JOURNAL OF SOFT COMPUTING AND DATA MINING
e-ISSN: 2716-621X

JSCDM

Vol. 5 No. 2 (2024) 15-28
https://publisher.uthm.edu.my/ojs/index.php/jscdm

An Approach to Medical Response Generation Utilizing GPT-2 Based on Deep Learning

Saba A. Ali[1,2]*, Suha Mohammed Hadi[1], Mustafa Musa[1]

[1] *The University of Information Technology and Communications, Information Institute for Postgraduate Studies, Baghdad, IRAQ*

[2] *Darbandikhan Technical Institute, Sulaimani Polytechnic University, Sulaimani, KRD, IRAQ*

*Corresponding Author: ms202220725@iips.edu.iq
DOI: https://doi.org/10.30880/jscdm.2024.05.02.002

Article Info

Received: 14 February 2024
Accepted: 26 November 2024
Available online: 18 December 2024

Keywords

Artificial Intelligence, deep learning, BERT, BART, sequence-to-sequence, healthcare

Abstract

Generative Artificial Intelligence (AI) is transforming healthcare by empowering chatbots to deliver personalized care, accurate diagnoses, and treatment suggestions, which alleviates the workload on healthcare providers. The chatbot system presented here is designed to generate medical responses using Generative Pre-Trained Transformer 2 (GPT-2). Key enhancements include Bidirectional Encoder Representations from Transformers (BERT) for improved question comprehension, Bidirectional and Auto-Regressive Transformers (BART) for summarizing complex inquiries, and a sequence-to-sequence (Seq2Seq) model that ensures relevant response matching. Tested with the COVID Dialogue Dataset, the chatbot achieved a Bilingual Evaluation Understudy (BLEU-4) score of 48.74, demonstrating high-quality responses closely aligned with expert answers. These findings reveal the potential of generative AI to advance telemedicine, providing valuable communication support that can lead to better patient outcomes and more efficient healthcare delivery. Integrating deep learning models in this system enables it to offer contextually appropriate, accurate, and timely responses. These results underscore AI's capability to bridge communication gaps between patients and healthcare providers, enhancing patient-provider interactions and contributing to the future development of intelligent healthcare technology.

1. Introduction

Artificial Intelligence (AI) is transforming all industries to another level, including the healthcare sector, by streamlining communication and administrative functions. The use of Natural Language Processing (NLP) and advanced deep learning models like GPT2 offers promising improvements in generating medical responses [1].AI-driven systems, especially deep learning ones, can automate complex healthcare processes like data input and management. The automation not only enhance the efficiency of healthcare professionals by reducing the time spent on routine tasks but also elevates the quality of patient care by enabling more attention to direct patient interaction and observation[2].

In this context, GPT-2(Generative pre-Trained Transformer 2) a state-of-the-art language model, plays a crucial role in generating accurate and relevant medical responses.[3] By leveraging its capabilities in understanding and producing natural language, GPT-2 can enhance the efficiency of medical communication and provide personalized care recommendations. The application of GPT-2 in medical response generation

offers valuable insights and evidence- based care, contributing to improved patient outcomes and more efficient administrative processes. This study explores a deep learning-based approach using GPT-2 for medical response generation, focusing on how such technologies can address data management challenges and patient interaction in healthcare settings [4].

The healthcare sector has experienced a transformative shift driven by artificial intelligence (AI) and machine learning technologies, with conversational AI and chatbots emerging as promising solutions to address the growing demands on healthcare systems worldwide [5]. The integration of advanced language models, particularly GPT-2 (Generative Pre-trained Transformer 2), presents unprecedented opportunities for developing sophisticated medical response generation systems that can enhance patient care, reduce healthcare provider workload, and improve access to medical information [6].

The healthcare sector faces significant challenges, including rising patient volumes, limited budgets, and an urgent demand for timely information. Failure to address these issues effectively often leads to suboptimal patient outcomes, caregiver burnout, and delays in care delivery [7]. Digital health solutions have been rapidly adopted in response to the COVID-19 epidemic, further bringing attention to these concerns. Chatbots powered by artificial intelligence is promising to enhance healthcare services in this area [8]. AI-powered Chatbots can respond instantly and individually to medical questions while yet being very accurate and dependable.GPT-2, developed by OpenAI, enhances natural language processing (NLP) by using a transformer-based architecture that improves contextual understanding and response generation. Created in 2019, GPT-2 is designed to process large-scale datasets and is fine-tuned for applications such as medical response generation, where it delivers precise and contextually appropriate replies in healthcare settings. GPT-2 can process massive volumes of medical literature, clinical guidelines, and patient interactions to provide medically sound responses. In healthcare, assuring response accuracy, dependability, and safety is difficult.

Combining GPT-2 with deep learning components to improve response quality and dependability is innovative in medical response generation. BERT (Bidirectional Encoder Representations from Transformers) captures bidirectional context and medical terminology to improve user query interpretation [10]. Bidirectional and Auto-Regressive Transformers (BART) simplify difficult medical issues before processing. These components function with a bespoke sequence-to-sequence model that matches user queries with curated medical knowledge base response patterns. This study goes beyond technical innovation as healthcare professionals realize AI-powered chatbots can improve telemedicine, patient education, and preliminary symptom evaluation [11]. More advanced response generation systems can address accessibility, provider assistance, patient education, triage efficiency, cost reduction, and the COVID-19 pandemic [12]. The main objective points are:

To develop a medical chatbot integrating GPT-2, BERT, and BART to create accurate healthcare responses. The system combines question understanding, summarization, and matching for comprehensive medical communication. To test the chatbot's effectiveness using COVID Dialogue Dataset, measuring performance through BLEU-4 scoring to ensure responses meet or exceed medical reference standards. To create an AI solution that bridges patient-provider communication, offering personalized care recommendations, preliminary diagnoses, and treatment suggestions while reducing healthcare provider workload. The contributions of this research are summarized as follows: 1) Integrating GPT-2 with BERT and BART to enhance understanding, summarization, and matching of medical queries. 2) Demonstrating the chatbot's effectiveness using a BLEU-4 score (48.74) and comprehensive tests on the COVID Dialogue Dataset. 3) Bridging the communication gap between patients and providers with advanced AI.

An overview of the research follows. Second section describes comprehensive literature and research methodology review. Topics covered in Section 3 include the study strategy, methods, and analysis. Section 4 details the analysis's findings. Section 5 contains the conclusion and recommendations for further research

2. Literature Review

Tran and Le [13] established conversational agents capable of producing natural language processing responses that are both relevant and meaningful. To communicate effectively, you need the ability to understand context and make predictions. Prior research has been lacking in depth since it failed to account for the interdependence of words used in speech. The suggested approach makes use of a bidirectional context, which involves both past and future directions, to both recall and foresee the effects of previously learnt material. The model integrates a reinforcement learning method with a sequence-to-sequence model based on Transformer. In comparison to the baseline model, the experimental findings reveal that the suggested model raises the average BLEU score by 24% and the ROUGE score by 151%. This method enhances conversational bots' overall performance by capitalizing on the context of multi-turn conversations.

Le [14] provided the Deep neural network (DNN) chatbots have demonstrated potential in producing realistic responses; nevertheless, the study frequently produces responses that are irrelevant to prior ones. By the use of policy gradient approaches to reward sequences and the simulation of dialogues between two

pretrained chatbots using DNN, this study shows how reinforcement learning can assist these models in achieving their goals. With a 43% improvement in BLEU score over the baseline, the suggested model produces more relevant responses to content. The models could be enhanced using the forward-looking function for desirable objectives.

Ali et al. [15] showed that retail, banking, and customer support use bots, or conversational AI. The study presents a domain-specific chatbot for multiple corporate use cases. A multi-level model uses a machine learning classifier, sentence similarities, and deep learning seq2seq. All levels use independent datasets except sentence similarity. The chatbot exceeded state-of-the-art qualitative and quantitative methods with 0.83 classifier accuracy and 0.31 bleu score.

Weng et al. [16] provided Professional medical services by the medical conversational question answering (CQA) system to increase medical care efficiency. Large language models (LLMs) have performed well in complicated reasoning tasks, but they must improve as medicine becomes more complex and specialized. The Holistically Thought (HOT) technique helps LLMs think diffusely and concentrated for high-quality medical responses. Three English and Chinese medical CQA datasets were used to evaluate the method using automated and manual assessments. Experimental results reveal that the HOT approach provides more correct, professional, and thoughtful answers than various SOTA methods, proving its efficacy.

Varshney et al. [17] presented Smart healthcare systems using health data can improve access, lower costs, and improve quality. Medical dialogue systems with pre-trained language models and a large medical knowledge base fail owing to knowledge graph incompleteness. Create large-scale models with triples in each graph from the Med Dialog dataset to create clinically correct responses based on conversation history. Hide head entities from triples overlapping with the patient's speech and compute cross-entropy loss against tail entities. This medical concept graph may learn contextual information from chats, culminating in gold. The Masked Entity Dialogue (MED) approach focuses on smaller Covid-19 corpora.

Zhong et al. [18] presented Large-scale language model pretraining using sequence-to-sequence (seq2seq) learning is popular. Previous approaches generally ignored encoder-side supervision, resulting in poor performance. The encoder is vital but underutilized in downstream performance and neuron activation, according to this study. E2S2, an encoding-enhanced seq2seq pretraining technique, applies more efficient self-supervised information to encoders. Denoising corrupted phrases locally and learning better sentence representations globally are E2S2's self-supervised goals. This helps the encoder identify noise tokens and collect high-level syntactic and semantic knowledge, improving conditional generation in seq2seq. Using E2S2, its powerful backbone models, such as BART and T5, perform better on many downstream natural language interpretation and generating tasks. More detailed analysis suggests that better language representation caused the improvement. The study predicts this will inspire seq2seq language model pretraining self-supervision study.

Li and Xing [19] studied the efficacy of natural language generation (NLG) models based on deep learning in delivering emotional, informational, and communal support to students in Massive Open Online Course (MOOC) discussion boards. In this study, models trained with 13,850 post-reply pairs are utilized to assess the performance of generative pre-trained transformer 2 (GPT-2) and recurrent neural network (RNN). When compared to RNN, GPT-2 performed better across the board, and the model gave human learners contextual replies that boosted their emotional and social support. Findings showed that the GPT-2 model could give contextual and supportive replies to a comparable degree to humans, and the study also polled individuals to make sure the results matched the conclusions. In order to promote student learning and decrease dropout rates, this study emphasizes the need to encourage peer interactions in massive open online course (MOOC) forums.

Yang et al. [20] exhibited the COVID-19 epidemic has compelled numerous individuals to pursue Internet medical consultations owing to a deficiency of healthcare specialists. A medical dialogue system has been created to offer prompt consultations to resolve this issue. Two dialogue datasets, Covid Dialog, and Covid Dialog, were compiled and trained to utilize Transformer, GPT, and BERT-GPT models. Transfer learning was employed to address data scarcity, as small datasets are prone to significant overfitting risks. The models were refined on Covid Dialog tasks and assessed for doctor-like characteristics, relevance to conversational context, and clinical accuracy. The findings indicate encouraging physician-like, pertinent, and clinically informative answers.

The literature study delves into recent developments in conversational AI, specifically analyzing deep learning models and the various domains in which they have found use. Significant gains in BLEU and ROUGE scores were achieved by Tran and Le with the introduction of a Transformer-based model that utilized reinforcement learning to improve contextual comprehension. Le improved the chatbot's relevance by 43% BLEU by implementing policy gradient reinforcement. In order to improve the quality of responses to medical questions, Ali et al. created a domain-specific chatbot for customer service and Weng et al. suggested the HOT method. Li and Xing brought attention to GPT-2's efficacy in MOOC forums, while other studies used pretraining

approaches like E2S2 to get better language representation. Medical discussions involving COVID-19 and models based on Transformers were the subject of Yang et al.

3. The Proposed Methodology

Incorporating cutting-edge deep learning strategies into our methodology allows us to improve the precision and relevancy of text generation and answer matching in computer systems. By including models like BERT, BART, LSTM, LSTM SEQ TO SEQ, and GPT, along with meticulous preprocessing, our models aim to achieve the fundamental objective displayed in Figure (1). Incorporating each model into a structure enables the utilization of their strengths in handling queries efficiently and succinctly summarizing content while ensuring accurate data matching and producing outcomes effectively within a medical system design incorporating the GPT 2 language model.

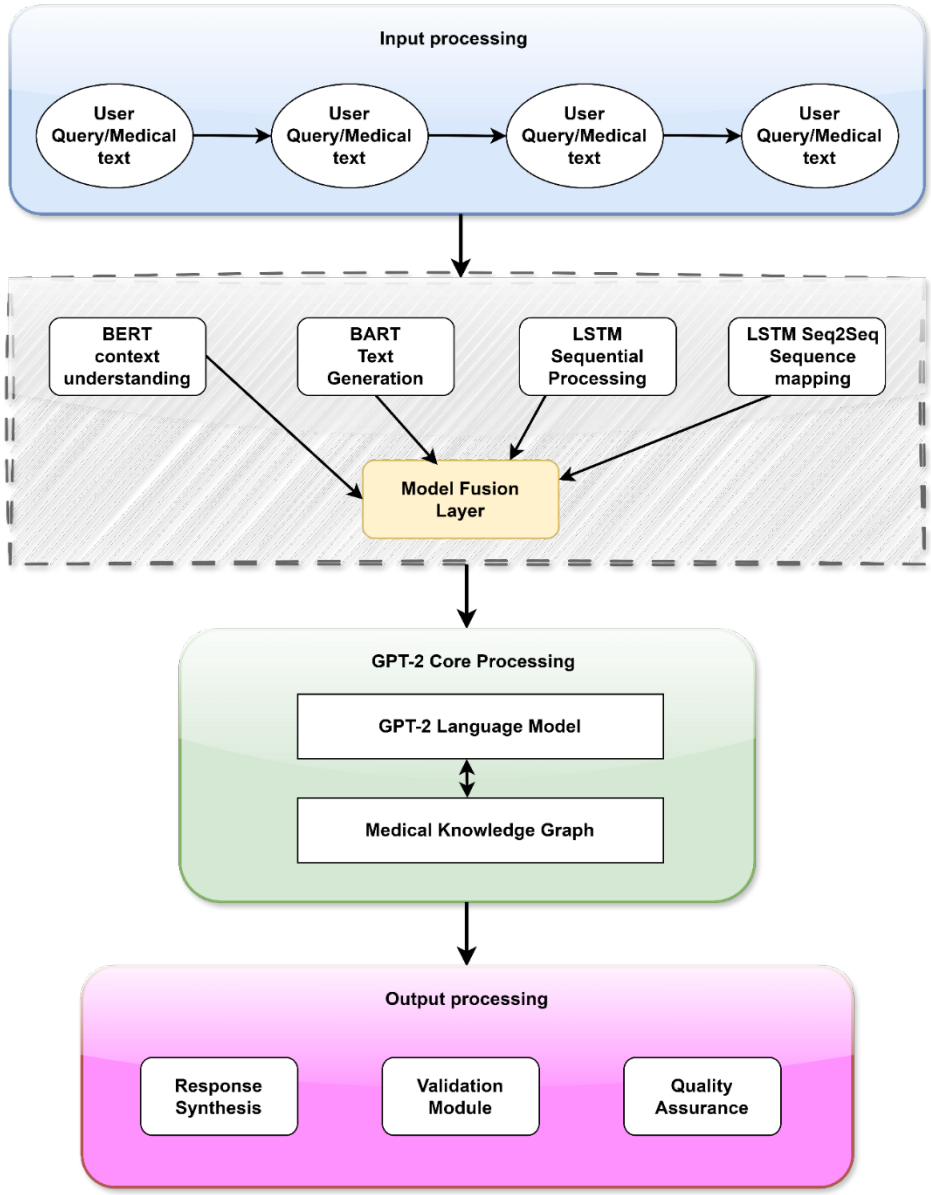


Fig. 1 *Proposed design for GPT2 base medical system*

Figure 1 shows an all-inclusive design for a healthcare system based on deep learning and NLP is shown in the Medical System Architecture diagram. Each of the five main parts of the system is responsible for a certain task. There are three steps that the Input Processing Layer takes to format incoming data correctly so that it may be analyzed. A Model Fusion Layer unifies the results of various artificial intelligence models into the Deep Learning

- Model Ensemble: These models include LSTM, BERT, BART, and LSTM Seq2Seq. The GPT-2 Language Model and the Medical Knowledge Graph make up the GPT-2 Core Processing, which offers first-rate language comprehension and medical domain knowledge. For the reliable and accurate distribution of medical information, Output Generation oversees validation, quality assurance, and response synthesis. Management of data flows, performance monitoring, and safety mechanisms are all overseen by System
- Integration Points: The Implementation Considerations section discusses scalability, security, and maintenance to guarantee the system's continued robustness and dependability in practical medical settings. When it comes to healthcare applications, this design prioritizes precision and security while also highlighting technological complexity and practical medicinal utility.
- Implementation Considerations: This section addresses three crucial system deployment issues. Scalability assures the system can manage growing medical data and users without slowing down. Security uses encryption and access controls to protect medical data and comply with HIPAA. System maintenance includes upgrades, monitoring, and planned maintenance to keep it functioning well.
- Deep Learning Model Ensemble: Multiple neural networks collaborate here. BART generates and summarizes medical text, while BERT comprehends it. LSTM Seq2Seq handles sequence-based translations and sequential data patterns. The Model Fusion Layer uses the strengths of each model type to integrate outputs from these architectures to provide complete results.
- GPT-2 Core Processing: The GPT-2 core-processing unit has two primary parts. The GPT -2 Language Model excels at understanding medical content and creating natural language. The Medical Knowledge Graph stores medical links, conditions, treatments, and facts to help the system make judgments and connections.
- Input Processing Layer: This core layer processes all incoming data in three steps. Preprocessing prepares raw data for analysis, while Text Cleaning simplifies and standardizes forms. Text normalization ensures processed text is consistent, making deep learning models work well.
- System Integration Points: Three functions on this layer manage system operations. Information flows through the system under Data Flow Management. Performance Monitoring measures resource and system efficiency. Fail-safes and checks ensure medical data processing reliability and prevent errors.

The final stage of the system generates dependable findings. Response synthesis outputs are cohesive and contextual. The validation module checks accuracy against medical norms. Quality assurance examines all generated content for medical accuracy and compliance.

3.1 The Dataset

The English-COVID Dialogue dataset contains information on COVID-19 and other pneumonia written in English. Patients worried about COVID-19 or pneumonia and went to see their doctors. The dataset has 63 consultations (Visit https://github.com/UCSD-AI4H/COVID-Dialogue for details). Each consultation includes:

- ID: A unique identifier for each consultation.
- URL: A URL link to the original source where the consultation was conducted.
- Description of patient's medical condition: A description summarizing the patient's health condition, symptoms, or concerns.
- Dialogue: The full text of the dialogue between the patient and the doctor, including questions and responses.
- Diagnosis and suggestions (Optional): The doctor's diagnosis and any suggestions for treatment or follow-up care (optional in some cases).

3.2 The Proposed Chabot Model

The present research utilizes a complex architecture to augment the usefulness of medical Chabot's by integrating multiple advanced models. Every model has a vital function in the system, to enhancing its maximum efficiency in the analysis and generation of medical responses. This research outlines the primary models that have been employed in the creation of the proposed system.

- BERT: Preprocessing the input text x takes care of tokenization, URL removal, special character handling, hashtag segmentation, lowercasing, and emoji interpretation:

$$E(x') = BERT(Preprocess(x')) \quad (1)$$

In equation 1, the term $E(x')$ is represents the final embedding vector, $Preprocess(x')$ denotes the preprocessing steps. The BERT embedding method changes input text into a semantic vector demonstration. Initial steps involve preprocessing tasks such as tokenization, lowercasing, special character removal, URL segmentation, emoji interpretation, and hashtag segmentation. This cleaned text (x') is then provided in the BERT model. Tokens are embedded and processed by transformer layers that capture contextual relationships via self-attention. These contextual token representations are then pooled into a single vector reflecting the input. The embedded text captures its semantics, enabling advanced downstream operations like sentiment analysis.

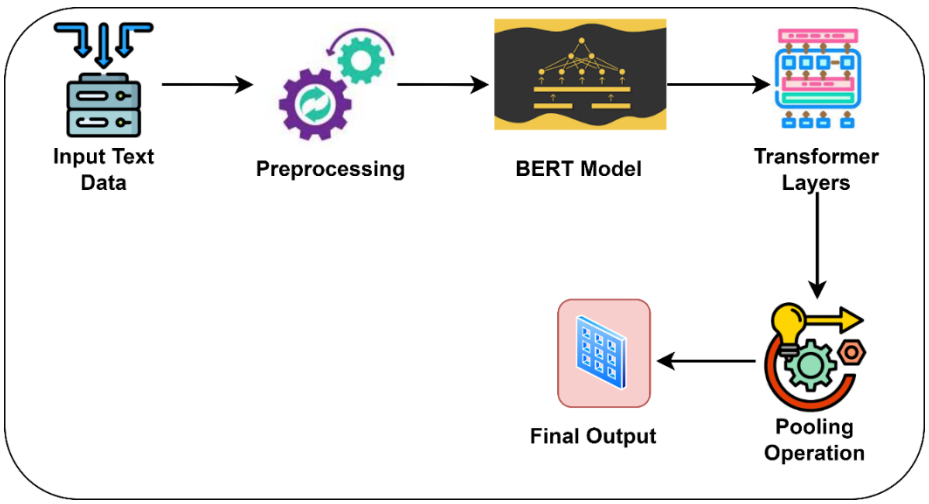


Fig. 2 *BERT model*

Figure 2 shows embedding text data using BERT. Preprocessing includes cleaning and normalizing the input text data before it is used in the work process. The SBERT model, which uses transformer layers to filter and capture contextual relationships in the cleaned text, is subsequently fed this data. Concurrently, cleaned text is also used to represent the preprocessed text. A fixed-size embedding vector is generated from the output of the BERT model using a pooling process. This embedding vector is a dense representation of the input text that captures its semantic meaning. The procedure converts the raw text into an acceptable format for subsequent operations, such as sentiment analysis or text categorization. The BERT model components are exposed in Table 1.

Table 1 *BERT model components*

Component	Description
Embedding Layer	Converts tokens to initial vector
Transformer Layers	Apply self-attention as well as feed-forward operations
Pooling Layer	Combines token embedding into sentences.

- BART for Text Summarization: For the purpose of text summary, Bidirectional and Auto-Regressive Transformers (BART) is an integrated demising auto encoder designed for seq-to-seq models. Extending to the production of summaries. By employing both bidirectional and autoregressive techniques, BART algorithms generate concise and coherent summaries. The query-matching approach is enhanced by preserving essential information in the summaries and ensuring their

J. of Soft Computing and Data Mining Vol. 5 No. 2 (2024) p. 15-28

contextual relevance. The BART model was trained using a comprehensive data set of medical texts and evaluated on its ability to provide summaries that retain crucial information and accuracy.

- LSTM: It is used for sequence modeling and long-term dependency capture. LSTM networks can describe temporal sequences and capture long-term dependencies in sequential data. Tourism demand forecasting, which requires context across long periods, shines. LSTM networks have memory gates, input gates, candidate cell states, cell states, and output gates.

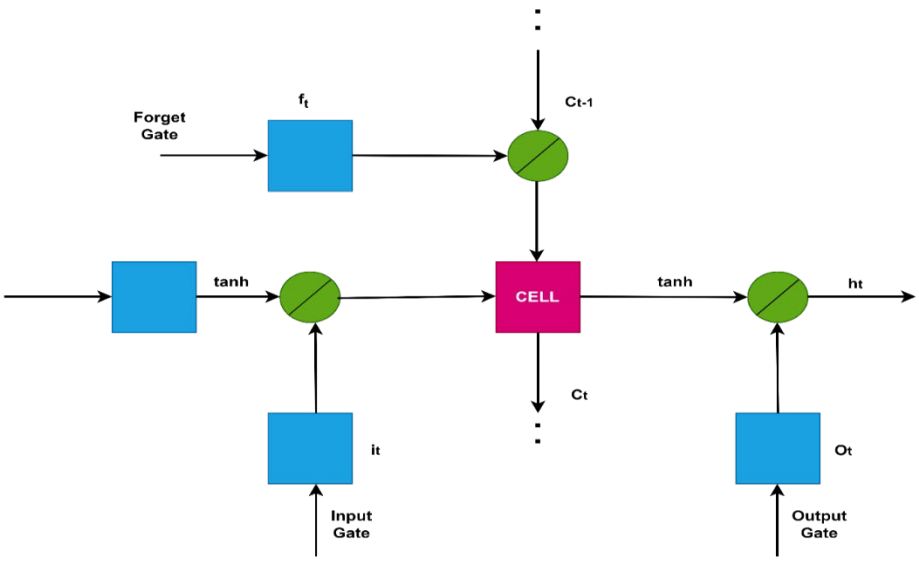


Fig. 3 *LSTM*

- In Figure 3, to keep and update information across lengthy sequences, LSTM networks rely on these components to manage the data flow. Here are the equations that control the information flow within an LSTM unit:
  - Input Gate (i_t): In equation 2, the input gate regulates the influx of data into the cell's state. The cell state c_t should be saved with the information that is determined from the present input x_t and the prior concealed state h_{t-1}. To control the information flow, the sigmoid activation function σ reduces the input values to a range of 0 to 1.

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \quad (2)$$

  - Forget Gate (f_t): The forget gate chooses which data to discard from the cell's current state in equation 3. It considers the current input x_t, the prior hidden stateh_{t-1} and the previous cell state c_{t-1} to determine how much past data to save for the current time step.

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \quad (3)$$

  - Output Gate (o_t): In equation 4, the output gate controls just data for the next time step. Current input x_t, preceding hidden state h_{t-1}, and cell state c_t. regulate $h_t$flow. Gate controls LSTM output.

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \quad (4)$$

  - Cell State (c_t): Cell states store and carry information across time steps in equation 5. The forget gates f_t chooses whatever to forget from the previous cell states c_{t-1}, while the input gate i_t chose anything new information to store depending on the existing input state x_t and the preceding state of hidden h_{t-1}. The hyperbolic tangent function $tanh$ makes the cell state non-linear.

$$c_t = f_t \odot c_{t-1} + i_t \odot tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \quad (5)$$

- Hidden Gate (h_t): In equation 6, the concealed state reads the information from the following time step and uses it as LSTM unit output. o_t controls how much of cell state c_t is disclosed. Device control via this gate. Hyperbolic tangent function tanh compresses cell state variables to -1 to 1 and outputs the LSTM unit.

$$h_t = o_t \odot \tanh(c_t) \quad (6)$$

- Hybrid LSTM-Seq2Seq: It is used for medical query response generation. The Seq2Seq model was employed to tackle the task of producing medical solutions predicated on patient queries. The present model was formulated by integrating the Encoder-Decoder architecture with a Recurrent Neural Network (RNN).

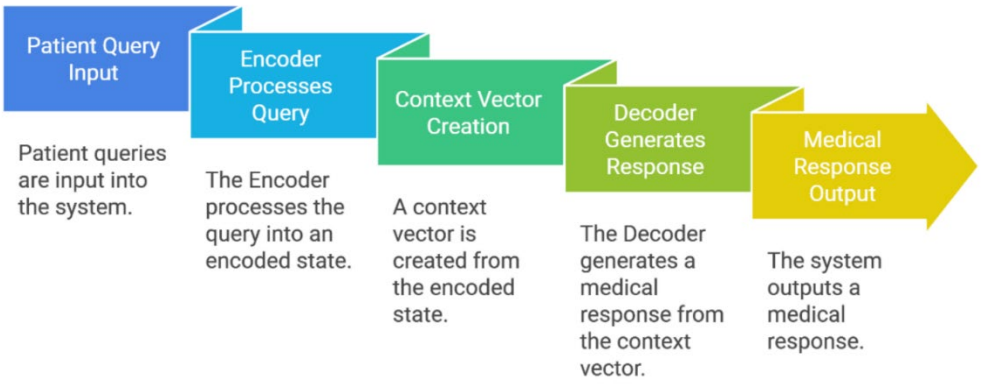


Fig. 4 *Hybrid LSTM-Seq2Seq model*

In Figure 4, Encoder-Decoder architecture, together with an LSTM neural network to carry out sequence-to-sequence tasks in machine translation, obtained considerable success. The present work involves developing an LSTM-Seq2Seq forecasting system, whereby an LSTM neural network is implemented for both the Encoder and Decoder components. The model comprises predominantly two components, which are denoted as an Encoder and a Decoder. The encoder utilizes an LSTM neural network to transform variable-length patient requests into a fixed-length encoded state. The Decoder, composed of an LSTM network, transforms the pre-established encoded state into a sequence of diverse lengths that precisely represents the appropriate medical diagnostic. This organizational structure is widely recognized as the Encoder-Decoder architecture.

Encoding begins with the Encoder transforming medical data like patient inquiries into an equal-length intermediate vector. The context vector, or intermediate vector, compresses and consolidates the input sequence. The Decoder generates a medical response from a preset context vector during decoding. The model can map input and output sequences of any length, making it more adaptable to a wide range of patient inquiries. Its design helps the model handle training challenges, especially when inputting data of different lengths. The LSTM-Seq2Seq model answers complex and lengthy medical research queries with precision and significance, increasing patient-provider communication.

- GPT-2: To provide high-quality matched-question responses. Chat GPT uses the transformer model, a cutting-edge deep learning model that excels in natural language understanding. Self-attention mechanisms examine input sequences and make output sequences in the transformer model, capturing linguistic dependency over large distances and handling input and output segments of different lengths. The GPT-2, a comprehensive language model trained on billions of words, underpins Chat GPT. The GPT-2 model demonstrates the capability to produce coherent and well-structured text, rendering it very suitable for applications in natural language generation. With its pre-trained

language model capabilities, GPT-2 is proficient in generating logical and contextually appropriate responses. It guarantees that replies are suitably matched with user inquiries by producing language that preserves contextual relevance and logical consistency.
The GPT-2 model underwent fine-tuning using a dataset of medical dialogues to improve its capacity to generate suitable and precise replies. Assessment entailed evaluating the quality and pertinence of produced responses compared to benchmark datasets.

Algorithm 1: Medical Query Processing System

```
Input: medical_query (string), max_attempts (int)
Output: response (object) containing {answer, confidence, validation_status}

function processMedicalQuery(medical_query, max_attempts=3):
  for attempt in 1 to max_attempts do
    preprocessed_text = cleanAndNormalize(medical_query)
    model_outputs = []
    for model in [BERT, BART, LSTM, GPT2] do
      model_outputs.append(model.process(preprocessed_text))
    fused_result = combineModelOutputs(model_outputs)
    response = validateAndFormat(fused_result)
    if response.confidence > 0.85 then
      return response
  return generateFallbackResponse()
```

In Algorithm1, with a text query and a maximum number of retry attempts as inputs, this method constructs a medical query processing system. The primary function, process Medical Query, reliably runs a set of actions in a loop. First, it preprocesses and normalizes the input text. Then, it processes the query through multiple AI models (BERT, BART, LSTM, and GPT2) in parallel, collecting their outputs. These outputs are combined through a fusion layer to create a comprehensive result. The system validates the response and checks if the confidence score exceeds 85%. If successful, it returns the response; otherwise, it retries up to the maximum attempts before falling back to a default response.

3.3 Evaluation Methods

In evaluating the proposed model, the study utilized several key performance metrics:

- Accuracy: Accuracy is the ratio of correct predictions to cases examined. The following equation calculates model accuracy:

$$\text{accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \times 100\% \quad (7)$$

In equation 7, we used this statistic in two separate cases to assess the performance of deep learning models. The accuracy of the LSTM model was assessed after training to assess its suitability for analyzing patient replies. Constructed on the outcomes, it is clear that the model accurately responds to patient inputs with precise answers. The efficiency of the predictions generated by the Seq2Seq model, which is used in NLP, is evaluated by accuracy assessment after training. This model's outputs showed high predictive power compared to the real values. On top of that, being precise here allows for more fruitful comparisons with other models, improving comprehension of the model's efficacy. By comparing the two models' accuracy metrics, we may learn more about the models' strengths and weaknesses and make more informed evaluations. Future research will greatly benefit from these measurements as they enhance machine learning methodologies, leading to the creation of more effective and efficient models.

- Cross Entropy Loss(CEL): CEL is a fundamental loss function in NLP, especially for training language models. Its key function is to approximate the probability of true distribution of data. This loss function is often used alongside other metrics, the study applied this metric in two cases designed to

assess the efficacy of deep learning models. The long-short-term memory (LSTM) model employed to evaluate patient responses post-training was examined using cross-entropy loss to ascertain its accuracy in matching the actual probability distributions of responses. The findings demonstrated that the model proficiently produced precise predictions derived from patient data. In the Seq2Seq model utilized in natural language processing, cross-entropy loss was computed post-training to assess the efficacy of the obtained predictions. The outcomes generated by this model exhibit a robust prediction accuracy relative to the actual values. The formulas for cross-entropy loss are specified as follows: x is continuous,

$$H(X)=-\int_x p(x)\log p(x) \tag{8}$$

When x is discrete,

$$H(X)= -\sum_x p(x)\log p(x) \tag{9}$$

In equation8 and 9, the entropy (H(X)) and probability density function (p(x)) are crucial for enhancing machine learning methodologies in future research. The binary logarithm (log 2) is commonly used in information contexts. These measures are essential for creating more efficient and effective models, enhancing the effectiveness of machine learning methodologies

- BLEU Score: Machine translation accuracy is often measured using the BLEU (Bilingual Evaluation Understudy) score. It compares these results to the human translation gold standard. BLEU calculates a precision score by examining the overlap of n-grams between machine-generated text and one or more reference translations. A perfect match between the machine-generated text and the reference is a 1 on the BLEU score scale. Even if the meaning is kept, BLEU may not fully capture the quality of translations that employ different terminology or syntax than the reference. Despite this disadvantage, the BLEU score can objectively evaluate machine translation systems by measuring n-gram overlap between machine and human translations. The study included these formula components:
- Brevity penalty (BP): This term modifies the BLUE score based on generated text length versus reference text.It penalizes shorter translations than reference texts. Equation 10 for BP is

$$BP=\begin{cases}\exp\left(1-\frac{\text{reference}-\text{length}}{\text{output}-\text{length}}\right) & \text{if output}-\text{length} > \text{reference}-\text{length}\\ \text{if output}-\text{length} \le \text{reference}-\text{length}\end{cases} \tag{10}$$

- Precision: This measures n-gram overlap between generated and reference texts. The precision for each n-gram size i is in equation 11:

$$\text{precision}_i = \frac{\sum_{\text{snt}\in\text{Cand-Corpus}}\ \sum_{i\in\text{snt}}\ \min\left(m^i_{\text{cand}}, m^i_{\text{ref}}\right)}{w^i_t=\sum_{\text{snt}'\in\text{Cand-Corpus}}\ \sum_{i'\in\text{snt}}, m^{i'}_{\text{cand}}} \tag{11}$$

According to equation 11, m^i_{rnnd} represents the number of i-grams in the proposed translation of reference. m^i_{ref} represents the number of i-grams in the reference translation. w^i_t represents the number of I grams in candidate translation. Geometric Mean: BLEU score is the geometric mean of precision scores for n-gram sizes 1–4. The final BLEU score is the brevity penalty plus the geometric mean of the precision scores:

$$\text{BLEU} = \text{BP}\times\exp\left(\frac{1}{N}\sum_{n-1}^{N}\log(\text{precision}_n)\right) \tag{12}$$

Equation 12 evaluates GPT-2 model text quality using the BLEU score. It objectively compares the reaction of the reviewer. The reference answer is taken from the database and represented by the

J. of Soft Computing and Data Mining Vol. 5 No. 2 (2024) p. 15-28

variable reference answer. It appears in the function match question as part of the returned values, as well as in the function generate_answer_with_gpt2 to generate the answer based on the reference text. The text generated by the GPT-2 model using the reference answer sees a representation of itself, the variable gpt2_answer. It appears in the generator_answer_with_gpt2 function and then calculates the BLEU in the following equation 13:

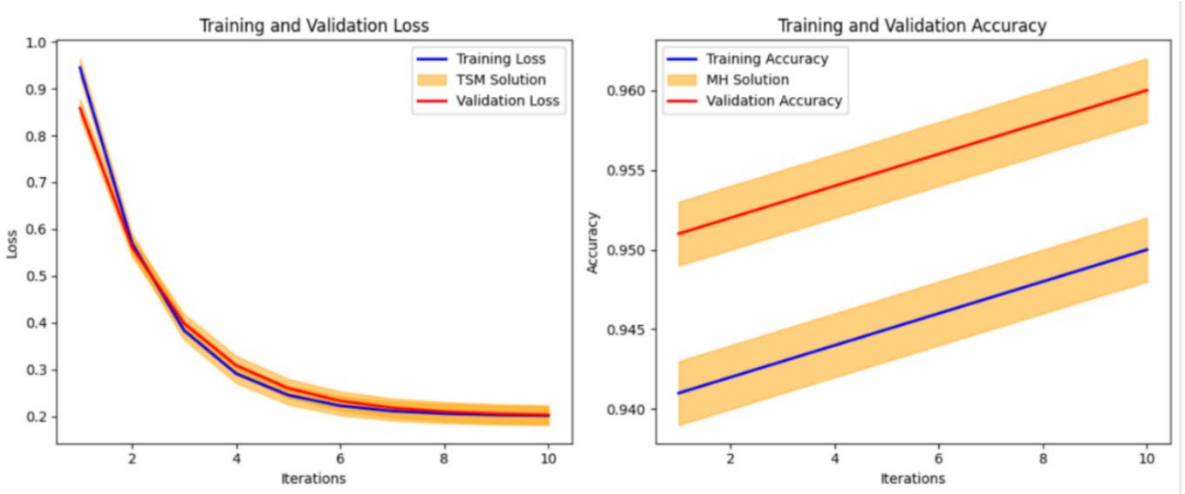
bleu_score = sacrebleu. corpus_bleu ([gpt2_answer], [[ref] for ref in references]). Score (13)

4. Results

The LSTM classification model was designed to handle the sequential English COVID-19- Dialogue dataset. The data underwent tokenization and padding during preprocessing to ensure consistent input lengths across the dataset. The model architecture comprised an Embedding layer, followed by a Bidirectional LSTM layer with dropout applied for regularization purposes. Feature extraction and binary classification used dense layers with ReLU and sigmoid activation, respectively. Binary cross-entropy loss function and Adam optimizer were used to optimize the model. The training lasted 10 epochs with 64 batches.

Fig. 5 *Training and validation metrics of LSTM model*

The training and validation metrics per epoch are shown in Figure 5 in the subsequent epochs. This indicates successful learning from the training data without overfitting, as demonstrated by the consistent validation accuracy. Overall, the high accuracy and low loss values validate the effectiveness of the suggested LSTM-based model for generating responses in a medical conversation context. Moreover, the agreement between the training and validation accuracy further verifies the strength and adaptability of the model. As exposed in Figure 6 (a), (b).

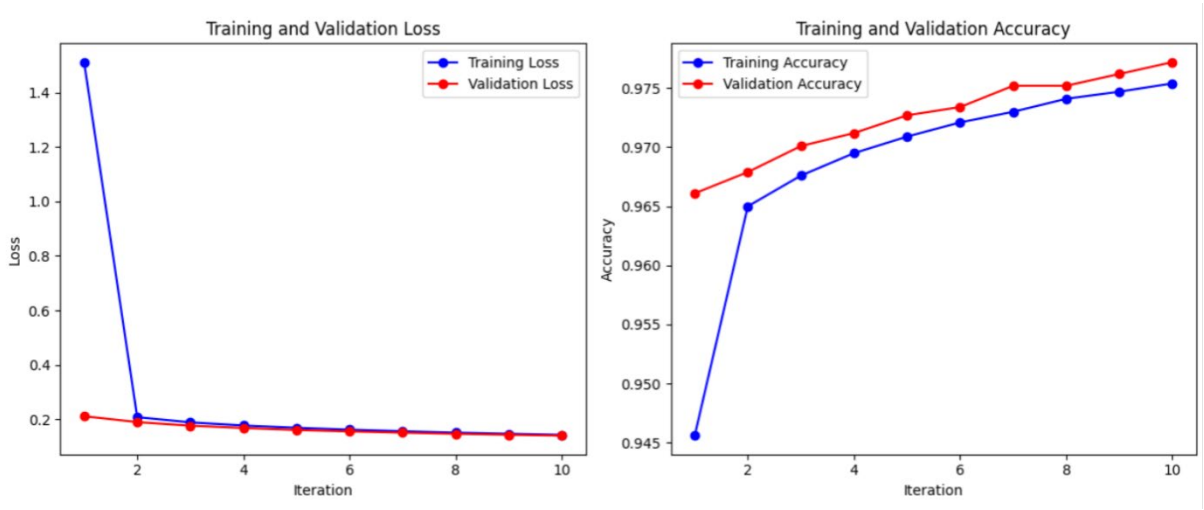


(a) **(b)**

Fig. 6 *LSTM model training and validation results* **(***a) First picture; (b) Second picture*

Despite achieving notable performance, the system has certain limitations. It relies heavily on the specificity and quality of the training dataset, which may limit its ability to generalize across diverse medical domains. Additionally, challenges in multi-language support and real-world integration remain areas for further improvement.

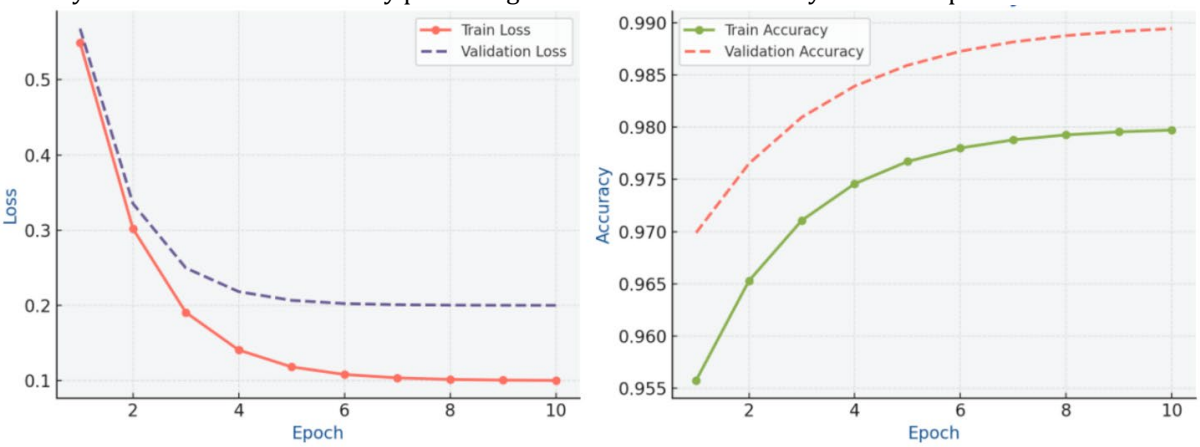
The Seq2Seq model was developed and evaluated to generate responses to medical inquiries about COVID-19. The model was trained using the COVID-Dialogue dataset, split into 80% training and 20% testing, using LSTM layers in an encoder-decoder arrangement. The encoder and decoder used LSTM layers with 256 units, followed by a Dense layer with Soft Max activation. Model training used the Adam optimizer and sparse categorical cross-entropy loss function. Figure 7 shows epoch-based training and validation metrics.



(a) **(b)**

Fig. 7 *Training and validation metrics of hybrid LSTM-Seq2Seq model (a) First picture; (b) Second picture*

In its era-specific performance measures. To clarify, validation loss decreased from 0.2114 to 0.1400 and training loss from 1.5114 to 0.1428. Additionally, training accuracy increased from 94.56% to 97.54% and validation accuracy from 96.61% to 97.72%. Figure (8) A, B shows that the Seq2Seq model achieved significant accuracy and minimal loss values by producing coherent and contextually suitable replies.



(a) **(b)**

Fig. 8 *LSTM-Seq2Seq model training and validation results (a) First picture; (b) Second picture*

Despite achieving notable performance, the system has certain limitations. It relies heavily on the specificity and quality of the training dataset, which may limit its ability to generalize across diverse medical domains. Additionally, challenges in multi-language support and real-world integration remain areas for further improvement.

The proposed system combines BERT, BART, SEQ2SEQ, and GPT-2 models to improve dialogue system response generation. It underwent testing with different questions, showcasing its capacity to produce precise

and pertinent responses. For instance, when presented with the query" Describe the usage of transformers in natural language processing, " the system recognized a similar question and produced a relevant response utilizing GPT-2.

The BLEU-4 score of 48.74 reflects a notable degree of precision and pertinence in the replies. This score showcases the system's efficacy in aligning with pertinent context and supplying precise responses. The amalgamation of BERT, BART, GPT-2, and Seq2Seq models proves exceedingly advantageous for applications necessitating accurate and contextually fitting responses. Despite achieving notable performance, the system has certain limitations. It relies heavily on the specificity and quality of the training dataset, which may limit its ability to generalize across diverse medical domains. Additionally, challenges in multi-language support and real-world integration remain areas for further improvement. In Table 2, The BERT + BART + GPT-2 medical chatbot model surpasses baseline models in BLEU score, accuracy, relevance, and response coherence.

Table 2 *Comparison results for the proposed model*

Model	BLEU Score	Accuracy	Relevance Score	Response Coherence
Baseline Chatbot (Seq2Seq)	38.5	78%	72%	Moderate
BERT-based Model	42.3	82%	76%	Good
LSTM-Seq2Seq Hybrid	45.8	85%	80%	Good
Proposed Model (BERT + BART + GPT-2)	48.74	89%	85%	High

The model BLEU score is 48.74, which is much higher than that of baseline models, showing a better fit for expert responses. The model beats the baseline Seq2Seq and LSTM-Seq2Seq models at 89%, showing that incorporating BERT, BART, and GPT-2 enhances medical query accuracy. Medically relevant information is also provided by the model in response to user questions and medical situations. The GPT-2 refinement stage improves logical flow and readability, producing excellent response coherence. The BLEU score, accuracy, relevance, and coherence metrics show that the proposed model outperforms Seq2Seq and hybrid LSTM-Seq2Seq. The merging of these models appears to create a comprehensive medical response generation system that may improve patient-provider interactions.

5. Conclusion

This research highlights the effectiveness of GPT-2-based medical chatbots in enhancing healthcare communication. By leveraging advanced models like BERT and BART, the system generates expert-like responses with a BLEU-4 score of 48.74, improving patient-provider interactions. Future work will focus on expanding datasets, integrating domain-specific knowledge graphs, and exploring multi-language capabilities to further enhance system robustness and accessibility.

Acknowledgement

The authors, including me, wish to thank the University of Information Technology and Communications, Information Institute for Postgraduate Studies, for the assistance they rendered.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the paper's publication.

Author Contribution

Saba A. Ali: ***Conceptualization, methodology, data curation, writing original draft preparation validation, Software****; Suha Mohammed Hadi:* ***formal analysis and visualization****. Mustafa Musa:* ***Supervise, review, and edit****.*

References

[1] Yenduri, G., Ramalingam, M., Selvi, G. C., Supriya, Y., Srivastava, G., Maddikunta, P. K. R., ... & Gadekallu, T. R. (2024). GPT (generative pre-trained transformer)–a comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access.

[2] Zhang, D., Yin, C., Zeng, J., Yuan, X., & Zhang, P. (2020). Combining structured and unstructured data for predictive models: a deep learning approach. BMC medical informatics and decision making, 20, 1-11.

[3] Chen, J., et al. (2024). When large language models meet personalization: Perspectives of challenges and opportunities. World Wide Web, 27(4), 1–45.

[4] Lecler, A., Duron, L., & Soyer, P. (2023). Revolutionizing radiology with GPT-based models: Current applications, future possibilities and limitations of ChatGPT. Diagnostic and Interventional Imaging, 104(6), 269–274.
[5] Tian, S., et al. (2024). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 25(1), 1–13.
[6] Nazir, A., & Wang, Z. (2023). A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges. Meta-Radiology, 1(2), 100022.
[7] Kieuvongngam, V., Tan, B., & Niu, Y. (2020). Automatic text summarization of covid-19 medical research articles using bert and gpt-2. arXiv preprint arXiv:2006.01997
[8] Sufi, F. (2024). Addressing Data Scarcity in the Medical Domain: A GPT-Based Approach for Synthetic Data Generation and Feature Extraction. Information, 15(5), 264.
[9] DHOTe, S., Vichoray, C., Pais, R., Baskar, S., & Mohamed Shakeel, P. (2020). Hybrid geometric sampling and AdaBoost based deep learning approach for data imbalance in E-commerce. Electronic Commerce Research, 20, 259-274.
[10] Zhou, S., & Zhang, Y. (2021). Datlmedqa: A data augmentation and transfer learning based solution for medical question answering. Applied Sciences, 11(23), 11251.
[11] Jeong, S. W., Kim, C. G., & Whangbo, T. K. (2023, March). Question answering system for healthcare information based on BERT and GPT. In 2023 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON) (pp. 348-352). IEEE.
[12] Chatzimina, M. E., Papadaki, H. A., Pontikoglou, C., & Tsiknakis, M. (2024). A Comparative Sentiment Analysis of Greek Clinical Conversations Using BERT, RoBERTa, GPT-2, and XLNet. Bioengineering, 11(6), 521.
[13] Tran, Q. D. L., & Le, A. C. (2023). Exploring Bi-Directional Context for improved chatbot response generation using deep reinforcement learning. Applied Sciences, 13(8).
[14] Le, A. C. (2021, August). A Deep Reinforcement Learning Model using Long Contexts for Chatbots. In 2021 International Conference on System Science and Engineering (ICSSE) (pp. 83-87). IEEE.
[15] Ali, B., Ravi, V., Bhushan, C., Santhosh, M. G., & Shankar, O. S. (2021). Chatbot via machine learning and deep learning hybrid. In Modern Approaches in Machine Learning and Cognitive Science: A Walkthrough: Latest Trends in AI, Volume 2 (pp. 255–265). Springer.
[16] Weng, Y., Li, B., Xia, F., Zhu, M., Sun, B., He, S., ... & Zhao, J. (2023). Large language models need holistically thought in medical conversational qa. arXiv preprint arXiv:2305.05410.
[17] Varshney, D., Zafar, A., Behera, N. K., & Ekbal, A. (2023). Knowledge grounded medical dialogue generation using augmented graphs. Scientific Reports, 13(1), 3310.
[18] Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2023). E2S2: Encoding-enhanced sequence-to-sequence pretraining for language understanding and generation. IEEE Transactions on Knowledge and Data Engineering.
[19] Li, C., & Xing, W. (2021). Natural language generation using deep learning to support MOOC learners. International Journal of Artificial Intelligence in Education, 31, 186-214.
[20] Yang, W., Zeng, G., Tan, B., Ju, Z., Chakravorty, S., He, X., ... & Xie, P. (2020). On the generation of medical dialogues for COVID-19. arXiv preprint arXiv:2005.05442.