

Synthetic Medical Notes: Bridging the Gap in Healthcare Data

Project Team: Sawiya Aidarus, Connor Holden, August Moses, Shashank Sinha
Computer Science
Project Number CS 25-334

Faculty Advisor: Preetam Ghosh, Ph.D.

Sponsor: Intelligent Care Solutions

Sponsor Mentors: Ford Sleeman, Rishabh Kapoor, and Josh Braunstein

Advancements in technology have introduced the possibility of medical analysis of patient data; however, extracting discrete data values from real clinical notes is a time-consuming process prone to errors and data leaks. Synthetic Note Generator aims to generate simulated, yet realistic, clinical notes to facilitate this process. Our software can generate notes with randomized values, specific given values, and ranges of values, offering distributed results to enhance the believability of synthetic notes. To ensure linguistic variability, we utilized Meta™'s Llama model for rephrasing, testing multiple Llama parameters to increase speech diversity while maintaining the integrity of key values within each generated note. Additionally, our team has prepared an interactive Turing test to gather physician feedback regarding how realistic the generated notes appear. The finalized project results consist of a Flask-based generation web tool and an HTML site containing the Turing test in a quiz-like format with a feedback text box. Conclusively, Synthetic Note Generator is fully functional in terms of mass generating varied notes without data manipulation, and our project is ready to move onto the physician-assessed Turing test. In the future, these generated notes will be utilized for AI model training, ultimately enabling data extraction from real clinical notes.

Keywords: Large Language Models, prompt engineering, Turing test, web integration

