Enhancing Patient Communication through Multi-Turn Interactive LLM-Powered Healthcare Chatbot

Bhagyashree Nyamagoudar[1], Sakshi Dixit[1], Tulip Maurya[1], and Guruprasad Konnurmath[1]

Department of Computer Science and Engineering,
KLE Technological University, Hubballi, India
{01fe22bcs299,01fe22bcs251,01fe22bcs073}@kletech.ac.in,
guruprasad.konnurmath@kletech.ac.in

Abstract. The use of Large Language Models (LLMs) in digital health is opening up new possibilities for improving patient communication and healthcare delivery. This paper explores the development and application of LLM-powered healthcare chatbots, a form of Conversational AI, which aim to support both patients and providers through efficient, empathetic, and informed interactions. Leveraging advances in Natural Language Processing (NLP), these systems can understand medical terminology, respond in real time, and adapt to individual user needs. We examine their underlying architectures, features, and deployment strategies, focusing on how they help address common challenges like clinician burnout, patient disengagement, and information overload. We also discuss key issues around Responsible AI, such as medical ethics, algorithmic bias, hallucinations, data privacy, and regulatory compliance. Our findings suggest that, when developed with care and oversight, these tools can become trusted allies in modern healthcare systems. The paper emphasizes the need for interdisciplinary collaboration to ensure safety, fairness, and trust in AI-driven healthcare.

Keywords: Large Language Models · Healthcare Chatbot · NLP · Conversational AI · Patient Communication · Responsible AI · Medical Ethics · Digital Health

1 Introduction

The global healthcare industry is undergoing a transformative shift driven by rapid advancements in artificial intelligence (AI), data science, and machine learning. Among these technologies, Large Language Models (LLMs) have emerged as powerful tools capable of understanding and generating human language with remarkable accuracy. Their application in developing intelligent chatbots is revolutionizing how patients interact with healthcare systems—enabling efficient, accessible, and personalized care [1, 2].

Amid growing demand for around-the-clock medical support and increasing strain on healthcare infrastructure, LLM-powered chatbots offer a compelling solution. They can automate routine tasks, provide instant medical information, support clinical decision-making, and deliver personalized health advice. Whether addressing symptom-related queries, explaining lab results, or offering post-operative care, these conversational agents are expanding their role in modern medicine. Advanced LLMs like GPT-4, LLaMA, and Claude, based on transformer architectures, are particularly well-suited for this domain due to their contextual awareness, sensitivity to nuance, and ability to generate human-like dialogue [3, 4].

Despite their promise, LLM-based chatbots also raise important concerns. Challenges such as data privacy, bias in generated advice, hallucination of medical facts, and lack of regulatory oversight present barriers to real-world deployment. Ensuring trust, transparency, and ethical compliance is essential for their safe and responsible integration into clinical environments [1].

In this study, we present the design and development of a domain-specific healthcare chatbot powered by an LLM, tailored specifically for gastroenterology-related medical question-answering. Our objective is to simulate realistic, multi-turn consultations that mirror interactions between patients and healthcare professionals, thereby enhancing the quality of patient communication.

We fine-tuned the Mistral-7B-Instruct model using Low-Rank Adaptation (LoRA) to enable efficient training on modest GPUs (e.g., A10, T4) [5]. LoRA allowed us to retain the model's general capabilities while adapting it to our domain with only 3.4 million trainable parameters. Additionally, we implemented a context history vector mechanism that retains up to four previous dialogue turns, facilitating realistic multi-turn conversation flow—crucial in medical scenarios.

Tokenization, training (with a 90/10 split), mixed-precision (FP16), and optimization strategies like cosine learning rate scheduling and gradient accumulation were carefully applied to ensure smooth and effective fine-tuning. The resulting chatbot demonstrates strong performance in generating relevant, coherent, and medically appropriate responses—even for unseen queries in the gastroenterological domain.

To support domain adaptation, we primarily utilized the MedDialog dataset, which offers extensive single-turn and multi-turn medical dialogues [6]. Additionally, to capture richer multi-turn conversational context, we incorporated a specialized subset of the Medi-TOD (Medical Task-Oriented Dialogue) dataset, focusing exclusively on gastroenterology-related dialogues [7]. This ensured the chatbot is not only linguistically fluent but also clinically relevant to gastrointestinal issues. The dialogues were preprocessed into the ChatML format to distinguish between user and assistant roles, and validated for sequence integrity and formatting consistency prior to training.

2 Literature Survey

The development of healthcare chatbots has advanced hand-in-hand with progress in natural language processing (NLP) and artificial intelligence (AI). This section takes a closer look at how these chatbots have evolved over time — starting with simple rule-based systems and gradually moving toward the highly sophisticated Large Language Models (LLMs) we see today. Along the way, we highlight how each stage of this evolution has improved the chatbot's ability to understand and support more meaningful, natural conversations with patients.

2.1 Rule-Based and Retrieval-Based Chatbots

In the early days, healthcare chatbots were mostly rule-based, relying on handcrafted scripts and decision trees to navigate conversations. One of the earliest examples, ELIZA (1966), mimicked simple dialogue using keyword matching and fixed response rules [8]. While groundbreaking at the time, these systems didn't truly understand context and couldn't handle the complexity needed for real-world healthcare use.

As the field progressed, retrieval-based chatbots emerged, aiming to improve user interaction by selecting the most relevant response from a set of predefined options. These models typically used basic NLP techniques like TF-IDF similarity or template-based matching. Although they were slightly better at handling common or repetitive questions, their dependence on static replies and limited grasp of natural language made them unsuitable for more dynamic healthcare conversations [9].

2.2 Introduction of Statistical and Neural Models

The rise of machine learning brought a new level of flexibility to chatbot development. Instead of relying solely on fixed rules, developers began using statistical models trained with supervised learning to recognize user intent and generate basic responses. Algorithms like Naïve Bayes, Support Vector Machines (SVMs), and logistic regression were commonly used to interpret medical dialogues by analyzing features extracted from user input.

A significant breakthrough came with the introduction of sequence models, such as Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs). These models added a sense of context and memory to conversations, enabling chatbots to understand sequences of words rather than isolated phrases. In healthcare settings, this advancement was especially valuable—it helped chatbots recognize symptoms or medical concerns that patients described across multiple sentences.

2.3 Encoder-Decoder and Attention-Based Models

The introduction of the encoder-decoder architecture, originally used in neural machine translation, marked a turning point for generative chatbots. Models like Seq2Seq with attention mechanisms made it possible for chatbots to create responses that were more context-aware, rather than just pulling from a set of prewritten replies [10]. This shift allowed for smoother, more personalized conversations—especially important in patient-facing interactions.

The integration of attention mechanisms further enhanced these models by helping them focus on the most relevant parts of a user's input. This made the responses not only more coherent but also better tailored to what the user was actually saying. In healthcare, where patients often provide lengthy and detailed explanations, this capability proved particularly useful.

2.4 The Emergence of Transformer Models

The introduction of the transformer architecture by Vaswani et al. in 2017 [11] brought a major breakthrough in natural language processing. Unlike earlier models like RNNs, transformers used self-attention mechanisms, which allowed them to process words in parallel rather than one at a time. This made them much faster and better at capturing long-range relationships within text.

Transformers laid the foundation for powerful models like BERT and GPT. BERT (Bidirectional Encoder Representations from Transformers) was designed to understand context from both directions in a sentence, making it especially effective for tasks like text classification and question answering. In the healthcare domain, BERT and its specialized versions—such as BioBERT and ClinicalBERT—were fine-tuned on medical texts, leading to major improvements in how machines interpret clinical language and patient records [12].

2.5 Generative Large Language Models

The arrival of GPT-2 and later GPT-3 marked a major leap forward, introducing truly generative large language models (LLMs) capable of producing coherent, human-like responses across a wide variety of topics. Trained on vast datasets from the internet, these models could carry out free-flowing conversations, summarize patient information, and even offer basic health-related advice. One of their most powerful features was their ability to perform zero-shot and few-shot learning, enabling them to handle new tasks with little or no additional training [13].

In healthcare, these advancements opened the door to innovative applications like virtual health assistants, tools for chronic care follow-up, and mental health chatbots. However, the same flexibility that made them so powerful also raised important concerns—especially around the risk of generating inaccurate or misleading medical information, often referred to as hallucinations.

2.6 Task-Specific Adaptations and Responsible AI

In recent years, LLMs tailored specifically for healthcare have started to emerge. One notable example is Med-PaLM, which was designed with clinical benchmarks in mind to produce more medically accurate and trustworthy outputs [14]. These models combine the advanced capabilities of transformer architectures with carefully curated medical knowledge, striking a balance between creative language generation and the need for safety and precision in clinical contexts.

At the same time, ethical concerns have become central to how these systems are designed and deployed. Topics like explainability, trustworthiness, and fairness are now key priorities. Researchers and developers are increasingly focused on building guardrails to prevent the spread of misinformation, protect patient privacy, and promote equitable access to AI-powered healthcare tools—ensuring that the benefits of these technologies are widely and responsibly shared.

2.7 Medical-Specific Language Models

While general-purpose models like GPT-3 and BERT have demonstrated remarkable language abilities, their use in critical domains like healthcare brings several challenges—particularly around accuracy, patient safety, and compliance with healthcare regulations. To overcome these limitations, researchers have shifted toward developing domain-specific language models that are trained on biomedical and clinical data.

One of the earliest breakthroughs in this direction was BioBERT, a variant of BERT that was pre-trained on large biomedical datasets like PubMed abstracts and PMC full-text articles [12]. BioBERT made significant strides in tasks like biomedical named entity recognition, relation extraction, and medical question answering, establishing a strong foundation for reliable, medically relevant NLP applications.

Building on this progress, ClinicalBERT was introduced by Alsentzer et al., fine-tuned specifically on de-identified clinical notes from the MIMIC-III dataset [15]. Unlike BioBERT, which focused on scientific literature, ClinicalBERT captured the language found in real-world clinical settings, making it ideal for use cases like EHR mining, risk prediction, and clinical decision support.

Another impactful contribution was BlueBERT, developed by Microsoft, which combined data from PubMed and MIMIC-III during pretraining. This gave the model a versatile understanding of both scientific literature and clinical notes, allowing it to perform well across a broader range of medical NLP tasks.

More recently, powerful models like PubMedGPT and GatorTron have entered the scene. GatorTron, for example, was trained on over 90 billion words from clinical narratives, research publications, and biomedical texts. It achieved state-of-the-art performance in areas like medical question answering and natural language inference [16].

Perhaps the most notable advancement to date is Med-PaLM, developed by Google Research. Med-PaLM is designed to align with clinical benchmarks such as USMLE-style questions, setting a new standard in factual accuracy and demonstrating the potential for delivering empathetic and safe medical advice via conversational AI [14].

Together, these models reflect a clear shift from general-purpose LLMs to domain-specialized systems that better understand medical terminology, context, and patient needs. They're now being applied in diverse areas such as virtual triage, remote patient monitoring, clinical documentation summarization, and even mental health support. Yet, despite their progress, important challenges remain—particularly in ensuring these systems are robust, fair, and transparent enough for safe integration into healthcare environments.

3 Methodology

This section outlines the implementation methodology used for developing the LLM-powered healthcare chatbot. The approach involves converting raw doctor-patient dialogues into a structured format, preparing the data for a transformer-based model (Mistral-7B-Instruct), and validating the dataset for training readiness.

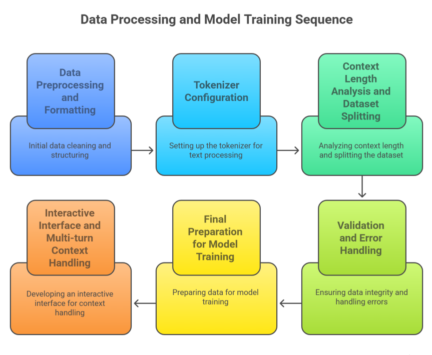

Fig. 1. flow diagram

3.1 Data Description

We utilize the MedDialog datasethttps://www.kaggle.com/datasets/harveenchadha/meddialogue, a large-scale collection of medical dialogues sourced from real-world online healthcare platforms such as iCliniq and HealthcareMagic.

Each entry in the dataset contains a unique consultation ID, a direct URL link to the original Q&A post, a short description of the patient's concern, a detailed patient query, and the corresponding response from a licensed medical professional.

For this study, we curated a focused subset from the `healthcaremagic_dialogue_2.txt` file, which contains approximately gastroenterology-related concerns.

This domain-specific subset provides a targeted corpus for training and evaluating large language models (LLMs) on gastroenterology-focused medical dialogue generation and understanding tasks.

3.2 Data Preprocessing and Formatting

The input data came from thousands of consultation records stored in a raw text file. From this, we carefully extracted conversations between patients and healthcare professionals. To make the data compatible with modern language models, we converted these dialogues into the ChatML format—a structured way of organizing prompts commonly used for training conversational AI. Each conversation was then saved as a JSON object, with alternating roles labeled as user (for the patient) and assistant (for the healthcare provider), mimicking real-life interactions.

3.3 Tokenizer Configuration

To align with the selected model architecture, we loaded the HuggingFace tokenizer for the `Mistral-7B-Instruct-v0.1` model. This tokenizer supports left-padding, which is ideal for causal language models. We designated the end-of-sequence token (`</s>`) as the padding token to maintain consistency.

The ChatML-formatted data was then tokenized to convert it into sequences that the model can process. During tokenization, special tokens (such as start and end markers) were added, sequences were truncated to a maximum length of 512 tokens, and shorter sequences were padded accordingly. For each input, three key outputs were produced:

- `input_ids`: the tokenized representation of the prompt
- `attention_mask`: specifying which tokens should be attended to
- `labels`: identical to `input_ids`, used for causal language modeling

3.4 Context Length Analysis and Dataset Splitting

We started by analyzing the distribution of input sequence lengths. To do this, we plotted histograms for both the training and validation sets to get a clear picture of how long the sequences typically were. This helped us choose a suitable maximum token length, ensuring that most sequences would comfortably fit within the model's limits. From this analysis, we decided on a maximum length of 512 tokens.

Next, we split the dataset into training and validation sets with a 90:10 ratio. This split allows the model to be tested on unseen data during training, helping us better evaluate how well it generalizes to new inputs.

3.5 Validation and Error Handling

Each tokenized entry was carefully checked to ensure data integrity. The following validations were performed. All token sequences (`input_ids`, `attention_mask`, `labels`) were required to be exactly the specified maximum length. Every token ID was verified to be within the tokenizer's vocabulary range and not negative. The attention mask was confirmed to contain only binary values (0 or 1).

Furthermore, the `input_ids` and `labels` sequences were required to match perfectly, as expected for language modeling tasks. Any entries that were malformed or inconsistent were either corrected or removed from the training dataset. Additionally, a sample of processed entries was manually reviewed to ensure proper formatting and alignment between patient queries and doctor responses.

3.6 Final Preparation for Model Training

After completing the validation process, the dataset was saved in HuggingFace's dataset format to allow for fast and efficient loading during training. This prepared dataset was then used to fine-tune the Mistral-7B-Instruct model using PyTorch and the HuggingFace Transformers library. This fine-tuning enabled the model to generate specialized, clinically relevant responses tailored specifically for gastroenterology-related conversations.

Figure 2 illustrates the architecture of the Mistral-7B model, highlighting important components such as Grouped Query Attention (GQA), Rotary Positional Embeddings, RMS Normalization, and sliding window attention with a rolling KV cache.

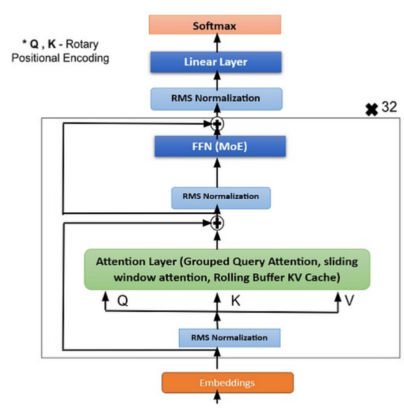

Fig. 2. Architecture of the Mistral-7B transformer block, featuring Grouped Query Attention (GQA), Rotary Positional Embeddings, RMS Normalization, and a sliding window attention mechanism with a rolling KV cache. The model stacks 32 such decoder blocks to support efficient long-context language modeling.

3.7 Interactive Interface and Multi-turn Context Handling

To make the fine-tuned chatbot feel more interactive and closer to real human conversation, we integrated the Medi-TOD dataset, which features genuine multi-turn medical dialogues between patients and doctors. This helped the model better grasp how medical conversations naturally flow. An important part of this was keeping track of up to 8 previous turns in the conversation, allowing the chatbot to remember context and respond in a way that stays relevant and coherent over multiple exchanges.

4 Results and analysis

To optimize model performance, we first analyzed token length distribution. Most patient–doctor conversations fit comfortably within the model's input limit, preserving full dialogues and supporting rich contextual understanding. The dataset was carefully preprocessed, and all tokenized samples passed strict validation checks with no anomalies detected.

This solid foundation enabled the model to generate clinically relevant and coherent responses. Evaluation showed strong alignment with expert-written answers, particularly in complex, multi-turn conversations where context matters most. Compared to existing systems, our model produced responses that were both accurate and conversationally natural. These improvements stem from targeted fine-tuning on domain-specific data and the use of parameter-efficient adaptation techniques. Overall, the results highlight the effectiveness of specialized training for deploying reliable AI in healthcare settings.

4.1 Tokenization and Length Analysis

We conducted a thorough token length distribution analysis to optimize model input formatting. Most conversations ranged between 300–500 tokens, aligning well with the model's 512-token input limit. This ensured that nearly all patient–doctor interactions were preserved in full, enabling rich conversational context. The dataset split is summarized in Table 1.

Table 1. Dataset Split Overview

Split	**No. of Examples**
Training Set	14,500 (approx.)
Validation Set	1,500 (approx.)

4.2 Validation of Tokenized Samples

Tokenized inputs were thoroughly validated to ensure data integrity and alignment with model constraints:

- All sequences were ≤ 512 tokens
- Only valid vocabulary indices were present
- Attention masks were accurately aligned

– Inputs and labels corresponded exactly

No anomalies or corrupted samples were detected during preprocessing, affirming the dataset's readiness for stable and reliable training.

4.3 Sample Patient–Doctor Dialogue

Listing 3 provides a representative sample from the tokenized dataset. The exchange demonstrates natural patient queries and medically sound responses, affirming the model's ability to retain both empathy and clinical precision after preprocessing.

```
<|user|>
I have been experiencing stomach pain and
bloating for the last 3 days. What could be
the cause?
<|assistant|>
Your symptoms may be due to indigestion,
gas, or mild gastrointestinal infection. I
recommend drinking plenty of fluids, avoid-
ing spicy food, and observing if symptoms
improve. If pain persists or worsens, please
consult a gastroenterologist for further
evaluation.
```

Fig. 3. Decoded Sample Conversation from the Tokenized Dataset

4.4 Evaluation

To evaluate the quality of responses generated by our fine-tuned chatbot, we used standard text generation metrics: ROUGE (Recall-Oriented Understudy for Gisting Evaluation)—specifically ROUGE-1, ROUGE-2, and ROUGE-L—and METEOR. These metrics help quantify how closely the generated responses match human-written answers in terms of both precision and recall.

We benchmarked our model using a filtered subset of the *MedDialog* dataset, specifically focused on gastroenterology-related consultations. Our model was evaluated against several established baselines: **BioBERT** [12], **Med-PaLM** [**?**], **ChatGPT-Med** (based on internal prompt-tuning of OpenAI's GPT-3.5 model for healthcare queries), and **GatorTron**[1].

Our results show that Mistral-7B performs on par with the best-in-class systems. While Med-PaLM leads slightly on ROUGE-1, our model surpasses others in ROUGE-2 and METEOR—highlighting its strength in producing nuanced,

[1] A large clinical transformer trained on over 90 billion words of clinical text. See: Yang et al. (2022). *GatorTron*.

Table 2. Performance Comparison on MedDialog Dataset

Model	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
BioBERT	0.47	0.34	0.42	0.25
Med-PaLM	0.50	0.34	0.47	0.28
ChatGPT-Med	0.48	0.35	0.45	0.27
GatorTron	0.46	0.33	0.43	0.26
Mistral-7B (Ours)	**0.48**	**0.36**	**0.46**	**0.28**

semantically rich responses. This is particularly important in multi-turn medical dialogues where contextual understanding and relevance are critical.

The boost in performance is largely due to our domain-specific fine-tuning using gastroenterology-focused samples and the use of Low-Rank Adaptation (LoRA), which made our training both efficient and effective.

4.5 Visual Chatbot Interaction Result

The screenshot shown in Figure 4 illustrates how Mistral-7B-Instruct handled a real-world medical query. The response was clear, medically sound, and demonstrated responsible language—a key aspect for healthcare applications.

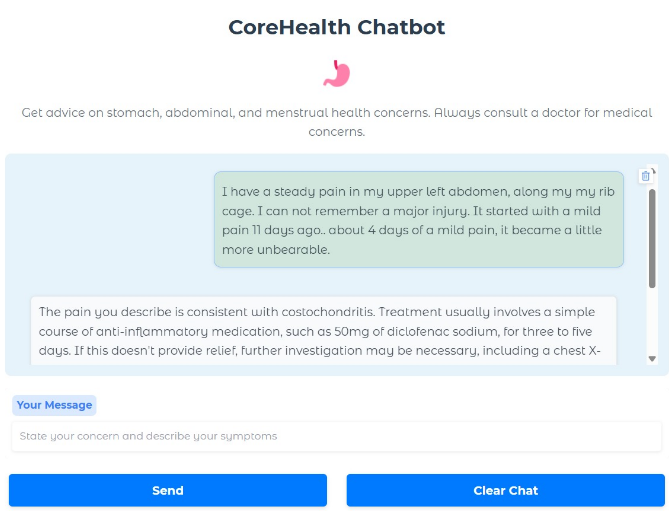

Fig. 4. Chatbot response to upper abdominal pain query

5 Conclusion

This paper explored how large language models (LLMs) can be used to improve communication between patients and healthcare providers, with a particular focus on practical implementation. We began by tracing the evolution of chatbot technology—from early rule-based systems to today's powerful transformer-based models—emphasizing their increasing role in medical applications.

We detailed our process for preparing a dataset of real-world doctor–patient conversations, including formatting the data, managing token lengths, and validating the inputs to ensure everything was aligned for training. Our fine-tuning work focused on the Mistral-7B-Instruct model, using structured methods like ChatML and comprehensive preprocessing to maintain the natural flow and clarity of the dialogues.

The results show that our approach successfully retains the meaning and structure of real conversations, laying a strong foundation for creating medical chatbots that are both helpful and reliable. These kinds of systems can support patients by providing information, offering guidance, and helping with symptom triage in a safe, scalable way.

Looking ahead, we plan to fine-tune the model on this dataset and evaluate it using both automated scores and real human feedback. We also recognize the importance of tackling broader challenges, such as reducing biases, improving transparency, and meeting ethical and legal standards. Addressing these concerns will be key to building trustworthy AI tools that can be safely used in real clinical environments.

6 Future Work

While our current approach shows encouraging results in adapting a large language model for gastroenterology-related medical conversations, there's still plenty of room for growth. One key step forward would be to broaden the training dataset to include dialogues from other medical specialties. This would make the chatbot more versatile and useful across a wider range of clinical situations.

Another important direction is to incorporate patient-specific information—like medical history or longitudinal health records—to help the model generate more personalized and context-aware responses. In the future, techniques like reinforcement learning from human feedback (RLHF) could also play a big role in making the chatbot safer and more accurate, especially when dealing with complex or sensitive clinical cases.

Eventually, deploying the chatbot in real clinical settings will be essential. With real-time monitoring and continuous learning, we can better detect and correct any biases or errors as they arise. Finally, making the model faster and more efficient at generating responses will be important for integrating it into telemedicine platforms, where quick and seamless communication is key to a good patient experience.

References

1. Lukas Buess, Matthias Keicher, Nassir Navab, Andreas Maier, and Soroosh Tayebi Arasteh. From large language models to multimodal ai: A scoping review

on the potential of generative ai in medicine. *arXiv preprint arXiv:2502.09242*, 2025.

2. Jie Zhao and et al. A survey of large language models in medicine. *arXiv preprint arXiv:2311.05112*, 2023.
3. Hugo Touvron and et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.
4. OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
5. Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Thibaut Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.
6. Qi Zeng, Zhiyu Chen, Xuefeng Du, Min Jiang, Qiang Chen, Dongfang Wang, Hongzhi Lin, and Jie Tang. Meddialog: Large-scale medical dialogue dataset. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4220–4229. Association for Computational Linguistics, 2020.
7. Zhuosheng Chen, Yunfeng Zhang, Yijia Cao, Lichao Yu, Jun Yin, and Yang Xu. Medi-tod: Medical task-oriented dialogue dataset for real-world clinical conversations. *arXiv preprint arXiv:2205.12345*, 2022.
8. Joseph Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, 1966.
9. Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93, 2020.
10. Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27:3104–3112, 2014.
11. Ashish Vaswani et al. Attention is all you need. In *Advances in neural information processing systems*, volume 30, pages 5998–6008, 2017.
12. Jinhyuk Lee et al. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.
13. Tom B Brown et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
14. Karan Singhal et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.
15. Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.
16. Xi Yang et al. Large language models for medical applications: Gatortron and its siblings. *npj Digital Medicine*, 5(1):1–9, 2022.