Synthetic Medical Notes: Bridging the Gap in Healthcare Data

Team members: Connor Holden, Sawiya Aidarus, August Moses, Shashank Sinha | **Faculty adviser:** Preetam Ghosh, Ph.D. | **Sponsor:** Intelligent Health Solutions | **Mentor:** Ford Sleeman, Rishabh Kapoor, Josh Braunstein

Background

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. Our project addresses this challenge by creating a tool that generates synthetic medical notes – realistic but entirely artificial patient records.

These synthetic notes mimic the structure and content of real medical documents, particularly focusing on radiation oncology consults for prostate cancer patients. By providing a source of "fake" but medically accurate data, our tool enables:

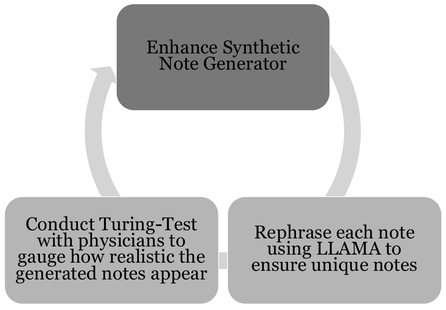
- Training of medical professionals without risking patient privacy
- Development and testing of healthcare software systems
- Medical research studies that require large datasets

Our innovative approach combines large language models with carefully crafted templates to produce notes that are indistinguishable from real ones yet contain no actual patient information. This project aims to accelerate medical research and improve healthcare practices while maintaining the highest standards of patient confidentiality.

Objective

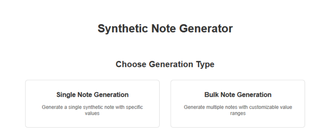
Our primary objective is to work with the pre-existing Synthetic Note Generator code to create a functional web-tool capable of generating realistic notes. Our note-generator will be able to rephrase sections of text and offer a wider variety of generated notes by utilizing *Meta*'s *Llama 3-7b* chatbot. After achieving consistently realistic and varied notes, our team has completed a clinical Turing-test to gauge how comparable our generated notes are to physician notes.

By enhancing the generated notes to be as believable and realistic as possible, we were able to later fine-tune a Large Language Model (LLM) to extract data from generated clinical notes on a large scale while not compromising fake patient data.



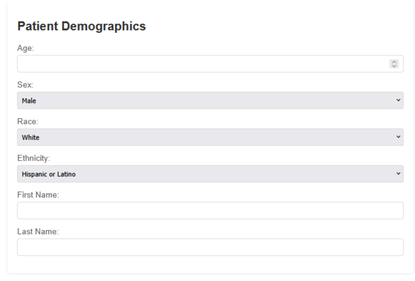
Note Generator & Web Tool

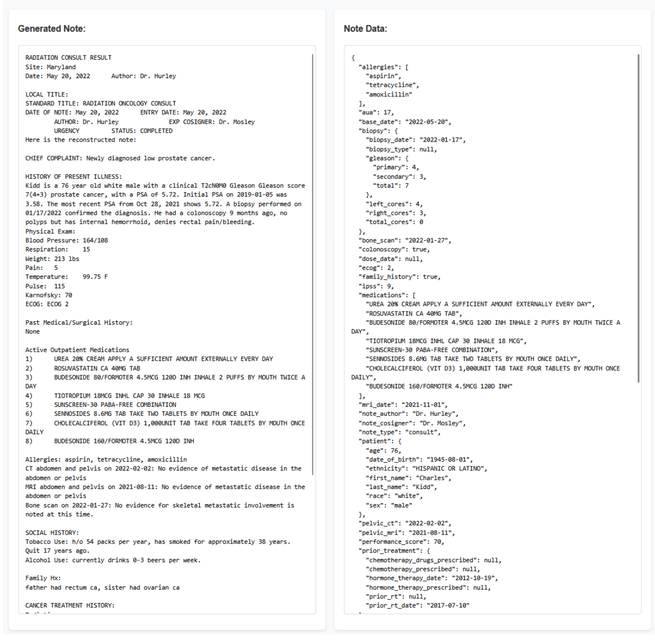
- **Web tool**: Developed and refined the web-based interface, ensuring usability and easy data entry for users.
  - Landing Page
  - Single Note Generation
  - Bulk Note Generation
    - Export functionality, LLM Rephrasing, Note Display, and Section inclusion/exclusion.
- **Note Types**: Implemented four new note types, incorporated internal feedback, and ensured seamless functionality throughout the app.
  - Initial Consultation, Follow Up, On-Treatment Visit, & Treatment Summary



Landing Page with Single or Bulk note Generation

Sample fields on Web Tool. Allows user to input specific values to be in the note.





LLM Rephrasing

- **LLM integration**: Focused on smooth integration with *Groq* API, creating a mapping system, and implementing validation tests to ensure that original note variables haven't been altered. This validation system promises that *Llama* keeps original "patient" information while increasing note variability.

Original Text:
Mr. {11} is a {1} year old {13} {2} with {18} risk prostate cancer, stage {17}. Initial PSA was {28} on {19}, most recently {3} on {14}. Biopsy on {16} showed Gleason {5}. {10}

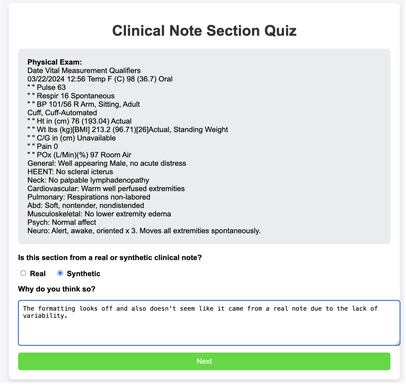
AI Rephrased Text:
Patient {11} is a {1} year old {13} {2} with high-risk prostate cancer, stage {17}.

Initial PSA: {28} on {19}.
Latest PSA: {3} on {14}.
Biopsy on {16} revealed Gleason {5}.
{10}

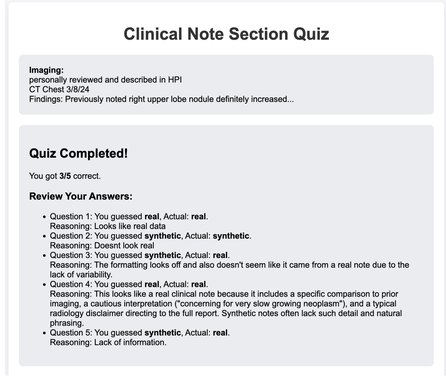
Turing Test

Turing Test: Implemented updates and testing cycles to enhance the tool's ability to pass a clinical Turing Test by generating realistic and coherent notes.

Test Screen with user selection and reason text box



User sees the results of their quiz at the end



Future Direction

In the future our team plans to share our code with physicians, request them to return Turing Test surveys, and analyze the results. Gaining direct feedback from physicians will allow us to reach our goal of creating clinical notes that are accurate, contextually relevant, and indistinguishable from both AI-generated and physician-authored notes. Additionally, our project will focus on expanding the synthetic note generator to support a broader range of cancer types, allowing for more comprehensive use across medical scenarios.