From GPT to Open Models: Building Trustworthy AI Chatbots for Hypertension Care

Researchers from leading Italian institutions developed and tested two chatbot architectures using LLMs to support hypertensive patient self-management while ensuring data privacy. Their findings show that while GPT-based models offer superior performance, open-source alternatives like Mixtral and Alfred show strong potential for privacy-preserving applications.

CoE-EDP, VisionRI | Updated: 04-08-2025 10:07 IST | Created: 04-08-2025 10:07 IST

Representative Image.

SHARE

A multidisciplinary team from the University of Urbino, University of Bologna, University of Milano-Bicocca, and the Istituto Auxologico Italiano IRCCS has taken a significant step toward safer AI deployment in digital health. Their recent study, published in *Smart Health*, investigates how large language models (LLMs) can be effectively and ethically integrated into hypertensive patient self-management via chatbots. Led by researchers Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Michelangelo Ungolo, Martino Francesco Pengo, Gianluca Aguzzi, and Matteo Magnini, the work focuses on solving one of the biggest roadblocks to AI in healthcare: maintaining patient privacy without sacrificing conversational quality. They propose and test two alternative chatbot architectures, one using a hybrid GPT-assisted model and the other based on local deployment of open-source LLMs, each designed to protect sensitive data while delivering user-centric support for chronic disease management.

Designing for Empathy, Accuracy, and Security

The system developed by the researchers is built around four main modules: a chatbot interface for patient interaction, a natural language processing and generation (NLP&NLG) module to interpret and respond to user queries, a secure database for storing patient information, and a data processing unit that generates visuals and statistical summaries. A key feature of the architecture is its ability to categorize incoming messages into four intent types: Insertion, Request, Mood, and General, each corresponding to a specific interaction pathway. Insertion-type messages, containing sensitive medical or personal data, are never sent to third-party services. Requests, which ask for visual summaries or data insights, are filtered accordingly. Mood messages require emotional sensitivity and support, while General messages typically seek broad information and are routed carefully to avoid misinformation.

To test these workflows, the team developed a Telegram chatbot prototype named AI 4 HyperTension, tailored to Italian-speaking users. The chatbot engages patients by prompting them for regular blood pressure entries, visualizing trends, and offering motivational support. The study also emphasizes that the interface must be simple and intuitive, particularly given the diverse digital literacy levels among patients.

Comparing GPT with Open Models: A Tale of Two Approaches

In the first architecture, a hybrid solution was built using ML.NET 2.0, a machine learning library for C#, to classify user inputs and detect sensitive information. This classifier decides whether a message can be safely processed by OpenAI's GPT-3.5 Turbo. If the message is benign, it is sent to the model via API for generating responses. This strategy benefits from GPT's conversational strengths while filtering sensitive data before transmission. However, it depends heavily on the classifier's accuracy. A misclassified input, such as a health-related concern mistakenly labeled as a general question, can still lead to information leakage.

The second architecture avoids this risk altogether by running open-source LLMs locally on servers using platforms like *ollama*. Models tested include Mixtral, Mistral, and various versions of Llama2. This setup prevents any data from leaving the user's device or institutional infrastructure, making it more privacy-friendly. However, these open models require careful prompt engineering and often struggle with intent detection, parameter extraction, and multilingual performance.

How the Models Performed in Real-World Simulations

The study evaluated both approaches using a dataset of 128 simulated patient interactions. Results showed that ML.NET 2.0, when paired with GPT-3.5, outperformed all open-source alternatives in intent recognition, achieving an impressive 96% overall accuracy. Open models like Mixtral and Llama2 hovered around 74%, with noticeable limitations in processing Italian-language inputs. In handling requests, GPT-3.5 also demonstrated higher accuracy in identifying parameters like data type, time range, and desired output format. The open models were less consistent, particularly when interpreting ambiguous terms like “visualize,” which could imply either a graph or a text-based list.

Nevertheless, when responses were evaluated semantically using BERTScore, the differences narrowed. Open models like Alfred and Mixtral showed strong similarities to GPT-generated answers. These semantic results were further validated by internal medicine physicians, who reviewed 210 anonymized chatbot responses. The domain experts scored GPT-3.5 Turbo highest, but Alfred and Mixtral followed closely behind, receiving praise for clarity, empathy, and medical relevance. Their favorable reviews suggest that open models, while less precise in parsing, are not far behind in delivering meaningful and contextually appropriate interactions.

Balancing Privacy, Performance, and Scalability

The paper's findings highlight a crucial tension in AI-driven healthcare: balancing performance with privacy. The GPT-assisted solution, while superior in accuracy and user experience, comes with concerns about data leakage, subscription costs, API limits, and dependence on third-party services. Even with low misclassification rates, only one error out of 83 for insertion and one in 19 for mood, privacy breaches are considered unacceptable in medical contexts.

On the other hand, open-source models offer a scalable and privacy-compliant alternative that can operate without external dependencies. Although they currently require improvement in classification and data extraction, their performance in human-reviewed dialogue suggests they can become viable substitutes with further tuning and task-specific fine-tuning. The researchers recommend leveraging techniques like Retrieval Augmented Generation (RAG) and chain-of-thought prompting, as well as exploring larger, more capable models like LLaMA 3.1.

Looking ahead, the team plans to conduct clinical trials and usability studies to evaluate the chatbot's real-world effectiveness in improving patient adherence and engagement. Their work sets the stage for a new generation of AI tools that not only converse fluently but also care responsibly, offering privacy-preserving support to millions of patients managing chronic conditions.

FIRST PUBLISHED IN: Devdiscourse