Create a Chatbot from Scratch

Daniel Evans, Rohan Ashraf Hashmi, Jim Huynh, Alexander Robinson

# Introduction

## Background

AI chatbots are seen almost everywhere in today’s world offering human-like conversations without the need for a person to write the messages. They have the ability to get user input, generate contextually aware conversations, and maintain a dialog that feels natural to the user. These chatbots are used to help users with simple questions immediately. AI based Chatbots have huge potential in healthcare settings because they can answer a patient's question quickly and hopefully help direct them if the question is outside of the area of expertise of the chatbot.

## Aim of Project

The aim of this project is to design and implement an AI based chatbot that can help patients in a hospital setting. The chatbot will assist patients and visitors by answering basic health related questions, giving directions, scheduling appointments, and checking prescription statuses. The chatbot will also have a personality by using emojis which will make it seem more human-like when conversing with people. The chatbot will also be able to understand misspelled words and questions that aren’t directly in the dataset.

# **Related Work**

Med-PaLM: Google’s State-of-the-Art Medical AI Chatbot  
  
Med-PaLM is a state-of-the-art medical large language model (LLM) developed by Google Research, designed specifically for answering health-related questions with high factual accuracy and safety. It is built on top of Google’s PaLM architecture and fine-tuned using a diverse dataset called MultiMedQA, which includes U.S. medical licensing exam (USMLE) questions, consumer health FAQs, and clinical datasets. Med-PaLM leverages instruction-tuning techniques with expert-curated examples to generate helpful, evidence-based responses that are aligned with medical best practices.

What makes Med-PaLM state-of-the-art is its benchmark-topping performance on key medical QA tasks. It was the first AI model to exceed the USMLE passing score (~60%) with a 67.6% accuracy rate on the MedQA benchmark. Its successor, Med-PaLM 2, improved this to 86.5%, approaching human-level performance. Additionally, over 90% of Med-PaLM’s long-form answers were rated by clinicians as medically accurate and safe — significantly outperforming generic LLMs like GPT-3 or base Flan-PaLM. These evaluations reflect not just raw performance but also the model’s reliability in high-stakes healthcare scenarios.

Med-PaLM’s specialization, trustworthiness, and fine-tuning for the healthcare domain make it one of the most advanced AI chatbot systems currently available for medical applications. While it operates at a much larger scale than our project, it serves as a benchmark and inspiration for developing safe, useful chatbots in healthcare settings.

# Methodology

## Data Generation

Generalized Data Generation

1. **Prompt engineering (1)**: Used ChatGPT-4o to generate a prompt to start generating data. Supplied the given *intents.json* as an example.
2. **LLM Prompt (1)**: Fed the generated prompt to ChatGPT-o3 generate a skeleton *intents.json* file (generate only tags and contexts)
3. **Python (1)**: Separated the skeleton of intents into batches to help the LLM focus on data generation
4. **Prompt engineering (2)**: Used ChatGPT-4o to generate a prompt to generate some patterns and responses
5. **LLM Prompt (2)**: Fed that prompt with each batch to ChatGPT-o3 to generate data for each batch
6. **Prompt engineering (3)**: Used ChatGPT-4o to generate a prompt to expand on each intent with more patterns/responses
7. **LLM Prompt (3)**: Fed that prompt with each batch to ChatGPT-o3 to expand on each batch
8. **Python (2)**: Recombined all the batches into a unified *.json*
9. **Python (3)**: Removed or replaced special characters (like non-breaking space/hyphen, em-dashes, accented vowels) with simple counterparts

Fine Tune Data Generation

1. Created a set of over 40 intent categories tailored to healthcare scenarios, such as appointment booking, billing inquiries, department directions, greetings, and farewells.
2. For each intent, wrote a variety of user input examples (patterns) that reflect natural, conversational ways people might ask for that service. These included polite requests, direct commands, and casual phrasing to capture real-world variation.
3. Designed multiple response options for each intent that express an extroverted and friendly personality. Responses were written with emojis and a warm tone to make the chatbot feel approachable and human-like.
4. Organized the data in a structured format that supports fine-tuning, ensuring each intent was clearly associated with a rich set of example patterns and matching personality-driven replies.
5. The resulting dataset was used as a foundation for aligning the chatbot’s behavior with real-world hospital communication needs, emphasizing both clarity and personality in interactions.

Final Model  
The final model was created by taking both the general data and fine-tuned data then creating every possible combination of those pattern/response pairs and turning them into conversations. Then the conversations had their emojis replaced with placeholders. 6.1 million conversations were then tokenized and saved for later loading. The pretrained GPT2 tokenizer was loaded. The model was then initialized with the GPT2 architecture with randomized weights to ensure the model was trained from scratch. The model was then trained for ~40 hours on the generalized data and got through 256,000 conversations. The model was then saved to a .pth file. The .pth file was loaded with the fine-tuning data and then trained with the same hyper parameters except for the batch size, which was scaled down by a factor of four, and so was the learning rate. This model was also trained for ~40 hours using the Colab T4 GPU and got through 640,000 conversations. The final .pth file was saved and then the model was tested.

# Discussion and Results

## Discussion

Data Generation  
I created 48 unique intents to provide broad coverage of hospital-related queries while keeping the structure manageable and logically organized. This number struck a balance between flexibility and clarity, allowing me to handle a wide range of patient interactions without overwhelming the model. ChatGPT and Grok were used to diversify phrasing while maintaining a consistent tone. Personality, emojis, and follow-up flows were intentionally added to make the chatbot feel more natural and engaging in patient-facing scenarios.

Final Model  
 I decided to take every combination of the possible conversations because we had a limited data size, and more samples would lead to less overfitting. The emojis were replaced with placeholders because the GPT2 pre-trained tokenizer doesn't have support for emojis, so the placeholders can translate them back into emojis during inference. The tokenized dataset was saved to a. parquet file so that the dataset didn't need to be created and tokenized every time we were to resume training. The existing GPT2 architecture was chosen because it has been proven to work with a large data corpus, so we wanted to see how well it would work with a lot less information. The ~40 hours training time was chosen because the loss seemed to flatline around step 2,000 with the generalized data. The training was stopped after ~40 hours for the fine-tuned data for the same reason. The batch size had to be reduced from the generalized model to the fine-tuning model because the VRAM size decreased from 64Gb with an NVIDIA Orin to 15Gb on a T4 GPU.

Model Evaluation  
The final model combined generalized and fine-tuned data to generate all pattern/response pairs. After replacing emojis with placeholders, 6.1 million conversations were tokenized. The GPT-2 model was trained from scratch for ~40 hours on generalized data (256,000 conversations) and fine-tuned for another ~40 hours (640,000 conversations) with reduced batch size and learning rate. The final model showed improved contextual understanding and coherence for healthcare chatbot use.

Results  
The fine-tuned GPT-2 model outperformed the Chatterbot model, achieving a Perplexity of 13.66, ROUGE-1 of 0.5, ROUGE-2 of 0.15, and ROUGE-L of 0.5, while Chatterbot had a 20% Fallback Rate and no accuracy or precision. This highlights the GPT-2 model's superior fluency and contextual accuracy for healthcare conversations.

Conclusion and Future Work  
We successfully developed an AI chatbot for hospital settings, designed to answer patient inquiries efficiently while maintaining a human-like, friendly tone. By combining generalized and fine-tuned data, we achieved a model capable of handling diverse healthcare-related questions. The chatbot’s ability to understand varied phrasing and respond naturally was a key outcome.

For future improvements, we aim to enhance error handling, integrate speech-to-text for accessibility, and increase model interpretability to build user trust. Expanding the dataset with more complex medical queries and deploying them on a cloud platform would further boost performance and reliability.

References

<https://sites.research.google/med-palm>

<https://www.digitalhealth.net/2023/01/google-research-and-deepmind-develop-ai-medical-chatbot/>