LLM Chatbots for Enhancing Customer Service

Pradeep Pujari
Standford AI Specialization
EleutherAI
USA
pkpujari@acm.org

Arvapalli Sai Susmitha
Department of Computer Science
IIT Kanpur
India
arvapallisaisusmitha@gmail.com

Abstract

This paper presents the development of a natural language processing (NLP) chatbot aimed at enhancing customer service by providing human-like interactions. The primary objective of this project is to create a chatbot that improves the personalized experience for customers across various industries. To achieve this, a sequence-to-sequence model was utilized, with three large language models—BART, LLaMA, and Cosmo—being analyzed for their effectiveness in this task. The models were trained on over 1.4 million curated data points from customer interactions within the healthcare sector, specifically from Kaiser Permanente. The comparison of these models demonstrates significant potential, with each model generating coherent and contextually relevant responses to input text. Incorporating the most effective chatbot model into existing customer service platforms is expected to enhance the overall customer experience, streamlining communication and providing timely, accurate assistance.

1 Introduction

The rapid advancements in natural language processing (NLP) have given rise to numerous applications in various domains, with healthcare being a rapidly growing field for the implementation of intelligent chatbot systems. Chatbots have the potential to streamline customer service, enhance patient engagement, and support healthcare professionals in their daily tasks (Laranjo et al., 2018). Despite this, existing application scenarios of chatbots in the medical domain have been very limited (Reis et al., 2020). This is likely due to the anticipated risk of having such unregulated conversation in the healthcare sector - a wrong or misleading communication result can directly affects a patient's health.

This paper presents the development of a conversational chatbot for Kaiser Permanente, a leading healthcare and coverage provider, aiming to improve the personalized experience for its clients.

Kaiser Permanente (KP) is a healthcare consortium that operates hospitals, clinics, and a large network of healthcare professionals to provide comprehensive care to more than 12.4 million members. As part of their commitment to personalized experiences, the organization offers a 24/7 chat service on their website, where clients can ask questions related to appointments, billing, membership, and other concerns. This project's objective is to create a chatbot that emulates natural staff conversation, thereby enhancing the efficiency and quality of the chat service.

In order to achieve this goal, a sequence-to-sequence model using the BART transformer (Lewis et al., 2019) is trained to create a context-sensitive conversational chatbot. Two large language models (LLMs), known as LLaMA (Touvron et al., 2023) and COSMO (Kim et al., 2023) were also trained. Our models are trained and evaluated using data given to us from KP's website chat service (Section 3). This paper details the development process, from data acquisition and cleaning to model training and evaluation, as well as the challenges encountered along the way.

It's important to note that our focus is adjacent to the medical domain, as it lies more in the realm of client-company interactions (for example, "When is my next appointment"), rather than handling specifically medical queries (such as "I have a headache, can I take Advil even though I'm pregnant?").

This work makes two main contributions as follows:

1. Reconstruction of data for customer service conversation dialogue

2. Chatbot specially designed to address administrative questions.

3. Analysis of three LLM based chatbots.

To the best of our knowledge, this is the first work that uses LLMs for training on the administrative dialogue domain. This work will facilitate the development of future chatbots not only for the medical domain, but all administrative domains.

2 Related Work

The burgeoning field of medical chatbots has been explored from different angles, varying from disease-specific advice and triaging to administrative tasks. Notably, the work by Hsu and Yu (year) titled "A Medical Chatbot using Machine Learning and Natural Language Understanding" deserves attention Hsu and Yu (2022). Their study delineates the construction and use of a chatbot in the medical domain utilizing machine learning techniques, albeit not the Transformer models, which our research centers around. Their approach, however, remains firmly within the ambit of direct healthcare advice and disease diagnosis. This differentiates from our study, which is focused on the administrative side of the healthcare sector.

Specifically, their chatbot employed a decision tree algorithm for disease prediction based on the input symptoms from patients. Although this model performed satisfactorily within its scope, it lacked the capacity to handle more complex conversation dynamics or provide highly personalized responses to the user (Hsu and Yu, 2022). This limitation underscores the need for more advanced models like the Transformer, which we harness in our work, to better comprehend and reply to user inquiries.

Significant progress in the field of Natural Language Processing (NLP) has been driven by the introduction of large language models. Among them, the Transformer model proposed by Vaswani et al. has been particularly influential due to its ability to capture complex contextual relationships within text (Vaswani et al., 2017). Subsequent large language models, such as BERT (Bidirectional Encoder Representations from Transformers), LLaMA (Large Language Model Meta AI), and COSMO (COnverSation MOdel) have built upon this architecture, demonstrating significant improvements in various tasks including translation, question-answering, and conversation generation (Devlin et al., 2019; Touvron et al., 2023; Kim et al., 2023).

Although these large language models have been extensively explored in general domains, their application in the medical chatbot space, particularly for administrative tasks, remains less explored. Our study seeks to bridge these gaps by applying Transformer models and large language models to the design of administrative medical chatbots. We anticipate that the enhanced capabilities of these models will significantly improve the efficacy and personalization of medical chatbots within healthcare administration.

3 Approach

3.1 Data Acquisition and Preprocessing

Kaiser Permanente gave us access to 291,519 website chat dialogue transcripts from April 2022 to January 2023.

Working with Kaiser Permanente's data required addressing certain data limitations. Initially, the company attempted to anonymize the data using basic pattern matching, which proved to be problematic, since words like "benefit" or "purchased" would become anonymized due to them having the name "Ben" or "Chase" in the word. The models were trained without anonymizing the data.

The data are preprocessed to create a suitable dataset for training the various models. In our preprocessing, URLs are tokenized, and replaced with *"<URL>"*. Several tokens that weren't transcribed properly in the data are replaced. For example, after the word *"with"*, there is always a *"-"*, making *"with -"* instead of *"with"*. In addition, all quotes are marked as *"???"*. These, along with several other examples of such issues, are fixed in preprocessing. Any repeated utterances in the data are concatenated, and became one singular utterance. Figure 1 shows an example of the cleaning of a conversation, with repeated utterances concatenated and incorrect characters fixed.

Statistics about the original dataset provided by Kaiser Permanente are shown in Table 1.

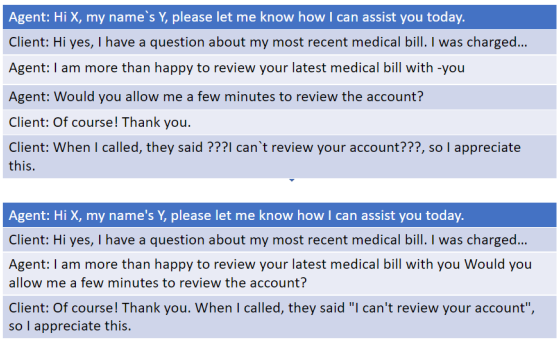

Figure 1: Example of cleaning a conversation.

Number of Conversations	291,519
Average Length	9.58
Standard Deviation of Length	7.19
Median Length	8.0
Mode Length	7
Longest Length	131
Total Agent Utterances	1.4 Million

Table 1: Dataset Statistics. All Lengths are in terms of number of utterances.

3.2 Dataset Construction

After cleaning and preprocessing, we have approximately 1.4 million individual datapoints. Each of these datapoints corresponds to an agent utterance within the conversation data. To provide contextual information necessary for generating appropriate responses, each datapoint also includes the dialogue history leading up to the corresponding agent utterance. This dialogue history comprises of the preceding utterances by both the user and agent in the conversation. Incorporating this history allow the models to understand the progression of the conversation and the context in which the agent utterance is made.

It is important to note that the length and complexity of the included dialogue history varies from one datapoint to another, depending on the position of the agent utterance within the conversation. Early utterances, for instance, will have shorter histories, while later utterances will include more prior dialogue.

3.2.1 BART

BART is well-suited for sequence-to-sequence modeling, which involves taking an input sequence and generating an output sequence. This can be used for conditional generation, generating text based on a given condition or prompt.

For BART, the input is a reformatted dialogue history H_{wBART}, where w is the history window size. The output is the agent utterance U. The dialogue history H_{wBART} was built as follows:

For the most recent w user and agent utterances u_i and u_j in the dialogue history, the appropriate speaker ID (*USER:* or *AGENT:*) was prepended. Each utterance was then separated by a separator token </s>. As such,

$$H_{wBART} = \text{AGENT: } u_{n-w}\text{</s>...</s>USER: } u_n$$

where n is the total number of previous utterances. If $w > n$, meaning there weren't w previous utterances, we used a token <*START*> instead of an utterance after using all the utterances available. Figure 2 shows an example of a short conversation consisting of 10 utterances being converted to datapoints for training, with a history window size $w = 1$.

3.2.2 COSMO

(Kim et al., 2023) COSMO (Kim et al., 2023) is not an auto-encoding model, so it needs a prompt.The prompt *"A customer is speaking to an agent from Kaiser Permanente, a healthcare company. <sep> Imagine you are the agent and respond to the customer <sep> "* was used to give the model context to what the task was. This prompt was then prepended to the dialogue history H_{wCOSMO}. This dialogue history is almost identical to that of BART, but <*turn*> was used as a separator token instead:

$$H_{wCOSMO} = \text{AGENT: } u_{n-w}\text{<turn>...<turn>USER: } u_n$$

where n is again the total number of previous utterances. However, if $w > n$, meaning there weren't w previous utterances, we simply included all of the history there was, and prepended a <*START_CONVERSATION*> tag to the dialogue history. This setup was used previously by Kim et al., 2023, when they introduced COSMO.

3.2.3 LLaMA

As opposed to COSMO, which is a sequence to sequence LLM, LLaMA is a Causal LLM (Touvron et al., 2023). Thus, there is no output to predict, but rather next tokens to predict. As such, a very similar prompt to COSMO was used, but *" <sep>*

Response: " was appended, along with the agent response, during training. For LLaMA, the same dialogue history as BART (H_{wBART}) was used.

4 Experiments

4.1 Models

4.1.1 BART

Three BART-Large (Lewis et al., 2019) models, with history ranging from $w = 1$ to $w = 3$,were fine-tuned on the constructed dataset, and a data pipeline for testing the models' performance was created. During the initial testing, we encountered a few issues with the generated responses, such as repetitive tokens or spaces. To address these problems, a repetition penalty of 5.0 is introduced, in order to discourage the model from generating the same token twice in a row. These models are trained with a learning rate of 2×10^{-5} for 5 epochs, with a batch size of 32.

4.1.2 COSMO

The large COSMO-11B (Kim et al., 2023) model was trained with a history hyperparameter of $w = 3$. This uses the dataset described in section 3.2.2. The model is trained with a learning rate of 3×10^{-4} for 1 epoch, with a batch size of 32 and weight decay of 0.01. We employed a Low-Rank adaptation (LoRA) to train this model (Hu et al., 2021; Dettmers et al., 2022), with a dropout value of .05, for the task *SEQ_2_SEQ_LM*.

4.1.3 LLaMA

The small LLaMA-7B (Touvron et al., 2023) model was trained with a history hyperparameter of $w = 3$. This uses the dataset described in section 3.2.3. The model is trained with a learning rate of 3×10^{-4} for 1 epoch, with a batch size of 32 and weight decay of 0.01. We also employed a Low-Rank adaptation (LoRA) to train this model (Hu et al., 2021; Dettmers et al., 2022), with a dropout value of .05, for the task *CAUSAL_LM*.

All experiments are conducted on a single NVIDIA A100 GPU, which takes 1/4/8 days to train for the BART, COSMO, and LLaMA experiments, respectively.

4.2 Results

All models were automatically evaluated using ROUGE-1, ROUGE-2 Ganesan, 2015, and ROUGE-L metrics. The scores shown in Table 2 provide an overview of the model's ability to accurately generate summaries, capturing not just unigram (ROUGE-1) (Lin, 2004) and bigram (ROUGE-2) overlaps, but also the longest matching sequence of words (ROUGE-L).

5 Analysis

Although the BART model with History of 3 seems to be the best model according to our automatic evaluation metric, a qualitative analysis was also performed, in which 100 examples of model responses were evaluated manually. BART with history of 3, COSMO, and LLaMA models were evaluated using this metric. The results are shown in figure 3.

LLaMA had extremely poor results in the automatic evaluation. The manual evaluation confirmed this. The responses from LLaMA were entirely nonsensical - it often generated *"USER: <START>"* and *"AGENT: <START>"*, and would continue generating that until it ran out of tokens. Sometimes it would reiterate the history of the dialogue, but it never responded in any concise way that made sense.

The BART model performed much better than the LLaMA model. It often would respond in ways that made sense, but there were definitely some significant issues. A very common occurrence in the training data was that the agent would need some time to verify a claim or answer a question, and would ask the user to wait a minute. So, the BART model also has an affinity for responding with a request for some time to verify a claim. These were marked as "time". In addition, every time the user said *"thank you"*, the bot would immediately respond ending the conversation, even if the user had asked a follow-up question. These were marked as "premature end".

The COSMO model fixed a lot of the issues the BART model had. It never asked for more time to verify a claim or to answer a question. However, it didn't adopt the exact wording that was used very often in the training data. For example, at the end of almost every conversation, the Agent in the training data would respond with *"Thank you for contacting Kaiser Permanente, have a nice day."* The BART model mimicked this word for word, while the COSMO model said something with the same intent - *"No problem, have*

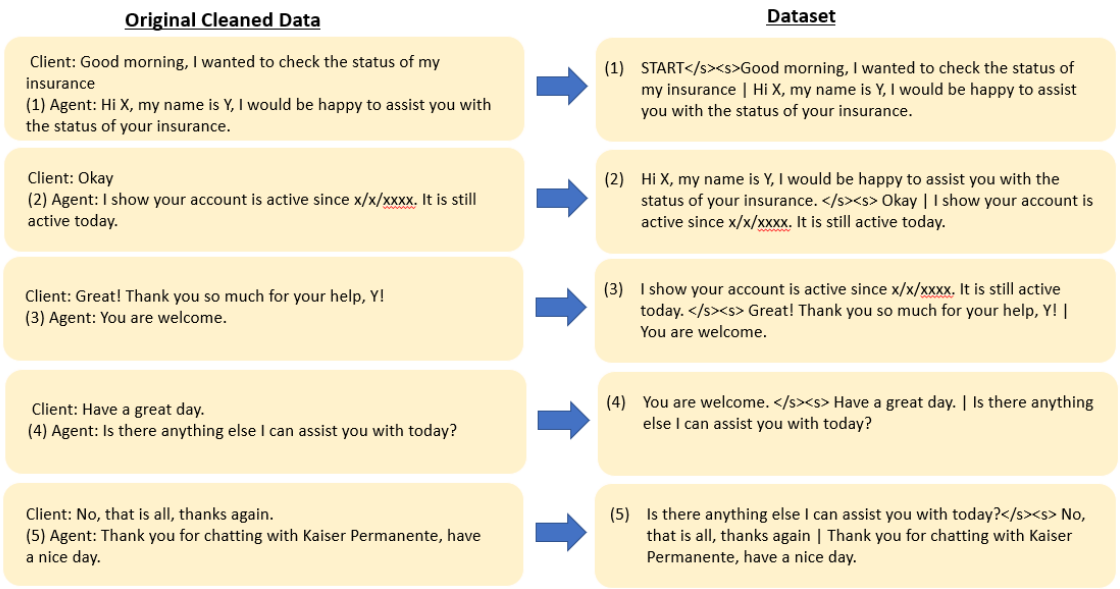

Figure 2: Dataset Construction for BART (History=1)

	BART			COSMO	LLaMA
History	**History=1**	**History=2**	**History=3**	**History=3**	**History=3**
ROUGE-1	30.3	30.5	33.9	33.1	5.3
ROUGE-2	16.1	18.0	21.0	7.9	1.3
ROUGE-L	27.4	27.9	31.2	27.4	4.1

Table 2: Comparative Analysis of F-scores for ROUGE-1, ROUGE-2, and ROUGE-L Metrics Across Different Models. This table illustrates the performance evaluation of various models based on the F-scores obtained through the ROUGE-1, ROUGE-2, and ROUGE-L evaluation metrics.

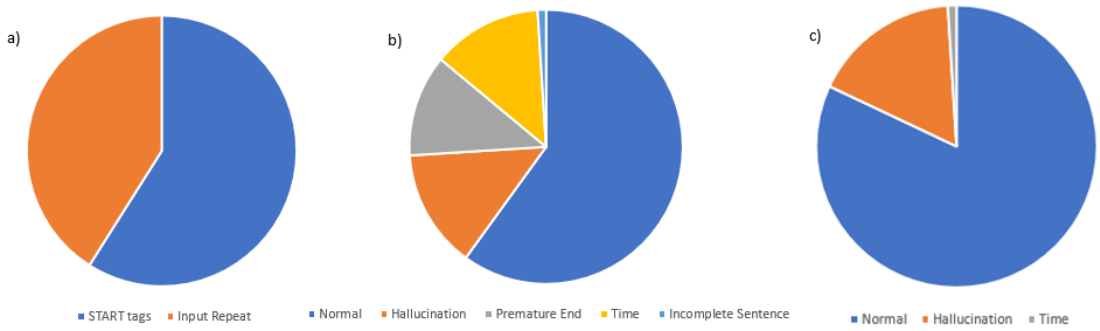

Figure 3: Qualitative analysis of 100 chatbot responses a) LLaMA, b) BART with history of 3, and c) COSMO models.

a good day." Thus, it performed much worse on the automatic evaluation which is looking at similarity between sequences of words.

Both models that had responses that were contextually relevant (BART and COSMO) had a major flaw - hallucinations. Since there was no data cleaning process in which the model would know to look up certain attributes about the user, the bot would often use the wrong name, or hallucinate an appointment that was coming up, or something similar. This can only be fixed by annotating the training data with this information.

6 Conclusion

This paper details the development of an NLP chatbot for Kaiser Permanente, aiming to enhance the personalized experience for its clients and improve overall customer satisfaction. By leveraging state-of-the-art NLP techniques, this paper aims to create a chatbot that could provide a natural and interactive experience for users, offering assistance with various tasks such as scheduling appointments, canceling memberships, and clarifying billing information.

The results of our work demonstrate the potential of transformer and LLM-based chatbots to support healthcare organizations like Kaiser Permanente in enhancing patient engagement, streamlining customer service, and providing personalized assistance to clients. Our approach offers a foundation for future research and development in the healthcare domain, particularly in the context of large-scale, comprehensive healthcare providers.

6.1 Future Work

6.1.1 Fine-Tuning the LLMs

While our LLM-based chatbots showed promising results, further optimization of these models could improve its performance. Due to constrained time and memory, LLMs were only trained with a history hyperparameter of 3. Training on different history amounts could enhance the chatbot's ability to generate more accurate and contextually relevant responses.

6.1.2 Evaluation in Real-World Settings

Conducting evaluations of the chatbot in real-world settings with actual users can provide valuable insights into its effectiveness in meeting client needs. User feedback and performance metrics can be collected and analyzed to refine the chatbot's capabilities and tailor it to better serve Kaiser Permanente's clients.

6.1.3 Integration with Health Records

Integrating the chatbot with Kaiser Permanente's electronic health record system would allow it to access patient data securely and provide more accurate and personalized healthcare advice. Augmentation of our training data with patient health records will also likely strengthen the model.

By addressing these areas of future work, our research can contribute to the ongoing efforts in using NLP technologies to improve healthcare experiences and outcomes for patients worldwide.

Acknowledgements

To be filled.

References

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Kavita A. Ganesan. 2015. Rouge 2.0: Updated and improved measures for evaluation of summarization tasks. *ArXiv*, abs/1803.01937.

I-Ching Hsu and Jiun-De Yu. 2022. A medical chatbot using machine learning and natural language understanding. *Multimedia Tools and Applications*, 81(17):23777–23799.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, Maarten Sap, and Yejin Choi. 2023. Soda: Million-scale dialogue distillation with social commonsense contextualization.

L. Laranjo, A. G. Dunn, H. L. Tong, A. B. Kocaballi, J. Chen, and R Bashir. 2018. Conversational agents in healthcare: a system review. *American Medical Informatics Association*, 25(9).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Lea Reis, Christian Maier, Jens Mattke, and Tim Weitzel. 2020. Chatbots in healthcare: Status quo, application scenarios for physicians and patients and future directions. In *European Conference on Information Systems*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A Appendix

To be filled.