Synthetic Medical Notes: Bridging the Gap in Healthcare Data

Team members: Connor Holden, Sawiya Aidarus, August Moses, Shashank Sinha | **Faculty adviser:** Preetam Ghosh | **Sponsor:** College of Engineering | **Mentor:** Rishabh Kapoor

Background

In the world of medical research and healthcare improvement, access to patient data is crucial. However, real patient information is often restricted due to privacy concerns. 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 advanced 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 3-8b* chatbot. After achieving consistently realistic and varied notes, our team plans to complete 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 can 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.

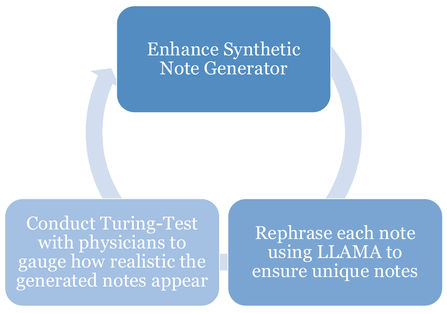


Figure 1: Ideal cycle of Synthetic Note Generator development

Areas of Work

We worked on broadening the tool's capabilities by adding support for new disease sites and types of medical notes. This includes integration and testing with Groq's LLM

- Web tool: Developed and refined the web-based interface, ensuring usability and easy data entry for users.

- Groq & LLM integration: Focused on smooth integration with Groq's LM, updating Python code and testing responses from the LLM to meet our requirements for generating synthetic notes.

- 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.

- Python code changes: Made necessary updates to new note types, incorporated internal feedback, and ensured seamless functionality throughout the app.

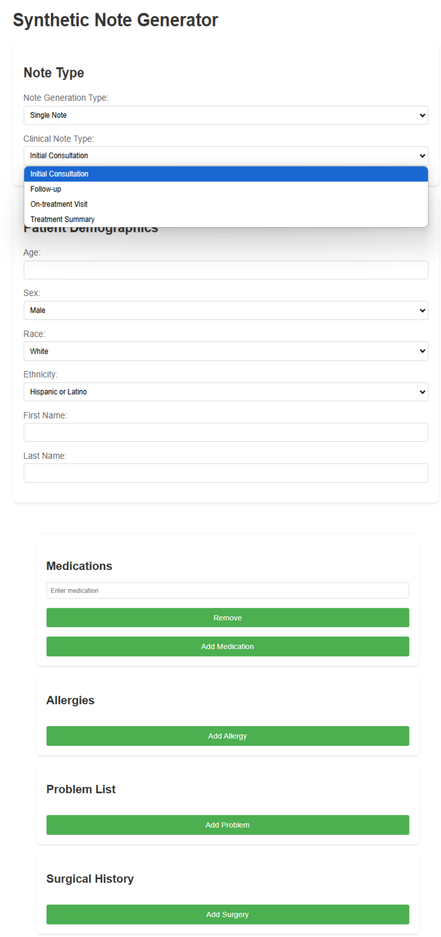
Generated Note:

```
RAD THERAPY CONSULT VISIT
Site: Nevada
Date: 2021-10-13        Author: Dr. Kline

LOCAL TITLE:
STANDARD TITLE: RADIATION ONCOLOGY CONSULT
DATE OF NOTE: 2021-10-13        ENTRY DATE: 2021-10-13
       AUTHOR: Dr. Kline              EXP COSIGNER: Dr. Allen
       URGENCY        STATUS: COMPLETED
A 58-year-old male, Mr Sturdivant, presented to discuss his newly diagnosed prostate cancer. He had undergone a PSA
screening, which revealed a PSA level of 9.0, leading to a referral to urology. A transrectal-guided biopsy was
performed on 01/15/2021, and the pathology report revealed a Gleason score 7(4+3).The left base was found to be the
site of the 2/6 left cores positive. Mr. Sturdivant scheduled a future appointment to discuss treatment options.
Physical Exam:
TEMP    PULSE    RESP    BP              PAIN    WT
96.13 F 68               21              137/90  1               294 lbs
Karnofsky: 90
ECOG: ECOG of 3

PAST MEDICAL
Computerized Problem List is the source of the following:
1: Sleep apnea
2: Atrial fibrillation
3: Depression
4: Glaucoma suspect
5: Malignant Neoplasm of the Prostate
6: Coronary arteriosclerosis
7: GERD
8: LT KNEE ARTHROSCOPY W / DEBRIDEMENT
9: Chronic back pain
10: Pelvic pain syndrome
11: Benign prostatic hyperplasia

Active Outpatient Medications
BUDESONIDE 160/FORMOTER 4.5MCG 120D INH
```



Future Direction

Looking ahead, 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. We also plan to refine and enhance the user interface to ensure it meets the needs and preferences of our client and end-users. A key objective in our development is to conduct a Turing-style test, gathering direct feedback from physicians who interact with the generated notes. This 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.

VCU College of Engineering