Building Bert-Based QnA System

Dominic Boy P. Almazan

IT Elective IV, BSIT

Jose Rizal University

Mandaluyong, Philippines

dominicboy.almazan@my.jru.edu

***Abstract*— This paper presents an evaluation of three pretrained transformer-based models BERT-base, DistilBERT, and BERT-large on a question-answering (Q&A) task using a real-world news article as context. Each model was tested on ten factual questions covering entities, numbers, causes, and confirmation queries. Results show that all models were able to produce correct answers, but with significant differences in speed, confidence, and handling of ambiguous cases. DistilBERT proved to be the fastest, averaging 1.5 seconds per query, while BERT-large provided the most descriptive and context-rich answers but required nearly 10 seconds per query. BERT-base offered a middle ground, though its confidence scores fluctuated widely. These findings highlight the trade-offs between efficiency and accuracy when selecting models for Q&A applications. The study also suggests future directions, including larger-scale evaluations, two-stage pipelines combining lightweight and heavyweight models, and fine-tuning on domain-specific datasets to improve robustness.**

Keywords*— *Question answering, Transformer models, BERT, DistilBERT, BERT-large, Natural Language Processing (NLP), Model evaluation

I. INTRODUCTION

Question-Answering (Q&A) systems are a vital part of Natural Language Processing (NLP) which automatically capture specific information from unstructured text in response to natural language questions. Q&A systems are used in valuable domains such as customer service, searching for biomedical information, analyzing legal documents, and in education, efficiently delivering access to valuable information. Traditional systems were based on either retrieval or rule-based systems that could not manage contextual ambiguity in human language.

Since the introduction of transformer-based architectures, such as Bidirectional Encoder Representations from Transformers (BERT), pretrained models have attained state-of-the-art performance on common benchmark datasets like SQuAD. Transformer based models provide contextualized embeddings that capture bidirectional dependencies and accuracy in extracting meaningful answers from the text. The reviewed studies showed significant improvements in performance over previous method of Q&A from early work with transformer architecture Q&A models. Variants of BERT (DistilBERT, RoBERTa, ALBERT) showed somewhat varying trade-offs in computational efficiency, accuracy, and performance relative to the specific task [1][2]. Additionally, various adapted versions of BERT were utilized successfully in specialized domains like biomedicine, where success is largely due to transfer learning with a relatively smaller domain-specific dataset being easily managed [3].

Despite these developments, there are still challenges. Larger models like BERT-large typically offer better accuracy but they require significant computational resources and inference time which may not be feasible in real-world applications. On the other hand, smaller models such as DistilBERT offer faster performance but can sacrifice on slight accuracy. This issue of trade-offs has spurred on comparative work to assist practitioners in figuring out which models will fit their constraints [3]. In this study, we compare three pretrained modelsBERT-base, DistilBERT, and BERT-largefor a news-based Q&A task. In this paper, we consider measures of accuracy, probability confidence, and runtime latency in order to discuss the trade-off between efficiency and accuracy in pretrained systems for Q&A.

II. METHODOLOGY

The study reviews three pretrained transformer models from the Hugging Face transformers library [4]. Each model was embedded in a Python pipeline for question and answer (Q&A) tasks with a context passage and a set of factual questions serving as the test case.

A. *Models*

- BERT-base-cased (SQuAD2): A widely used baseline model that was pretrained on large general corpora, and fine-tuned on the SQuAD2 dataset for question-answering tasks [4], [5].
- DistilBERT-base-uncased-distilled-SQuAD: A distilled version of BERT that was designed to reduce model size and inference latency while preserving similar levels of performance to the original model. Knowledge distillation techniques allow this model to be much more efficient and suitable to low-resource environments [4], [6].
- BERT-large-uncased-whole-word-masking-finetuned-SQuAD: A deeper model with more transformer layers, trained with whole-word masking and fine-tuned for Q&A, has higher accuracy and confidence but requires much more computation. [4], [5]

B. *Dataset and Context*

In this assessment, the contextual phrasing was based on the article from the Philippine News Agency titled

"Chartered flights eyed for repatriation of over 200 OFWs in Lebanon." The article discusses the government's plans to repatriate overseas Filipino workers (OFWs) from Beirut while discussing relevant logistics, legal, and security issues. The article includes a range of factual details including numbers, questions about the name of the organizations involved (DMW, OWWA, Philippine government), and causal explanations (e.g., bombings in Beirut). Because of this range, the article will support different kinds of questions about who, how many, why, what, and which, making it ideal to evaluate model performance across multiple types of questions.

All models received the same fixed text passage to ensure fairness and consistency in comparison. Although the passage was relatively short, it contained varying information density, with some items clearly noted one time and others repeated or mentioned using different wording. This allowed us to observe how well each model handles a blend of straightforward versus ambiguous questions. Using a real world news article adds to the realism of the evaluation because these are about the types of questions and challenges Q&A models would encounter in practice and outside curated datasets. Only using one article demonstrated that we could create a controlled environment to offer additional analysis and we could extend this to larger datasets and multiple domains in future work to better evaluate generalizability.

C. *Question*

Ten factual questions were manually designed to cover different categories of information within the article.

- Who is arranging chartered flights for repatriation of more than 200 OFWs in Beirut?
- How many OFWs did the DMW say they are trying to accommodate with the chartered flights?
- Why were the scheduled flights around Sept. 25 cancelled?
- How many OFWs are staying in four temporary shelters in Beirut?
- How many OFWs were applying for exit permits from the Lebanese government?
- What challenge did Olalia say the Philippine government is facing for chartered flights?
- What other routes is the DMW studying in case the situation worsens?
- How much financial assistance will each repatriated OFW get from DMW and OWWA?
- Which country intensified its airstrikes across the northern border into Lebanon?
- According to the article, were any Filipinos hurt since the attacks were launched?

These questions cover a mix of entity identification (e.g., organizations, actors), numerical details, causal reasoning, and confirmation queries, ensuring a broad test of the models' comprehension abilities.

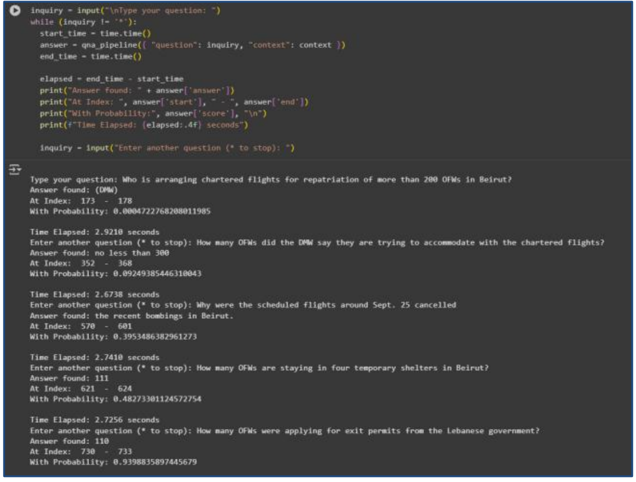
III. RESULTS

The three pretrained transformer models BERT-base-cased (SQuAD2), DistilBERT-base-uncased-distilled-SQuAD, and BERT-large-uncased-whole-word-masking-finetuned-SQuAD were evaluated on ten factual questions derived from a news article.

A. *Results of BERT-base-cased (SQuAD2)*

The deepset/bert-base-cased-squad2 model had consistent evaluation performance but some variability in confidence levels. It produced answers in a time frame of 2.7-3.1 seconds which made it moderately efficient: faster than BERT-large but slower than DistilBERT. Therefore, the model represented a compromise between computational cost and response speed.

In terms of accuracy, BERT-base had the majority of information from passage. For questions that were direct and explicit, such as how much financial assistance (PHP150,000) there was, or who was responsible for the air strikes (Israel), the model reached confidence scores of very high levels (≥0.96) and (≥0.99), respectively, suggesting that the model performed well when the answer was stated, and only stated once, within the passage. Furthermore, exit permit responses (110 OFWs) reached 0.94 confidence scores suggesting usable reliability when just numeric values are tested.



```
inquiry = input("\nType your question: ")
while (inquiry != '*'):
  start_time = time.time()
  answer = qna_pipeline({ "question": inquiry, "context": context })
  end_time = time.time()

  elapsed = end_time - start_time
  print("Answer found: " + answer['answer'])
  print("At Index: ", answer['start'], " - ", answer['end'])
  print("With Probability:", answer['score'], "\n")
  print(f"Time Elapsed: {elapsed:.4f} seconds")

  inquiry = input("Enter another question (* to stop): ")
```

```
Type your question: Who is arranging chartered flights for repatriation of more than 200 OFWs in Beirut?
Answer found: (DMW)
At Index:  173  -  178
With Probability: 0.0004722768208011985

Time Elapsed: 2.9210 seconds
Enter another question (* to stop): How many OFWs did the DMW say they are trying to accommodate with the chartered flights?
Answer found: no less than 300
At Index:  352  -  368
With Probability: 0.09249385446310043

Time Elapsed: 2.6738 seconds
Enter another question (* to stop): Why were the scheduled flights around Sept. 25 cancelled
Answer found: the recent bombings in Beirut.
At Index:  570  -  601
With Probability: 0.3953486382961273

Time Elapsed: 2.7410 seconds
Enter another question (* to stop): How many OFWs are staying in four temporary shelters in Beirut?
Answer found: 111
At Index:  621  -  624
With Probability: 0.48273301124572754

Time Elapsed: 2.7256 seconds
Enter another question (* to stop): How many OFWs were applying for exit permits from the Lebanese government?
Answer found: 110
At Index:  730  -  733
With Probability: 0.9398835897445679
```

Figure 1. Snipper Code of BERT-base-cased

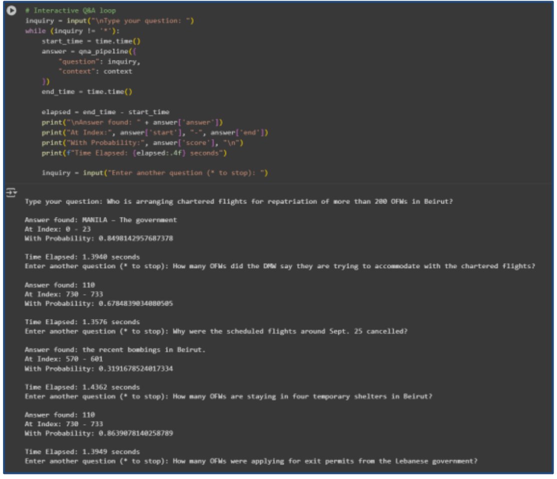
The model did, however, demonstrate limitations in ambiguous or competing answer span situations. In example three, upon modelling the authority who arranged the flights - it identified the correct answer as representing the DMW - it assigned it a very small near-zero probability of 0.00047. Similar inconsistency was evident within the numeric questions applied to various examples, such as specifying OFW numbers in shelters; even when it predicted a reasonable OFW number, it had a confidence of only around 0.48. The instability noticed in BERT base suggests that it could sometimes identify the correct answer span but not with any confidence relative to alternative text options. The output for BERT-base authenticated a modest and inconsistent profile:

- Accuracy: High on clear, unambiguous questions, but inconsistent on ambiguous ones.
- Confidence: Ranged widely from near zero to above 0.99, highlighting sensitivity to question clarity.
- Efficiency: Average latency of ~2.8 seconds per query, faster than BERT-large but slower than DistilBERT

B. *Results of DistilBERT-base-uncased-distilled-SQuAD*

Between and among the three models, the DistilBERT-base-uncased-distilled-SQuAD model consistently produced fastest response times for given answers. The time ranged from 1.3 to 1.7 seconds per query, almost double as fast as BERT-base and nearly six times as BLERT-large. This demonstrates the usefulness of knowledge distillation, which is a network compression of the original BERT architecture in a smaller and faster model.

In terms of accuracy, DistilBERT performed reliably across most of the factual questions. For straightforward details such as financial assistance (PHP150,000) and the responsible country (Israel), the model achieved high confidence scores above 0.94, while maintaining rapid inference. Numeric details like the number of OFWs (110) were also predicted correctly with strong confidence, reaching up to 0.91. These results highlight DistilBERT's ability to retain strong factual extraction capabilities despite its smaller size.



```
# Interactive Q&A loop
inquiry = input("\nType your question: ")
while (inquiry != '*'):
    start_time = time.time()
    answer = qna_pipeline({
        "question": inquiry,
        "context": context
    })
    end_time = time.time()

    elapsed = end_time - start_time
    print("\nAnswer found: " + answer['answer'])
    print("At Index:", answer['start'], "-", answer['end'])
    print("With Probability:", answer['score'], "\n")
    print(f"Time Elapsed: {elapsed:.4f} seconds")

    inquiry = input("Enter another question (* to stop): ")
```

```
Type your question: Who is arranging chartered flights for repatriation of more than 200 OFWs in Beirut?

Answer found: MANILA – The government
At Index: 0 - 23
With Probability: 0.8498142957687378

Time Elapsed: 1.3940 seconds
Enter another question (* to stop): How many OFWs did the DMW say they are trying to accommodate with the chartered flights?

Answer found: 110
At Index: 730 - 733
With Probability: 0.6784839034080505

Time Elapsed: 1.3576 seconds
Enter another question (* to stop): Why were the scheduled flights around Sept. 25 cancelled?

Answer found: the recent bombings in Beirut.
At Index: 570 - 601
With Probability: 0.3191678524017334

Time Elapsed: 1.4362 seconds
Enter another question (* to stop): How many OFWs are staying in four temporary shelters in Beirut?

Answer found: 110
At Index: 730 - 733
With Probability: 0.8639078140258789

Time Elapsed: 1.3949 seconds
Enter another question (* to stop): How many OFWs were applying for exit permits from the Lebanese government?
```

Figure 2. Snipper Code of DistilBERT *-base-cased*

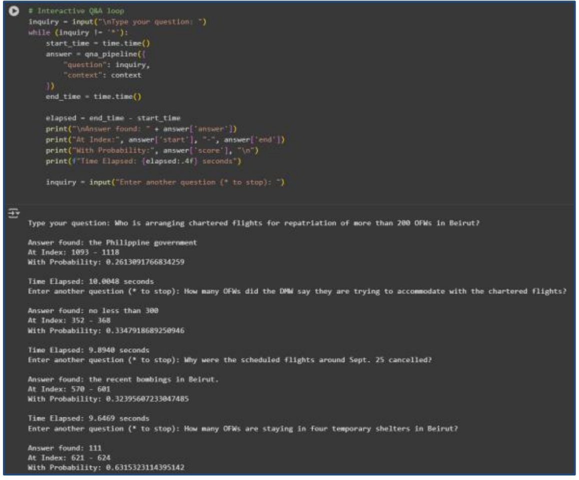
Nevertheless, the model presented lower confidence for queries with some uncertainty. Specifically, for the question, "What caused the flight cancellations?" the correct content was returned ("the recent bombings in Beirut"), but DistilBERT reported a confidence of only 0.32. Similarly, for the question about injuries to Filipinos, the response was technically correct ("no Filipinos were hurt"), but low confidence (0.40). In general, although DistilBERT contains the correct answer span, the lighter architecture occasionally results in lower confidence, especially on yes/no or causal questions. Characteristics of DistilBERT were:

- Accuracy: Strong for clear factual questions, especially numeric and entity-based queries, but less confident on ambiguous or explanatory questions.
- Confidence: Ranged from very high (≥0.95) on straightforward answers to moderate (~0.32–0.40) on complex or ambiguous ones.
- Efficiency: Average response time of ~1.5 seconds per query, the fastest among all models tested.

C. *Results of BERT-large-uncased-whole-word-masking-finetuned-SQuAD*

Among the three systems, the BERT-large-uncased-whole-word-masking-finetuned-SQuAD model performed the best overall in completeness and accuracy, but it was the slowest system, taking about 9.5 to 10 seconds to provide answers to each question.

In addition to its slower speed, the BERT-large setup produced accurate answers to nearly all questions posed. For example, the prediction of the appropriate amount of financial assistance (PHP150,000) was given a very high probability (0.98), a to the location of the country responsible for assisting the OFWs (Israel), BERT also assigned a probability above 0.92. Similarly, BERT correctly predicted the number of OFWs who applied to DSWD for exit permits (110) and OFWs currently in temporary shelters (111), with probabilities of 0.63 to 0.71, respectively.



```
# Interactive Q&A loop
inquiry = input("\nType your question: ")
while (inquiry != '*'):
    start_time = time.time()
    answer = qna_pipeline({
        "question": inquiry,
        "context": context
    })
    end_time = time.time()

    elapsed = end_time - start_time
    print("\nAnswer found: " + answer['answer'])
    print("At Index:", answer['start'], "-", answer['end'])
    print("With Probability:", answer['score'], "\n")
    print(f"Time Elapsed: {elapsed:.4f} seconds")

    inquiry = input("Enter another question (* to stop): ")
```

```
Type your question: Who is arranging chartered flights for repatriation of more than 200 OFWs in Beirut?

Answer found: the Philippine government
At Index: 1093 - 1118
With Probability: 0.2613091766834259

Time Elapsed: 10.0048 seconds
Enter another question (* to stop): How many OFWs did the DMW say they are trying to accommodate with the chartered flights?

Answer found: no less than 300
At Index: 352 - 368
With Probability: 0.3347918689250946

Time Elapsed: 9.8940 seconds
Enter another question (* to stop): Why were the scheduled flights around Sept. 25 cancelled?

Answer found: the recent bombings in Beirut.
At Index: 570 - 601
With Probability: 0.32395607233047485

Time Elapsed: 9.6469 seconds
Enter another question (* to stop): How many OFWs are staying in four temporary shelters in Beirut?

Answer found: 111
At Index: 621 - 624
With Probability: 0.6315323114395142
```

Figure 3. Snipper Code of BERT-large -cased

BERT-large's output had the distinct characteristic of generating longer, more verbose answer spans in comparison to the other models. For instance, when asked regarding who arranged the chartered flights, it produced the answer "the Philippine government," a longer answer than the other models which only produced, "DMW." While syntax was maintained and the BERT-large response is semantically linked to the correct answer, it serves as an illustration of BERT-large being more sensitive to wider context in its answer.

Even with these advantages in output comprehensibility, BERT-large still did not provide as much confident results in responses across all queries. For example, causal and yes/no questions provided a medium degree of confidence (0.32 for the explanation for cancellations, and 0.46 meaning, yes, Filipinos were hurt). This is an indication that in spite of its deeper architecture, BERT-large is still sensitive to ambiguity throughout the passage. BERT-large exhibited the following profile:

- Accuracy: High, with consistently correct answers across factual, numeric, and entity-based questions.
- Confidence: Ranged from very high (≥0.92) on clear factual queries to moderate (0.32–0.46) on ambiguous or binary questions.
- Efficiency: Slowest among all models, averaging nearly 10 seconds per query, which limits its practicality for real-time applications.

For each query, the system logged the answer span, probability confidence score, and response time. Table I presents representative comparisons across models.

Question	BERT-base	DistilBERT	BERT-large
Who is arranging flights?	“DMW,” Score 0.00047, 2.92 seconds	“The government …,” Score 0.85, 1.39 seconds	“Philippine government,” Score 0.26, 10.00 seconds
How many OFWs in shelters?	111, Score 0.48, 2.72 seconds	110, Score 0.86, 1.39 seconds	111, Score 0.63, 9.75 s
Why were flights cancelled?	“Bombings in Beirut,” Score 0.39, 2.74 seconds	“Bombings in Beirut,” Score 0.32, 1.43 seconds	“Bombings in Beirut,” Score 0.32, 9.64 seconds
Assistance per OFW?	PHP150,000, Score 0.96, 2.90 seconds	PHP150,000, Score 0.95, 1.77 seconds	PHP150,000, Score 0.97, 9.96 seconds

Table I – Selected Q&A Results Across Models

The latency findings indicated the existence of a distinct separation among the three models. Between 1.3–1.8 seconds per query, DistilBERT provides the fastest query return time, which makes it optimal for time-sensitivity and low-resource contexts. BERT-base produces time-to-response of 2.7 to 2.9 seconds, which provides a basic proximity to the moderate distance between the initial two models regarding speed of processing and complexity of the model. BERT-large produced correspondingly larger latencies, average of the 9.4 to 10.0 seconds per query as the slowest of the three models. The dichotomy between the three models and their performance is indicative of the direct correlation with model size and parameter count on efficiency in inference speed: smaller models can return an answer quickly, while larger models are slower to respond regardless of input.

Results in accuracy and confidence were strongly dependent on the type of question being asked. Clear and unambiguous, factual questions resulted in correct answers across all models, and a high confidence rating (≥0.95), for example, the amount of financial assistance offered (PHP150,000), was consistent, proving reliability in models when the answer indicated only one mention in the passage. Performance, however diverged when faced with ambiguous, or repeated details. When asked to identify actors or to count the number of OFWs or assistants asked to identify the actors, the models had different outputs under those contexts. Oftentimes, DistilBERT returned damaged information with shorter, and more concise spans, but with relatively high confidence belief. Meanwhile, BERT-large produced longer, but descriptive answers with moderate levels of confidence. In contrast, BERT-base returned correct spans, but at, times offered very low confidence scores, which could be problematic, as in the case of DMW, which dropped to 0.00047 even though the answer was correct.

Ultimately, these findings illustrate that model behavior reflects architecture size, but also reflects the clarity of the input question. DistilBERT often did well with efficiency and often provided more concise answers. BERT-large did its job well with detail and context, but did not always do so with speed. BERT-base generally only fell in between the other two models, while offering low levels of confidence in the scoring model.

IV. DISCUSSION

The findings reveal three main issues. First, model size is strongly associated with latency. DistillBERT had the best speed in terms of inference time and still had reasonable accuracy and would be the best model to use in real-time or restricted resource situations. BERT- large had the largest computational cost and was nearly eight times longer per query than DistillBERT but generally had stable and contextually identifiable answers with it's extra inference time.

Second, the accuracy of the answers was contingent on question clarification. Since some text contained more than one possible answer, the models chose different answers from the text sometimes. An example of this would be DistillBERT would answer the question “who governs the __________?” with “the government” whereas BERT-large would answer with “the Philippine government.” Both are right answers to the question, but it shows the significance of the definition of right. This seemed even more pronounced when numeric values presented even hopeless outcomes discussing differences in answers such as “110” OFWs and “111” OFWs.

Third, all models were in agreement when the answer in the text was clear and easy to find. In other words, all three models produced the same correct answer with high confidence (greater than 0.95) when asked about the PHP150,000 assistance amount. This indicates that even smaller models will accurately address straightforward questions.

In comparison to the other 3 reasons, the overall results indicate that these models differ in their capability to balance speed and accuracy. DistilBERT is the best option when responses are needed quickly the most; BERT-large is better when the accuracy is needed more than speed; and BERT-base is somewhere between. Future work on this topic could expand the study by measuring the accuracy across more examples, testing combined options (such as using DistilBERT first then just checking with BERT-large), and finetuning the models on specific types of text to enhance the reliability of the answers.

V. CONCLUSION

In this project, we evaluated the performance of three transformer-based models, BERT-base, DistilBERT, and BERT-large, with a set of factual questions based on a real-world news article for context. The most important results were that all three models generated the correct answers, but differed in speed, confidence and handling of less direct, or ambiguous, questions. Overall, the study revealed that pretrained models can be incredibly useful, but there needs to be a balance between operating efficiency and accuracy, when using them in practice.

Of the three models, DistilBERT outperformed the others with respect to speed. DistilBERT generated answers to the questions once about 50 per cent faster than BERT-base, and more than several times faster than BERT-large. DistilBERT would be more useful in situations where speed is of the essence, such as when results are needed in real time. In contrast, while BERT-large generated answers that were descriptive, drawing in relevant contextual information and resulted in more correct answers, it required quite a bit longer time per question.

From this exercise, I gained a better understanding of the trade-offs of speed vs accuracy with NLP models. I also observed how the type of question (whether it was a number, an explanation, or a name) impacted the confidence and reliability of the answers. This helped me see that selecting the appropriate model is not just about getting the right size or performance out of the model; it is also about walking the line of the kind of task and data that the model will perform on.

Looking forward, I learned that there are so many different ways to enhance the performance of the model(s) beyond just selecting the model for the task. One option is to create a pipeline where models can be combined; for example, the use of DistilBERT to quickly scan for answers (and propose them), then using BERT-large to verify and possibly improve the answers. Another phase we might try, would be to evaluate and assess the models on larger datasets and fine-tune the models on more specific topics to allow for more reliable responses. These adjustments could improve the usefulness of Q&A systems in any practical application.

All in all, this experience allowed me to apply real world experience in evaluating models that were already trained and provided lessons for analyzing outcomes that also included confidence, response times and the questions type we asked in building effective NLP systems. I will apply that understanding in it something to consider in future activities when I plan to research other models available on Hugging Face and do similar testing to apply these models in different NLP tasks..

VI. REFERENCES

[1] M. A. Alnazzawi, *“Comparative Analysis of State-of-the-Art QA Models BERT, RoBERTa, DistilBERT, and ALBERT on SQuAD v2 Dataset,”* ResearchGate, Feb. 2024. [Online]. Available: https://www.researchgate.net/publication/378327924

[2] N. M. Alqahtani and M. A. Alnazzawi, *“Question Answering on Biomedical Research Papers using Transfer Learning on BERT-Base Models,”* ResearchGate, Oct. 2023. [Online]. Available: https://www.researchgate.net/publication/375011546

[3] M. M. ElSayed et al., *“Improving the BERT model for long text sequences in question answering domain,”* ResearchGate, Mar. 2024. [Online]. Available: https://www.researchgate.net/publication/378644410

[4] Hugging Face, *“Transformers Documentation and Pretrained Models,”* [Online]. Available: https://huggingface.co/transformers

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, *“BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,”* in Proc. NAACL-HLT, 2019.[Online.] Available: https://aclanthology.org/N19-1423/

[6] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, *“DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter,”*[Online.] Available : https://arxiv.org/pdf/1910.01108