# Comparison of Pre-trained Models for Optimized Transformer Based Question Answering System

Tunahan Gokcimen
*Department of Software Engineering*
*Firat University*
Elazig, Turkey
tunahangokcimen@gmail.com

Bihter Das
*Department of Software Engineering*
*Firat University*
Elazig, Turkey
bihterdas@firat.edu.tr

***Abstract*— This study delves into the evaluation and optimization of transformer-based models for question-answering systems, focusing on health-related inquiries. Utilizing a specialized dataset extracted from Wikipedia articles, transformer models, namely Bert-base-cased, Electra-base, Deberta-base, Xlm-roberta-base, Distilbert-base, and Albert-base, were scrutinized based on their F1 scores and exact match accuracy. Electra-base and Deberta-base exhibited notable performance, showcasing the significance of models equipped with denoising mechanisms and disentangled attention. The outcomes highlight the critical role of tailored model selection in specific domains, particularly within health-related contexts. Future research avenues may explore fine-tuning strategies and optimizations for health datasets, addressing challenges in medical information extraction and question-answering. This study contributes valuable insights to the natural language processing field, guiding advancements in transformer-based question-answering systems, especially in the health domain.**

***Keywords—question-answering system, transformer models, semantic search, medical information***

## I. Introduction

In recent years, the field of natural language processing (NLP) has witnessed remarkable advancements, particularly in the domain of question-answering systems [1]. These systems play a pivotal role in transforming raw textual data into meaningful and actionable insights, with applications ranging from virtual assistants to information retrieval platforms. Amidst these advancements, the utilization of transformer-based models has emerged as a cornerstone, showcasing unparalleled performance in various NLP tasks [2,3].

This paper delves into the realm of question-answering (QA) systems, specifically focusing on the application of an optimized transformer model on a ready-made data set produced from Wikipedia articles. The quest for efficient and accurate question-answering mechanisms within the different domain holds immense significance, as it directly contributes to improved information accessibility, decision-making processes, and overall healthcare, education, commerce outcomes [4].

The use of transformer models, fine-tuned to the intricacies of big data, presents a promising avenue for enhancing the performance of question-answering systems [5,6]. This study explores the nuances of transformer optimization and its impact on the resolution of questions posed on a dataset. By scrutinizing the interplay between model architecture, training strategies, and dataset characteristics, we aim to contribute valuable insights that advance the state-of-the-art in question-answering systems tailored for a dataset including Wikipedia articles [7].

The remainder of this paper is organized as follows: Section 2 provides a comprehensive review of related work in the field of question-answering systems and transformer models. Section 3 details the methodology employed, including the dataset description, model architecture, training process and pre-trained models. Section 4 presents the experimental results and their implications. Finally, Section 5 concludes the paper with a summary of findings and outlines potential avenues for future research.

Through this exploration, we aspire to deepen our understanding of how optimized transformers can elevate question-answering capabilities, particularly in the context of health-related inquiries.

## II. Literature Review

In recent years, the exploration of question-answering (QA) systems has garnered significant attention from researchers and practitioners alike. Various approaches and models have been proposed to enhance the precision and efficiency of QA resolution, with a focus on leveraging advanced transformer architectures. The advent of transformer models, as introduced by Vaswani et al. [8], revolutionized natural language processing tasks, including QA systems. Transformer architectures, such as BERT (Devlin et al. [9]) and GPT (Radford et al. [10]), have demonstrated remarkable capabilities in capturing contextual information and semantic relationships within textual data. The effectiveness of transformers in understanding complex linguistic structures has led to their widespread adoption in QA applications. To further enhance the performance of transformer-based QA systems, researchers have delved into optimization techniques tailored for health datasets. Fine-tuning transformer architectures specifically for the intricacies of medical and health-related information has become a focal point [11]. The optimization process involves adapting pre-trained transformer models to the specific nuances of the health domain, resulting in improved question resolution accuracy. Question-answering on health datasets presents unique challenges and opportunities. Prior research, such as that by Tsatsaronis et al. [12] and Morante et al. [13], has emphasized the importance of biomedical semantic indexing and QA challenges. These efforts have provided valuable insights into the complexities of extracting medical knowledge from diverse datasets, laying the foundation for subsequent work on optimized transformer-based QA systems. Despite the advancements, several challenges persist in the realm of QA resolution on health datasets. The diversity of medical information, the need for precise entity recognition, and the incorporation of domain-specific

2024 12th International Symposium on Digital Forensics and Security (ISDFS) | 979-8-3503-3036-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/ISDFS60797.2024.10527306

979-8-3503-3036-6/24/$31.00 ©2024 IEEE

Authorized licensed use limited to: TUBITAK ULAKBIM. Downloaded on July 02,2024 at 06:50:25 UTC from IEEE Xplore. Restrictions apply.

knowledge bases pose ongoing challenges. Recent innovations, including attention mechanisms [14] and domain-adaptive techniques [15], aim to address these challenges and further refine the performance of QA systems in the health domain. Several studies have explored the optimization of transformer models for specific domains.

In conclusion, the literature surrounding question-answering systems on different datasets underscores the critical role of transformer architectures and their optimization for domain-specific challenges. The proposed work seeks to contribute to this growing body of knowledge by presenting an optimized transformer-based approach tailored for efficient question resolution in the scientific domain.

## III. Materials and Method

The materials and methods of this study involve the introduction of a customized Transformer-based Question-Answering (QA) approach on a data. Figure 1 illustrates the system architecture. A pre-existing dataset prepared for health data [16] is utilized as the dataset. This dataset has undergone question generation and answer creation processes. Subsequently, the previously generated QA data undergoes a fine-tuning process for a customized QA approach, involving different Transformer-based models. The models used in this step include Bert-base-cased, Electra-base, Deberta-base, Xlm-roberta-base, Distilbert-base, and Albert-base. The model demonstrating the best performance is determined through a comparative analysis.

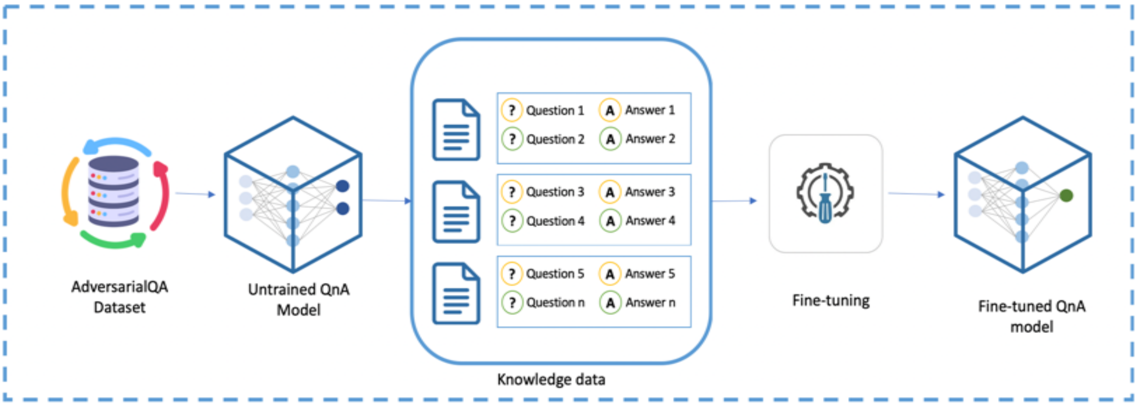

Fig 1. The flowchart of the proposed QA system

### A. *Dataset*

The data set used in the experimental application was derived from a reading comprehension data set created by Stanford University and named Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). This dataset contains over 100,000 questions generated by crowd workers working on Wikipedia articles. The answer to each question is a text segment from the relevant reading passage. In the study, 80% of the data was used for training and 20% for testing. Figure 2 shows a context from the dataset.

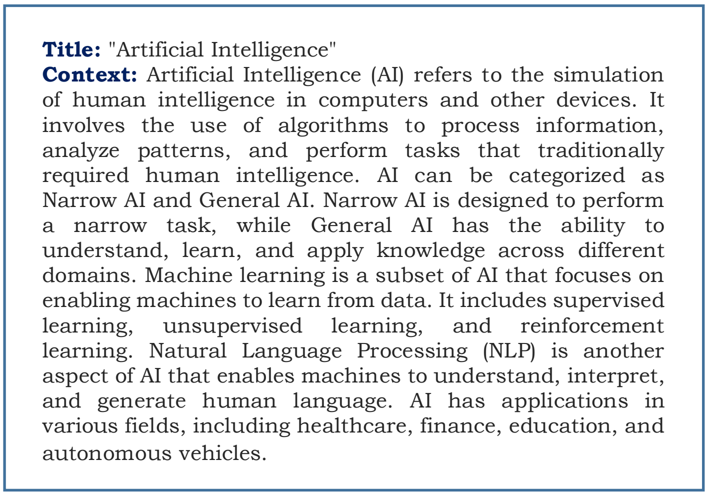
**Title:** "Artificial Intelligence"
**Context:** Artificial Intelligence (AI) refers to the simulation of human intelligence in computers and other devices. It involves the use of algorithms to process information, analyze patterns, and perform tasks that traditionally required human intelligence. AI can be categorized as Narrow AI and General AI. Narrow AI is designed to perform a narrow task, while General AI has the ability to understand, learn, and apply knowledge across different domains. Machine learning is a subset of AI that focuses on enabling machines to learn from data. It includes supervised learning, unsupervised learning, and reinforcement learning. Natural Language Processing (NLP) is another aspect of AI that enables machines to understand, interpret, and generate human language. AI has applications in various fields, including healthcare, finance, education, and autonomous vehicles.

Fig 2. A context from the dataset

### B. *Model architecture and Training process*

In the proposed QA system encompasses several key steps within its architecture and training process. After completing the data collection phase, the system initiates the question generation and data analysis steps, crafting a diverse set of representative questions that span various topics and products. Through a meticulous question generation process, the dataset is classified, and specific queries are formulated based on the categorized data. Subsequently, the system employs a Knowledge Graph, a structured model representing interconnected data, to create meaningful relationships and enhance the contextual alignment of questions [17]. This facilitates the generation of richer and contextually aligned questions [18]. In the final stages, the system focuses on answer generation by conducting similarity inferences on the Q&A data, performing popularity assessments, and executing validation operations. These processes contribute to data labeling for answer generation, ensuring the production of reliable and accurate answers. The Q&A data is then partitioned into training, testing, and validation sets for model training. Validation processes are crucial for accurately evaluating the training data, enhancing the overall robustness and effectiveness of the QA system. In last section of the study, a crucial fine-tuning process has been conducted on a customized question-answer dataset to enhance the performance of the proposed QA system. The objective of this process is to update the parameters of the pre-trained model in a dataset-specific manner, aiming to achieve a tailored and optimized QA model.

### C. *Transformer Models*

The evaluation of transformer-based models for customized question answering includes assessing the performance of various models. The transformer models utilized in the experiment comprise Bert-base-cased, Electra-base, Deberta-base, Xlm-roberta-base, Distilbert-base, and Albert-base.

Authorized licensed use limited to: TUBITAK ULAKBIM. Downloaded on July 02,2024 at 06:50:25 UTC from IEEE Xplore. Restrictions apply.

- BERT-Base (Bidirectional Encoder Representations from Transformers): BERT is a powerful bidirectional language model that incorporates an attention mechanism. With the ability to understand context from both directions, BERT has become a cornerstone in various natural language processing tasks, providing comprehensive insights into language understanding and semantics [18].
- ELECTRA-Base (Efficiently Learning an Encoder that Classifies Token Replacements Accurately): ELECTRA is distinguished by its focus on the task of token replacements. By efficiently training models through the creation of real and fake input pairs, ELECTRA achieves enhanced learning capabilities and improved performance, making it a valuable choice for general natural language processing applications [19].
- DeBERTa-Base (Denoising-enhanced BERT with disentangled attention): DeBERTa represents an advancement over BERT, incorporating features such as a

- denoising mechanism and disentangled attention. This model aims to reduce noise in the learning process and enhance attention mechanisms, contributing to superior performance in various natural language processing tasks [20].
- XLM-RoBERTa-Base (Cross-lingual Language Model - RoBERTa Base): XLM-RoBERTa is tailored for cross-lingual applications, excelling in modeling languages across diverse linguistic backgrounds. Trained on extensive multilingual data, it serves as an effective solution for tasks requiring understanding and processing in multiple languages [21].
- DistilBERT-Base (Distilled BERT): As a distilled version of BERT, DistilBERT offers a smaller and faster alternative. While having fewer parameters, it maintains the essence of the original BERT model, making it suitable for resource-constrained environments where efficiency is crucial [22].
- ALBERT-Base (A Lite BERT): ALBERT stands out as a lightweight version of BERT, featuring parameter reduction strategies for efficient learning. This model is particularly useful in scenarios demanding rapid inference and reduced computational overhead, without compromising on performance [23].

Table 1 shows the parameters of the pre-trained models.

TABLE I. THE PARAMETERS OF THE MODELS

| Models | num_attention_heads | num_hidden_layers | hidden_size | hidden_act | hidden_dropout_prob | max_position_embeddings | layer_norm_eps | intermediate_size |
|---|---|---|---|---|---|---|---|---|
| **Bert-base** | 12 | 12 | 768 | gelu | 0.1 | 512 | 1e-12 | 3072 |
| **Electra-base** | 12 | 12 | 768 | gelu | 0.1 | 512 | 1e-12 | 3072 |
| **Deberta-base** | 12 | 12 | 768 | gelu | 0.1 | 512 | 1e-7 | 3072 |
| **Xlm-roberta-base** | 12 | 12 | 768 | gelu | 0.1 | 512 | 1e-5 | 3072 |
| **Distilbert-base** | 12 | 6 | 768 | gelu | 0.1 | 512 | 1e-12 | 3072 |
| **Albert-base** | 12 | 12 | 768 | gelu | 0.1 | 512 | 1e-12 | 3072 |

## IV. EXPERIMENTAL RESULTS

In this section, we present the comprehensive findings of our experimental evaluation, focusing on the performance of transformer-based models for customized question-answering on a health-related dataset. The rigorous experimentation sheds light on the nuances of each model's capabilities, providing valuable insights into their effectiveness in addressing the intricacies of health-related inquiries. To assess the performance of the models—Bert-base-cased, Electra-base, Deberta-base, Xlm-roberta-base, Distilbert-base, and Albert-base. We conducted a thorough examination using quantitative metrics such as F1 score and exact match. The outcomes offer a nuanced understanding of each model's strengths, revealing their distinctive characteristics in the context of health-related question-answering tasks. The observed F1 scores and exact match values for the transformer-based models—Bert-base-cased, Electra-base, Deberta-base, Xlm-roberta-base, Distilbert-base, and Albert-base—provide nuanced insights into their respective performances in the context of health-related question-answering on the utilized dataset. Figure 3 shows the experimental results for each transformer model.

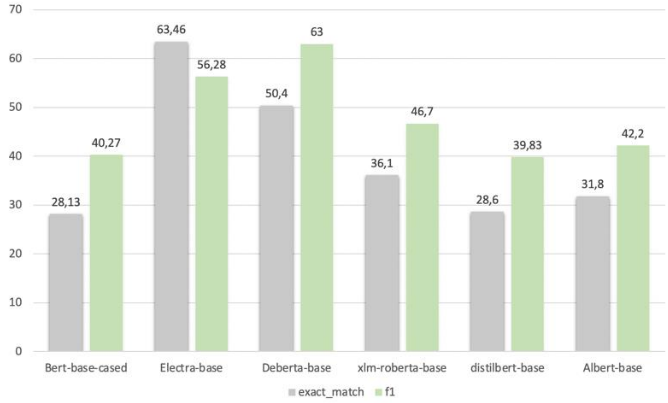

Fig 3. The experimental results for transformer models

The F1 scores and exact match values for the transformer-based models [24,25]—Bert-base-cased, Electra-base, Deberta-base, Xlm-roberta-base, Distilbert-base, and Albert-base—illuminate their performances in health-related question-answering on the dataset. Electra-base and Deberta-base stand out with impressive F1 scores, showcasing their efficacy in processing and comprehensively understanding

Authorized licensed use limited to: TUBITAK ULAKBIM. Downloaded on July 02,2024 at 06:50:25 UTC from IEEE Xplore. Restrictions apply.

health-related inquiries. Electra-base's exceptional exact match accuracy indicates precise answer generation, while Deberta-base maintains commendable accuracy despite the complexities introduced by denoising and disentangled attention. Bert-base-cased, Xlm-roberta-base, Distilbert-base, and Albert-base exhibit varying degrees of performance, emphasizing the trade-offs between model size, efficiency, and accuracy. The findings underscore the influence of model architecture, training strategies, and dataset characteristics on overall performance. Future research could explore fine-tuning tailored for health datasets, advancing question-answering capabilities in the medical domain. These insights benefit practitioners and researchers navigating the deployment of transformer-based models in question-answering systems, emphasizing considerations for specific domain requirements and challenges.

## V. Conclusion

This study delved into the optimization of transformer-based models for health-related question-answering, utilizing a customized dataset derived from Wikipedia articles. The diverse set of transformer models—Bert-base-cased, Electra-base, Deberta-base, Xlm-roberta-base, Distilbert-base, and Albert-base—were evaluated, revealing varying performances in terms of F1 scores and exact match accuracy. The experimental results highlight the importance of selecting appropriate transformer models tailored to specific domain requirements. Electra-base and Deberta-base exhibited good performance, emphasizing the potential benefits of models designed with denoising mechanisms and disentangled attention.

Future directions in research could explore further optimizations for health datasets, considering the unique challenges posed by medical information extraction and question-answering. Fine-tuning strategies specific to health-related contexts may contribute to enhanced model performance and accuracy. This study contributes valuable insights for practitioners and researchers in the field of natural language processing, particularly those deploying question-answering systems in the health domain. The findings underscore the significance of thoughtful model selection and customization for optimal performance, paving the way for advancements in transformer-based question-answering systems.

## Acknowledgment

This work is supported by the Republic of Turkey, Ministry of Science, Technology and Industry project named "AI Based Smart Digital Assistant Customer Dialog Bot project" and project code AR-22-087-0001. It is funded by R&D project within the scope of law 5746 by the Arçelik Digital Transformation, Big Data and Artificial Intelligence R&D Center.

## References

[1] R. H. AlMahmoud and M. Alian, “The effect of clustering algorithms on question answering,” Expert Systems with Applications, vol. 243, p. 122959, Jun. 2024, doi: 10.1016/j.eswa.2023.122959.

[2] T. Huai, S. Yang, J. Zhang, J. Zhao, and L. He, “Debiased Visual Question Answering via the perspective of question types,” Pattern Recognition Letters, Jan. 2024, doi: 10.1016/j.patrec.2024.01.009.

[3] Y. Sun, Z. Zhu, Z. Zuo, K. Li, S. Gong, and J. Qi, “DSAMR: Dual-Stream Attention Multi-hop Reasoning for knowledge-based visual question answering,” Expert Systems with Applications, vol. 245, p. 123092, Jul. 2024, doi: 10.1016/j.eswa.2023.123092.

[4] A. Ben Abacha and P. Zweigenbaum, “MEANS: A medical question-answering system combining NLP techniques and semantic Web technologies,” Information Processing & Management, vol. 51, no. 5, pp. 570–594, Sep. 2015, doi: 10.1016/j.ipm.2015.04.006.

[5] T. T. Aurpa, R. K. Rifat, M. S. Ahmed, Md. M. Anwar, and A. B. M. S. Ali, “Reading comprehension based question answering system in Bangla language with transformer-based learning,” Heliyon, vol. 8, no. 10, p. e11052, Oct. 2022, doi: 10.1016/j.heliyon.2022.e11052.

[6] L. Zhu, L. Peng, W. Zhou, and J. Yang, “Dual-decoder transformer network for answer grounding in visual question answering,” Pattern Recognition Letters, vol. 171, pp. 53–60, Jul. 2023, doi: 10.1016/j.patrec.2023.04.003.

[7] J. F. Ruma, T. T. Mayeesha, and R. M. Rahman, “Transformer based Answer-Aware Bengali Question Generation,” International Journal of Cognitive Computing in Engineering, vol. 4, pp. 314–326, Jun. 2023, doi: 10.1016/j.ijcce.2023.09.003.

[8] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, 2017, pp. 5998-6008.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv, May 24, 2019, doi: 10.48550/arXiv.1810.04805.

[10] A. Radford, K. Narasimhan, T. Salimans, "Improving language understanding by generative pre-training," BibSonomy. Accessed: Jan. 14, 2024. [Online]. Available: https://www.bibsonomy.org/bibtex/273ced32c0d4588eb95b6986dc2c8147c/jonaskaiser

[11] T. B. Brown et al., "Language models are few-shot learners," arXiv preprint arXiv:2005.14165, 2020.

[12] G. Tsatsaronis et al., "An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition," BMC Bioinformatics, vol. 16, no. 1, p. 138, Apr. 2015, doi: 10.1186/s12859-015-0564-6.

[13] R. Morante, M. Krallinger, A. Valencia, W. Daelemans, "Machine reading biomedical texts about Alzheimer's disease," Proceedings of the BioNLP Shared Task 2012 Workshop, 2012, pp. 26-34.

[14] Y. Zhang, W. Zhang, W. Che, T. Liu, Z. Chen, "ERNIE: Enhanced representation through knowledge integration," arXiv preprint arXiv:1904.09223, 2020.

[15] Y. Li, J. Shen, J. Liu, M. Sun, "Learning entity and relation embeddings for knowledge graph completion," Proceedings of the AAAI conference on artificial intelligence, 2019, pp. 5902-5909.

[16] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, "SQuAD: 100,000+ Questions for Machine Comprehension of Text." arXiv, Oct. 10, 2016. doi: 10.48550/arXiv.1606.05250.

[17] J. Yang et al., "BERT and hierarchical cross attention-based question answering over bridge inspection knowledge graph," Expert Systems with Applications, vol. 233, p. 120896, Dec. 2023, doi: 10.1016/j.eswa.2023.120896.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.” arXiv, May 24, 2019. doi: 10.48550/arXiv.1810.04805.

[19] K. Clark, M.T. Luong, Q.V. Le, and C.D. Manning, “Electra: Pre-training text encoders as discriminators rather than generators,” arXiv preprint arXiv:2003.10555, 2020.

[20] P. He, X. Liu, J. Gao, and W. Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv:2006.03654, 2020.

[21] A. Conneau et al., "Unsupervised cross-lingual representation learning at scale," ArXiv, 2020. https://doi.org/10.48550/arXiv.1911.02116

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, “DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter.” arXiv, Feb. 29, 2020. doi: 10.48550/arXiv.1910.01108.

[23] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, “ALBERT: A Lite BERT for Self-supervised Learning of Language Representations.” arXiv, Feb. 08, 2020. doi: 10.48550/arXiv.1909.11942.

Authorized licensed use limited to: TUBITAK ULAKBIM. Downloaded on July 02,2024 at 06:50:25 UTC from IEEE Xplore. Restrictions apply.

[24] P. Savci and B. Das, “Comparison of pre-trained language models in terms of carbon emissions, time and accuracy in multi-label text classification using AutoML,” Heliyon, vol. 9, no. 5, p. e15670, May 2023, doi: 10.1016/j.heliyon.2023.e15670.

[25] P. Savci and B. Das, “Prediction of the customers’ interests using sentiment analysis in e-commerce data for comparison of Arabic, English, and Turkish languages,” Journal of King Saud University - Computer and Information Sciences, vol. 35, no. 3, pp. 227–237, Mar. 2023, doi: 10.1016/j.jksuci.2023.02.017.

Authorized licensed use limited to: TUBITAK ULAKBIM. Downloaded on July 02,2024 at 06:50:25 UTC from IEEE Xplore. Restrictions apply.