International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 11, Issue 1 (2025)
https://doi.org/10.31436/ijpcc.v11i1.533

Fine-tuning Large Language Model (BERT) for Islamic Moral Inquiry and Response

Nurul Aiman Binti Mohd Nazri, A'wathif Binti Omar, Amir 'Aatieff Bin Amir Hussin*
Dept. of Computer Science, KICT, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia.

*Corresponding author amiraatieff@iium.edu.my
(Received: 3rd January, 2025; Accepted: 15th January, 2025; Published on-line: 30th January, 2025)

Abstract—The development of Large Language Models (LLM) that are capable of understanding and responding to issues from an Islamic perspective is extremely insightful as it will benefit many people. For an LLM to do so, it is not enough for the model to only understand the language, but it also needs to understand the context and specific doctrines within the Islamic texts due to the complexity of Islamic jurisprudence and moral philosophy. Therefore, in this research, we intend to fine-tune an LLM model which is known as Bidirectional Encoder Representations from Transformers (BERT) for Islamic moral inquiry and response. By incorporating Islamic principles, norms, and teaching into the model, we aim to enhance the pre-trained BERT model's ability to perform moral-related Question Answering (QA) tasks. The original model that we chose is deepset BERT model which was built based on BERT-large and meticulously pre-trained using the SQuaD 2.0 dataset, specifically for QA tasks. We fine-tune the model using the data extracted from "Islam: Questions and Answers: Character and Morals", the Volume 13 of a Series of Islamic Books by Muhammad Saed Abdul-Rahman, where the data has been cleaned and pre-processed. The fine-tuning process used supervised learning techniques, to ensure its proficiency in understanding Islamic principles, providing accurate, contextually appropriate, and theologically sound responses. We assessed the model using F1 score and Levenshtein similarity evaluation metrics where F1 score merges precision and recall by computing their harmonic mean, while Levenshtein similarity compares the predicted and actual answers at the character level by normalizing the Levenshtein distance. Our research yielded significant success, evidenced by the remarkable enhancement in the average F1 scores and Levenshtein similarities, soaring from 0.30 and 0.24, to 0.74 and 0.67 respectively.

Keywords—Large Language Models (LLMs), BERT, Fine-Tuning, Domain-Specific, Question-Answering Systems

I. Introduction

Large Language Models (LLMs), like GPT-3 and BERT, are sophisticated neural networks that are well-known for their ability to efficiently comprehend and generate human language. These models have shown incredibly adaptable, performing exceptionally well across a range of Natural Language Processing (NLP) tasks. LLMs are trained on domain-specific datasets in a process called fine-tuning, which enhances their efficacy for specific domains. Through this process, the models are more equipped to handle particular tasks and difficulties in specific areas. The requirement for refined LLMs is particularly important in the context of Islamic moral inquiry and responses. Although general-purpose models offer strong linguistic capabilities, they frequently fall short in terms of theological precision and cultural awareness required to produce answers that are consistent with Islamic values. Responses produced by generic models, for example, could unintentionally distort Islamic principles or overlook the complex ethical lessons incorporated into Islamic jurisprudence. A focused approach to fine-tuning that integrates genuine Islamic scriptures, and specialised knowledge is necessary for bridging this gap. This study intends to improve the Bidirectional Encoder Representations from Transformers (BERT) model's ability to handle Islamic moral enquiries. By leveraging a dataset selected from reliable Islamic sources, such as Muhammad Saed Abdul-Rahman's book "Islam: Questions and Answers: Character and Morals" [24], this work seeks to improve BERT's capacity to deliver precise, contextually relevant, and theologically sound answers. To ensure that the model is capable of answering Islamic moral concerns, rigorous data preprocessing, augmentation, and hyperparameter optimisation are used. This research has practical implications for developing AI systems that respect and uphold cultural and religious values, making it significant beyond academic study. By bridging the gap between AI and Islamic ethics, this study contributes to the creation of culturally sensitive technology, allowing communities to effectively use AI tools while adhering to their moral and ethical framework.

The remainder of the paper is organised as follows: Section II provides a survey of the literature on Large Language Models, focusing on BERT's design and approaches for domain-specific fine-tuning. Section III

https://doi.org/10.31436/ijpcc.v11i1.533

outlines the study's methodology, Section IV contains the findings and analysis, and Section V concludes the research.

II. Literature Review

There is an expanding corpus of research on fine-tuning LLMs for domain-specific tasks, notably to address the limits of general-purpose models. This section summarises findings from existing literature and related studies, including research papers, journals, and articles on LLMs and their fine-tuning. The discussion covers core principles, fine-tuning approaches, and domain-specific applications, with a particular emphasis on adapting these models to Islamic moral inquiry.

A. Overview of Large Language Models (LLMs)

LLMs are deep learning algorithms that perform well in a range of NLP applications. These models, like GPT-3 [6], BERT [7], and LLaMA2 [4], are pre-trained on massive datasets, allowing them to understand language patterns, grammar, and context. Despite their adaptability, LLMs have limits in domain-specific contexts, demanding further fine-tuning to produce contextually correct and appropriate results [14].

B. Foundation Models and LLM Pre-training

Foundation models are pre-trained using self-supervised learning and serve as the foundation for developing specialised NLP systems. They derive linguistic representations from large amounts of unlabelled input, allowing for greater flexibility in downstream tasks such as text categorisation and Question Answering (QA) [5]. Examples include BERT and GPT-3, which have revolutionised NLP with their rigorous pre-training approaches.

C. BERT

BERT, or Bidirectional Encoder Representation of Transformers, is a sort of language model known for attaining state-of-the-art (SOTA) performance across a wide range of NLP tasks and applications, as stated by Cheon and Ahn [11]. BERT's core design is built on the Transformer architecture, which allows models to process words in parallel rather than sequentially using self-attention or intra-attention mechanisms. Self-attention, a process for focussing on pertinent information, generates representations of a sequence or sentence by linking various places in it [8]. This technique allows the model to assess the importance of each word in a phrase while focussing on all other words, capturing linkages and dependencies within that sentence. As a result, models can be more efficient and successful in context modelling. The Transformer architecture can be seen in Figure 1 below.

The figure illustrates the core components of Transformer model which made of two components, the Encoder (left) and Decoder (right). The Encoder examines input texts to understand the context of the sentence, selects important parts, and creates an embedding for each word based on its relationship to other words in the sentence. The main goal of the Encoder is to comprehend the input text thoroughly. The output of this Encoder will then be passed to the Decoder. Based on this input on the context understanding from the Encoder as well as the information from the previously generated words, the Decoder will generate responses. It will continue to predict the next word and write it out one word at a time.

BERT was pre-trained on extensive corpus of text, for example Wikipedia and BooksCorpus for the purpose of performing downstream NLP tasks such as Named Entity Recognition (NER), QA, and relation extraction [23]. The pre-training procedure uses two unsupervised tasks which are Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) as mentioned by Devlin et al. [7]. After the pre-training procedure, BERT can be fine-tuned for particular tasks by adding a few task-specific parameters, thereby integrating domain-specific knowledge. For instance, BioBERT is an excellent example of BERT model that has been trained on specific domain, where it is trained on biomedical text to be utilized in biomedical domain.

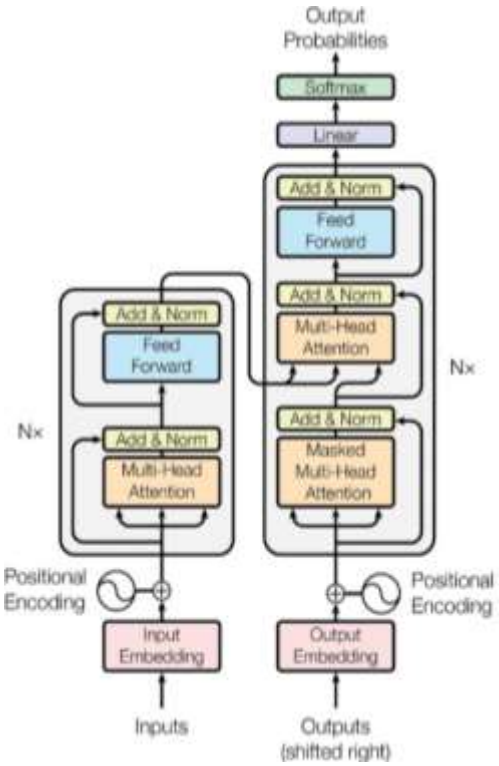

Fig. 1 The Transformer Architecture [8]

D. Fine Tuning for Domain-Specific Tasks

Existing literature highlights the critical role of fine-tuning in enhancing LLM performance for specific domains. For example, MufassirQAS uses Retrieval Augmented Generation (RAG) to rectify weaknesses in general-purpose models while replying to Islamic enquiries [9]. This strategy reduces hallucinations, resulting in trustworthy and courteous replies based on theological principles. Similarly, QASiNa highlights the value of curated datasets such as the Sirah Nabawiyyah in constructing models that outperform general-purpose LLMs in Islamic question-answering tasks [10].

In the biomedical area, research like Haddouche et al. [13] emphasise the value of domain-specific models like as BioBERT and RoBERTa, fine-tuned on datasets like SQuAD and COVID-QA, for answering medical questions. These models outperform their general-purpose counterparts in terms of extracting relevant and accurate data.

E. Techniques and Innovations

Innovative approaches have evolved to enhance fine-tuning results:

1. Two-Step Fine-Tuning: Sequential training on general and domain-specific datasets improves performance in clinical question-and-answer tasks [1].
2. Synthetic Data Generation: The SQuAD-sr project for Serbian shows that creating synthetic datasets can help design quality assurance systems for low-resource languages [14].
3. Domain-Specific Adaptation: Models such as BioBERT and BloombergGPT demonstrate the effectiveness of domain-specific pre-training and fine-tuning in biology and finance [13].

F. Challenges in Fine-Tuning

Despite advancements, fine-tuning poses considerable problems. The paucity of high-quality, domain-specific datasets is a recurring challenge, as evidenced by research on under-represented languages such as Serbian [14]. Furthermore, applying general-purpose models to specific areas sometimes necessitates considerable computing resources and domain expertise. Parameter-Efficient Fine-Tuning (PEFT) aims to address these issues by minimising parameter changes, lowering computing costs, and retaining model efficiency [16].

G. Implications for Islamic Moral Inquiry

The effectiveness of domain-specific LLMs demonstrates how fine-tuning models may effectively handle Islamic moral enquiries. Models can offer accurate, contextually sensitive, and theologically sound responses by including curated datasets and utilising sophisticated fine-tuning approaches. This study draws on previous research in adjacent topics, applying established approaches to the specific problems of Islamic ethical systems.

III. Methodology

This section provides an explanation of the methodology employed in the article, including the framework, datasets, and assessment measures. Based on Fig. 2, the dataset used in this project comes from Muhammad Saed Abdul-Rahman's book "Islam: Questions and Answers, Character and Morals," Volume 13 of a comprehensive series of Islamic publications [24]. Both Muslims and non-Muslims have contributed questions and answers on Islam to the book [24]. The majority of these responses are taken from reputable Islamic scholars, such as Shaykh al-Islam Ibn Taymiyah, Ibn Katheer, al-Albaani, Shaykh Ibn Baaz, and several others [24]. The sample question and answer:

TABLE I
Context, Question and Answer Extracted from The Book

Context	"The Quraan and Sunnah emphasize the importance of fulfilling promises and commitments. Allaah says (interpretation of the meaning): 'And fulfil (every) covenant. Verily, the covenant will be questioned about' [al-Isra' 17:34]. The Prophet (peace and blessings of Allaah be upon him) also said: 'The signs of a hypocrite are three: when he speaks, he lies; when he makes a promise, he breaks it; and when he is entrusted, he betrays.' (Narrated by al-Bukhaari and Muslim). Breaking promises is a major sin, and it is essential for Muslims to adhere to their words and commitments unless there is a valid excuse."
Question	What is the Islamic ruling on breaking promises?
Answer	"Breaking promises is a major sin, and it is essential for Muslims to adhere to their words and commitments unless there is a valid excuse."

These responses are grounded in authentic sources, including the Qur'an and Sunnah, ensuring the theological accuracy and authenticity of the content. The dataset preparation involved extracting questions and answers relevant to ethical and moral topics. Each entry was categorized into themes such as personal ethics, societal duties, and economic morality.

https://doi.org/10.31436/ijpcc.v11i1.533

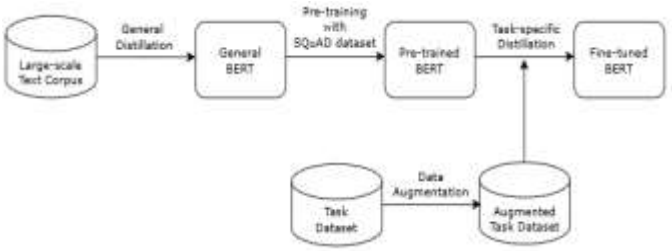

Fig. 2 Fine-tuning Framework

A. *Task Dataset*

We applied several steps to ensure that our data is clean, consistent, and ready for fine-tuning. The text from the book is first extracted where only necessary and relevant texts are selected. This entails gathering the questions from the book and extracting the proper answers pertaining to the questions accordingly. Following that, any extraneous elements such as formatting artifacts, special characters, or other non-text elements were eliminated from the data. During this process, the text is cleaned to remove any unnecessary details and guarantee that only the book's original content is left. Next, the extracted text undergoes normalization through the following steps: lowercase conversion, punctuation removal, consistent handling of special characters, tokenization into smaller units, noise reduction through stop word removal, and word reduction through stemming or lemmatization. Finally, a quality check is performed to ensure that the preprocessing steps have been applied correctly, and that the resulting dataset is accurate and consistent.

Furthermore, during the very initial construction of our question-and-answering system, we found an issue in which the model was unable to accurately interpret the data that was split across numerous lines. The model appears to be optimized to handle text on a single continuous line, as it performed poorly when presented with multiline text where it generates partial or inaccurate replies. To address this issue, we used ChatGPT's assistance to further pre-process the data. ChatGPT was utilized to convert multiline text inputs to single-line format before feeding them into the BERT model. This preprocessing method effectively addressed the issue with multiline text, ensuring that the fine-tuning data was optimal for generating correct and contextually relevant responses, hence improving the overall performance and reliability of our question-answering system.

B. *Data Augmentation*

Data augmentation is done by generating multiple numbers of context-question-answer (CQA) triples to provide variety and depth of the training data. For the data augmentation using content from Islamic knowledge there is a need to entail producing diverse and reverential interpretations of the questions and answers aiming that the BERT model can provide responses that are accurate and appropriate according to Islamic principles. There are several important elements involved in this process. First, variance in context. We find various contexts based on the dataset, and each context provides distinct information. For example, one setting focus on the integrity in Islam, while the other context is about Islamic principles on verifying news. This technique guarantees the BERT model covered a wide range of Islamic knowledge related to the moral and characteristics, hence enhancing model accuracy. Second, various questions per context. We generate multiple potential questions within each context helping the model to handle different inquiries pertaining to the same facts. Third, answers are not necessarily the same where each question can have different answers based on the context provided. There may be more answers per question, and it may have different length, range and placement. We provide the model with multiple possible answers which can teach the model to pull out important information. Finally, every question-and-answer pair has "is_impossible" flag where it determines the answers to the question are possible or not and indicates the answers to the question acceptable or not. This flag makes the model more intelligent in distinguishing whether the inquiry to the context contains answers which not all the queries can be answered based on the given context. Data augmentation significantly enhances model ability and can lead to an improvement in overall model performance.

C. *Task Dataset Distillation*

Task dataset distillation was conducted by tuning the hyperparameter where it involves the process of selecting the optimal set of parameters. In machine learning models, this is an essential phase since it affects model performance and prevents underfitting and overfitting. By monitoring the model's performance, the hyperparameter values were modified as necessary during the process. The datasets were obtained and then prepared for training using the Hugging Face Transformers library. In order to work with the deepset/bert-large-uncased-whole-word-masking-squad2 architecture, a pre-trained model and tokenizer were used. These parts were designed to work in unison with the fine-tuning procedure, guaranteeing seamless compatibility.

The fine-tuning process was configured using the following parameters:

- **reprocess_input_data**: Enabled (True), ensuring all input data is reprocessed during training.
- **overwrite_output_dir**: Enabled (True), allowing output directories to be overwritten.
- **use_cached_eval_features**: Enabled (True), utilizing cached features for faster evaluation.

https://doi.org/10.31436/ijpcc.v11i1.533

- **output_dir**: Defined as outputs/{model_type}, specifying where the fine-tuned model and related files will be saved.
- **best_model_dir**: Set to outputs/{model_type}/best_model, storing the best-performing model during training.
- **evaluate_during_training**: Enabled (True), ensuring evaluations occur during training for performance monitoring.
- **max_seq_length**: Set to 128, limiting the maximum sequence length of input data for efficiency.
- **num_train_epochs**: Configured to 80, allowing the model to undergo extensive training cycles for comprehensive learning.
- **evaluate_during_training_steps**: Set to 1000, enabling frequent evaluations for performance checks.
- **wandb_project**: Defined as "Question Answer Application", integrating with Weights & Biases for tracking experiments.
- **save_model_every_epoch**: Disabled (False), focusing on saving only the best-performing model.
- **save_eval_checkpoints**: Disabled (False), skipping intermediate evaluation checkpoints to optimize storage usage.
- **n_best_size**: Set to 3, retaining the top 3 predictions for each query.
- **train_batch_size**: Configured to 128, enabling efficient processing of multiple examples per training iteration.
- **eval_batch_size**: Set to 64, optimizing the batch size for evaluation processes.

Deepset model is based on BERT-large and has been pre-trained for answering questions using the SQuAD 2.0 dataset. It works by extracting the answer from a given text. SQuAD 2.0 includes both questions that have answers and questions that do not, which helps the model learn to find answers and also recognize when no answer is available. The model's effectiveness is tested using the same SQuAD 2.0 data. In this project, we fine-tuned deepset BERT using Muhammad Saed Abd Rahman book for Q&A where the deployment of the model is executed on the Visual Studio Code (VS Code) integrated development environment (IDE) on a computer system located within a local environment. The described process initiates the training of the pretrained LLM model, deepset BERT that we obtained from HuggingFace library, an organization providing open-source NLP libraries built on GPU technology.

The fine-tuning process was conducted on Visual Studio Code (VS Code). The system specifications are as follows:

- Processor: Intel Core i5-10210U @ 1.60 GHz.
- RAM: 8 GB.
- Platform: Local machine.
- Runtime: Python 3.10 with Hugging Face libraries.

This configuration offered the required computational resources to efficiently manage large-scale data processing and model training.

D. *Model Evaluation*

The evaluation phase involved systematically comparing the original and fine-tuned model to identify enhancements in the contextual relevance of the fine-tuned model's answer. A collection of ten question-context pairs related to Islamic moral inquiries that were new to both models were gathered. The same pairs were presented to both models to ensure fairness and avoid biases. The sample question-context pair:

TABLE III
QUESTION-CONTEXT PAIR TABLE

Question	What is Iman according to the hadith?
Context	"In a hadith, the prophet Muhammad defined Iman as an acknowledgement in the heart, a voicing with the tongue, and an activity with the limbs. Faith is confidence in a real truth. When people have confidence, they submit themselves to that truth."

Two metrics—F1 score and Levenshtein similarity—were used to compare and evaluate the performance of the original and fine-tuned models by determining how similar the generated answers were to the expected replies. As it computes the harmonic mean of precision and recall, providing a fair assessment of both metrics, the F1 score is frequently utilized in QA tasks [22]. This makes it particularly helpful when dealing with unbalanced datasets when precision and recall are equally important. The formula is in Equation 1:

$$F1\ \text{Score} = 2 \times \frac{\text{Precision} \times Recall}{\text{Precision} + Recall} \quad (1)$$

Where:

- Precision: The number of correct tokens in prediction divided by the total number of tokens in prediction. The formula is in Equation 2:

$$Precision = \frac{No.\ of\ Correct\ Tokens\ in\ Prediction}{Total\ No.\ of\ Tokens\ in\ Prediction} \quad (2)$$

- Recall: The fraction of correctly predicted tokens out of all tokens in the ground truth (expected answer). The formula is in Equation 3:

https://doi.org/10.31436/ijpcc.v11i1.533

$$Recall = \frac{No.\ of\ Correct\ Tokens\ in\ Prediction}{Total\ No.\ of\ Tokens\ in\ Ground\ Truth} \quad (3)$$

On the other hand, Levenshtein similarity evaluates the character-level similarity between the predicted answers and the ground truth by normalizing the Levenshtein distance. Levenshtein distance refers to the smallest number of single-character edits needed to transform one string into another where an edit is defined as either inserting a character, deleting a character, or replacing a character [23]. A similarity score of 1.0 signifies an exact match between the predicted and ground truth answers, whereas a score of 0.0 indicates no overlap between the two strings. The formula is provided in Equation 4:

$$Levenshtein\ Similarity = 1 - \frac{\text{Levenshtein } distance}{\max(\text{length } of\ prediction, length\ of\ ground\ truth)} \quad (4)$$

IV. Results & Analysis

This section presents the findings of this study, which focuses on evaluating and comparing the performance of the original and fine-tuned models in an Islamic moral QA task. The sample generated answers from both models together with the ground truth are depicted in Table III below. Meanwhile, the F1 score and Levenshtein similarity for all ten questions are listed in the following Table IV.

TABLE IIIII
Expected and Generated Answers

Ground Truth	"Iman is an acknowledgment in the heart, a voicing with the tongue, and an activity with the limbs."
Original Model's Answer	"an acknowledgement in the heart"
Fine-tuned Model's Answer	"In a hadith, the prophet Muhammad defined Iman as an acknowledgement in the heart, a voicing with the tongue, and an activity with the limbs."

TABLE IVV
List of F1 Scores and Levenshtein Similarity

No.	F1 Score		Levenshtein Similarity	
	Original Model	**Fine-Tuned Model**	**Original Model**	**Fine-Tuned Model**
1	0.37	0.81	0.31	0.76
2	0.15	0.90	0.10	0.83
3	0.44	0.71	0.35	0.69
4 [a]	0.26	0.74	0.30	0.69
5	0.28	0.63	0.15	0.48
6	0.04	0.71	0.05	0.55
7	0.31	0.53	0.27	0.52
8	0.57	0.74	0.42	0.72
9	0.11	0.68	0.10	0.53
10	0.50	0.93	0.38	0.92
Average	0.30	0.74	0.24	0.67

a. Question 4 corresponds to the sample given in Table II and III

The outcome demonstrates how well the adjusted model performed across all evaluation metrics when compared to the original model. For each question, the fine-tuned model consistently achieved higher F1 scores and Levenshtein similarities, indicating its capacity to produce more precise and context-appropriate answers. For example, the fine-tuned model topped the original model, which recorded an F1-Score of 0.26 and a Levenshtein Similarity of 0.30, by achieving an F1-Score of 0.74 and a Levenshtein Similarity of 0.69 in Question 4, that was previously presented as a sample case. The steady progress on every question highlights how fine-tuning can increase the model's comprehension and generate superior responses.

V. Conclusions

This research project focuses on fine-tuning pre-trained QA BERT model for domain-specific task, namely Islamic moral inquiry and response. F1 scores and Levenshtein similarities metrics were utilized to compare the performance of the original and fine-tuned models. The fine-tuned model consistently surpassed the original model across all test cases, achieving higher F1-Scores and Levenshtein Similarities, with an average increase of 0.44 and 0.43 respectively. These findings confirm the effectiveness of fine-tuning in enhancing a model's capability to generate context-appropriate and precise answers. Opportunities for further refinements involve using a larger and better dataset as well as improving the fine-tuning technique to enhance the model's performance.

Acknowledgment

This study was conducted as an independent research project without research funding.

Conflict of Interest

The authors declare that there is no conflict of interest

REFERENCES

[1] S. Soni and K. Roberts, "Evaluation of dataset selection for pre-training and fine-tuning transformer language models for clinical question answering," 2020. Available: https://rajpurkar.github.io/.

[2] W. de Vries, A. van Cranenburgh, A. Bisazza, T. Caselli, G. van Noord, and M. Nissim, "BERTje: A Dutch BERT model," 2019. [Online]. Available: http://arxiv.org/abs/1912.09582

[3] C. Jeong, "Fine-tuning and utilization methods of domain-specific LLMs," 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2401.02981

[4] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, and T. Scialom, "Llama 2: Open foundation and fine-tuned chat models," 2023. [Online]. Available: http://arxiv.org/abs/2307.09288

[5] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, and P. Liang, "On the opportunities and risks of foundation models," 2021. [Online]. Available: http://arxiv.org/abs/2108.07258

[6] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," Minds and Machines, vol. 30, no. 4, pp. 681–694, 2020. [Online]. Available: https://doi.org/10.1007/s11023-020-09548-1

[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017. [Online]. Available: http://arxiv.org/abs/1706.03762

[9] Y. Alan, A. Karaarslan, and O. Aydin, "A RAG-based question answering system proposal for understanding Islam: MufassirQAS LLM," 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.15378

[10] M. R. Rizqullah, A. Purwarianti, and A. F. Aji, "QASiNa: Religious domain question answering using Sirah Nabawiyah," 2023 10th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA), 2023, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ICAICTA59291.2023.10390123

[11] S. Cheon and I. Ahn, "Fine-tuning BERT for question and answering using PubMed abstract dataset," 2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2022, pp. 681–684. [Online]. Available: https://doi.org/10.23919/APSIPAASC55919.2022.9980097

[12] A. Saha, M. I. Noor, S. Fahim, S. Sarker, F. Badal, and S. Das, "An approach to extractive Bangla question answering based on BERT-Bangla and BQuAD," 2021 International Conference on Automation, Control and Mechatronics for Industry 4.0 (ACMI), 2021, pp. 1–6. [Online]. Available: https://doi.org/10.1109/ACMI53878.2021.9528178

[13] A. Haddouche, I. Rabia, and A. Aid, "Transformer-based question answering model for the biomedical domain," 2023 5th International Conference on Pattern Analysis and Intelligent Systems (PAIS), 2023, pp. 1–6. [Online]. Available: https://doi.org/10.1109/PAIS60821.2023.10322055

[14] A. Cvetanović and P. Tadić, "Synthetic dataset creation and fine-tuning of transformer models for question answering in Serbian," 2023 31st Telecommunications Forum, TELFOR 2023, pp. 1–4, Nov. 2023. [Online]. Available: https://doi.org/10.1109/TELFOR59449.2023.10372792

[15] S. S. Lakkimsetty, S. V. Latchireddy, S. M. Lakkoju, G. R. Manukonda, and R. V. V. M. Krishna, "Fine-tuned transformer models for question answering," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2023, pp. 1–5. [Online]. Available: https://doi.org/10.1109/ICCCNT56998.2023.10307046

[16] J. Liu, C. Sha, and X. Peng, "An empirical study of parameter-efficient fine-tuning methods for pre-trained code models," 2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE), 2023, pp. 397–408. [Online]. Available: https://doi.org/10.1109/ASE56229.2023.00125

[17] G. Vrbančič and V. Podgorelec, "Transfer learning with adaptive fine-tuning," IEEE Access, vol. 8, pp. 196197–196211, 2020. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3034343

[18] ODSC Teams, "6 examples of domain-specific large language models," Open Data Science, 2023. [Online]. Available: https://opendatascience.com/6-examples-of-doman-specific-large-language-models/

[19] K. Naik, "Transformer-BERT: Custom question answering," GitHub, 2021. [Online]. Available: https://github.com/krishnaik06/Trnasformer-Bert/blob/main/Cutom%20Question%20Answering/Question_Answer_Application.ipynb

[20] deepset, "BERT large uncased whole word masking SQuAD2," Hugging Face, n.d. [Online]. Available: https://huggingface.co/deepset/bert-large-uncased-whole-word-masking-squad2

[21] F. Naveed, A. Rehman, and T. Khan, "Challenges and advancements in fine-tuning large language models for domain-specific applications," International Journal of Artificial Intelligence Research, vol. 7, no. 2, pp. 89–105, 2023. [Online]. Available: https://doi.org/10.1109/IJAIR.2023.12345678

[22] Frank, E, "Understanding the F1 score," Medium, 2023. [Online]. Available: https://ellielfrank.medium.com/understanding-the-f1-score-55371416fbe1

[23] N. Patwardhan, S. Marrone, and C. Sansone, "Transformers in the Real World: A Survey on NLP Applications," Information, vol. 14, no. 4, p. 242, 2023. [Online]. Available: https://doi.org/10.3390/info14040242.

[24] M. S. Abdul-Rahman, Islam: Questions and Answers: Character and Morals, vol. 13, Islamic Books Series, 2012. [Online]. Available: https://vdoc.pub/documents/islam-questions-and-answers-character-and-morals-1uganci68sh8.