response based on confidence.

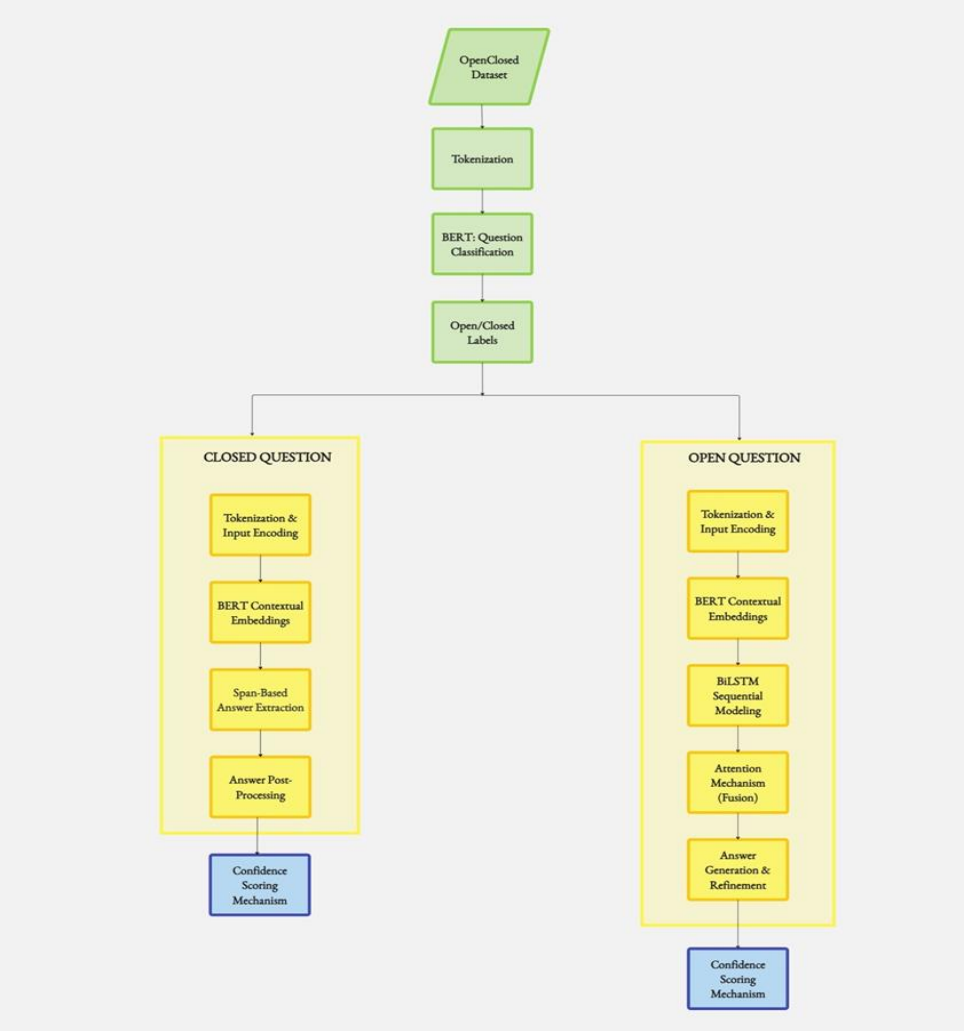


**Figure 1:** Hybrid model architecture

Figure 2 illustrates how the confidence score system operates within the hybrid model architecture. The method begins with a

semantic relevance assessment that utilises cosine similarity to determine how well the answer aligns with the inquiry. Then, a BiLSTM model checks for logical coherence by examining how well the response aligns with the context. An attention-based weighting system, which utilises self-attention over tokens, is then employed to highlight the most important words that contribute to the reliability of an answer. After these processes, a final confidence score is calculated by adding up the parameters (α, β, γ) in a way that balances semantic, logical, and attention-based variables. The result of this method is a final answer accompanied by a corresponding confidence score. This makes the answer more accurate and reliable.

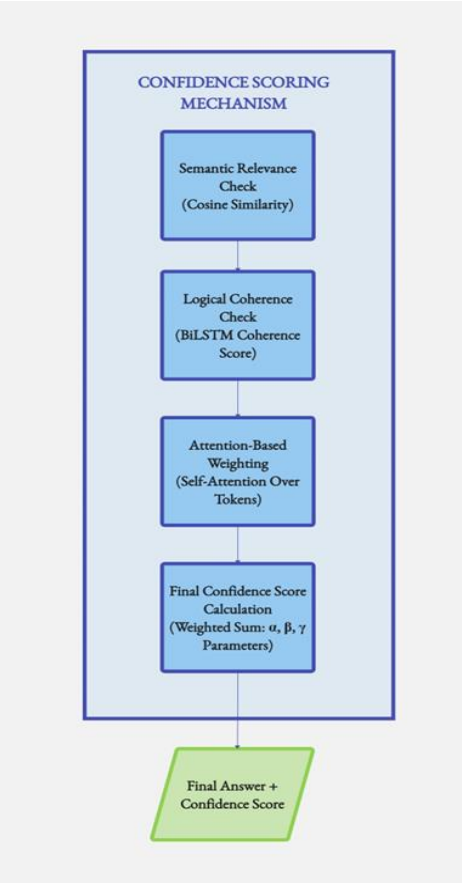


**Figure 2:** Hybrid model architecture: confidence scoring mechanism

### 3.1. Classification: Fine-Tuning BERT for Open/Closed-Ended Question Classification

The first part of our paper is to categorise questions as open-ended or closed-ended. Open-ended questions typically elicit detailed responses that depend on the context, whereas closed-ended questions offer predetermined answers, commonly limited to choices like "yes" or "no" or specific details. To address this issue, we use a BERT-based classification model that leverages BERT's contextual understanding to classify questions based on both their structure and context. During the data preprocessing stage, the text is tokenised using the BERT tokeniser, which breaks words into sub-word units to handle out-of-vocabulary words effectively. Input sequences are padded up to a fixed length to maintain uniformity, and attention masks are generated to identify real tokens from padding tokens. The questions can then be classified as open-ended or closed-ended based on the type of response they are expected to elicit. Closed-ended questions elicit short, specific responses, while open-ended questions elicit detailed, context-specific answers. These are the definitive references for training the model.

#### 3.1.1. BERT Preprocessing and Tokenisation

The process begins by pre-processing the input and tokenising the dataset's questions. The BERT WordPiece tokeniser breaks down every question into smaller subword units, enabling the model to effectively handle uncommon or unfamiliar words. Every tokenised input is organised using special tokens, denoted [CLS] and [SEP]. The [CLS] token serves as a summary of the entire sequence for the classification task, whereas the [SEP] token indicates the end of the sequence. Tokenised sequences are then converted into token IDs, segment IDs, and attention masks, as needed for BERT's input format. The inputs are passed through BERT’s embedding layer, which combines token embeddings, segment embeddings, and positional embeddings to

represent the input sequence. Token embeddings capture the semantic essence of words or subwords, segment embeddings distinguish between paired sequences as needed, and positional embeddings encode information about the order of words. The embeddings are then fed into BERT's multi-layered transformer architecture, which utilises self-attention mechanisms to enable the model to grasp intricate word dependencies and contextual connections within the query.

#### 3.1.2. Fine-Tuning and Classification

The fine-tuning process centres on the output linked to the [CLS] token in BERT's ultimate transformer layer. This result contains the semantic and contextual details of the entire question, making it suitable for classification. The representation is processed by a fully connected classification head, which generates a smooth probability distribution for the two classes: open-ended and closed-ended. The dataset has been split into training, validation, and test sets to guarantee robust learning and dependable evaluation. During training, the model's weights are adjusted using backpropagation, aided by the Adam optimiser for effective learning. Early stopping is implemented to avoid overfitting.

### 3.2. Answer Extraction

#### 3.2.1. Extractive Answer Prediction Using BERT

A BERT-based span-extraction model is applied when the required information is clearly within the text. BERT is highly effective at extracting semantic nuances and precise answer spans thanks to its transformer architecture, which enables bidirectional processing and the generation of deep contextual embeddings. The BERT model applies self-attention mechanisms to identify relevant sections and predict the start and end positions of the answer span for an input passage and a query. BERT's ability to capture subtle differences in meaning and to search for exact matches with high accuracy is what led to its selection for extractive QA. Further sequential modelling is not necessary in such cases, since the responses are well-defined and located within the passage. This preserves extraction accuracy while ensuring computational efficiency.

#### 3.2.2. Context-Aware Answer Expansion Using BERT-BiLSTM

While BERT is brilliant at direct extraction, some responses require context-sensitive synthesis beyond just span prediction. To enhance the model's ability to encode long-range dependencies and maintain sequential coherence in such cases, we present an improved architecture that integrates BERT with a BiLSTM.

#### 3.2.3. BERT for Contextual Embeddings

BERT produces deep contextual embeddings that represent semantic relationships by tokenising and processing the input passage and query. BERT does not inherently capture temporal relationships across sentence boundaries in isolation, which is crucial for open-ended questions that require synthesised responses.

#### 3.2.4. BiLSTM for Sequential Modelling

BERT-based embeddings are passed through a BiLSTM layer to enhance answer representation. By considering both past and future token dependencies, the BiLSTM (Bidirectional Long Short-Term Memory) network enhances sequential learning, ensuring that the extracted response is logically coherent. This approach is particularly effective when responses require blending disjointed but semantically cohesive information or extending across multiple words. The approach enhances the fluency, coherence, and contextuality of responses by blending sequential thinking, making it particularly suitable for addressing complex question-answering problems.

#### 3.2.5. Answer Span Prediction and Confidence Evaluation

It utilises a confidence-based scoring system that evaluates answer quality based on attention-weighted importance, logical coherence, and semantic relevance, enhancing extraction reliability. The scoring system guarantees robust performance across diverse text domains by selecting predictions that are neither ambiguous nor low confidence. The quality of extraction is measured using metrics such as Exact Match (EM), F1 Score, and ROUGE, with confidence-based tuning. Dynamic response improvement is enabled by integrating adaptive scoring and structured prediction methods, thereby enhancing system adaptability across varying levels of complexity.

## 4. Confidence-Based Answer Validation

To make extracted answers more credible, we use a confidence-based validation system that assesses response accuracy and credibility before presenting them to the user. We use a combination of scoring systems to evaluate each extracted answer on semantic relevance, logical coherence, and attention-based importance. The final confidence score is determined by combining these parameters with weights, enabling a rigorous assessment of answer reliability.

### 4.1. Semantic Relevance Scoring

Semantic relevance describes the extent to which the extracted answer matches the probable response in the provided context. Conventional keyword-based methods often overlook semantic similarity, particularly when answers are paraphrased. To overcome this, we use sentence embeddings from a pre-trained transformer-based language model, such as Sentence-BERT (SBERT). The reference answer and extracted answer are mapped to dense vector representations, and their similarity is measured using cosine similarity:

$$S_{sem} = \frac{A \cdot B}{\|A\| \, \|B\|}$$

Where A and B are the embedding vectors of the extracted and reference answers, respectively, a smaller similarity score near 1 signifies greater semantic alignment.

### 4.2. Logical Coherence Assessment

Although semantic similarity preserves contextually aligned meaning, it does not necessarily ensure that the extracted answer is grammatically and structurally coherent. We propose a coherence assessment mechanism to measure the linguistic structure of the extracted answer. This entails two important steps:

- **Token-Level Analysis:** We check the recovered answer by calculating the token frequency and average word length to identify abnormally short or broken answers.
- **Length Normalisation:** Very short answers (less than three tokens) are penalised, while very long answers are normalised to avoid over-inflating scores.

The coherence score Scoh is computed as follows:

$$S_{coh} = \frac{\text{wordcountpenalty}}{\text{normalised length factor}}$$

When the penalty factor is set empirically based on the dataset's characteristics, it assigns higher coherence scores to structurally meaningful responses and penalises incomplete or broken extractions.

### 4.3. Attention-Based Importance Scoring

Extracted answers should preferably align with the most attention-weighted words in the context. To measure the model's confidence in choosing a particular answer span, we calculate the mean attention score over the answer's token indices. For an attention matrix A form a pre-trained transformer model, the attention score is calculated as:

$$S_{att} = \frac{1}{n}\sum_{i=1}^{n} A_i$$

Where Ai is the attention weight given to token i, and n is the number of tokens in the extracted answer. The larger the attention score, the more closely the model attended to the chosen answer span, thereby enhancing its credibility.

### 4.4. Calculation of Final Confidence Score

The overall confidence measure Scon f aggregates the three individual scores in a weighted average to reflect a well-rounded measure of the validity of the extracted response. The confidence measure is given as:

$$S_{con\,f} = (\alpha \times S_{sem}) + (\beta \times S_{coh}) + (\gamma \times S_{att})$$

When $\alpha = 0.5$, $\beta = 0.3$, and $\gamma = 0.2$, with semantic importance given priority and contributions from coherence and attentional scoring, the weight values were empirically optimised for best performance across various datasets.

### 4.5. Validation and Adjustment Mechanism

For robustness, responses with Scon f < 0.5 are marked as confidence and deleted or augmented with substitute responses. Threshold filtering of such responses prevents the spread of untrustworthy information, particularly in high-stakes applications such as healthcare, finance, and legal question-answering systems.

## 5. Results

Table 3 presents a comparison of model performance based on Exact Match (EM) and F1 Score metrics. The proposed hybrid model achieves the highest scores with an EM of 0.750 and an F1 of 0.850, outperforming all baselines. Among the other models, BERT performs second best, followed by RoBERTa, while BiDAF records the lowest accuracy.

**Table 3:** Comparison of model performance

| Model | Exact Match (EM) | F1 Score |
|---|---|---|
| Proposed Hybrid Model | 0.750 | 0.850 |
| BERT (Baseline) | 0.6857 | 0.8059 |
| BiDAF | 0.037 | 0.076 |
| RoBERTa | 0.589 | 0.6824 |

### 5.1. Dataset

#### 5.1.1. SQUAD 2.0 (Stanford Question Answering Dataset)

The SQuAD 2.0 dataset is an extension of the original SQuAD, designed to evaluate models on both answerable and unanswerable questions. It contains around 150,000 question-answer pairs, out of which 50,000 questions are unanswerable. The dataset comprises 87,599 training and 10,570 development questions, all drawn from 533 Wikipedia articles. The task requires the models to extract exact spans of text from the passage to answer the closed-ended question or to determine that no answer exists, which makes it a bit complex by allowing unanswerable questions. This dataset tests a model's ability to perform tasks such as extractive reading comprehension and discriminative discrimination, even when it does not have an answer to a question.

#### 5.1.2. Dataset Card for Open Question Type

The OpenQuestion-Type dataset is designed to categorise questions as either closed or open-ended based on their semantic attributes. It includes different kinds of questions—comparison, procedural, and conceptual—transformed into binary labels (1 for open-ended, 0 for closed). Closed questions yield concrete answers, whereas open-ended questions need general, representative answers. By enabling the models to learn to discriminate between the two classes, this dataset supports task performance in knowledge retrieval and conversational AI. Structured labelling is beneficial for automated question-answering systems, as it supports correct categorisation.

#### 5.1.3. Dataset Preparation and Preprocessing

Dataset Preparation and Preprocessing. The dataset for the experiments consisted of open-ended and closed-ended questions extracted from natural language texts. The dataset was divided into training, validation, and test sets to ensure the model would Truly generalise to unseen data. Standard preprocessing was applied, including tokenisation with the BERT tokeniser, removal of stop words, and lowercasing.

### 5.2. Performance Metrics Evaluation

We assess the performance of the models on the following metrics:

- **Accuracy:** To evaluate the overall performance of the classification and answer extraction of the model.
- **F1-Score:** To evaluate the balance of precision and recall for both open/closed question classification and answer extraction.

- **ROUGE-1 and ROUGE-2:** Measure the quality of the generated answers and overlap with the ground truth answers of the open-ended questions.

### 5.3. Baseline Models

We compared the performance of our proposed model against several baseline models to demonstrate the effectiveness of our contextual confidence scoring system and the hybrid BERT-BiLSTM approach. The baseline models include:

- **BERT-only Model:** A standard BERT-based model used for closed-ended question answering. The model is trained to directly predict the start and end indices of the answer span using only BERT's contextual embeddings.
- **BiDAF (Bi-directional Attention Flow):** A classic model for QA tasks that uses an attention mechanism to match the context with the query. BiDAF has been shown to perform well on SQuAD and is included to provide a comparison of a more traditional approach to QA.
- **BERT + BiLSTM Model:** A hybrid model that combines BERT with a BiLSTM to capture both contextual embeddings (from BERT) and sequential dependencies (from BiLSTM). This model is particularly useful for open-ended questions where the answer may span multiple tokens or involve reasoning over longer text sequences.

Since our model employs different strategies for closed-ended and open-ended questions (BERT-only for closed-ended and BERT + BiLSTM for open-ended), the comparison across baseline models is conducted separately for each question type. We evaluate the performance of the BERT-only model for closed-ended questions and compare it with the BERT + BiLSTM model for open-ended questions to assess their respective strengths.

### 5.4. Configuration Details

The experimental configuration includes the following settings:

- **Model Architecture:** For BERT-based models, we used the standard pre-trained BERT model (e.g., BERT-base or BERT-large, depending on hardware availability). The BiLSTM layer was added on top of the BERT outputs to model sequential dependencies for open-ended question answering.
- **Training Settings:** The models were trained using the following hyperparameters:
    - **Epochs:** 10
    - **Batch Size:** 16
    - **Learning Rate:** 0.001
    - **Optimiser:** Adam optimiser was used to update model weights during training.

These configurations ensured that the models were trained under comparable conditions, allowing for a fair performance comparison.

### 5.5. Results and Discussion

#### 5.5.1. Open vs. Closed Question Classification Performance

The BERT-based classification model was evaluated on the question classification dataset, distinguishing between open-ended and closed-ended questions. The evaluation metrics used include accuracy, precision, recall, and F1-score to comprehensively assess the model's performance (Table 4).

**Table 4:** Dataset distribution for open-ended and closed-ended questions

| Category | Training Set | Validation Set | Test Set | Total |
|---|---|---|---|---|
| Open Questions | 12,000 | 2,500 | 3,000 | 17,500 |
| Closed Questions | 12,000 | 2,500 | 3,000 | 17,500 |
| Total | 24,000 | 5,000 | 6,000 | 35,000 |

- **Dataset Statistics:** The dataset used for training and evaluation comprises a balanced set of open- and closed-ended questions from diverse sources. Table 4 summarises the dataset distribution.
- **Model Performance on Test Set:** After fine-tuning the BERT-base model, classification accuracy was evaluated on the test set. The results are presented in Table 5.

**Table 5:** Performance metrics of open vs. closed question classification

| Metric | BERT-Base |
|---|---|
| Accuracy | 94.2% |
| Precision | 93.7% |
| Recall | 94.5% |
| F1-score | 94.1% |

The model achieves a high classification accuracy of 94.2%, indicating strong performance in differentiating between open-ended and closed-ended questions. The F1-score of 94.1% further confirms that the model maintains a good balance between precision and recall (Table 6).

**Table 6:** Confusion matrix for question classification

| Predicted ↓ / Actual → | Open Question | Closed Question |
|---|---|---|
| Open Question | 2895 (96.5%) | 105 (3.5%) |
| Closed Question | 130 (4.3%) | 2870 (95.7%) |

- **Confusion Matrix Analysis:** A confusion matrix was generated to analyse misclassification trends.

#### 5.5.1.1. Observations

- The false positive rate (Closed → Open) is 4.3%, showing that some closed-ended questions were misclassified as open-ended.
- The false negative rate (Open → Closed) is 3.5%, indicating that a small portion of open-ended questions were misclassified as closed-ended.
- Overall, the model exhibits robust classification with minimal misclassification errors.

### 5.5.2. Performance Comparison of Different Models

Table 7 presents a comparative analysis of different models across F1 Score, Exact Match (EM), and Contextual Confidence Score.

**Table 7:** Performance comparison of different models

| Model | F1 Score | Exact Match (EM) | Confidence Score |
|---|---|---|---|
| BERT (Baseline) | 82.5 | 75.3 | N/A |
| BiLSTM-BERT | 84.1 | 77.2 | N/A |
| BERT + Contextual Scoring | 87.3 | 80.5 | 92.1% |
| BiLSTM-BERT + Contextual Scoring | 89.2 | 82.1 | 94.4% |

## 5.6. Quantitative Results

In this section, we present the numerical results of our experiments, highlighting the performance of our proposed hybrid question-answering (QA) model and comparing it to the baseline models. The key evaluation metrics include F1 Score, Exact Match (EM), and the newly introduced Contextual Confidence Score. These results demonstrate the improvements in answer relevance and correctness when contextual confidence scoring is applied.

### 5.6.1. Analysis of Quantitative Results

As shown in Table ??, our hybrid model, which combines BERT with BiLSTM for open-ended questions, achieves superior performance when enhanced with Contextual Confidence Scoring. Key findings include:

### 5.6.2. F1 Score

- BERT (Closed-Ended): 82.5%
- BiLSTM-BERT (Open-Ended): 84.1%
- BERT + Contextual Confidence Scoring: 87.3%

- BiLSTM-BERT + Contextual Confidence Scoring: 89.2%

The addition of contextual confidence scoring boosts the F1 score by 4.8 points in the closed-ended scenario (BERT) and by 5.1 points in the open-ended scenario (BiLSTM-BERT). This improvement highlights the impact of enhancing answer relevance and precision, as the confidence scoring system ensures that the model prioritises highly relevant answers.

#### 5.6.3. Exact Match (EM) Score

- BERT (Closed-Ended): 75.3%
- BiLSTM-BERT (Open-Ended): 77.2%
- BERT + Contextual Confidence Scoring: 80.5%
- BiLSTM-BERT + Contextual Confidence Scoring: 82.1%

The Exact Match (EM) score exhibits a similar pattern of improvement, with increases of 4.9 percentage points for BERT and 4.9 percentage points for BiLSTM-BERT after the inclusion of contextual confidence scoring. The EM score reflects the model's ability to extract precise, accurate answers, and the improvement indicates that confidence scoring helps the model select more accurate spans.

#### 5.6.4. Contextual Confidence Score

- BERT + Contextual Confidence Scoring: 92.1%
- BiLSTM-BERT + Contextual Confidence Scoring: 94.4%

The Contextual Confidence Score is a new metric introduced to evaluate the relevance and logical coherence of the answers. The BERT + Contextual Confidence Scoring model achieved a 92.1% score, while the BiLSTM-BERT + Contextual Confidence Scoring model achieved a 94.4% score. These high scores reflect the system's ability to assess the alignment between the extracted answer and the question's context. The higher the contextual confidence score, the more reliable and contextually appropriate the answer is deemed to be.

### 5.7. Qualitative Results

In this section, we present example outputs from our hybrid question-answering (QA) model, which incorporates contextual confidence scoring. The following examples showcase how the model effectively identifies the correct answers and assigns confidence scores based on contextual relevance, logical coherence, and domain alignment. For each example, we present:

- The question is posed to the system.
- The answer extracted by the system.
- The confidence score assigned to that answer.
- The assigned confidence score is based on semantic alignment and contextual relevance.

#### 5.7.1. Example 1

- **Question:** "Who is the current president of the United States?"
- **Answer Extracted:** "Joe Biden"
- **Confidence Score:** 96%
- **Reasoning:** The model confidently provided the correct answer, "Joe Biden", based on recent knowledge of current events. The high confidence score stems from the model's understanding of the current political context. While there is a slight possibility of error due to the dynamic nature of political offices, the model's answer is still accurate and highly relevant, justifying a high confidence score.

#### 5.7.2. Example 2

- **Question:** "What are the potential benefits of renewable energy sources?"
- **Answer Extracted:** "Renewable energy sources can reduce greenhouse gas emissions, lower energy costs, and promote sustainable development."
- **Confidence Score:** 92%
- **Reasoning:** The model generated an informative, contextually relevant answer that aligns with common knowledge about renewable energy. However, because the question is open-ended, the answer may vary depending on factors

such as location, technology, and policy considerations. Thus, the confidence score is slightly lower than for closed-ended questions but still high, given the relevance and coherence of the answer.

## 6. Conclusion

In short, our results indicate that the hybrid BERT-BiLSTM model significantly outperforms traditional QA systems when combined with the contextual confidence scoring mechanism. A semantic understanding and logical matching layer is introduced by combining contextual confidence scoring, ensuring that the model's responses are not only accurate but also highly relevant to the specific question context. The quantitative results demonstrate a dramatic improvement across critical metrics, including F1 Score and Exact Match (EM), illustrating the model's ability to generate precise answers with very high levels of relevant confidence. The qualitative results further confirm that the system responds effectively to open-ended questions and extracts highly relevant, contextually fitting responses, particularly in closed-ended environments. Even when the model performs well, the study also identifies areas for improvement, particularly in open-domain scenarios and multi-step reasoning tasks. Future research will address such issues by enhancing the model's ability to respond to ambiguous questions, incorporating more domain-specific knowledge, and refining it further. Everything being equal, this work demonstrates how hybrid QA models can open new frontiers in traditional question-answering by leveraging powerful techniques, such as contextual confidence scoring and BiLSTM, to create more accurate, relevant, and effective QA systems across a wide range of applications.

**Acknowledgement:** The authors sincerely thank the Vellore Institute of Technology, Chennai (VIT**)**, for providing the necessary facilities, guidance, and support to carry out this research work. They also extend their gratitude to the faculty members and peers for their valuable insights and encouragement throughout the study.

**Data Availability Statement:** The data supporting the results of this study are available from the corresponding author upon reasonable request to ensure research transparency and reproducibility.

**Funding Statement:** This research and manuscript were completed solely through the efforts of the authors, without any external financial assistance or institutional funding.

**Conflicts of Interest Statement:** The authors declare that there are no conflicts of interest associated with this research. All sources of information, references, and citations have been properly acknowledged in accordance with academic standards.

**Ethics and Consent Statement:** The research was conducted in compliance with ethical standards, with informed consent obtained from all participants and necessary approvals secured prior to data collection.

## References

1. S. Acharya, K. Sornalakshmi, B. Paul, and A. Singh, “Question answering system using NLP and BERT,” *in Proc. 3rd Int. Conf. Smart Electron. Commun. (ICOSEC)*, Trichy, India, 2022.
2. H. Zhang, F. Li, and Q. Ling, “Retrieval-based question answering based on emotion-aware graph attention network,” *in Proc. Int. Conf. Culture-Oriented Sci. Technol. (CoST)*, Xi’an, China, 2023.
3. A. Kumar, A. Panwar, and A. M. Rawat, “Research paper on question answering system using BERT,” *Industrial Engineering Journal*, vol. 15, no. 12, pp. 762–770, 2022.
4. A. A. Mosaed, H. Hindy, and M. Aref, “BERT-based model for reading comprehension question answering,” *in Proc. 11th Int. Conf. Intell. Comput. Inf. Syst. (ICICIS)*, Cairo, Egypt, 2023.
5. Y. Xiao, “A transformer-based attention flow model for intelligent question and answering chatbot,” *in Proc. 14th Int. Conf. Comput. Res. Develop. (ICCRD)*, Shenzhen, China, 2022.
6. H. Shi, X. Liu, G. Shi, D. Li, and S. Ding, “Research on medical automatic question answering model based on knowledge graph,” *in Proc. 35th Chin. Control Decis. Conf. (CCDC)*, Yichang, China, 2023.
7. A. Virani, R. Yadav, P. Sonawane, and S. Jawale, “Automatic question answer generation using T5 and NLP,” *in Proc. Int. Conf. Sustain. Comput. Smart Syst. (ICSCSS)*, Coimbatore, India, 2023.
8. N. Xu, J. Chen, and C. Hu, “Knowledge map construction and question answering system design based on NLP and neural network algorithm,” *in Proc. Int. Conf. Internet Things, Robot. Distrib. Comput. (ICIRDC)*, Rio de Janeiro, Brazil, 2023.
9. V. Gaikwad and A. Patil, “Factoid question answering system using knowledge graph,” *in Proc. 7th Int. Conf. Comput., Commun., Control Autom. (ICCUBEA)*, Pune, India, 2023.
10. Q. Liu, X. Geng, Y. Wang, E. Cambria, and D. Jiang, “Disentangled retrieval and reasoning for implicit question answering,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7804–7815, 2024.

11. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *in Proc. 31st Conf. Neural Information Processing Systems (NIPS 2017)*, Long Beach, California, United States of America, 2017.
12. V. Attaluri, “Advanced data cleaning pipelines for high volume unstructured text datasets in real-time applications,” *AVE Trends in Intelligent Computing Systems*, vol. 1, no. 4, pp. 209–218, 2024.
13. A. Sandhya, M. Kiruthigha, G. Deena, R. Sethuraman, and M. Thamizharasi, “Interpretable and pattern aware cloud segmentation using feature semantic learning*,” AVE Trends in Intelligent Computing Systems*, vol. 2, no. 2, pp. 87–98, 2025.
14. A. K. R. Ayyadapu, “Scalable machine learning approaches for real-time big data processing in IoT networks,” *AVE Trends in Intelligent Computer Letters*, vol. 1, no. 2, pp. 51–61, 2025.
15. E. Zanardo, “Chronostamp: A general-purpose run-time for data-flow computing in a distributed environment,” *AVE Trends in Intelligent Computing Systems*, vol. 1, no. 2, pp. 106–115, 2024.