A Comparative Analysis of Transformer-Based Models for Question-Answering

Juan Carlos Miguel Timoteo
Computer Studies & Engineering
José Rizal University
Mandaluyong City, Philippines
juancarlosmiguel.timoteo@my.jru.edu

***Abstract*—Artificial Intelligence (AI) chatbots have emerged as influential tools for natural language interaction, particularly with the rise of large language models (LLMs) trained on massive corpora. While their fluency and conversational capabilities are remarkable, chatbots often suffer from inaccuracies or "hallucinations," producing outputs that are factually incorrect yet delivered with confidence. This paper reviews the limitations of such chatbots, introduces Retrieval-Augmented Generation (RAG) as a methodological solution, and explores how LangChain serves as a practical framework for building RAG-based applications. The study situates these findings in the larger context of trustworthy AI development, emphasizing the need for credibility, transparency, and modular frameworks when deploying conversational systems across domains.**

***Keywords*—Transformer-Based Models, Question Answering, BERT, DistilBERT, RoBERTa, BERT-large, Retrieval-Augmented Generation, Natural Language Processing, Model Comparison, Performance Evaluation, Inference Speed, Accuracy in AI, Knowledge Distillation, Stanford Question Answering Dataset, Large Language Models, Transfer Learning, AI Transparency**

I. Introduction

In the field of Natural Language Processing (NLP), the development of effective question-answering (Q&A) systems remains a significant area of research. The ability of a machine to comprehend a given text and provide accurate, concise answers to user queries has wide-ranging applications, from search engines and chatbots to data analysis and virtual assistants. The advent of transformer architectures, particularly the Bidirectional Encoder Representations from Transformers (BERT) model, has revolutionized this domain [1].

By pre-training on vast amounts of text data, these models develop a deep contextual understanding of language, which can then be fine-tuned for specific tasks like Q&A. This paradigm of transfer learning has made it possible to achieve state-of-the-art performance without needing to train a model from scratch. However, the original BERT model is computationally intensive, leading to the development of various optimized and modified architectures. This study aims to compare the performance of four popular pre-trained, transformer-based models on a single, controlled Q&A task to evaluate the trade-offs between model size, speed, and accuracy.

II. Methodology

The experiment was designed to provide a direct comparison of the models' performance under identical conditions. The setup involved a consistent context, a fixed set of questions, and a uniform method for measuring performance.

1. Models: Four pre-trained models, all fine-tuned on the Stanford Question Answering Dataset (SQuAD) 2.0 [2], were selected from the Hugging Face model hub:

- BERT-base (deepset/bert-base-cased-squad2): A standard BERT model serving as the baseline [1].
- DistilBERT (distilbert-base-cased-distilled-squad): A distilled, lighter, and faster version of BERT [3].
- RoBERTa (deepset/roberta-base-squad2): A model based on BERT but optimized with an improved pre-training methodology [4].
- BERT-large (deepset/bert-large-uncased-whole-word-masking-squad2): A larger version of BERT with more parameters, intended to be more powerful.

- Dataset: A single news article regarding the repatriation of Filipino workers from Lebanon was used as the context. This text provided the sole source of information from which the models could derive their answers.

Experimental Procedure: For each of the four models, a question-answering pipeline was initialized using the transformers library in Python, as shown in Fig. 1. An interactive loop was executed where the same ten questions were posed to each mode based on the context article[5] (see Fig. 2). These questions were designed to test the models' ability to extract various types of information, including names, numbers, locations, and reasons..

```
# Import necessary libraries
from transformers import pipeline, BertForQuestionAnswering, BertTokenizer
import textwrap
import time
```

Fig. 1. Model and Tokenizer Initialization using the Hugging Face `pipeline`

- Performance Metrics: For each question, three metrics were recorded:
    1. Answer: The text extracted by the model from the context.
    2. Confidence Score**:** A numerical value indicating the model's confidence in its answer.
    3. Inference Time: The wall-clock time in seconds required for the model to generate an answer.

```
Context Article:

MANILA – The government is arranging chartered flights for the repatriation of more than 200
overseas Filipino workers in Beirut, Lebanon, the Department of Migrant Workers (DMW) said
Wednesday. “We are trying to provide for chartered flights. We're talking to airline companies so
that the chartered flights would be able to accommodate for example, no less than 300 overseas
Filipino workers from Beirut,” DMW Undersecretary Bernard Olalia said in a Palace press briefing.
This was after the scheduled flights of around 15 OFWs on Sept. 25 were cancelled because of the
recent bombings in Beirut. Olalia said around 111 OFWs are staying in four temporary shelters in
Beirut and waiting for their repatriation. An additional 110 OFWs are applying for exit permits from
the Lebanese government, Olalia said. “Apart from the documented OFWs, we have undocumented OFWs who
need to secure travel documents and once they're given travel documents, we will help them in
securing also exit visas or exit permits from the Immigration of the Lebanese government,” he said.
Olalia, however, said the Philippine government is facing several challenges, including securing
landing rights for chartered flights. He said land and sea routes are being considered, in case the
situation escalates and makes it “impossible” to take the air route. “The DMW is also studying the
possibility of other routes. Apart from air route, we will be assessing the sea and the land route,
should the case or the situation there worsen,” Olalia said. He said the DMW, the Overseas Workers
Welfare Administration (OWWA), and other concerned agencies will adopt a “whole-of-government
assistance” upon the directive of President Ferdinand R. Marcos Jr. He said each repatriated OFW
will get PHP150,000 in financial assistance from the DMW and OWWA, as well as psychosocial services.
Israel has intensified its airstrikes across the northern border into Lebanon, targeting the Iran-
backed militant group Hezbollah. Iran fired ballistic missiles in Israel on Tuesday night, following
the deadly attacks on Gaza and Lebanon and the recent killings of Hamas, Hezbollah, and Islamic
Revolutionary Guard Corps leaders. Olalia said no Filipinos were hurt since the attacks were
launched. “We have men on the ground. They work around the clock. At 'yung mga staff po natin,
dinagdagan na po natin (And we augmented our staff) both in Lebanon at (and) nearby posts to be able
to provide safest route, to evacuate and ultimately to facilitate the repatriation of our OFWs both
either in Lebanon or in Israel,” he said. (PNA)
```

Fig. 2. The Context Article Used for Testing.

III. Results

To mitigate inaccuracies, Retrieval-Augmented Generation (RAG) has emerged as a promising strategy. Conceptually, RAG adds an external retrieval component to the generative process [5]. Instead of depending solely on model memorization, an LLM enriched with RAG queries a knowledge base in real time, grounding its outputs in retrieved documents.

The performance of the four models varied significantly across the key metrics of accuracy and speed. A summary of the results is presented below.

- **Model 1: bert-base-cased-squad2 (Baseline)**
    - **Accuracy:** This model answered 9 out of 10 questions correctly. It incorrectly identified "Bernard Olalia" as the source of a directive that came from "President Ferdinand R. Marcos Jr.," as shown in Fig. 3.
    - **Inference Speed:** The average inference time was approximately 3.7 to 4.6 seconds per question.

```
Type your question (or enter '*' to stop): Whose directive led to the "whole-of-government assistance"?

Answer: Bernard Olalia
Score (Confidence): 0.9843
Inference Time: 3.9887 seconds
```

Fig. 3. Sample Output from Baseline Model (BERT-base), Highlighting an Incorrect Answer with a High Confidence Score.

- **Model 2: distilbert-base-cased-distilled-squad**
    - **Accuracy:** This model was the top performer in terms of accuracy, correctly answering all 10 questions. Notably, it correctly identified "President Ferdinand R. Marcos Jr." for the question the baseline model failed (see Fig. 4).
    - **Inference Speed:** This was the fastest model, with inference times ranging from 1.2 to 2.6 seconds, making it roughly 2-3 times faster than the baseline

```
Type your question (or enter '*' to stop): How much financial assistance will each repatriated OFW receive?

Answer: PHP150,000
Score (Confidence): 0.8716
Inference Time: 1.2829 seconds

Type your question (or enter '*' to stop): Which Iran-backed militant group was targeted by Israel's airstrikes?

Answer: Hezbollah
Score (Confidence): 0.9844
Inference Time: 1.1987 seconds

Type your question (or enter '*' to stop): According to Undersecretary Olalia, were any Filipinos hurt in the attacks?

Answer: no Filipinos were hurt
Score (Confidence): 0.4247
Inference Time: 1.9354 seconds

Type your question (or enter '*' to stop): Whose directive led to the "whole-of-government assistance"?

Answer: President Ferdinand R. Marcos Jr
Score (Confidence): 0.5624
Inference Time: 1.1969 seconds
```

Fig. 4. Sample Output from DistilBERT, Demonstrating Correctness and Significantly Improved Inference Speed Compared to the Baseline.

- **Model 3: deepset/roberta-base-squad2**
    - **Accuracy:** This model also achieved 100% accuracy, correctly answering all questions.
    - **Inference Speed:** Inference times ranged from 2.2 to 3.3 seconds, placing it between the baseline and DistilBERT in terms of speed.

- **Model 4: deepset/bert-large-uncased-whole-word-masking-squad2**
    - **Accuracy:** This model answered 9 out of 10 questions correctly. It made a significant error on a numerical question, answering "PHP15" when the correct answer was "PHP150,000," despite reporting a very high confidence score (0.99), as shown in Fig. 5.
    - **Inference Speed:** As expected, this was the slowest model by a significant margin, with inference times ranging from 8 to 13 seconds.

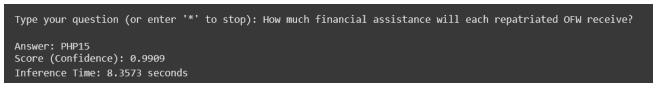

Fig. 5. Critical Error by the BERT-large Model on a Numerical Question, Answering "PHP15" Instead of "PHP150,000".

IV. Discussion

The results highlight the critical impact of pre-training methodology and model architecture on performance. The baseline BERT-base model served as a competent foundation but was surpassed by more recent or optimized architectures.

The standout performer, DistilBERT, demonstrates the success of knowledge distillation [3]. It retains the high accuracy of its larger parent model while being significantly smaller and faster. Its ability to correctly answer a question that the baseline model failed (Fig. 3 vs. Fig. 4) suggests that the distillation process may help in generalizing better on certain tasks.

RoBERTa's perfect accuracy and balanced speed underscore the benefits of its robustly optimized pre-training approach [4]. It offers a compelling alternative to the baseline, providing higher reliability with a moderate speed improvement.

The most surprising result came from the BERT-large model. Its failure on a simple numerical extraction task (Fig. 5), despite its larger size and high confidence, serves as a crucial reminder that "bigger is not always better." This type of error could stem from tokenization artifacts or other model-specific weaknesses and proves that high confidence scores should not be taken as an absolute measure of correctness. The substantial increase in computational cost for this model did not translate to superior performance in this experiment.

V. Conclusion

This comparative study evaluated four transformer-based models on a controlled question-answering task. The key finding is that the distilled model, DistilBERT, provided the best overall performance, delivering the highest accuracy and the fastest inference speed. RoBERTa also proved to be a highly reliable and well-balanced model. The performance of the largest model, BERT-large, which was the slowest and made a critical error, indicates that model selection requires a careful balance of size, speed, and task-specific accuracy rather than defaulting to the largest architecture. For practical applications where efficiency is a key consideration, optimized models like DistilBERT represent an excellent and often superior choice.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[2] P. Rajpurkar, R. Jia, and P. Liang, "Know What You Don't Know: Unanswerable Questions for SQuAD," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 784–789.

[3] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," presented at the 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS, 2019.

[4] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv preprint arXiv:1907.11692, 2019.

[5] Ruth Abbey Gita-Carlos, Chartered flights eyed for repatriation of over 200 OFWs in Lebanon, https://www.pna.gov.ph/articles/1234579