# Synthetic Note Generation and LLM Based Discrete Data Extraction Project

## Introduction

Although there is a significant amount of historical patient data in Electronic Medical Record (EMR) systems, most of this information is stored in free text notes. Getting discrete data values (demographics, diagnosis, outcomes, etc.) out of these notes currently requires a manual process that is time consuming and error prone. Tracking treatment outcomes for a specific patient population, contributing to registries or performing research is very difficult to do at scale without already having access to this embedded data. In this project, we plan to address some of these data extraction challenges with Large Language Models (LLM) and to improve inter-facility data sharing with a novel method of generating synthetic clinical notes.

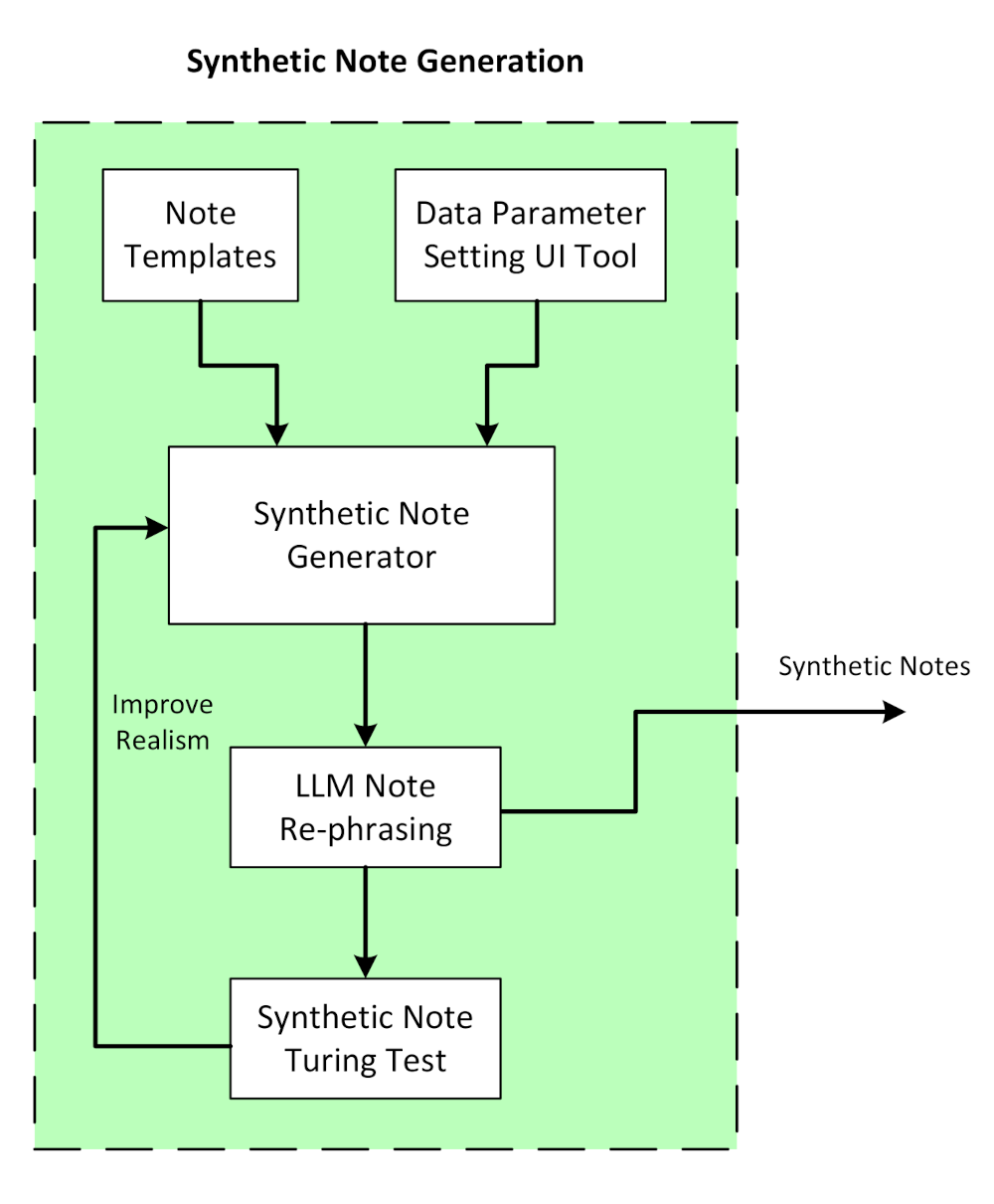
Popular foundation models like GPT-4 (OpenAI) or Llama 2 (Meta) have been shown to work well on many tasks but may fail to perform at a high level on more specific medical scenarios. Better results are often achieved when fine-tuning models but this requires a lot of training data which can be difficult to acquire. Individual healthcare providers may not have a lot of data that could be used for training, especially after the number of patients is filtered by the individual cancer, specific demographic parameters or clinical trial enrollment status. One solution for the lack of training data is to combine patient notes from multiple facilities, but concerns of Protected Health Information (PHI) exposure makes it difficult for healthcare providers to collaborate. Some of the privacy issues could be addressed by anonymization but doing it manually is too time consuming and automated tools often miss some PHI or strip out critical information.

Although synthetic clinical notes have been created using LLMs, they can still leak PHI if trained on real patient data. To avoid the risks with PHI, synthetic notes could be used for training instead. Cutting edge LLMs have shown success with medical text but can still produce unrealistic text, become repetitive or introduce fake information which reduces their utility. There have been some recent works on generating completely PHI-free synthetic notes but so far have only focused on individual sections or paragraphs of notes instead of the complete note text. These works have not presented approaches for ensuring that the underlying data is clinically appropriate.

In our initial work, we fine-tuned the ClinicalBioBERT model using 100,000 of our own synthetic radiation oncology notes to extract 27 discrete features. While the base model failed to extract any of these values, the fine-tuned model provided a 95% weighted F1 score across the 27 features. Because the synthetic notes were based on the format and content of real notes, the generated notes were representative of the real clinical notes. The main goals of this project is to significantly expand the synthetic note generation tool and develop an web application so that users can customize how notes are generated.

## Main Project Goals

As shown in Figure 1, this project is broken into three main phases. Although this appears to be a serial process, this work should be done in parallel as part of a cyclical training/validation process.

Figure 1: High level diagram of the note generation process

### Phase 1: Basic Synthetic Note Generation

#### Initial Work

Our process to generate synthetic notes starts with manually curated templates from radiation oncology encounter notes which we have access to through our association with the the Veterans Health Administration (VHA). In a radiation oncology consult note, there is almost always a History of Present Illness (HPI) section which provides a basic summary of the patient’s demographics information, initial staging of the cancer and why they are being seen in the clinic that day. Here is an example of a HPI section with the patient specific information swapped with placeholders:

*Mr. {last\_name} is a {age} y/o {race} {sex} who presented to Urology with elevated PSA of {psa\_score} drawn on {psa\_date}. He underwent a {biopsy\_type} with pathology on {biopsy\_date} showing {gleason}. {colonoscopy} Diagnosis: {tnm}, {risk} prostate cancer. The patient is now referred for evaluation for definitive radiation*

*therapy.'*

We initially collected ten such HPI patterns to represent a wider variability of how VHA physicians write their notes. By combining the HPI with other common sections such as vitals, social history, medical list, physical exam and imaging, we can construct the overall structure of a radiation oncology consult note. Each synthetic note is then created by randomizing which version of each note section template to include, therefore providing a very large distribution of overall note structures. Finally, the placeholders in each note are replaced with randomized values, completing the note creation process. All of the discrete placeholder values are also stored with the free text note which can then be used for the LLM fine-tuning validation.

#### Current Tasks

The initial synthetic note generation code has been developed using Python and is available on GitHub <https://github.com/bnetlab/synthetic-note-generator>. Our goal for the Phase 1 work is to greatly expand the functionality of the system as described in these steps:

1. **Support Multiple Types of Clinical Notes**

The initial synthetic note template system only used radiation oncology prostate consult notes but will need to support other types of clinical notes used during cancer treatment such as the on treatment visit, treatment summary, and follow up. This will require creating additional templates that will have new sections but will also use some of the ones already created (vitals, medication, etc.).

1. **Support an Additional Disease Site**

The initial system only covered prostate cancer but we will need to include at least lung cancer. This will require additional sub-templates for lung cancer and should be designed to handle more cancer types in the future.

1. **Incorporate LLM to Re-Phrase Text**

One limitation of the static template system is that it will still lack the level of variability likely found across a large number of clinical notes. To address this issue, we propose using existing pre-trained models (GPT-4/Llama 2) to rephrase sections of synthetic note. However, this rephrasing must still contain the original meaning and keep all of the inserted discrete feature values. These pre-trained models can be prompted to generate custom styled tuning to mimic the note-writing styles of specific medical provider or institution, ensuring that rephrasing looks consistent with how clinical notes are structured across various healthcare settings. We can work on building of control parameters that can provide the users with the ability to control the degree of rephrasing, allowing for slight or significant changes based on their preference for variability.

1. **Prepare Data for a Clinical Turing Test**

To validate that the generated synthetic notes are truly representative of real clinical notes, we would like to run a Turning Test style experiment. Both real and synthetic notes will be shown to physicians and see if they can tell the difference. In cases where they correctly detect synthetic notes, we will ask the physicians to provide feedback on what tipped them off so the synthetic note generation process can be improved.

1. **Web Based Configuration Tool**

Right now, the ranges or valid values that could be inserted in the notes is hard coded. To make this process more extendable, we need a tool that users can specify the data ranges they want to include in their notes and what kind of disease sites to target.

1. **Chronological Synthetic Note Generation for an episode of care**

Using the synthetic note data generation program, we can create a series of sequential notes that follow the progression of care for a typical cancer patient (e.g. prostate cancer). Starting with the initial consult, the system will generate a note capturing demographics, diagnosis, patient's history of present illness, prognosis and plan of care. As treatment progresses, the program will generate on-treatment visit notes that document changes in the patient's condition, including updates to vitals, side effects, and responses to radiation therapy. After treatment, the program can produce a treatment summary and follow-up notes, reflecting the patient's recovery and long-term management plan. By randomizing values and rephrasing sections, the synthetic notes provide a realistic, coherent narrative of the patient's care while maintaining clinical relevance and consistency across notes.