International Journal on Engineering Technology and Infrastructure Development (InJET-InDev)
www.lec.edu.np/journal
Volume 2, Issue no. 2, Page 62-71
Received Date: 11/15/2025
Revised Date: 5/22/2026
Accepted Date: 5/30/2026
ISSN: 3021-999X (Print)
https://doi.org/10.3126/injet-indev.v2i2.95701

Fine-Tuning DialoGPT on Common Diseases in Rural Nepal for Medical Conversations

Birat Poudel[1], Satyam Ghimire[2], Prakash Chandra Prasad[3]

[1]*Department of Electronics and Computer Engineering, Thapathali Campus, IOE, TU, Kathmandu, Nepal, poudel.birat25@gmail.com*
[2]*Department of Electronics and Computer Engineering, Thapathali Campus, IOE, TU, Kathmandu, Nepal, satyamghimirestar@gmail.com*
[3]*Department of Electronics and Computer Engineering, Pulchowk Engineering Campus, IOE, TU, Kathmandu, Nepal, prakash.chandra@pccampus.edu.np*

Abstract

Conversational agents are increasingly being explored to support healthcare delivery, particularly in resource constrained settings such as rural Nepal. Large-scale conversational models typically rely on internet connectivity and cloud infrastructure, which may not be accessible in rural areas. In this study, we fine-tuned DialoGPT, a lightweight generative dialogue model that can operate offline, on a synthetically constructed dataset of doctor–patient interactions covering ten common diseases prevalent in rural Nepal, including common cold, seasonal fever, diarrhea, typhoid fever, gastritis, food poisoning, malaria, dengue fever, tuberculosis, and pneumonia. Despite being trained on a limited, domain specific dataset, the fine-tuned model produced coherent, contextually relevant, and medically appropriate responses, demonstrating an understanding of symptoms, disease context, and empathetic communication. The model achieved a perplexity score of 5.9632 and accuracy score of 0.1633, a major improvement compared to the baseline score of 0.0372. These results highlight the adaptability of compact, offline capable dialogue models and the effectiveness of targeted datasets for domain adaptation in low resource healthcare environments, offering promising directions for future rural medical conversational AI.

Keywords: Medical Dialogue, DialoGPT, Rural Healthcare, Offline Conversational AI, Common Diseases

1. Introduction

Rural healthcare in Nepal faces significant challenges due to limited medical infrastructure, scarcity of trained personnel, and constrained access to reliable information. Patients in these areas frequently seek guidance on common illnesses such as the common cold, seasonal fever, diarrhea, typhoid fever, gastritis, food poisoning, malaria, dengue fever, tuberculosis, and pneumonia. Traditionally, these queries are addressed by local healthcare workers or visiting doctors, but the increasing population and remoteness of some communities make timely, personalized support difficult to provide. This gap highlights the need for scalable, automated solutions that can assist patients with accurate and contextually relevant medical advice.

Pre-trained dialogue models such as DialoGPT (dialogue generative pre-trained transformer) can generate coherent responses in open domain settings, but their performance in specialized medical contexts is limited without domain specific adaptation. Moreover, rural healthcare contexts pose additional challenges: internet connectivity and electricity may be unreliable, making cloud-based solutions difficult to deploy. Lightweight, offline capable models offer a practical alternative for such low resource environments.

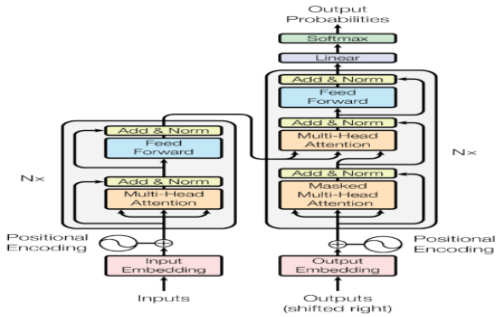

Figure 1. The Transformer-model architecture

Our approach to overcoming the scarcity of medical dialogue data with rural healthcare context is the generation of synthetic datasets. By constructing question–answer pairs that reflect typical doctor–patient interactions for common rural illnesses, it is possible to simulate realistic medical conversations. Synthetic data generation is cost effective, scalable, and enables the development of domain adapted dialogue systems even in settings with limited access to annotated corpora.

In this study, we fine-tune DialoGPT on a synthetically constructed dataset representing medical conversations about ten common diseases in rural Nepal. The goal is to enable the model to generate empathetic, medically accurate, and context aware responses while remaining suitable for offline deployment. Our work demonstrates the feasibility of adapting pre-trained conversational models to low resource healthcare environments, providing a pathway toward accessible and scalable medical dialogue systems in rural communities.

2. Literature Review

Recent progress in medical conversational AI systems focuses on the generation of specialized medical datasets supported by domain specific modeling techniques. A widely used foundation model for conversation is DialoGPT, a large scale generative pre-trained transformer trained on 147 million Reddit conversational exchanges; DialoGPT has become a common starting point for fine-tuning conversational systems because of its strong conversational fluency and context modeling (Zhang et al. 2019).

At the same time, a number of large medical dialogue corpora have been released that provide the domain coverage necessary for clinical fine-tuning and evaluation. MedDialog collected hundreds of thousands to millions of doctor-patient exchanges in English and Chinese and has been extensively used for training and transfer learning in medical dialogue research (Zeng et al. 2020). Complementing those broad collections, MedDG provides an entity centric dataset for gastrointestinal consultations with detailed entity annotations (symptoms, tests, medicines), and shows that explicitly modeling entities helps produce more medically relevant responses (Liu et al. 2022).

Methodological advances in medical dialogue modelling reflect the specific challenges of clinical conversations: limited labeled data, the need for explicit state tracking, and the requirement for factual correctness, etc. For instance, VRBot introduces semi-supervised variational reasoning with latent variables representing patient state and physician actions to improve reasoning under scarce labels (Li et al. 2021). ReMeDi offers multi-domain, multi-service dialogues with fine-grained annotations designed to benchmark task-oriented medical systems across services (e.g., diagnosis, triage, prescription, etc.) (Yan et al. 2022).

For downstream clinical tasks that draw directly on conversation context, datasets such as DialMed (dialogue-based medication recommendation) demonstrate how dialogue data can be used for concrete clinical decision tasks (He et al. 2022). The Dual-Flow / DFMed family of models argues for modeling both entity transitions and dialogue act flow to better capture how medical dialogues evolve turn by turn (Xu et al. 2023). Another active direction is knowledge grounded generation, where medical knowledge graphs or augmented knowledge sources are integrated with generative models to reduce hallucination and boost factuality (an approach shown to be effective in recent work on augmented-graph grounding) (Varshney et al. 2023). Parallel work on domain-specific pretraining has shown clear benefits for medical language tasks. Generative biomedical pretraining such as BioGPT improves biomedical text generation and QA relative to general models by pretraining on PubMed literature (Luo et al. 2022).

On the encoder side, models like ClinicalBERT and BioBERT, pretrained respectively on clinical notes and biomedical corpora, consistently improve downstream clinical NLP tasks (entity extraction, relation extraction, predictive tasks, etc.) and are commonly used as components or baselines in medical dialogue research (Huang et al. 2019; Lee et al. 2019). Finally, several surveys and meta-analyses synthesize progress across these areas and emphasize recurring evaluation and safety challenges, particularly the need for human-centered clinical evaluation, domain adaptation to low resource languages, and rigorous assessment of factuality and risk when models are deployed in real clinical or community settings (Valizadeh et al. 2022).

Extending beyond these foundational works, recent datasets such as RealMedDial have collected real tele-medical dialogues from short-video clips, enabling research on realistic doctor–patient interaction in online video settings (Xu et al. 2022). Low resource modelling methods such as Graph-Evolving Meta-Learning (GEML) enable

medical dialogue generation by transferring diagnostic experience via evolving disease–symptom graphs, addressing situations with limited labelled dialogue data (Lin et al. 2021). Knowledge enhanced frameworks like the work titled Knowledge Grounded Medical Dialogue Generation using Augmented Graphs integrate pre-trained language models with medical knowledge graphs to produce responses that are both fluent and clinically appropriate (Varshney et al. 2023). Terminology aware frameworks such as Terminology Aware Medical Dialogue Generation incorporate domain-specific terminology representation and recognition tasks into the generative process to bridge the gap between general language models and medical domain specificity (Tang et al. 2023).

Recent developments in multi-modal and enriched knowledge generation frameworks include KIMMDG: Knowledge Infused Multi-modal Medical Dialogue Generation, which integrate visual cues (e.g., skin lesions) and conversation context together with medical knowledge graphs for improved diagnosis relevant dialogue generation (Tiwari et al. 2024). Additional recent datasets such as MediTOD provide doctor–patient dialogues with annotated turns for task-oriented dialogue, enabling systematic research on medical history taking and physician–patient interaction (Saley et al. 2024). Knowledge enhanced extraction works such as A Knowledge Enhanced Two Stage Generative Framework for Medical Dialogue Information Extraction address the challenge of extracting term status pairs from medical dialogues using staged generative frameworks to better model relation and status inference (Hu et al. 2024). And newer works like MedKP: Medical Dialogue with Knowledge Enhancement and Clinical Pathway Encoding integrate external knowledge modules and clinical pathway encoding to reduce hallucinations and improve domain-specific generation performance (Wu et al. 2024).

3. Methodology

3.1. Synthetic Dataset Generation

Due to the lack of publicly available medical dialogue datasets in the Nepali rural context, a synthetic dataset was constructed to simulate realistic doctor–patient interactions. The dataset focused on ten common diseases frequently encountered in rural Nepal: common cold, seasonal fever, diarrhea, typhoid fever, gastritis, food poisoning, malaria, dengue fever, tuberculosis, and pneumonia.

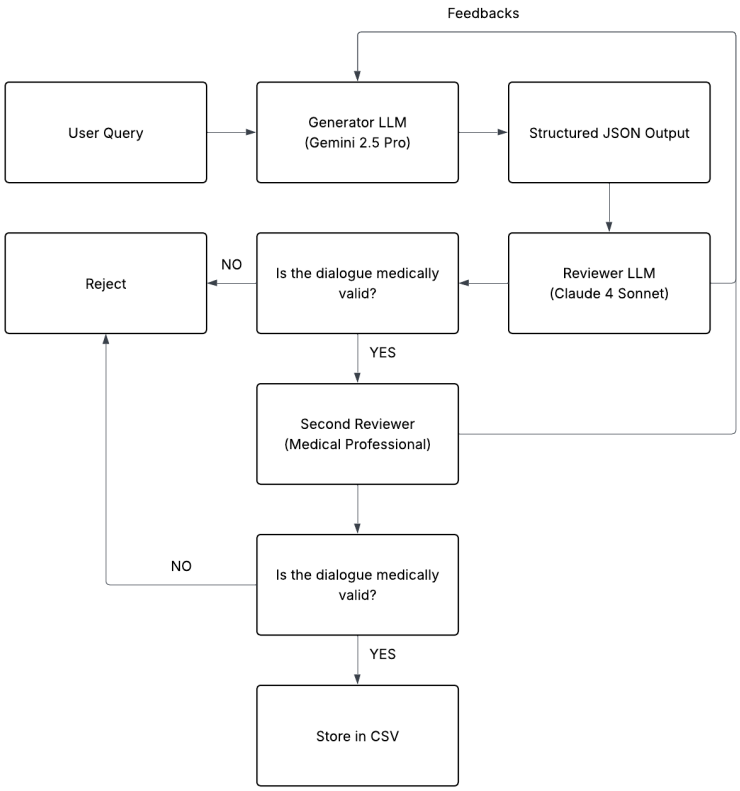

Figure 2. Synthetic Dataset Generation Block Diagram

A multi-stage feedback pipeline was developed for generating and validating synthetic doctor–patient dialogues. The process begins with user queries that are sent to the Generator LLM (Gemini 2.5 Pro), which produces structured JSON formatted dialogues. These outputs are then evaluated by the Reviewer LLM (Claude 4 Sonnet) to determine whether each dialogue is medically valid.

If a dialogue fails the medical validity check, it is rejected. Valid dialogues are forwarded to a second reviewer (a licensed medical professional) for further expert validation. Only dialogues confirmed as medically accurate by both the Reviewer LLM and the medical professional are stored in a CSV dataset.

Importantly, feedback loops are integrated at two stages. Both the Reviewer LLM and the medical professional provide feedback to the Generator LLM, enabling continuous refinement of generation quality and reduction of medically invalid outputs.

This architecture combines automated and human evaluation in a closed feedback system, ensuring that the final dataset is clinically accurate, consistent, and reliable.Each dialogue represents typical symptom descriptions and medically appropriate advice. The dialogues emphasized:

- Clear symptom reporting (e.g., “I have been experiencing fever and body aches for three days”).
- Providing empathetic medical responses while encouraging users to seek professional care when needed.

While the proposed system is intended for rural healthcare contexts in Nepal, the current dataset was developed in English primarily as a proof-of-concept for domain-specific medical dialogue generation. English was selected because DialoGPT is pretrained predominantly on English-language corpora, allowing more stable fine-tuning and evaluation within the limited scope of this study. However, the phrasing and symptom descriptions were intentionally adapted to reflect the communication patterns and healthcare concerns commonly observed in rural Nepali communities.

Additionally, the lack of large-scale publicly available Nepali medical dialogue datasets motivated the use of English synthetic data in this study. Future work will focus on developing multilingual or Nepali-language datasets and evaluating the model in more realistic community healthcare interaction settings. The dataset was divided into training (80%) and validation (20%) splits to fine-tune and evaluate model generalization.

3.2. Fine-tuning DialoGPT

We used DialoGPT-medium, a pre-trained conversational transformer model based on GPT-2 architecture, as the foundation for fine-tuning. The model was fine-tuned using the Hugging Face Transformers library. Each dialogue sample was formatted with alternating user and system turns to align with DialoGPT’s expected conversational input.

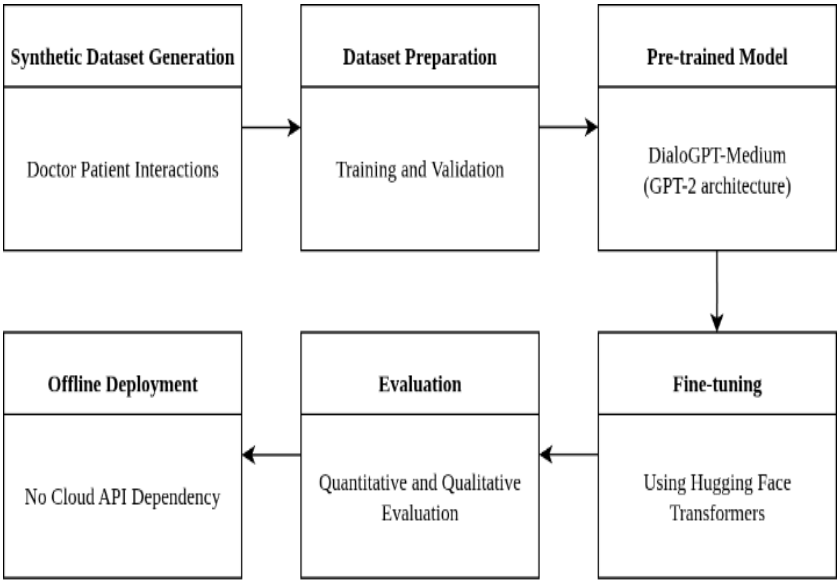

Figure 3. Fine-tuning DialoGPT Block Diagram

The training process aimed to adapt DialoGPT's general conversational ability to domain-specific medical dialogue, improving its contextual understanding of disease related symptoms and responses. Fine-tuning was conducted on a Google Colab T4 GPU, and the resulting model was exported for offline deployment, removing dependency on cloud APIs.

3.3. Evaluation Metrics

Model evaluation was conducted using a combination of quantitative and qualitative approaches to assess both linguistic quality and medical relevance of generated responses.

3.3.1. Quantitative Evaluation

Evaluation Loss: The evaluation (validation) loss, computed as the average cross-entropy between predicted and target tokens on the validation set, was monitored throughout training. A lower loss indicates improved learning and better model generalization.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N} logP(x_i \mid x_{<i}) \qquad \text{(Equation 1)}$$

Here Loss is the average cross-entropy loss per token, N is the total number of tokens in the sequence, x_i is the i^{th} target token, $x_{<i}$ are all previous tokens before x_i, and $P(x_i \mid x_{<i})$ is the predicted probability of the correct token.

Perplexity (PPL): Perplexity was derived from the exponential of the average cross-entropy loss, serving as a measure of the model's fluency and predictive confidence. Lower perplexity corresponds to more coherent and natural responses.

$$Perplexity = e^{Loss} \qquad \text{(Equation 2)}$$

Accuracy, Precision, Recall, and F1-Score: During training, these metrics were computed at the token level to evaluate the model's predictive performance in the causal language modeling framework.

Accuracy: Represents the proportion of tokens correctly predicted by the model compared to the reference response.

Precision: Measures the fraction of generated tokens that were correct, helping assess how often the model produced relevant words without unnecessary additions.

Recall: Indicates the fraction of reference tokens the model successfully predicted, reflecting its completeness in reproducing contextually appropriate responses.

F1-Score: The harmonic mean of precision and recall, providing a balanced evaluation of both correctness and coverage in token prediction.

These metrics collectively assess how effectively the fine-tuned DialoGPT model learns to generate accurate, fluent, and contextually aligned responses within medical dialogues.

3.3.2. Qualitative Evaluation

A human evaluation comprising 100 dialogue samples was conducted by healthcare professionals to assess the overall conversational quality of the fine-tuned model. The evaluation focused on three key dimensions:

Medical Appropriateness: The accuracy, reliability, and safety of the medical information or advice provided.

Empathy and Tone: The model's ability to express understanding, reassurance, and supportive communication.

Contextual Relevance: The extent to which responses appropriately addressed the patient's described symptoms and situation.

Each dimension was rated on a 5-point Likert scale, and the average scores were calculated to determine the model's overall conversational effectiveness.

3.4. System Methodology

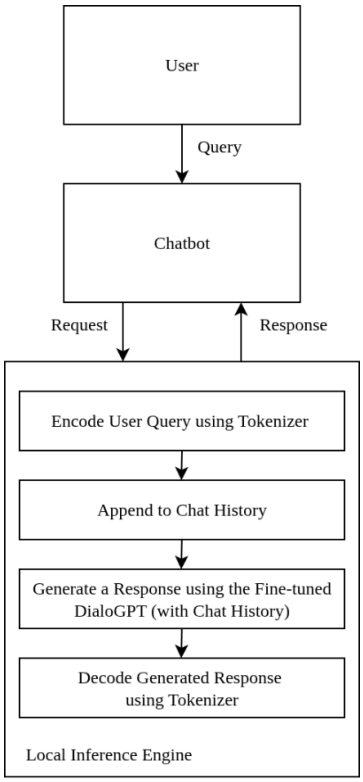

Figure 4. System Block Diagram

The proposed system is an offline conversational chatbot designed to generate human-like responses without requiring an internet connection. The overall architecture consists of two main components: the user interface and the offline inference engine. When a user inputs a query through the chatbot interface, the system processes the request locally. First, the input text is encoded using a pretrained tokenizer, which converts the query into a sequence of numerical tokens understandable by the model. These tokens are then appended to the existing chat history to maintain conversational context across multiple interactions. The combined input sequence is passed to a fine-tuned DialoGPT model, which generates an appropriate response based on the conversation history.

The model outputs tokenized text, which is subsequently decoded back into natural language using the same tokenizer. The generated response is then displayed to the user through the chatbot interface. Since all components, including the tokenizer and the fine-tuned DialoGPT model, are locally hosted, the system can operate fully offline, ensuring data privacy, low latency, and independence from external servers.

4. Results and Analysis

4.1 Fine-tuning DialoGPT

DialoGPT was fine-tuned on a synthetically constructed dataset of doctor–patient dialogues covering ten common diseases prevalent in rural Nepal. The model was trained to generate contextually relevant and medically appropriate responses while remaining lightweight enough for offline deployment. Training was performed using the Causal Language Modeling (CLM) objective with cross-entropy loss.

4.1.1. Perplexity Score

The model achieved a perplexity score of 5.9632, indicating that it generates fluent and coherent responses with relatively low uncertainty in next token prediction. Perplexity, derived as the exponential of the cross-entropy loss, provides a measure of how confidently the model predicts the next token in a sequence. Lower perplexity values suggest better language modeling performance.

4.1.2. Accuracy, Precision, Recall and F1-Score

Token level metrics were calculated to evaluate the model's predictive performance within the domain-specific medical dialogue setting. These results are compared against the pretrained DialoGPT-Medium baseline model without any fine-tuning. Both the fine-tuned and baseline DialoGPT-Medium models were evaluated on the same validation set to ensure a consistent and unbiased comparison. The results are summarized below:

Table 1. Accuracy, Precision, Recall, and F1-Score

Metric	DialoGPT-Medium (Baseline)	DialoGPT-Medium (Fine-tuned)
Accuracy	0.0372	0.1633
Precision	0.1591	0.1561
Recall	0.0372	0.1496
F1-Score	0.0576	0.1633

Although the token-level accuracy and F1-score remain relatively low in absolute terms, this behavior is consistent with the characteristics of open-ended natural language generation tasks, particularly in specialized medical dialogue systems. Unlike classification problems with fixed outputs, conversational response generation allows multiple semantically valid responses for the same prompt. Consequently, exact token-level matching metrics tend to underestimate the practical quality and contextual relevance of generated responses.

Furthermore, the study was conducted using a comparatively small synthetic dataset consisting of approximately 1,000 dialogue entries distributed across 10 disease categories. In low-resource settings, large transformer-based language models such as DialoGPT are unlikely to achieve high token overlap scores because the model must generalize linguistic structure, medical terminology, and conversational context from limited training examples. Despite this limitation, the fine-tuned model demonstrated substantial improvement over the pretrained baseline across all evaluation metrics, with accuracy increasing from 0.0372 to 0.1633 and F1-score increasing from 0.0576 to 0.1633. These improvements indicate that the model successfully adapted to the target medical dialogue domain.

Importantly, qualitative evaluation of generated responses showed that the fine-tuned model maintained conversational coherence and produced medically relevant and contextually appropriate replies for common rural healthcare scenarios. Therefore, while the numerical token-level metrics are modest, the observed relative improvement over the baseline and the model's ability to generate domain-consistent responses support the technical validity of the proposed approach.

4.1.3. Training and Evaluation Loss

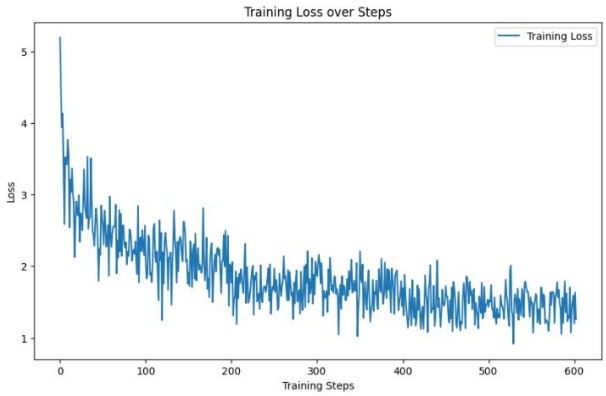

Figure 5. Training Loss

The above plot shows the model's training loss over 600 training steps. The loss is a measure of how wrong the model's predictions are on the data it's being trained on. As the model sees more data (i.e., as training steps increase), it adjusts its internal parameters to minimize its errors, causing the loss to decrease. The fluctuations (the spiky nature of the line) are normal, as the loss is typically calculated for each small batch of data.

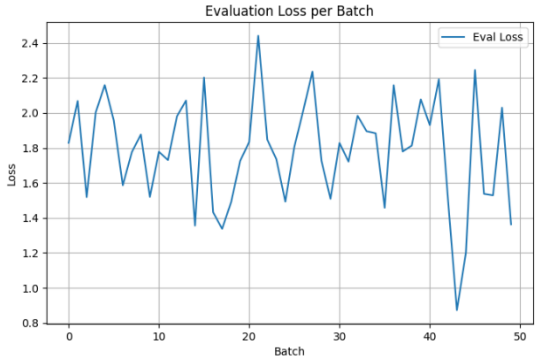

Figure 6. Evaluation Loss

This above plot shows the model's evaluation loss, calculated for each batch of a separate evaluation dataset that the model has not been trained on. This metric is crucial for understanding how well the model generalizes to new, unseen data. The loss fluctuates significantly from batch to batch, which is common for generative conversational models like DialoGPT.

4.1.4. Hyperparameters

The following hyperparameters were used during fine-tuning DialoGPT:

Table 2. Model Hyperparameters

Hyperparameter	Value
Learning Rate	5e-5
Optimizer	Adam
Batch Size	4
Epochs	3
Loss Function	Cross-Entropy Loss

4.2. Human Evaluation

A human evaluation was conducted on 100 dialogue samples by healthcare professionals to assess conversational quality across three dimensions: medical appropriateness, empathy and tone, and contextual relevance. Each aspect was rated on a 5-point Likert scale, and average scores were calculated. The evaluation confirmed that the model generates empathetic, context aware, and medically reasonable responses, highlighting its potential utility as a conversational assistant in rural healthcare settings.

Table 3. Human Evaluation Metrics

Metric	Mean Score
Medical Appropriateness	3.8
Empathy and Tone	4.1
Contextual References	3.9

5. Discussion

The findings of this study demonstrate the potential of lightweight domain-adapted conversational AI systems for healthcare assistance in low-resource settings. Although the quantitative token-level metrics remain modest, the fine-tuned DialoGPT model showed clear improvement over the pretrained baseline across all evaluation measures, indicating successful adaptation toward the medical dialogue domain.

These results align with broader trends in medical AI research, where transformer-based language models often require large, high-quality clinical datasets to achieve strong quantitative performance. In low-resource contexts such as rural Nepal, obtaining sufficiently large and ethically curated medical conversation datasets remains a

significant challenge. The use of synthetic data in this study therefore represents a practical compromise for early-stage experimentation and feasibility analysis.

Furthermore, this work highlights an important issue in healthcare AI for developing regions: the gap between state-of-the-art language models and real-world deployment constraints. Many existing medical conversational systems rely on cloud-based infrastructure, large computational resources, and English-centric datasets, which limit their applicability in rural and offline environments. In contrast, the proposed approach explores the feasibility of adapting a comparatively lightweight conversational model for localized healthcare assistance scenarios. The study also demonstrates that traditional token-level evaluation metrics alone may not fully capture the practical usefulness of medical dialogue systems. In conversational AI, especially within healthcare contexts, semantic relevance, contextual appropriateness, and linguistic coherence are often more important than exact token overlap. As a result, qualitative improvements in response relevance may not always correspond directly to high token-level accuracy scores.

However, several limitations remain. The dataset size was relatively small, consisting of only 1,000 synthetic dialogue entries across 10 disease categories. Additionally, the dataset was created in English due to the lack of publicly available Nepali medical dialogue corpora and the English-centric pretraining of DialoGPT. Consequently, the current system should be considered a proof-of-concept rather than a clinically deployable solution. Future work should focus on expanding the dataset, incorporating multilingual or Nepali-language dialogue generation, applying semantic evaluation metrics such as BLEU or BERTScore, and validating responses with healthcare professionals to improve reliability and practical usability.

6. Conclusion and Future Works

We fine-tuned DialoGPT on a synthetically constructed dataset representing medical conversations about ten common diseases in rural Nepal. The fine-tuned model demonstrated the ability to generate contextually relevant, medically appropriate, and empathetic responses, despite being trained on a limited dataset. Quantitative evaluation metrics, including perplexity, token level accuracy, precision, recall, and F1-score, alongside qualitative assessments by healthcare professionals, confirmed the model's potential to support rural healthcare delivery.

Since an English-only conversational interface would not be directly accessible to a large portion of the target rural population, the present work should be interpreted as an initial technical feasibility study rather than a fully deployable end-user healthcare solution. In practical deployment scenarios, the system would require integration with Nepali-language support, such as multilingual fine-tuning, transliteration, or speech-to-text and translation layers to improve accessibility for local users.

Acknowledgements

We would like to thank the medical professionals who generously dedicated their time and expertise to review and validate the synthesized doctor–patient dialogues used in this study. Their invaluable feedback ensured the clinical accuracy, safety, and reliability of the conversational data and model responses.

References

Abhisek Tiwari, Shreyangshu Bera, Preeti Verma, Jaithra Varma Manthena, Sriparna Saha, Pushpak Bhattacharyya, Minakshi Dhar, Sarbajeet Tiwari. 2024. "Seeing Is Believing! towards Knowledge-Infused Multi-modal Medical Dialogue Generation." Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) 14513–14523.

Bo Xu, Hongtong Zhang, Jian Wang, Xiaokun Zhang, Dezhi Hao, Linlin Zong, Hongfei Lin, Fenglong Ma. 2022. "RealMedDial: A Real Telemedical Dialogue Dataset Collected from Online Chinese Short-Video Clips." Proceedings of the 29th International Conference on Computational Linguistics 3342–3352.

Chen Tang, Hongbo Zhang, Tyler Loakman, Chenghua Lin, Frank Guerin. 2023. "Terminology-Aware Medical Dialogue Generation." IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1-5.

Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera & Asif Ekbal. 2023. “Knowledge-grounded medical dialogue generation using augmented graphs.” Scientific Reports 13: 3310.

Deeksha Varshney, Aizan Zafar, Niranshu Kumar Behera, Asif Ekbal. 2023. “Knowledge graph assisted end-to-end medical dialog generation.” Artificial Intelligence in Medicine 139.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen. 2021. “Semi-Supervised Variational Reasoning for Medical Dialogue Generation.” SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, Pengtao Xie. 2020. “MedDialog: Large-scale Medical Dialogue Datasets.” Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) 9241-9250.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin. 2022. “ReMeDi: Resources for Multi-domain, Multi-service, Medical Dialogues.” Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval 3013-3024.

Jiageng Wu, Xian Wu, Yefeng Zheng, Jie Yang. 2024. “MedKP: Medical Dialogue with Knowledge Enhancement and Clinical Pathway Encoding.” arXiv:2403.06611 .

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, Jaewoo Kang. 2019. “BioBERT: a pre-trained biomedical language representation model for biomedical text mining.” Bioinformatics 36 (4): 1234-1240.

Kaishuai Xu, Wenjun Hou, Yi Cheng, Jian Wang, Wenjie Li. 2023. “Medical Dialogue Generation via Dual Flow Modeling (DFMed).” Findings of the Association for Computational Linguistics: ACL 2023 6771-6784.

Kexin Huang, Jaan Altosaar, Rajesh Ranganath. 2019. “ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission.” Preprint arXiv.

Mina Valizadeh, Natalie Parde. 2022. “The AI Doctor Is In: A Survey of Task-Oriented Dialogue Systems for Healthcare Applications.” Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) 6638-6660.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, Tie-Yan Liu. 2022. “BioGPT: Generative Pre-trained Transformer for Biomedical Text Generation and Mining.” Briefings in Bioinformatics 23 (6).

Shuai Lin, Pan Zhou, Xiaodan Liang, Jianheng Tang, Ruihui Zhao, Ziliang Chen, Liang Lin. 2021. “Graph-Evolving Meta-Learning for Low-Resource Medical Dialogue Generation.” Proceedings of the AAAI Conference on Artificial Intelligence 13362-13370.

Vishal Vivek Saley, Goonjan Saha, Rocktim Jyoti Das, Dinesh Raghu, Mausam. 2024. “MediTOD: An English Dialogue Dataset for Medical History Taking with Comprehensive Annotations.” EMNLP.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, Xiaodan Liang. 2022. “MedDG: An Entity-Centric Medical Consultation Dataset for Entity-Aware Medical Dialogue Generation.” Natural Language Processing and Chinese Computing: 11th CCF International Conference 447-459.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. “DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation.” Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations 270-278.

Zefa Hu, Ziyi Ni, Jing Shi, Shuang Xu & Bo Xu. 2024. “A Knowledge-enhanced Two-stage Generative Framework for Medical Dialogue Information Extraction.” 21: 153-168.

Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, Jian Wu. 2022. “DialMed: A Dataset for Dialogue-based Medication Recommendation.” Proceedings of the 29th International Conference on Computational Linguistics 721-733.