scientific reports

Check for updates

OPEN

# Resolving passage ambiguity in machine reading comprehension using lightweight transformer architectures

Adnan Nawaz[1], Muzamil Ahmed[2], Hikmat Ullah Khan[3]✉, Ali Daud[4]✉, Bader Alshemaimri[5] & Tassawar Iqbal[1]

**Machine Reading Comprehension (MRC) refers to generating precise responses to the users' queries from text content using natural language processing. The exponential growth and complexities of online content have made it difficult to surf the required information by navigating through several web pages to retrieve precise and accurate answers to the users' questions. Therefore, MRC has emerged as an active and growing research area in recent years. The existing studies highlight the significance of deep learning models yet lack in resolving ambiguity, especially in complex passages. Bidirectional encoder representations from transformers have addressed passage ambiguity resolution, but their complexity results in the demand for high computational resources and a large volume of data for better text comprehension. To address passage ambiguities and reduce computational costs, this study fine-tunes the DistilBERT model for the MRC task. The resulting model termed Distil-BERT-MRC uses a reduced architecture ensuring efficiency while maintaining competitive performance. The results of the detailed analysis demonstrate that Distil-BERT-MRC attained up to 90.23% exact match and 91.42% F1 score on the WikiQA dataset. Moreover, to assess the generalizability and resource utilization, extensive experiments were performed on SQuAD 2.0, NewsQA, and Natural Questions using recent transformer models, including RoBERTa and XLNet. Overall, our findings confirm that distilled transformer models provide a resource-efficient and effective approach for MRC tasks.**

**Keywords** Deep learning, Machine reading comprehension, Natural language processing, Question answering, transformers model

Natural Language Processing (NLP), a subset of artificial intelligence, enables machines to understand, interpret, and generate human language through a blend of linguistics and computational techniques. NLP's capacity to automate text-based interactions has led to a considerable increase in importance in recent years, which is essential in applications such as question answering (QA) systems, chatbots, voice assistants, and intelligent personal assistants[1]. When a user asks a natural language question, QA is used to deliver the best possible response. QA is different from search engines because search engines return a list of items, products, or services relevant to the query, while QA provides a direct answer to the question. There are four main types of QA tasks based on the structure as shown in Fig. 1: Text-based, Knowledge-based, Visual-based QA, and Hybrid QA. Specifically, text-based QA is the most common form of answering questions that source unstructured text from Wik ipedia articles. Subsequently, the automated text-based QA systems are further classified into four different types based on their applications including MRC-based QA, information retrieval, knowledge-based, and rule-based[2]. In MRC-based QA, the model aims to extract the response to a query from a specific set of contexts. The MRC QA and Open-domain QA tasks are part of text-based QA.

The MRC QA task can achieve better interaction between humans and computers. MRC QA helps the machine to retrieve information more efficiently[3]. MRC QA task works on textual data, and this technique

[1]Department of Computer Science, COMSATS University Islamabad, Wah Campus, Wah Cantt 470040, Pakistan. [2]Department of Computer Science, Namal University, Mianwali 42210, Pakistan. [3]Department of Information Technology, University of Sargodha, Sargodha 40100, Pakistan. [4]Faculty of Resilience, Rabdan Academy, Abu Dhabi, United Arab Emirates. [5]Software Engineering Department, College of Computing and Information Sciences, King Saud University, Riyadh, Saudi Arabia. ✉email: dr.hikmat.niazi@gmail.com; alimsdb@gmail.com

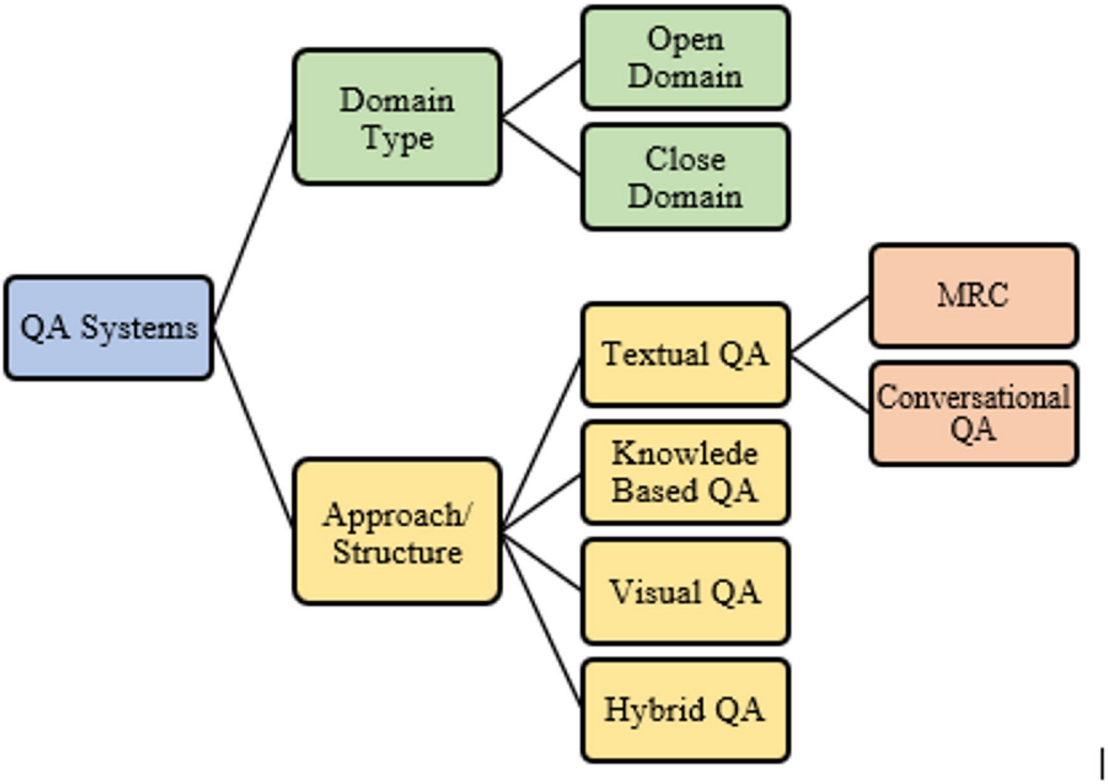


**Fig. 1**. Hierarchical classification of QA systems.

can also generate a complex answer to a question[4]. MRC-based QA is further classified into 4 different types including cloze style, multiple choice, span prediction, and free from answer. Moreover, the MRC task is further divided into four types including cloze style MRC system, multiple choice-based MRC, span prediction-based MRC, and free form answer. In all forms of MRC tasks, the machine tries to obtain or generate the answer that is most appropriate from the provided context or question. These processes are used as a mechanism to probe the machine's ability to understand query statements expressed in natural language and provide appropriate responses[5]. In real life, this technique is used to check the language proficiency of a student. Multiple Choice is also taken from the concept of testing a student's language proficiency. In this task, the choices are given with the answers, and the machine selects the most suitable choice to answer the question. Both cloze tests and multiple-choice questions are used to check a machine's understanding level of natural language, and there are some limitations in these tasks. Sometimes the questions are to be answered, not just in words or events, or maybe the answer is not available in the context. In this situation, span extraction is used to answer the question. To complete this task, the machine must retrieve a specified amount of information from the relevant context[6]. Span extraction covers the limitations of cloze tests and multiple-choice questions. The span extraction allows the machine to make a more flexible decision in answering a question than cloze tests and multiple-choice answers. Figure 2 represents the sample passage, question, and answer from the WikiQA dataset. In free answering, the machine selects the answers based on multiple reasons. Free answering is mainly used in real life and is one of the most complicated forms of QA. Free answering, in contrast to the earlier challenges, eliminates some limits and places an emphasis on utilizing free-form natural language to deliver conclusive responses.

Several NLP systems such as search engines and conversation systems use MRC QA. MRC assists a user in locating the right response to the query in a simple manner and efficiently. MRC is essential for enhancing search engine performance. There are many other online QA platforms such as Quora, Brainly, Stack Exchange, etc., where the user who needs the information posts the question and the other users answer it by giving a reply to that question. In MRC-based QA, DL models such as Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), and Convolutional Neural Network (CNN), and variations of BERT are applied to datasets such as SQuAD 1.1 and SQuAD 2.0 and promising results are reported. In recent research different studies of DL models and a variant of BERT, called DistilBERT, have been used for QA and other NLP tasks and achieved remarkable accuracy despite having a smaller architecture than BERT[7]. Similarly, in[8,9] DistilBERT is applied in QA systems, for feature extraction, classification, and performance improvement.

Although transformer-based models such as BERT have delivered the best performance in for MRC tasks, their computational costs and model complexity pose barriers to many potential real-time applications. Additionally, many models are limited by their inability to properly understand ambiguous passages where multi-faceted interpretations can result in multiple incorrect or incomplete answers. In considering both efficiency and issues of comprehension, this study examines the performance of DistilBERT, a leaner and faster version of BERT,

**Passage:** The remaining band members recorded \"Independent Women Part I\", which appeared on the soundtrack to the 2000 film, Charlie's Angels. It became their best-charting single, topping the U.S. Billboard Hot 100 chart for **eleven** consecutive weeks. In early 2001, while Destiny's Child was completing their third album, Beyonc\u00e9 landed a major role in the MTV made-for-television film, Carmen: A Hip Hopera, starring alongside American actor Mekhi Phifer. Set in Philadelphia, the film is a modern interpretation of the 19th century opera Carmen by French composer Georges Bizet. When the third album Survivor was released in May 2001, Luckett and Roberson filed a lawsuit claiming that the songs were aimed at them. The album debuted at number one on the U.S. Billboard 200, with first-week sales of 663,000 copies sold. The album spawned other number-one hits, \"Bootylicious\" and the title track, \"Survivor\", the latter of which earned the group a Grammy Award for Best R&B Performance by a Duo or Group with Vocals. After releasing their holiday album 8 Days of Christmas in October 2001, the group announced a hiatus to further pursue solo careers.

**Question:** How many weeks did their single \"Independent Women Part I\" stay on top?"

**Answer:** eleven

**Fig. 2**. Illustration of the Span Extraction task in MRC, featuring a passage, question, and answers sourced from the 'WikiQA' dataset.

where DistilBERT largely preserves BERT's attributes in relation to language understanding, and reduces the model size and inference time. Moreover, DistilBERT's ability to handle ambiguous passages has not been fully investigated on benchmark datasets like WikiQA, SQuAD, NewsQA, and Natural Questions. This study sets out to address this gap by fine-tuning a Distil-BERT-MRC model to improve the overall efficiency and ambiguity resolution of complex textual passages. This research study involves the following research contributions:

- Fine-tuned the pre-trained Distil-BERT-MRC model to address the passage ambiguity issue with machine reading comprehension.
- Introduced a resource-efficient approach for MRC that maintains competitive accuracy while requiring fewer parameters and less training time compared to dense Transformer models.
- Conducted empirical analysis with state-of-the-art deep learning as well as transformers-based models over the WikiQA MRC dataset.

The remainder part of the paper is organized as follows: Sect. 2 presents the literature review, while Sect. 3 presents the proposed methodology with details on the features, dataset, algorithms applied, and performance evaluation metrics, the results are shown in Sect. 4, and the conclusion is presented in Sect. 5.

## Related work

This section covers a literature review on MRC-based QA systems including MRC-based Multiple Choice, Conversational MRC, and Span-based MRC.

### MRC-based multiple choice QA

The task of a QA system falls within the category of NLP, which is a subset of AI. In a QA system, the user poses questions using language, inputs them into the trained model, and receives the appropriate answer generated from the text corpus. There are two different kinds of QA tasks: closed-domain and open-domain. Open-domain QA tasks are developed on huge data and Closed-domain QA tasks are quite small systems compared to Open-domain QA tasks because Closed-domain QA tasks are domain-specific and require less data to be built. Recently the Open-domain QA system has been used in many fields[7–11]. This research[12] delves into examining narrative text produced by machine learning models with a focus, on elements such as interest and believability. By utilizing the GPT Neo transformer model that has been trained and human-generated prompts a collection of stories is created. These narratives are then evaluated using both assessment and automated measures like BERT Score and BERT Next Sentence Prediction (NSP). The goal of the study is to reveal connections between automated ratings and human opinions. A study on MRC[13], solves the problem of feature spatial independence for improving saliency detection by using deep neural networks to enhance feature relations. A self-supervised method improves the salient regions, and it achieves the best performance on benchmarks with 140 FPS efficiency on a GTX1080 GPU. Another research[14], has introduced STKD, which is a knowledge distillation method that seeks to enhance student models by taking advantage of multi-instance semantic similarity and mix-up techniques. Using similar correlations between instances, STKD achieves better performance on classification benchmarks and state-of-the-art distillation methods.

Findings show disparities between assessments and automated evaluations underscoring the necessity for further enhancements, in automated metrics to accurately gauge text quality as perceived by humans. The study[15] proposed the Parallel-DistilBERT-QA model to handle the issue of large model size with knowledge distillation and parameter sharing. The model adopts a dual-stream feed-forward neural network architecture with three parallel DistilBERTs to encode questions, joint question-document, and document inputs respectively. The experimental results indicate that Parallel-DistilBERT-QA with reduced parameters outperforms BERT on SQuAD with higher F1 and Exact Match scores. To understand how different approaches have been used in integrating the attention mechanism with MRC to solve different tasks based on the SQUAD QA tasks[16]. This research explores the approaches to word embeddings, feature extraction, attention mechanisms, and the process of selecting an answer. It also outlines the limitations and issues of the model's fairness and trustworthiness; this is a critical analysis of the difficulties in deploying attention-based models for MRC. In the study[17], the authors explore ways to conduct human-inspired approaches for MRC through external knowledge, linguistic task transfer from other reading tasks, and discourse-aware semantic structures. It is shown that task-enhanced neural models that utilize background knowledge and linguistic annotation structure offer improved performance in cloze-style and narrative reading comprehension approaches. Before the concept of NLP, the researchers used a series of notable QA, which was considered the most advanced work at the time, and the most advanced models were retriever-reader QA models[9,18,19].

## Conversational MRC

MRC, which encompasses providing an answer based on a context paragraph, is one of the core challenges in NLP. MRC is a challenging task for computers, but it has the potential for various enterprise applications, leading to the expansion of the field of study in this domain[20]. To replicate the reasoning process used by human readers, The Reasoning Network (ReasoNet) is a novel neural network architecture developed by Microsoft researchers and MRC QA tasks should simulate complicated interactions between the context and the inquiry. Although traditional MRC in the single-turn setting has been well-researched, there are ongoing efforts to improve MRC performance and address challenges such as multi-turn MRC and domain adaptation[21]. Another research[22], identifies the shortcomings of relying on plain common sense entities as heuristics in logical reasoning tasks and presents a solution by introducing a hierarchical approach. This approach introduces a formalization of knowledge as structured as "subject-verb-object" facts and builds a subgraph to capture the interactions between sentences and entities. The results reveal significant enhancements in the performance on the reading comprehension logical reasoning benchmarks as well as the dialogue datasets when tested with the backbone models.

Recent advancements in learning have significantly enhanced the field of NLP in the domain of MRC. This study[23], presents a QA model that boosts its efficiency by identifying entities and determining queries. Through the integration of knowledge enrichment and answer validation techniques this model surpasses its counterparts, in tasks related to reading comprehension. The MRC has been extensively developed after the release of some standard datasets by Stanford and Microsoft, e.g., NewsQA[24], WikiQA[25] so on, and ML techniques have improved, resulting in the growth of several effective DL models. This article[26] presents the upgraded Doc KG model, designed to convert documents into knowledge graphs (KGs) by creating local KGs and linking them to target KGs such, as Wikidata. By utilizing details, the model effectively identifies entities and relationships enhancing accuracy in generation. Assessment on the WebNLG dataset showcases performance achieving an 86.64% accuracy rate in extraction and associating 61.35% of local KG with Wikidata introducing 38.65% new information for KG enrichment. The model's quality is validated through examination providing a framework for organizing unstructured data semantically and progressing in KG development.

During the MRC's early beginnings, the Match-LSTM[27] and BiDAF[28] models employ some elementary neural networks and occasionally use attention mechanisms to retrieve the right answer from the question. Later on, the researchers introduced some complex structure models, which are R-NET[29], and QANET[30], to extract the complex relationship between questions and the context. Since they succeed at obtaining semantic information from massive amounts of unlabeled data, language models with prior training, particularly BERT[31] as well as variations, have recently demonstrated excellent results when used to the extractive MRC problem. For instance, ELMo embedding was developed by Clark and Gardner[33] to improve the BiDAF algorithm and gain an important enhancement on SQuAD. This research[32], investigates the difficulties associated with creating MRC systems, for languages with resources, such as Vietnamese and Urdu which have datasets. Unlike languages with resources like English and Chinese developing MRC systems for these low-resource languages presents challenges. Therefore, Hu et al.[33] proposed a Multi-Type Multi-Span Network[34] and RE$^3$QA[33], embedding representation with BERT. In the study[35], the authors propose a few-shot MRC approach for bridge inspection utilizing pre-trained language models with a pre-tuning process and domain-specific data augmentation to minimize manual annotation. The approach outperforms conventional fine-tuning and baseline few-shot methods, demonstrating potential applicability for intelligent systems for bridge maintenance.

## Span-based MRC

Span-based MRC QA is a method that focuses on getting a response straight from the paragraph or reading context. In span-based MRC QA, the objective is to locate the passage's start and finish points that correlate to the response to a certain question. To better understand how semantic matching paradigms can be evaluated in MRC. The researchers have used a two-stage framework with pre-trained language models[36]. The experiments done on different datasets show that semantic matching enhances performance in terms of effectiveness and efficiency, particularly in noisy or adversarial environments and for certain types of questions including who/what/where. However, the performance is decreased in the case of the 'why' type of questions which shows that it has its limitations and is context-dependent. Some researchers have introduced the Transformer model

in the context of QA systems with a focus on the encoder, the decoder, and the encoder-decoder framework[37]. The study[38], introduced SpanBERT, a BERT model that concentrates on span-level data during pre-training. SpanBERT enhances the way data is represented by training the model to anticipate text spans rather than specific tokens and span comprehension, resulting in enhanced efficiency in span-based MRC tasks. UniLMv2[39] is a pre-training unified language model technique that enhances the first UniLM model to deal with span-based MRC tasks effectively. The research proposes a pseudo-masked language model that masks consecutive text spans during initial training.

Similarly, Another researcher[40], has focused on gathering data from discourse, a process strongly connected to span-based MRC. GraphIE models the relationships between conversation statements using graph neural networks and retrieves important information using span-based predictions. The study shows that GraphIE outperforms earlier algorithms for extracting information from conversation datasets, demonstrating the usefulness of span-based approaches in such a situation. Another researcher has proposed PAL-BERT[41], a compact model that is based on the ALBERT model and is targeted at improving the performance of QA systems while at the same time reducing the size of the model and the cost of training it. To this end, PAL-BERT employs first-order network pruning and the Mish activation function to maintain performance while reducing the model size and training time. Comparing the results with other traditional models like TextCNN and BiLSTM, PAL-BERT is seen to reduce the training time and enhance the efficiency and performance of the NLP tasks.

To solve the problems of Textbook QA, specifically the weak reasoning capability and the difficulty of understanding contextual cues in long contexts. In this research article, the researchers have used the Retrieval Llama-2 LLM model[42]. This approach enhances the reasoning capabilities and deals with out-of-context scenarios. When fine-tuning the model in the supervised manner along with the RAG approach the accuracy on the validation set was enhanced by 4.12% and on the test set for non-diagram multiple choice questions it was enhanced by 9.84%. The researchers have shown interest in some pre-trained context-based Language Models (LMs), e.g., Embedding for Language Model (ELMO)[43], Bidirectional Encoder Representations from Transformers (BERT)[44], or several GPTs (Generative Pre-Trained models)[45] and have used these models for MRC according to its outstanding results on many criteria for Natural Language Understanding (NLU).

A study[46], explores the recent explosion of interest in DL, particularly Large Language Models (LLMs), which have revolutionized machine learning. It discusses DL's dominance across various domains like NLP and image analysis, aiming to elucidate DL fundamentals, applications in NLP, and advancements. Despite their remarkable performance, LLMs struggle with contextual nuances and social norms. Overall, the paper offers a concise analysis of DL advancements and implications, enhancing readers' understanding of the subject. In this research article, the researchers have implemented the Automatic Short Answer Grading system using DL-based sentence embedding approaches[47]. The multilayer perceptron regression layer is used in the proposed research work and the model is trained on a reading comprehension dataset of aphorisms to assess the performance of BERT and Skip-Thought models for grading short answer questions. When tested on answers provided by 199 undergraduate students in Spanish, the results indicated that BERT produced better results as compared to other models when it comes to Pearson correlation coefficient and root mean squared error for ASAG.

### Limitations of existing work

The literature review discusses progress and issues in using MRC-based QA systems as multiple choice, conversational, or span methods. For multiple-choice QA, it highlights the distinction between open-domain and closed-domain systems. Even though BERT-NSP and STKD showed improved accuracy in MRC tasks, they are mainly concerned with fact retrieving. These models lack the ability to handle fine semantic ambiguity resolution which reduces their usefulness in complex cases where alternative meanings are quite possible. Conversational MRC approaches such as ReasoNet and FollowNet can handle conversations well, they depend more on how people speak than on the real meaning of the words. They have trouble dealing with cases where the speaker or listener must tell apart close-by segments or understand references left unsaid earlier. Even though SpanBERT and UniLMv2 have greatly enhanced answer span prediction, they typically depend on simple overlaps in words. These approaches are limited in their ability to resolve ambiguous or multiple spans in contexts where deep reasoning or multiple interpretations are required. In each category, a common gap is the lack of a targeted approach for resolving ambiguities between multiple correct or near-correct answers. Existing models are optimized for accuracy on standard datasets but lack robustness when handling semantic ambiguity or query interpretation.

## Proposed research methodology

This section explains the proposed strategy, which includes the proposed framework, along with the detailed architecture of deep learning and transformers-based models, features engineering, description of the dataset, and utilized performance evaluation measures.

### Deep learning models

The proposed framework for DL methods is presented in Fig. 3. In this research study, we employed three DL methods for MRC tasks including RNN, Gated Recurrent Unit, and LSTM. The DL models are based on neural network architectures that effectively process the sequential input data while gradually establishing connections for textual data. GRU is the simplified version of the LSTM by combining the input and forgets gates to create an update gate that reduces costs without sacrificing the ability to recognize long-range dependencies. On the other hand, LSTM maintains memory cell structures and gating mechanisms to manage information flow allowing it to capture intricate relationships in sequential data. The hidden states of these models capture information and correlations between content and queries during question answering tasks. An attention mechanism allows the model to concentrate on a certain portion of the data input for comprehension and accurate responses to queries.

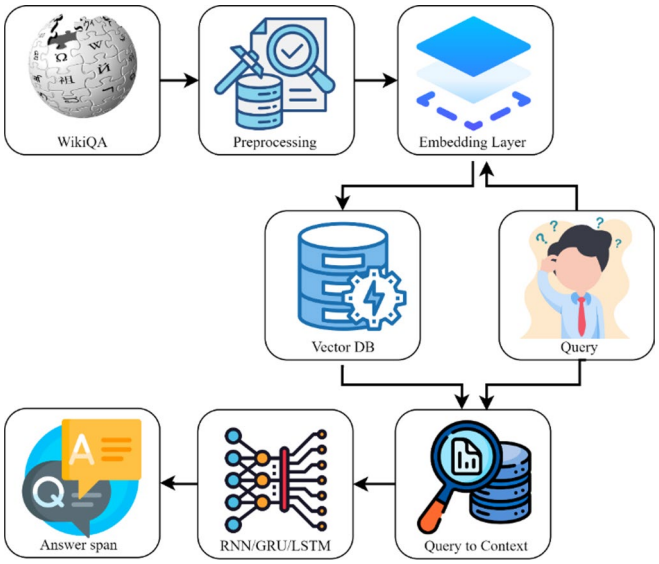


**Fig. 3**. The overview of the proposed framework of deep learning approach for MRC QA.

*Data preprocessing*

Data preprocessing is a crucial step for deep learning models. This initial step involves gathering the WikiQA dataset, which consists of pairs of questions and answers linked to passages. To prepare the data, for learning models both questions and passages are converted into tokens during the stage. Additionally, cleaning procedures are used to raise the dataset's quality by eliminating information, noise, and special characters. These procedures ensure that the MRC model receives input data laying a foundation for subsequent training and evaluation phases.

*Deep embeddings*

After the preprocessing step, the text is placed into embedded layers to transform the natural language text into a vector representation. Character embedding is used to maintain the details of the text at a scale, by converting each letter into a set length array. Subsequently, the word embedding techniques transform words into representations that hold onto their meanings. The embedding layer plays a role in preparing the text for the processing stages. Once the text input is incorporated, encoding techniques are applied to extract information for comprehension from both context and queries. The embedding layers support in creation of representations for passages and questions. Subsequently, self-attention mechanisms analyze context-related data in the passages. By utilizing this encoding technique, the model can extract details from queries aiding in comprehension and grasping relationships between terms and sentences, within the content.

*Model architecture for DL methods*

The use of attention is crucial in connecting the context with the question asked during understanding because it helps in concentrating on parts of the text. The attention from context to question assesses the significance of every word with respect to the question offering query examples that emphasize relevant information for answering the question at hand. On the other hand, by organizing the query terms within the context, attention from question to context enhances the representation of questions to better capture variations present in the context. The successful incorporation of query and context-specific details which enhances model prediction accuracy relies on these attention mechanisms. The system learns how to find the answer sections by adding layers and training on improved representations derived from attention processes. To ensure the reliability and efficiency of the MRC system variations, in hyperparameters, regularization techniques such as dropout and evaluation on a test set are utilized. By employing training methods, the system becomes adept at QA tasks by predicting answer sections based on the context and question provided. Metrics like accuracy are used to assess how well the trained model performs on a test set after training. This iterative approach guarantees that the system functions effectively in real-world scenarios, by providing answers and gathering pertinent information from passages.

To evaluate input data and produce pertinent outputs, MRC QA uses DL with several layers, as illustrated in Fig. 4. A separate layer that is generally employed in these types of structures is described as follows: In the architecture, the character embedding layer generates the fixed-length vector embeddings by utilizing the uncommon keywords in the input context by transforming individual word characters. This process helps in capturing finer details of meaning and complexities in the material to develop an understanding of the material. Subsequently, the attention layer assigns the attention scores for both the question and passage to identify the significant part of the text. As a result, the model can focus on the elements with higher attention scores. The modeling layer uses convolutional networks to generate comprehensive representations that support effective question answering by integrating contextual understanding along with attention mechanisms. Finally, the

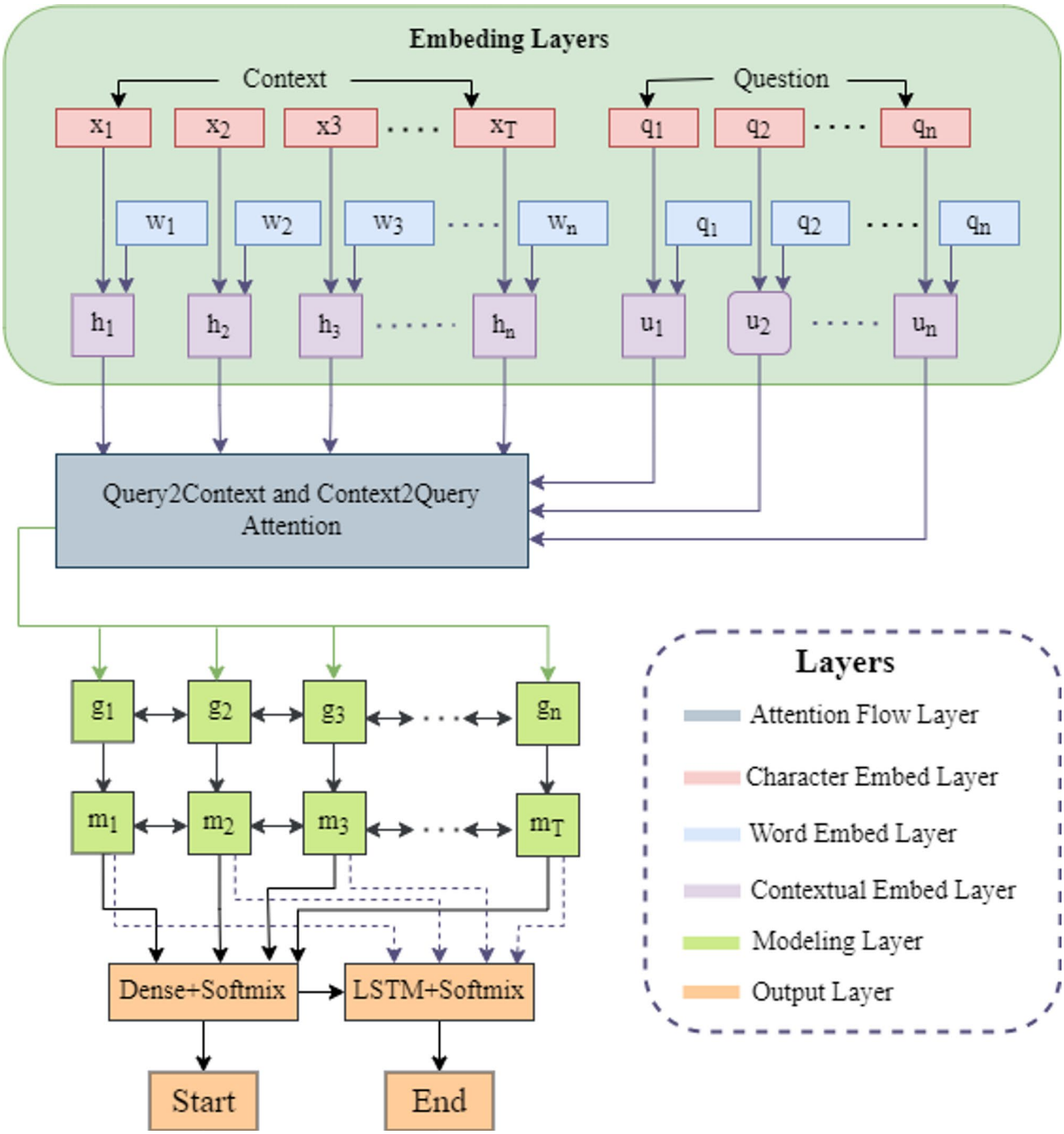


**Fig. 4**. The model architecture of deep learning-based methods.

output layer concludes with the pipeline by retrieving responses through representations. The output layer accomplishes this by providing answers to the given inquiries using approaches like span prediction and labeling.

### Transformers-based model

In this study, we utilize DistilBERT short from Distillation BERT transformers-based language model. The DistilBERT model is primarily designed for linguistic modeling, text classification, or the generation of sentences. Some of the key components of DistilBERT architecture include positional embeddings, a self-attention mechanism, and the ability to understand language at a deep context level.

This is because positional embeddings in DistilBERT preserve the order of the words in a sequence, and this is very important in determining the intended meaning of the words. The self-attention mechanism which forms the basis of transformers is the key component of DistilBERT as it allows the model to assign varying levels of importance to different parts of a sentence to understand complex relationships that may resolve the ambiguities

in each text. This is especially helpful for tasks like MRC since for a given question and a potentially confusing text, it is crucial to understand how the question is related to the text.

The attention mechanism in DistilBERT allows the model to directly extract key features from a passage, which becomes crucial in resolving any ambiguity in long-form passages. Input embeddings are passed through the layered structure of transformer encoders, where each relevant information or attention score is highly emphasized during the processing of the next word or token. The attention score $AS$ is computed using the following Eq. (1).

$$AS\ (Q_i, K_j) = \frac{Q_i\ .\ K_J^T}{\sqrt{d_k}} \tag{1}$$

Where $Q_i\ .\ K_J^T$ is the dot product of query $Q_i$ and key $K_J^T$ use to measure the compatibility and $\sqrt{d_k}$ denotes the vector dimension to avoid overly large values.

This enables DistilBERT to disambiguate words or phrases by considering the words that are in the passage around them because they help establish what the passage means, even in cases where the only way to size up the passage is ambiguous or has multiple interpretations. For instance, the word "bank" might be ambiguous in the sentence "The bank was closed after the flood".

However, the attention score of the word "flood" enables the model to infer the context-aware meaning of the word. Resultantly, based on this correct inference model can resolve ambiguities of complex passages where sentences have multiple interpretations.

To generate the answers to the question, take two phrases as input and assess how relevant they are. The DistilBERT model can combine two statements and send them to the input as a single sentence. This study combines the query and answer phrases into a single sentence, as (2) illustrates.

$$(Question,\ Answer)\underrightarrow{Convert}[CLS]\,Question\ [SEP]\,Answer \tag{2}$$

There are several unique tokens in the DistilBERT model, and each has a distinct function. The [CLS] token's function is to obtain a vector for categorization. Sentences are separated by the [SEP] token. This token's output vector has no specific value and is just utilized to divide sentences. When this phrase is fed into the DistilBERT model, each word produces an output vector$\in R^{768}$.

However, this analysis only uses the [CLS], QW, and EA resulting in vectors rather than all outputs. This part's performance may be represented as (3):

$$([CLS]\ ,\ QW,\ EA) = DB\,([CLS]\,Question\ [SEP]\,Answer) \tag{3}$$

The next part contains the Measurement Layer for similarity, which assesses the query's and answer's relevancy using the output from the phrase modeling section. As illustrated in (4), Concatenating the resulted vectors of [CLS], QW, and EA results in $I_v \in R^{2304}$.

$$Iv\ =\ concat\,([CLS]\ ,\ QW,\ EA) \tag{4}$$

A fully connected neural network uses this vector as its input vector. The $H_v \in R^{2048}$ is the network's hidden layer, and it can be calculated from (4). Here, $W_{h1} \in R^{2304\times 2048}$ and $b_{h1} \in R^{2048}$.

$$Hv\ =\ tanh\ (Wh1.Iv\ +\ bh1) \tag{5}$$

A vector with size vector$\in R^2$ also represents the network's output vector. The relevance is indicated by one element in this vector, and the irrelevance by the other. From (6), this vector is calculated. In this formula,

$$f(q,a)\ =\ softmax\ (Wh2.Hv\ +\ bh2) \tag{6}$$

Where $b^{h2} \in R^2$ and $W_{h2} \in R^{2\times 2048}$ and values can also be normalized using the softmax function. As illustrated by (7), this approach uses the cross-entropy technique to adjust the algorithm's parameters that can be trained. Where $a_i$ also denotes the $f(q,p)$ result and $y_i$ is the possible answer identification. To maximize the suggested approach's trainable parameters, we employ the Adam algorithm[48].

$$\mathcal{L}\ =-\sum\nolimits_{i=1}^{N}[y_i log a_i\ +\ (1-y_i)]\,log(1-a_i)] + \lambda\ ?W?^2 \tag{7}$$

The identified answer span is converted into natural language form, generating the final answer as illustrated Fig. 5.

## Experimental setup

This section presents the experimental setup of the conducted performance assessments along with comprehensive coverage of the datasets utilized in the study and focuses on precise evaluation procedures that guarantee accurate and important results.

*Dataset*

In this research study, the Microsoft-created WikiQA dataset is examined. There are statements and open-ended questions matched with relevant answers in it. Microsoft created this dataset for research purposes. This dataset

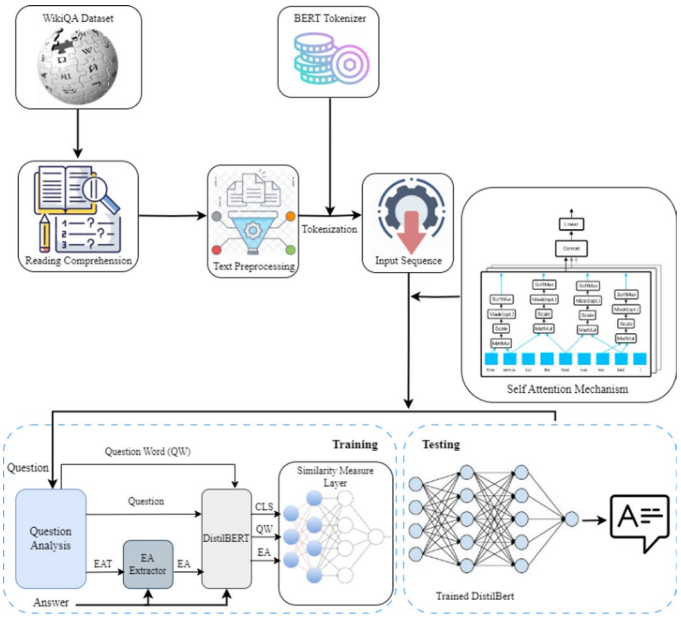


**Fig. 5**. Overview of the proposed framework utilizing a transformer-based DL approach for MRC in QA.

contains a pair of sentences with questions taken from the open domain Wikipedia articles. This data consists of 29,258 sentences with 3047 questions. Among these sentences, 1,473 sentences represent the answers. WikiQA is considered as a challenging and complex dataset for MRC tasks. However, the number of question-answers pairs is relatively small as compared with other MRC corpora, making it well suitable for DistilBERT evaluation.

*Performance evaluation measures*

In this part, performance evaluation metrics are discussed. This will be utilized in this research to evaluate the model's effectiveness. Accuracy evaluates the performance based on how well the model has extracted the right answer from the data. Equation 8 will be used to calculate accuracy.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

The symbols TP, TN, FP, and FN stand for True Positive, True Negative, False Positive, and False Negative respectively. F1-measures take both precision and recall evaluating the performance of the model. Equation 9 will be used to calculate the F-Measure.

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{9}$$

Where P stands for Precision, and R for Recall, respectively. Exact Match (EM) is a metric used to assess the integrity of the response chosen by the reader module in QA systems. EM measures how well the model's suggested response matches the true response. As the name suggests, EM calculates the percentage of the text corpus where the model's projected response exactly matches the true answer on a character-by-character basis. Equation 10 will be used to calculate the Exact Match.

$$EM = \begin{cases} 1, & if\ C_M = a_t \in A_t \\ 0, & otherwise \end{cases} \tag{10}$$

The letters *M*, *N*, and *EM* stand for the overall number of correct answers, total answers, and exact match, respectively.

*System specifications*

This research used a high-performance computing system which includes an NVIDIA RTX 3090 GPU with 24GB of VRAM to ensure that deep learning models and large datasets are processed efficiently. The CPU was an 8 core Intel Core i7, which gives strong computational support for parallel processing. The system also had 32 GB of DDR4 RAM and a 512 GB NVMe SSD that guaranteed smooth multitasking and fast data storage and retrieval. The experiments were carried out using the deep learning framework of PyTorch 2.0, after upgrading to CUDA 11.7 to take advantage of GPU acceleration for training and inference.

*Hyperparameter details*
To ensure optimal model training, carefully tuned hyperparameters were employed in the research. A learning rate of $5 \times 10^{-5}$ and a batch size of 16 were used to guarantee balanced convergence and memory efficiency. The Adam-W optimizer with a weight decay of 0.01 was used to train models for 15 epochs to avoid overfitting. To improve the training stability, a dropout rate of 0.1 and gradient clipping at 1 was used. A maximum sequence length of 128 tokens was chosen to be enough context, and 500 warmup steps was used to stabilize learning at the beginning of training. These hyperparameters were selected based on previous studies and initial tuning adjustments due to computing constraints for a full grid search. These settings supported solid and durable model performance.

## Results and discussion

The experiment's outcomes are thoroughly exploring, showcasing analysis through tables and figures. Visual presentation aids in comparing performance metrics and analyzing the outcomes. The section emphasizes the model's effectiveness and performance along with the techniques recommended in this research.

### Results of deep learning-based models

In this study the model's effectiveness is assessed by analyzing the training and validation loss curves to see how well it adapts to the training data and generalizes to validation data. A lower loss value indicates performance suggesting that the models' predictions are more precise. The LSTM model's performance can be gauged by reviewing the training and validation accuracy curves for both datasets as depicted in Fig. 6a. The trends of loss and accuracy can be used to analyze the performance of an LSTM model during training. The impressive validation accuracy of 89.60%. This is the performance of the LSTM model, on data that was not part of the training set and is 60% attained. This shows that the model can make predictions on data as evidenced by its validation accuracy score of 89.60%. As we can see, the training accuracy is usually higher than the validation accuracy. Therefore, it is critical to monitor performance throughout the training and validation phases to determine the model's accuracy.

The loss during training steadily dropped from 1.35% to 0.21% while the loss of validation decreased from 1.50% to 0.40%, which indicates that the model's error rates decreased over time as shown in Fig. 6b. The performance of the GRU model, on the dataset demonstrates enhancements in accuracy and loss metrics suggesting learning and adaptability over time. The training accuracy shows an increase from 49.52% to 90.25% as depicted in Fig. 7a whereas the validation accuracy also grows from 45.28, to 89.14% showcasing the model's ability to generalize effectively to the dataset. The training loss of GRU went down from 1.40%, to 0.30% as shown in Fig. 7b.

The validation loss dropped from 1.55% to 0.45%, which indicates an improvement in reducing errors. The consistent correlation between training and validation metrics without deviation implies that the GRU effectively grasps patterns while steering clear of overfitting issues. This leads to reliable performance, on tasks like QA. The RNN model's results show steady improvement in accuracy and reduction in loss, reflecting effective training and generalization as depicted in Fig. 8 (a). The training accuracy rises from 45.28% to 89.01%, while the validation accuracy increases from 40.25% to 87.39%, demonstrating the model's ability to generalize to unseen data. The training loss steadily decreases from 1.45 to 0.35%. The validation loss decreases from 1.60% to 0.49%, indicating a reduction in errors as shown in Fig. 8 (b). While RNN learns and generalizes effectively its performance stabilizes more slowly compared to more advanced models such as LSTMs and GRUs due to the complexities of managing long-term dependencies in sequential data.

### Performance of transformers-based models

We have trained DistilBERT on WikiQA dataset and its performance was evaluated based on five performance evaluation metrics: accuracy, recall, precision, F1 score, and exact match. This study also compares the results of DistilBERT with state-of-the-art transformers-based models such as RoBERTa and XLNet models. RoBERTa

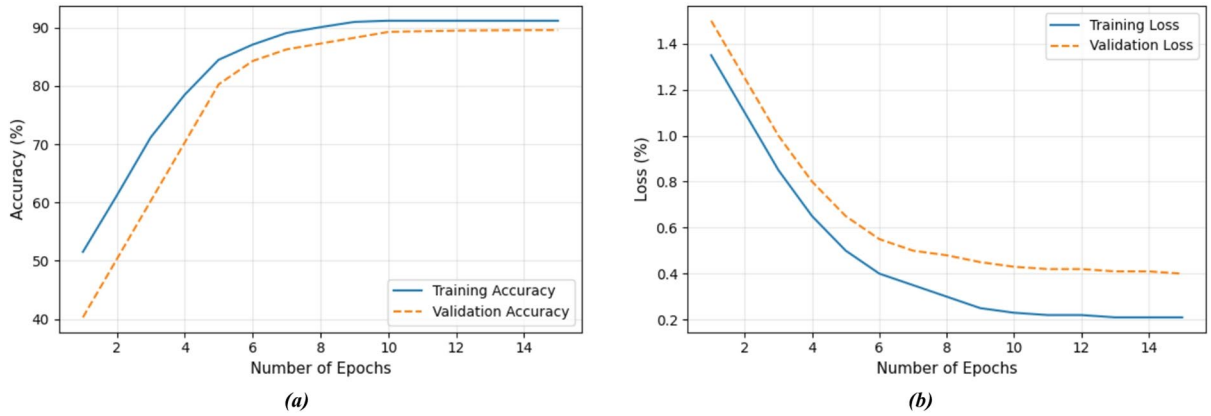


**Fig. 6**. The training and validation (**a**) accuracy and (**b**) loss of deep learning-based LSTM model.

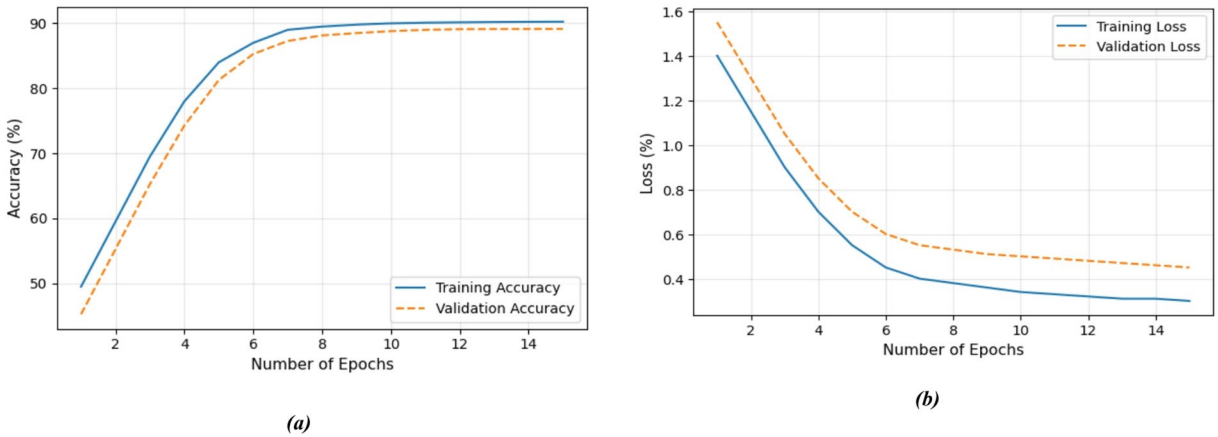


**Fig. 7**. The training and validation (**a**) accuracy and (**b**) loss of deep learning-based GRU model.

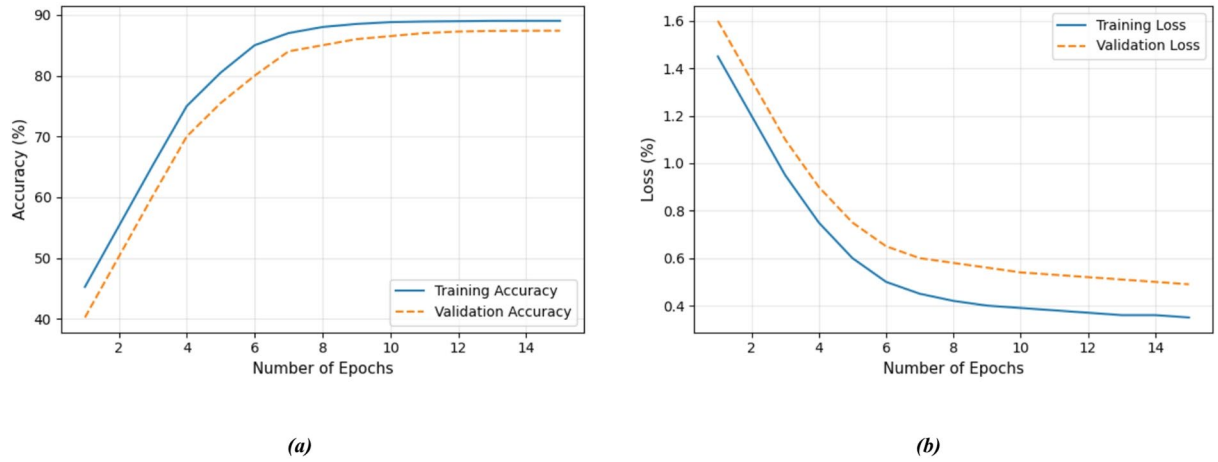


**Fig. 8**. The training and validation (**a**) accuracy and (**b**) loss of deep learning-based RNN model.

is another pre-training model which is developed by Facebook AI, which is an improved version of BERT. It builds on BERT by using more data, deleting the next-sentence prediction, and using the dynamic masking approach for its training. Similarly, XLNet is a transformer-based model developed by Google and Carnegie Mellon University. It employs a permutation-based training strategy to model contexts of both directions and incorporates several Transformer-XL improvements for dealing with long sequences.The performance of the DistilBERT model in terms of training and validation results shows a constant enhancement in accuracy as well as loss metrics. The training accuracy of the DistilBERT model starts at 75.52% and ends at 94.30% thus indicating that the model is capturing the training data properly as depicted in Fig. 9a. Also, the validation accuracy starts at 72.28% and goes up to 92.70% which proves that the model is well trained and can be easily applied to data that it has not seen before. The training loss on the dataset reduces from 1.20% to 0.26% which shows that the model is doing a better job at predicting the training data at each iteration.

The validation loss on the other hand starts at 1.35% and decreases to 0.36%, which means that the model is also making better predictions on data that it has not seen before as shown in Fig. 9b. The accuracy and loss curves show that there is a proper balance between the learning and generalization capabilities of the proposed model. The Roberta model's performance shows a steady enhancement in both training and validation and a well-established ability to learn and generalize. The metrics training indicating accuracy efficiency increases from a minimum of 72.50 to a maximum of 93.95%, which shows that the model is learning more about the training data. Likewise, the validation accuracy also boosts from 70.00% to 92.50% which shows that the model can perform well on the data as depicted in Fig. 10a. The training loss of RoBERTa decreases from 1.25% to 0.29%, and the validation loss drops from 1.40% to 0.37%, showing a continuous reduction in prediction errors as shown in Fig. 10b. The alignment between the training and validation trends, with no significant divergence, suggests that the model maintains a good balance between learning and generalization, avoiding overfitting and achieving robust performance. The XLNet model results demonstrate progressive learning and generalization, as evidenced by improving accuracy and decreasing loss values over training epochs. The training course begins at 70.00% and rises gradually to 93.50%, which shows that the model is becoming more and more capable of

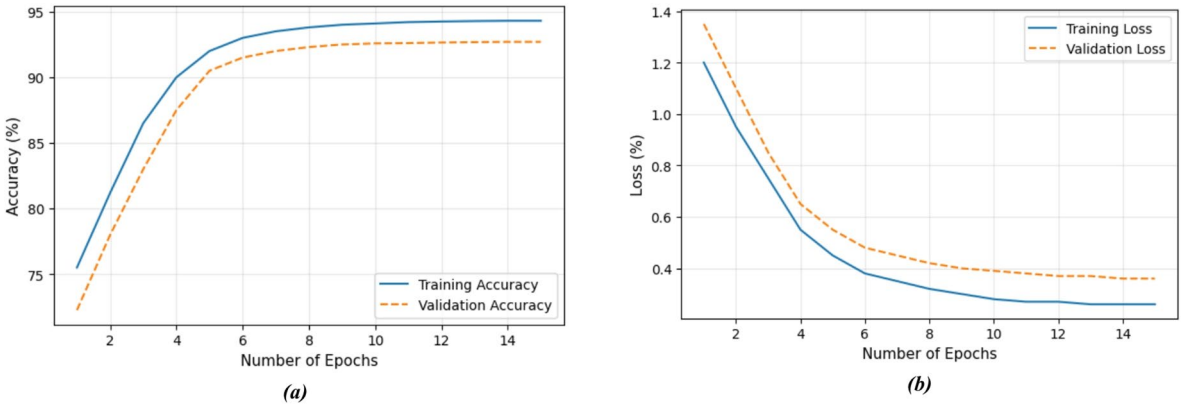


**Fig. 9**. The training and validation (**a**) accuracy and (**b**) loss of transformers-based Distil-BERT model.

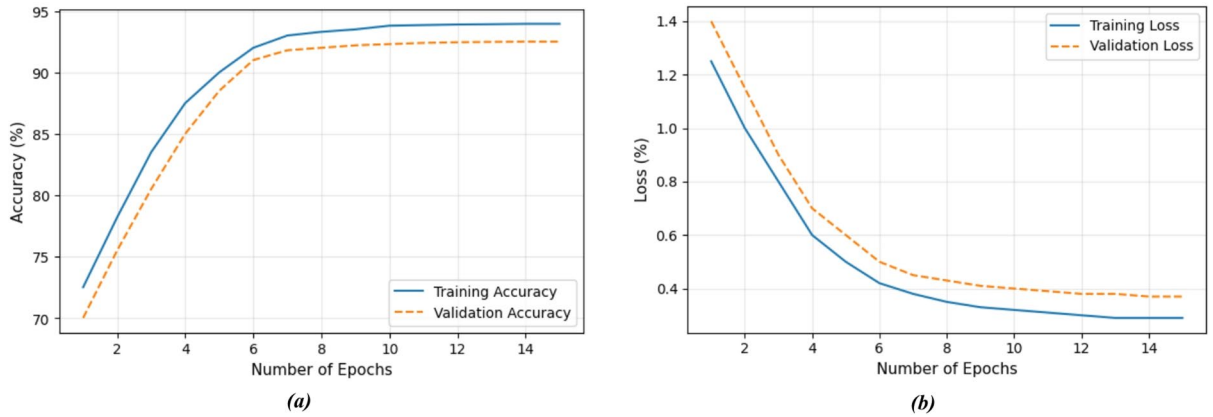


**Fig. 10**. The training and validation (**a**) accuracy and (**b**) loss of transformers-based RoBERTa model.

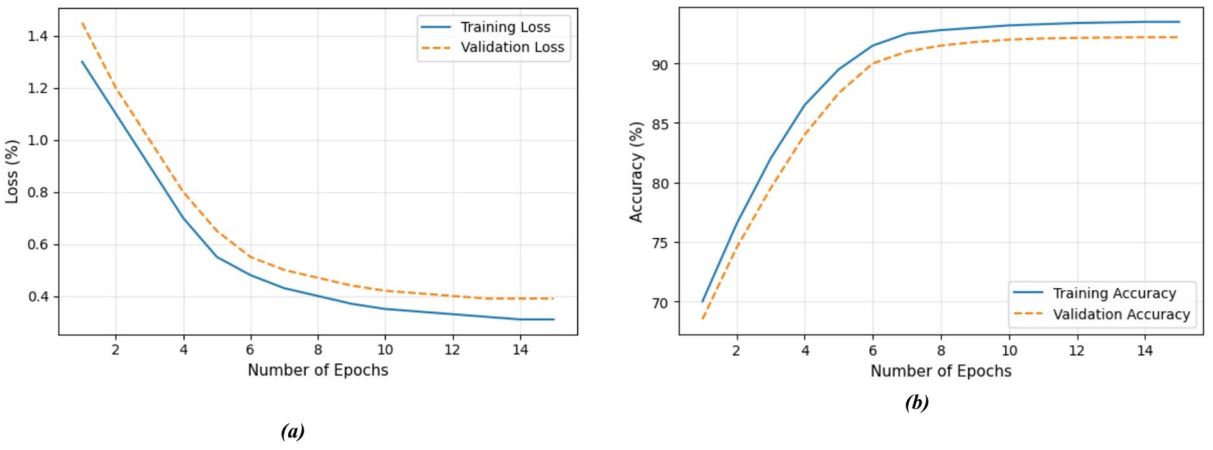


**Fig. 11**. The training and validation (**a**) accuracy and (**b**) loss of transformers-based XLNet model.

learning the training data as shown in Fig. 11a. Likewise, the validation accuracy also boosts from 68.50% to 92.20% which indicates that the model can perform well on the data.

The training loss on the first epoch decreases from 1.30% to 0.31% and the validation loss on the first epoch decreases from 1.45% to 0.39%, which shows the model effectively minimizing errors during both the training and validation process as depicted in Fig. 11b. The convergence of metrics between the training and validation sets without significant divergence suggests that the model avoids overfitting and achieves a robust balance

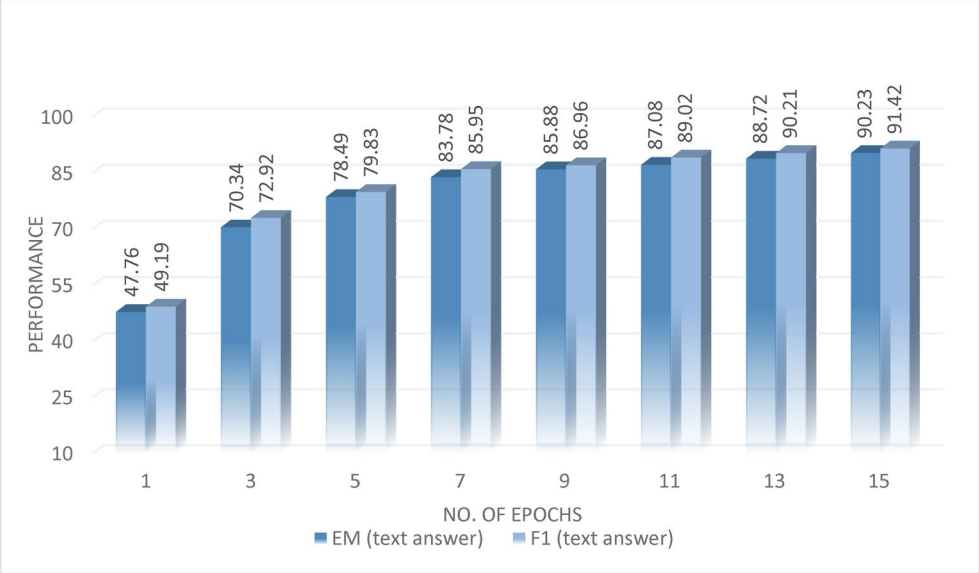


**Fig. 12**. Performance analysis of DistilBERT with respect to Epoch progression during training.

| Model | Dataset | Epochs # | Train Acc % | Valid Acc % | Train Loss % | Valid Loss % |
|---|---|---|---|---|---|---|
| Distil-BERT-MRC | SQuAD 2.0 | 1 | 76.10 | 68.85 | 1.18 | 1.41 |
| | | 5 | 91.30 | 89.20 | 0.49 | 0.60 |
| | | 10 | 93.50 | 91.80 | 0.33 | 0.44 |
| | | 15 | 94.10 | 93.25 | 0.27 | 0.38 |
| | NewsQA | 1 | 78.00 | 74.45 | 1.15 | 1.28 |
| | | 5 | 91.80 | 90.05 | 0.46 | 0.56 |
| | | 10 | 93.80 | 92.80 | 0.29 | 0.39 |
| | | 15 | 94.40 | 94.52 | 0.25 | 0.36 |
| | Natural Questions | 1 | 77.20 | 72.65 | 1.16 | 1.33 |
| | | 5 | 91.50 | 89.90 | 0.47 | 0.57 |
| | | 10 | 93.90 | 92.20 | 0.31 | 0.41 |
| | | 15 | 94.20 | 92.58 | 0.28 | 0.37 |

**Table 1**. Performance of the DISTIL-BERT-MRC model on squad 2.0, NewsQA, and natural questions datasets across different Epochs.

between bias and variance, delivering reliable performance. Figure 12 shows the performance of DistiBERT in terms of exact match and F1 score over the WikiQA dataset. Based on our experiments, we observed that the proposed model was superior to the rest in F1 score as well as the exact match. According to the results, DistilBERT achieved an F1 score of about 91. 42% and match scores of 90.23%. These findings indicate how well the model comprehends and extracts knowledge from the text paragraphs.

The resemblance is apparent in the similarity coefficient that equals 90%. 23% show that the response generated by DistilBERT is adequate with the response of 90. 23% of the questions. This is a clear manifestation of the model's capacity to understand the content to give an answer in this case.

## Results of transformers-based models on benchmark datasets

To assess the generalizability and robustness of the proposed architecture, we incorporated the DISTIL-BERT model on three diverse MRC datasets including SQuAD 2.0, NewsQA, and Natural Questions. These datasets differ in their complexity, type of context, and how questions and answers are formatted, ensuring that they serve as a thorough testbed for model validation. As shown in Table 1, we trained the model for different epochs capturing training accuracy, validation accuracy, and various loss metrics. For SQuAD 2.0, the model's validation accuracy increased consistently with epoch progression, achieving 68.85% in epoch 1 and reaching 93.25% by epoch 15. Simultaneously, training and validation loss metrics improved from 1.18 to 1.41 to 0.27 and 0.38, indicating the model was not only learning but also generalizing better across epochs. This reinforces the idea that greater training on the extractive question-answering frameworks enables the model to understand more sophisticated contextual dependencies. Subsequently, for the NewsQA dataset, where the model's validation accuracy started from 74.45% and peaked at 94.52% by epoch 15.

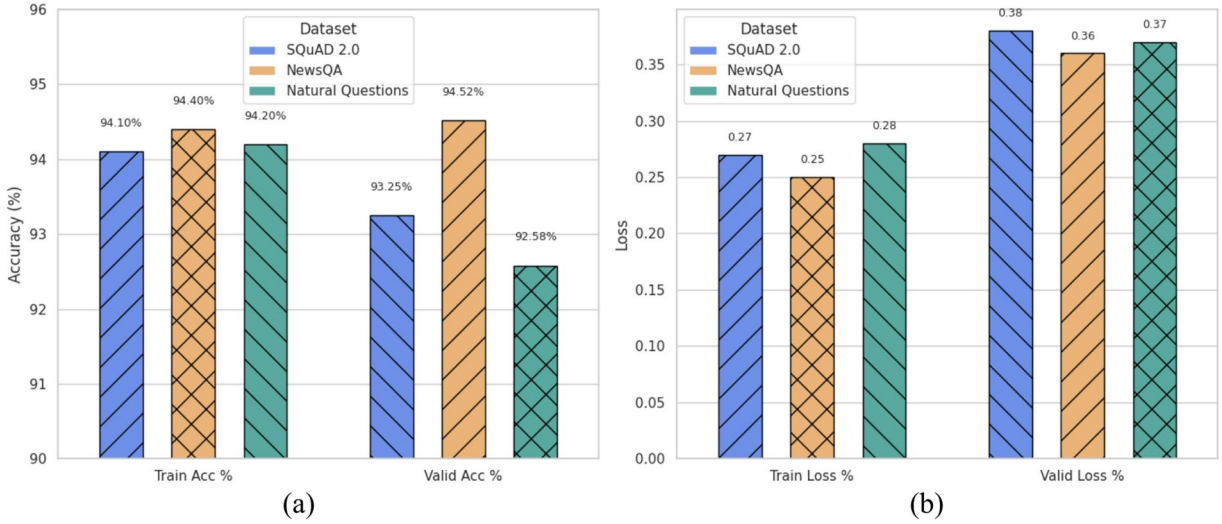


**Fig. 13**. Training and validation (**a**) accuracy and (**b**) loss of DISTIL-BERT on three datasets with 15 epoch.

| Model | Epochs # | Train Acc | Valid Acc | Train Loss | Valid Loss | Time (sec) |
|---|---|---|---|---|---|---|
| LSTM | 1 | 51.52 | 40.28 | 1.35 | 1.50 | |
| | 5 | 84.50 | 80.28 | 0.50 | 0.65 | 7223 |
| | 10 | 91.20 | 89.28 | 0.23 | 0.43 | |
| | 15 | 91.20 | 89.60 | 0.21 | 0.40 | |
| GRU | 1 | 49.52 | 45.28 | 1.40 | 1.55 | |
| | 5 | 84.00 | 81.28 | 0.55 | 0.70 | |
| | 10 | 90.00 | 88.80 | 0.34 | 0.50 | 2281 |
| | 15 | 90.25 | 89.14 | 0.30 | 0.45 | |
| RNN | 1 | 45.28 | 40.25 | 1.45 | 1.60 | |
| | 5 | 80.50 | 75.50 | 0.60 | 0.75 | 4225 |
| | 10 | 88.90 | 86.50 | 0.39 | 0.54 | |
| | 15 | 89.01 | 87.39 | 0.35 | 0.49 | |

**Table 2**. Performance comparison of deep learning models (LSTM, GRU, and RNN) over different Epochs.

It is important to highlight that this dataset reached the highest final validation accuracy out of the three, while maintaining a minimum validation loss of 0.36, which indicates that the model learned effectively, i.e., learned efficiently despite the dataset's longer passages and more complex question types. The Natural Questions dataset showed consistent progress as well, starting with 72.65% and reaching 92.58%. While the NewsQA dataset outperformed Natural Questions in terms of final validation accuracy, the model exhibited a well-behaved learning curve and consistent drop in loss, demonstrating robustness to noisy real-world inputs. Figure 13 depicts the training and validation accuracy and loss of the proposed architecture for all three datasets at the final epoch. The DISTIL-BERT model demonstrated strong scalability and effectiveness in different MRC datasets, supporting its use for MRC tasks.

## Comparative analysis

According to experimental results of the deep learning model over WikiQA dataset, the LSTM model performed better as compared with RNN and GRU. The LSTM model showed a validation accuracy of 89.60% and 91.20% accuracy during training, indicating its capability to respond to 89.60% of the questions based on the given texts, in the dataset. The complete results with each epoch and time taken during training and validation are presented in Table 2. The GRU model reached an 89.14% accuracy rate during validation and a 90.25% accuracy rate during training, outperforming the RNN model but not matching the accuracy of the LSTM model. This suggests that the GRU model correctly answered 89.14% of the questions in the dataset. Subsequently, the RNN model achieved a validation accuracy of 87.39% and 89.01% accuracy during training, indicating that it accurately answered 87.39% of the questions, in the dataset. Moreover, the performance of the.

DistilBERT model with other transformers-based models including RoBERTa and XLNet over WikiQA dataset. Table 3 shows a comparison of DistilBERT with RoBERTa and XLNet models during training, in terms of accuracy and loss rate.DistilBERT stood out as the model reaching a training accuracy of 94. 30% with 0.26% training loss. Subsequently, DistilBERT achieved a validation accuracy of 92. 70% with 0.36% loss over 15

| Model | Epochs # | Train Acc | Valid Acc | Train Loss | Valid Loss |
|---|---|---|---|---|---|
| Distil-BERT-MRC | 1 | 75.52 | 72.28 | 1.20 | 1.35 |
| | 5 | 92.00 | 90.50 | 0.45 | 0.55 |
| | 10 | 94.10 | 92.58 | 0.30 | 0.39 |
| | 15 | 94.30 | 92.70 | 0.26 | 0.36 |
| RoBERTa | 1 | 72.50 | 70.00 | 1.25 | 1.40 |
| | 5 | 90.00 | 88.50 | 0.50 | 0.50 |
| | 10 | 93.80 | 92.30 | 0.32 | 0.40 |
| | 15 | 93.95 | 92.50 | 0.29 | 0.37 |
| XLNet | 1 | 70.00 | 68.50 | 1.30 | 1.45 |
| | 5 | 89.50 | 87.50 | 0.55 | 0.65 |
| | 10 | 93.20 | 92.00 | 0.35 | 0.42 |
| | 15 | 93.50 | 92.20 | 0.31 | 0.39 |

**Table 3**. Performance comparison of transformer-based models across Epoch.

| Year | Ref | Model | Dataset | Exact match | F1 score | Key observations |
|---|---|---|---|---|---|---|
| 2022 | [49] | DistilBERT | Covid data | 80.6 | 87.3 | Achieved high accuracy on domain-specific data with efficiency. |
| 2023 | [50] | BERT | SQuAD 2.0 | 72.34 | 60.45 | Larger model; resource-intensive and slower inference. |
| 2023 | [51] | RoBERTa | MultiRC | 68.78 | 72.12 | Demonstrates strong contextual understanding in multi-sentence reasoning tasks. |
| 2024 | [52] | XLNet | NewsQA | 65.14 | 70.89 | Excels in paragraph-level QA but require significant computational resources. |
| 2024 | [53,54] | GPT-3 | TriviaQA | 85.23 | 88.67 | Outstanding performance demands high computational power. |
| 2024 | [15] | Parallel-DistilBERT | SQuAD 1.1 | 86.20 | 88.36 | Enhanced exact match and F1-score with 31% reduced parameters. |
| 2025 | [35] | Few-shot-Neural-MRC | Chinese comprehension corpus | 86.42 | 74.65 | Prepared Chinese MRC dataset and set the baseline for further MRC research studies. |
| - | Proposed | Distil-BERT-MRC | WikiQA | 90.23 | 91.42 | Achieved state-of-the-art performance with minimal computational cost. |

**Table 4**. Comparaison of Distil-BERT-MRC performance with State-of-the-Art Methods.

epochs. These results suggest that learning and generalization are effective in this system performance compared to RoBERTa which has a training accuracy of 93.95% and a validation accuracy of 92.50%, with elevated losses. A training loss of 0.29% and a validation loss of 0.37%. XLNet is decently effective. Falls, behind both DistilBERT and RoBERTa in terms of performance metrics with a training accuracy of 93.50% and a validation accuracy of 92.20% coupled with higher losses (training loss at 0.31% and validation loss at 0.39%).

After analyzing metrics, it was evident that our customized DistilBERT framework outperformed when tested using the WikiQA dataset. Moreover, Table 4 presents the performance of the proposed model against the recent transformers-based models over diverse datasets.

According to the table DistilBERT achieved a better exact match and F1 score comparatively. It emerged as the standout choice, being both efficient and effective as represented in Fig. 14. As stated in Table 5, RoBERTa and XLNet consist of 125 million and 117 million parameters, respectively, whereas the proposed model boasts only 66 million. Similarity DistilBERT also has a compact model size 240 megabytes, making it smaller than other models. Its inference time is just 60 milliseconds, making it computationally efficient compared to RoBERTa and XLNet which have double the inference time. Moreover, it requires four hours less training time than RoBERTa (eight hours) and XLNet (seven and a half hours). While DistilBERT is highly effective, achieving a training accuracy of 94.30% and validation accuracy of 92.70% outperforms RoBERTa and XLNet in balancing efficiency and effectiveness. While performing the experiments, we took care to minimize dataset, algorithmic, and evaluation biases by using diverse and widely accepted benchmark datasets and applying standardized evaluation metrics. In the experimental phase, we attempted to prevent data, evaluation and algorithmic biases, by employing different diverse benchmark datasets, evaluation metrics and recent models.

## Ablation study

We performed ablation study to assess the DistilBERT-MRC framework thoroughly. This study focuses on attempting to clarify over the constituents of architecture like token representation techniques including classification token (CLS), question word (QW) and entity aware (EA), the fully connected (FC) layer, SoftMax layer, selection of transformers original BERT against Distil-BERT-MRC. Results of ablation study experiments are presented in Table 6; Fig. 15. We evaluated the effect of each component's removal from the model and how it impacted text comprehension and ambiguity resolution performed by our system. Ablation experiments were conducted on the WikiQA dataset and were evaluated with accuracy and F1 score. Results of the complete model

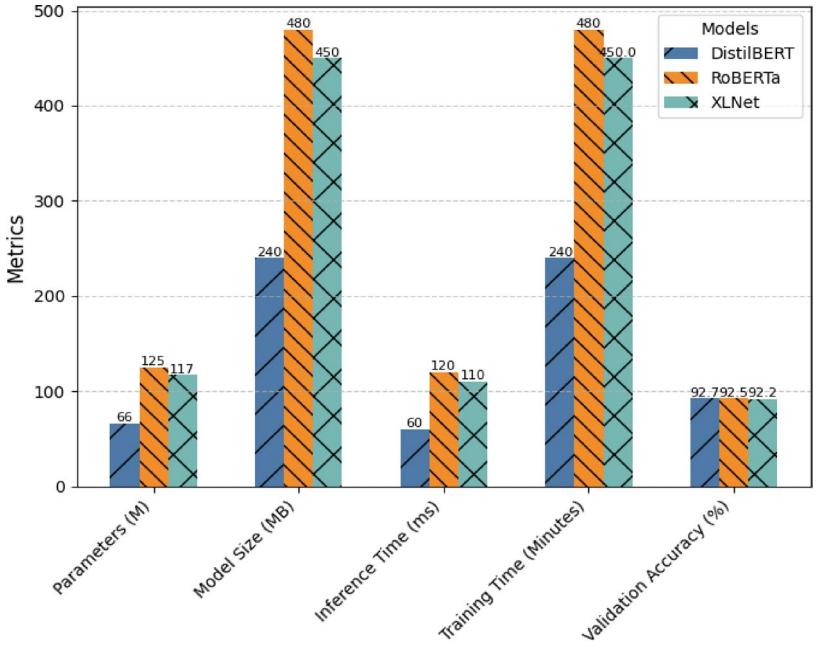


**Fig. 14**. A comparative analysis of efficiency and resource utilization.

| Metric | DistilBERT | RoBERTa | XLNet |
|---|---|---|---|
| Parameters | 66 M | 125 M | 117 M |
| Model size | 240 MB | 480 MB | 450 MB |
| Inference time | 60 ms | 120 ms | 110ms |
| Training time | 4 h | 8 h | 7.5 h |
| Training accuracy | 94.30% | 93.95% | 93.50% |
| Validation accuracy | 92.70% | 92.50% | 92.20% |

**Table 5**. Metric based comparison of distilbert with recent transformers-based models.

| Architecture variant | Validation accuracy (%) | F1 Score (%) |
|---|---|---|
| Proposed full architecture | 92.70 | 91.42 |
| Without QW + EA | 89.50 | 87.65 |
| Without [CLS] | 90.10 | 88.72 |
| Without FC Layer | 88.30 | 86.40 |
| Without softmax | 87.90 | 85.90 |
| Original BERT | 91.00 | 90.23 |

**Table 6**. Ablation study results showing the effect of removing individual components from the proposed model on validation accuracy and F1 Score.

are used as the first or baseline measurement. Specifically, the full model achieved 92.70% accuracy and an F1 score of 91.42% on validation. In the first ablation, we removed QW and EA vectors and classified using the [CLS] token. Resultantly, the system's accuracy dropped to 89.50% and the F1 score declined to 87.65%.

Thus, the inclusion of question-and-answer region data allows the [CLS] token to better capture intricate word relationships and distinguish words that are often confused. Subsequently, by excluding the CLS token and keeping QW and EA resulted in the second ablation which dropped accuracy to 90.10% and F1 to 88.72%. This suggests that although QW and EA vectors achieve their separate goals, the [CLS] token is critical for overarching summaries. The removal of the fully connected layer resulted in a drop in performance to a range of 88.30% accuracy and F1 score of 86.40% which demonstrates the FC layers importance in the collection and transformation of the concatenated vector into viable feature spaces prior to classification.

Furthermore, eliminating the Softmax layer yielded the worst performance among all ablation experiments with an accuracy of 87.90% and F1 of 85.90%. This loss further affirms the Softmax layer's integral function for

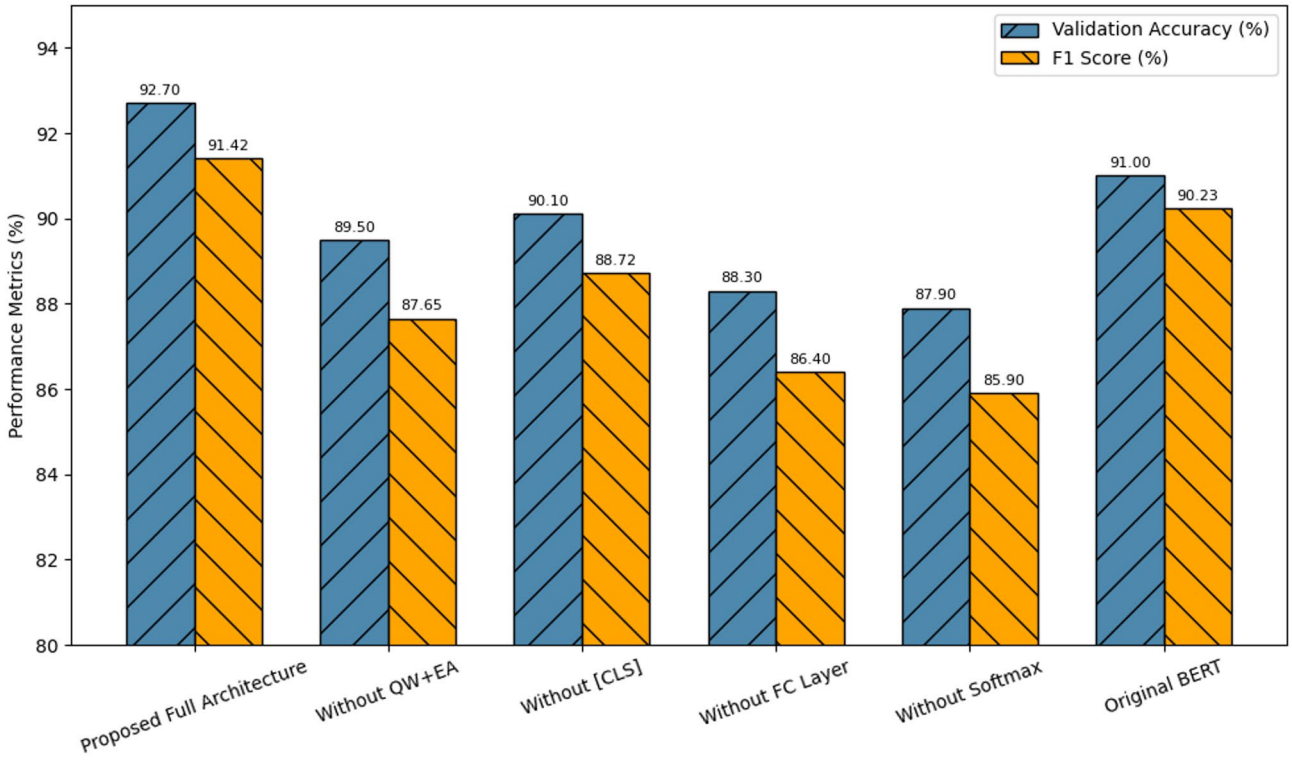


**Fig. 15**. Impact of removing individual components from the proposed architecture on model performance.

| Question | Passage snippet | Ambiguity type | Ground truth answer | RoBERTa prediction | XLNet prediction | DistilBERT-MRC prediction |
|---|---|---|---|---|---|---|
| When did Lincoln die? | Lincoln was shot on April 14 and died the next morning. | Temporal | April 15 | April 14 | April 14 | April 15 |
| Who founded Amazon? | Amazon was created by Jeff Bezos. He left his job at D. E. Shaw to pursue the idea. | Referential | Jeff Bezos | D. E. Shaw | D. E. Shaw | Jeff Bezos |
| What is the capital of Mercury? | Mercury is a planet, not a country. | Lexical | No answer | Mercury | Mercury | No answer |
| Who is the president of Apple? | Apple's CEO Tim Cook introduced the iPhone. The president of Apple Europe was also present. | Referential | Tim Cook | Apple Europe | Apple Europe | Tim Cook |
| When did WW2 start? | WW2 started in 1939, although some tensions had been rising since 1938. | Structural | 1939 | 1938 | 1938 | 1939 |

**Table 7**. Comparison of model predictions on ambiguous passages from the WikiQA Dataset.

defining class membership boundaries of a decision space and interpreting them probabilistically. To further investigate the impact of the underlying architecture in the transformer, we replaced DistilBERT with the full BERT model. This alternative offered a significant impact in accuracy to 91% and F1 score of 90.23%, but at the computation cost of BERT greatly increased model parameters, inference time, and memory requirements. Results from the ablation studies confirm the rationale behind the model architecture decisions. Each element from the token representations through neural layers holds a unique role that is crucial to achieving optimal performance and contextual understanding in the MRC tasks.

## Error analysis

We conducted a qualitative error analysis on a small sample taken from the WikiQA dataset to increase our understanding of how well the Distil-BERT-MRC model resolves ambiguities in passages. In the sample, we randomly selected 100 triples of questions-passages-answers studied the passage for occurrences of ambiguity. Each ambiguous example was defined into distinct ambiguity types, Lexical ambiguity (1 word can mean more than 1 thing), Referential ambiguity (reference by an unencyclopedic pronoun or noun), Temporal ambiguity (imprecise reference to time), and Structural ambiguity (complex nested). For each ambiguity type, we described the prediction of the Distil-BERT-MRC model and compared it to the predictions made by RoBERTa and XLNet. Table 7 presents five illustrative samples from our analysis, where the model correctly predicts the response despite ambiguity.

In each of these, either XLNet or RoBERTa fails to predict the correct span, typically misled by keyword matching or overlooking context nuances. DistilBERT-MRC, on the other hand, inferred the correct answer

through improved use of deeper contextual reasoning. This also establishes its ability to perform meaning ambiguity resolution irrespective of syntactic proximity or surface structure and supports our approach for improving the efficiency of comprehension while imposing low computation cost.

## Conclusion

In this study, we thoroughly evaluated types of network architectures for tasks regarding question answering including the recurrent neural networks RNN, LSTM, GRU, and Distil-BERT. The results showed that LSTM has attained a higher accuracy rate than RNN and GRU when used to test the WikiQA dataset since it efficiently identifies long-term dependencies. The LSTM model has achieved a validation accuracy of 89.60%. Subsequently, the proposed the Distil-BERT-MRC is compared with recent transformers-based architecture including RoBERTa and XLNet to assess algorithm biasness. However, the proposed model proved to be more effective than the one trained with the WikiQA dataset based on the F1 score and exact match assessments. This paper seeks to explore the efficiency of network structures and language models whilst answering questions with an F1 score of 91. 42% and the exact match is 90.23% based on the experiments conducted from the text snippets taken from the WikiQA dataset. These outcomes provided evidence of the successful understanding capabilities of the proposed Distil-BERT-MRC model. To assess the model's generalization, we further performed experiments on three additional benchmark datasets including SQuAD, NewsQA, and Natural Questions and the model maintained competitive performance across all datasets. These results emphasize that model adaptation and evaluation in contexts are essential for enhancing state-of-the-art in natural language processing and question answering. Furthermore, this work presents a resource-efficient adaptation of the Distil-BERT model to maximize the efficiency of MRC while actively minimizing training time and parameters without compromising accuracy. Despite effectively addressing passage ambiguity, the proposed architecture might be faced with reduced generalization with highly domain-specific or OOD text. Although, the attention mechanism provides a certain level of architecture understanding transformers-based Distil-BERT is indeed a black box since they do not have intuitive interpretability. In future work, we expect to apply domain-specific datasets; restate and continue to evolve the transformer architecture to better comprehend and process large and complex MRC passages.

## Data availability

The datasets used in this study include the Microsoft Research Wiki-QA corpus, SQuAD 2.0, NewsQA, and Natural Questions. All datasets are publicly available. The Wiki-QA corpus and NewsQA can be accessed from the official Microsoft website at [https://www.microsoft.com/en-us/download/details.aspx? id=52419](https:/www.microsoft.com/en-us/downloaddetails.aspx? id=52419) and [https://www.microsoft.com/en-us/research/project/newsqa-dataset/](https:/www.microsoft.com/en-us/research/project/newsqa-dataset) respectively, while SQuAD 2.0 is available at [https://rajpurkar.github.io/SQuAD-explorer/](https:/rajpurkar.github.io/SQuAD-explorer) , and Natural Questions at [https://ai.google.com/research/NaturalQuestions](https:/ai.google.com/research/NaturalQuestions) .

## Code availability

The source code used to implement the framework and conduct the experiments in this study is publicly available at GitHub: https://github.com/MuzamilAhmed0007/MRC_Transformers.git. This version corresponds to the implementation reported in the manuscript.

Received: 28 March 2025; Accepted: 29 October 2025
Published online: 27 November 2025

## References

1. Cho, S. et al. Multi-Paragraph machine reading comprehension with hybrid reader over tables and text. *Appl. Artif. Intell.* **38** (1), 2367820 (2024).
2. Zeng, C., Li, S., Li, Q., Hu, J. & Hu, J. A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets. *Appl. Sci.* **10** (21), 7640 (2020).
3. Liu, S., Zhang, X., Zhang, S., Wang, H. & Zhang, W. Neural machine reading comprehension: methods and trends. *Appl. Sci.* **9** (18), 3698 (2019).
4. Ahmed, M. et al. On solving textual ambiguities and semantic vagueness in MRC based question answering using generative pre-trained Transformers. *PeerJ Comput. Sci.* **9**, e1422 (2023).
5. Dai, Y., Fu, Y., Yang, L., Multiple-Choice, A. & Machine reading comprehension model with Multi-Granularity semantic reasoning. *Appl. Sci.* **11**, 7945. https://doi.org/10.3390/app11177945 (2021).
6. Luo, H., Shi, Y., Gong, M., Shou, L. & Li, T. MaP: A matrix-based prediction approach to improve span extraction in machine reading comprehension, Preprint at https://arXiv./org/abs/2009.14348, 2020. (2009).
7. Izacard, G. & Grave, E. Distilling knowledge from reader to retriever for question answering, Preprint at https://arXiv./org/abs/2012.04584, 2020.
8. Min, S., Chen, D., Zettlemoyer, L. & Hajishirzi, H. Knowledge guided text retrieval and reading for open domain question answering. Preprint at https://arXiv./org/abs/1911.03868, 2019.
9. Lee, K., Chang, M. W. & Toutanova, K. Latent retrieval for weakly supervised open domain question answering. Preprint at https://arXiv./org/abs/1906.00300, 2019
10. Izacard, G. & Grave, E. Leveraging passage retrieval with generative models for open domain question answering. Preprint at https://arXiv./org/abs/2007.01282, 2020
11. Guu, K., Lee, K., Tung, Z., Pasupat, P. & Chang, M. Retrieval augmented language model pre-training, in *International conference on machine learning*PMLR, 3929–3938. (2020).
12. Callan, D. & Foster, J. How interesting and coherent are the stories generated by a large-scale neural Language model? Comparing human and automatic evaluations of machine-generated text. *Expert Syst.* **40** (6), e13292. https://doi.org/10.1111/exsy.13292 (2023).

13. Chen, C., Sun, X., Hua, Y., Dong, J. & Xv, H. Learning deep relations to promote saliency detection, in *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 07, 10510–10517. (2020).
14. Zhao, H., Sun, X., Dong, J., Yu, H. & Wang, G. Multi-instance semantic similarity transferring for knowledge distillation. *Knowl. Based Syst.* **256**, 109832 (2022).
15. Li, B. A Study of DistilBERT-Based Answer Extraction Machine Reading Comprehension Algorithm, in *Proceedings of the 2024 3rd International Conference on Cyber Security, Artificial Intelligence and Digital Economy*, 261–268. (2024).
16. Dsouza, F., Bodade, A., Kolhe, H., Chaudhari, P. & Madankar, M. Optimizing MRC Tasks: Understanding and Resolving Ambiguities, in *2nd International Conference on Paradigm Shifts in Communications Embedded Systems, Machine Learning and Signal Processing (PCEMS)* (IEEE, 2023).
17. Mihaylov, T. B. Knowledge-enhanced neural networks for machine reading comprehension, (2024).
18. Chen, D., Fisch, A., Weston, J. & Bordes, A. Reading wikipedia to answer open-domain questions. Preprint at https://arXiv./org/abs/1704.00051, 2017.
19. Yang, W. et al. End-to-end open-domain question answering with bertserini. Preprint at https://arXiv./org/abs/1902.01718, (2019).
20. Zhu, C., Zeng, M. & Huang, X. Sdnet: Contextualized attention-based deep network for conversational question answering Preprint at https://arXiv./org/abs/1812.03593, (2018).
21. Ohsugi, Y., Saito, I., Nishida, K., Asano, H. & Tomita, J. A simple but effective method to incorporate multi-turn context with BERT for conversational machine comprehension, Preprint at https://arXiv./org/abs/1905.12848, 2019
22. Ouyang, S., Zhang, Z. & Zhao, H. Fact-driven logical reasoning for machine reading comprehension, In:*Proceedings of the AAAI Conference on Artificial Intelligence*. **38**, 17, 18851–18859. (2024).
23. Yang, Z., Sun, Y. & Kuang, Q. Question answering model based on machine reading comprehension with knowledge enhancement and answer verification. *Concurrency Computation: Pract. Experience*. **34** (12), e5828 (2022).
24. Trischler, A. et al. Newsqa: A machine comprehension dataset, Preprint at https://arXiv./org/abs/1611.09830, (2016).
25. Yang, Y., Yih, W. & Meek, C. Wikiqa: A challenge dataset for open-domain question answering, In: *Proceedings of the conference on empirical methods in natural language processing*, (2015).
26. Salman, M., Haller, A., Méndez, S. J. R. & Naseem, U. Doc-KG: Unstructured documents to knowledge graph construction, identification and validation with Wikidata, *Expert Syst.*. e13617. https://doi.org/10.1111/exsy.13617
27. Shi, T., Li, X., Liu, Z. & Wang, L. Research on Bi-LSTM machine reading comprehension algorithm based on attention mechanism, In: *Journal of Physics: Conference Series*, (IOP Publishing, 2022).
28. Seo, M., Kembhavi, A., Farhadi, A. & Hajishirzi, H. Bidirectional attention flow for machine comprehension, Preprint at https://arXiv./org/abs/1611.01603, (2016).
29. Wang, W., Yang, N., Wei, F., Chang, B. & Zhou, M. Gated self-matching networks for reading comprehension and question answering, In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 189–198. (2017).
30. Yu, A. W. et al. Qanet: Combining local convolution with global self-attention for reading comprehension, Preprint at https://arXiv./org/abs/1804.09541, (2018).
31. Chen, Z. & Wu, K. ForceReader: A BERT-based interactive machine reading comprehension model with attention separation, In: *Proceedings of the 28th International Conference on Computational Linguistics*, 2676–2686. (2020).
32. Kazi, S., Khoja, S. & Daud, A. A survey of deep learning techniques for machine reading comprehension. *Artif. Intell. Rev.* **56** (Suppl 2), 2509–2569 (2023).
33. Hu, M., Peng, Y., Huang, Z. & Li, D. Retrieve, read, rerank: Towards end-to-end multi-document reading comprehension, Preprint at https://arXiv./org/abs/04618, 2019. (1906).
34. Hu, M., Peng, Y., Huang, Z. & Li, D. A multi-type multi-span network for reading comprehension that requires discrete reasoning, Preprint at https://arXiv./org/abs/1908.05514, 2019.
35. Li, R. et al. Few-shot machine reading comprehension for Bridge inspection via domain-specific and task-aware pre-tuning approach. *Eng. Appl. Artif. Intell.* **147**, 110361 (2025).
36. Liu, Q., Mao, R., Geng, X. & Cambria, E. Semantic matching in machine reading comprehension: an empirical study. *Inf. Process. Manag.* **60** (2), 103145 (2023).
37. Nassiri, K. & Akhloufi, M. Transformer models used for text-based question answering systems. *Appl. Intell.* **53** (9), 10602–10635 (2023).
38. Joshi, M. et al. Spanbert: improving pre-training by representing and predicting spans. *Trans. Association Comput. Linguistics*. **8**, 64–77 (2020).
39. Bao, H. et al. Unilmv2: Pseudo-masked language models for unified language model pre-training, In: *International conference on machine learning* PMLR, 642–652. (2020).
40. Qian, Y., Santus, E., Jin, Z., Guo, J. & Barzilay, R. Graphie: A graph-based framework for information extraction, Preprint at https://arXiv./org/abs/1810.13083, (2018).
41. Zheng, W. et al. PAL-BERT: an improved question answering model. *Computer Model. Eng. & Sciences*, 1–10, (2023).
42. Alawwad, H. A., Alhothali, A., Naseem, U., Alkhathlan, A. & Jamal, A. Enhancing textbook question answering task with large language models and retrieval augmented generation, Preprint at https://arXiv./org/abs/2402.05128, (2024).
43. Sarzynska-Wawer, J. et al. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Res.* **304**, 114135 (2021).
44. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding, Preprint at https://arXiv./org/abs/1810.04805, (2018).
45. Radford, A., Narasimhan, K., Salimans, T. & Sutskever, I. Improving language understanding by generative pre-training, (2018).
46. Khan, W., Daud, A., Khan, K., Muhammad, S. & Haq, R. Exploring the frontiers of deep learning and natural Language processing: A comprehensive overview of key challenges and emerging trends. *Natural Lang. Process. J.*, 100026, (2023).
47. Mardini, I. D. et al. A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions. *Educ. Inform. Technol.* **29** (4), 4565–4590 (2024).
48. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization, Preprint at https://arXiv./org/abs/1412.6980, (2014).
49. Alzubi, J. A., Jain, R., Singh, A., Parwekar, P., & Gupta, M. COBERT: COVID-19 question answering system using BERT. *Arabian J. Sci. Eng.* pp. 1–11 (2021).
50. Wu, C.-S., Madotto, A., Liu, W., Fung, P., & Xiong, C. Qaconv: Question answering on informative conversations. arXiv preprint arXiv:2105.06912, (2021).
51. Caballero, E.Q., Rahman, M.S., Cerny, T., Rivas, P., & Bejarano, G. "Study of Question Answering on Legal Software Document using BERT based models," in LatinX in Natural Language Processing Research Workshop (2022).
52. Kumari, V., Keshari, S., Sharma, Y., & Goel, L., "Context-based question answering system with suggested questions," in 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, pp. 368–373, 2022.
53. Hosen, S., Eva, J. F., Hasib, A., Saha, A. K., Mridha, M., & Wadud, A. H. "HQA-Data: A Historical Question Answer Generation Dataset from Previous Multi-Perspective Conversation," Data in Brief, p. 109245 (2023).
54. Joshi, M., Choi, E., Weld, D. S., & L. Zettlemoyer, L. "TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension," arXiv preprint arXiv:1705.03551 (2017).

## Author contributions

A.N. and M.A. conceived the study and designed the methodology. M.A. implemented the experiments and performed data preprocessing. A.N. and T.I. contributed to model development and fine-tuning. H.U.K. and A.D. supervised the research and provided critical feedback on the study design. B.A. assisted in statistical analysis and evaluation metrics. M.A. and A.N. wrote the main manuscript text, while H.U.K. and A.D. reviewed and revised the manuscript. T.I. and B.A. contributed to the preparation of figures and tables. All authors reviewed and approved the final manuscript.

## Declarations

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to H.U.K. or A.D.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.