A picture containing graphical user interface

Description automatically generated

25-334: Large Language Models (LLM) for data extraction from clinical notes

Project Proposal

Prepared for

Rishabh Kapoor/Preetam Ghosh

VCU College of Engineering

By

Sawiya Aidarus, Connor Holden, Shashank Sinha, and August Moses

Under the supervision of

Preetam Ghosh

12/05/24

**Executive Summary**

This proposal outlines the Synthetic Note Generation and Large Language Model (LLM) Based Discrete Data Extraction project, which aims to address the challenge of extracting discrete data from free-text clinical notes. Current electronic medical records contain a wealth of unstructured data, but obtaining structured data (such as demographics, diagnosis, and treatment outcomes) from these notes is a manual, time-consuming process. The goal of this project is to build a robust tool for generating synthetic clinical notes, which can be used to train AI models for data extraction while avoiding issues related to protected health information (PHI).

This report covers the project's problem statement, engineering design requirements, scope of work, timeline, and team contract. Key deliverables include an expanded note generation system and a functional web tool for generating and customizing synthetic clinical notes. The milestones reflect a phased approach to development, and resources such as the Python-based system and available clinical note data are already in place.

With the successful completion of this project, we anticipate a significant improvement in the ability to train AI models for extracting data from clinical notes, while ensuring patient privacy and facilitating inter-facility collaboration.

**Table of Contents**

[Section A. Problem Statement 5](#_Toc176438097)

[Section B. Engineering Design Requirements 7](#_Toc176438098)

[B.1 Project Goals (i.e. Client Needs) 7](#_Toc176438099)

[B.2 Design Objectives 7](#_Toc176438100)

[B.3 Design Specifications and Constraints 8](#_Toc176438101)

[B.4 Codes and Standards 9](#_Toc176438102)

[Section C. Scope of Work 11](#_Toc176438103)

[C.1 Deliverables 11](#_Toc176438104)

[C.2 Milestones 12](#_Toc176438105)

[C.3 Resources 12](#_Toc176438106)

[Section D. Concept Generation 13](#_Toc176438107)

[Section E. Concept Evaluation and Selection 14](#_Toc176438108)

[Section F. Design Methodology 16](#_Toc176438109)

[F.1 Computational Methods (e.g. FEA or CFD Modeling, example sub-section) 16](#_Toc176438110)

[F.2 Experimental Methods (example subsection) 16](#_Toc176438111)

[F.5 Validation Procedure 16](#_Toc176438112)

[Section G. Results and Design Details 18](#_Toc176438113)

[G.1 Modeling Results (example subsection) 18](#_Toc176438114)

[G.2 Experimental Results (example subsection) 18](#_Toc176438115)

[G.3 Prototyping and Testing Results (example subsection) 18](#_Toc176438116)

[G.4. Final Design Details/Specifications (example subsection) 18](#_Toc176438117)

[Section H. Societal Impacts of Design 20](#_Toc176438118)

[H.1 Public Health, Safety, and Welfare 20](#_Toc176438119)

[H.2 Societal Impacts 20](#_Toc176438120)

[H.3 Political/Regulatory Impacts 20](#_Toc176438121)

[H.4. Economic Impacts 20](#_Toc176438122)

[H.5 Environmental Impacts 21](#_Toc176438123)

[H.6 Global Impacts 21](#_Toc176438124)

[H.7. Ethical Considerations 21](#_Toc176438125)

[Section I. Cost Analysis 22](#_Toc176438126)

[Section J. Conclusions and Recommendations 23](#_Toc176438127)

[Appendix 1: Project Timeline 24](#_Toc176438128)

[Appendix 2: Team Contract (i.e. Team Organization) 25](#_Toc176438129)

[Appendix 3: [Insert Appendix Title] 26](#_Toc176438130)

[References 27](#_Toc176438131)

### **Section A. Problem Statement**

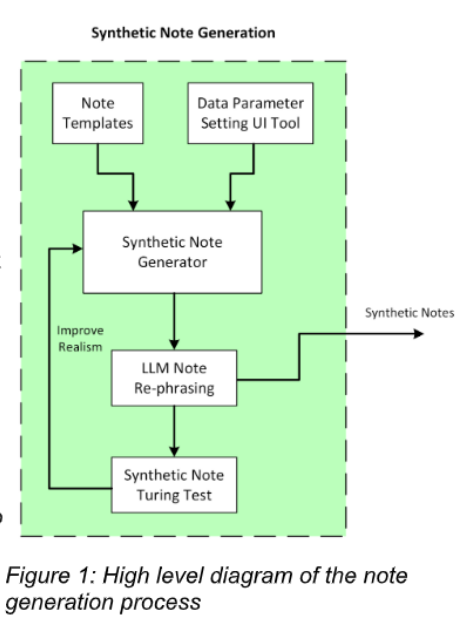
The Synthetic Note Generation and LLM-Based Discrete Data Extraction Project addresses a critical gap in the field of radiation oncology, where clinical documentation is both time-consuming and prone to errors. Electronic Medical Record (EMR) systems store vast amounts of patient data, much of which exists as free-text clinical notes. These notes, though essential for documenting patient encounters, present significant challenges when extracting discrete data—such as treatment outcomes and key clinical metrics—that are crucial for research, registries, and patient care improvement.

Manual data extraction from free-text notes is not only labor-intensive but also error-prone, leading to inefficiencies in healthcare delivery and difficulties in scaling data extraction across large healthcare institutions. According to a systematic review by Lee et al. (2020), errors in clinical notes, such as mis recorded dose calculations or incorrect treatment site documentation, can result in serious consequences, including treatment delays, suboptimal care, and even legal implications. These errors often arise from time pressures, lack of standardized documentation, and poorly designed EMR systems.

The project's overarching goal is to revolutionize clinical note generation in radiation oncology by developing a method to generate synthetic clinical notes that resemble real patient data but contain no Protected Health Information (PHI). This approach allows for safe and efficient fine-tuning of Large Language Models (LLMs) to automatically extract discrete data, bypassing the risks of PHI exposure. The synthetic notes will also enable cross-institutional data sharing without compromising patient privacy.

Current efforts to automate clinical note generation and data extraction have been promising but remain limited in scope. For instance, the fine-tuned ClinicalBioBERT model has achieved a 95% weighted F1 score in extracting 27 discrete features from synthetic radiation oncology notes. However, this project aims to push beyond existing limitations by expanding the synthetic note generation tool, fine-tuning LLMs with a broader range of clinical notes, and developing a web-based application for customizable note generation.

This project addresses the unmet engineering need for more accurate and efficient clinical documentation and data extraction, with significant implications for enhancing patient care, reducing provider burden, and improving research outcomes in radiation oncology. By advancing the use of machine learning in healthcare, this project aims to establish new standards in clinical documentation practices.



**Figure 1. High Level diagram of the note generation process.**

### **Section B. Engineering Design Requirements**

This section describes the goals and objectives of the project, as well as all **realistic constraints** to which the design is bound. It is meant to provide a structure that helps to formulate the problem. Design requirements are often derived from client or stakeholder needs. They may consider benchmarking against or improving on currently available solutions, providing novel techniques or design solutions, integration with existing components, systems, or equipment, required codes and standards, general observations of the problem space, etc. Describe how the requirements provided below were researched and decided upon. Common design requirements often include considerations of the design efficacy, cost, safety, reliability, usability, and risk, among others.roposal, but is required for all subsequent reports. This section should be comprehensive and thorough, requiring a significant research effort.

#### **B.1 Project Goals (i.e. Client Needs)**

Our project follows a typical software design project including steps such as developing functional code, creating a web-based interface, and data preparation for testing and feedback. Each of these steps has one or more associated goals, however, the primary goal of our work is to enhance the functionality of the preexisting synthetic note-generator system.

* To create additional note templates to support multiple types of clinical notes and disease sites
* To parameterize LLMs to increase the accuracy of text-rephrasing and generation
* To implement a user-friendly web-tool that can be used by our clients
* To prepare data for a Clinical Turing Test to provide feedback for higher product accuracy
* To implement code that will generate a sequential episode of care

#### **B.2 Design Objectives**

Our project objectives are as follows:

* The design will consist of a larger variety of patient note templates that can be generated online with the help of the Groq API.
* The design will contain more note templates than previously
* The design will possibly utilize a small amount of project funds for our Groq API key.
* The design will have a prototype web-tool utilizing the updated synthetic note-generator by the end of the semester

#### **B.3 Design Specifications and Constraints**

Our current specifications are generally focused around updating existing code to generate a larger variety of data and creating a fully functional web-tool that allows users to change note generation parameters. While unsure of how many note templates or web parameters will be ideal, we have decided on a few numbers to aim for.

Project specifications are as follows:

* Design should have at least 5 *new* templates within the Synthetic Note Generator code
* Design should have a functional web-tool that allows users to adjust at least *3 parameters* for note generation, including choosing the note template itself

Our most unavoidable constraint is working with pre-existing code. While the existing code is currently helpful, it may produce roadblocks later and lead us to work around annoyances that cannot be changed. Other primary constraints include factors such as funding and computational costliness. It’s essential for us to factor in the budget of our stakeholders and decide which elements they must pay for in the long run such as hosting a server or using an API subscription. The efficiency of our design can increase long-term costs in certain scenarios; therefore, we may need to format our design around an outside budget.

Project constraints include:

* Design must not let web users over-use Groq API since it is priced at tokens per second. Funds must not exceed $1,000 for our development process.
* Design for our web-tool must be user-friendly and allow for user specification.
* Design, if implemented locally, should work efficiently on a variety of machines such as Windows, Linux, and macOS.

#### **B.4 Codes and Standards**

The design and development of the Synthetic Note Generation Tool for clinical notes must adhere to several relevant codes and standards to ensure quality, safety, privacy, and compliance with best practices. These codes and standards will guide the technical and ethical aspects of the project, particularly in relation to web development, data privacy, and the handling of synthetic clinical data.

##### **Standards**

1. **IEEE Standard 1012-2016 – Systems and Software Verification and Validation**

This standard ensures that the software development process includes proper verification and validation steps, ensuring that the web tool functions as intended and meets the project's requirements. We will adopt this standard to guide our software testing and validation procedures, ensuring a reliable and bug-free user experience.

1. **W3C Standards for Web Development (HTML5, CSS, and JavaScript)**

The World Wide Web Consortium (W3C) sets global standards for web development to ensure accessibility, interoperability, and performance. Our web-based note generation tool will comply with the latest W3C standards for HTML5, CSS, and JavaScript to ensure cross-browser compatibility and adherence to best practices in web design. This will help ensure that the tool can be used across different platforms and devices without compatibility issues.

1. **ISO/IEC 27001:2013 – Information Security Management Systems (ISMS)**

Since the tool handles synthetic clinical data, it is essential to adhere to information security standards. ISO 27001 provides guidelines for maintaining security best practices, including risk management and access control. Though the data is synthetic and non-identifiable, this standard will ensure that our system is secure and protected from unauthorized access, safeguarding sensitive medical formats and structures.

1. **ACM Code of Ethics and Professional Conduct**

As members of the computing profession, our team will adhere to the ethical standards outlined by the ACM Code of Ethics. This involves ensuring that our design respects privacy, acknowledges data sensitivity, and promotes transparency in how synthetic notes are generated and managed. We will ensure that no synthetic notes resemble real patient data and maintain high ethical standards in the use of LLMs and data generation.

1. **NIST SP 800-53 – Security and Privacy Controls for Federal Information Systems and Organizations**

Since we are working with synthetic clinical notes and may need to interface with sensitive healthcare systems in the future, following the NIST guidelines ensures that our tool meets robust privacy and security controls. This is especially important as we design the system with privacy in mind, preventing PHI exposure even with synthetic data.

##### **Codes**

1. **HIPAA (Health Insurance Portability and Accountability Act) Privacy Rule**

Although the tool generates synthetic notes, it must still respect the regulations outlined in HIPAA, ensuring that the system does not accidentally expose any real PHI (Protected Health Information). Any data-handling process in the project must account for the legal and ethical constraints of HIPAA, particularly since our work could inform future developments involving real patient data.

1. **OSHA 1910.120 – Safety and Health Regulations for IT Environments**

While this code primarily deals with workplace safety, it is relevant to ensuring that the team follows safe IT practices while developing the tool. This includes ergonomic safety, electrical safety for computing environments, and proper handling of electronic components involved in the project.

Incorporating these codes and standards into the design ensures that the project aligns with industry best practices, legal requirements, and ethical standards. It also helps to guide the development process, ensuring that our software is secure, reliable, and designed with a focus on data integrity and privacy.

### **Section C. Scope of Work**

The Synthetic Note Generation and LLM-Based Discrete Data Extraction project focuses on building a system for generating synthetic clinical notes and extracting discrete data using large language models (LLMs). This project will expand the existing note generation tool to support various types of clinical notes and disease sites, initially focusing on prostate and lung cancer. LLMs will be integrated to rephrase notes, adding variability while maintaining clinical accuracy. Additionally, a web-based interface will be developed to allow users to customize note generation parameters. The project will follow an Agile methodology, allowing for flexibility and continuous validation. Clear milestones and deliverables will be set to prevent scope creep, with regular communication with sponsors and advisors to ensure objectives are met on time and that the project stays within the budget. The key deliverables include the note generation system, an LLM-based extraction tool, and clinical validation through a Turing test.

#### **C.1 Deliverables**

In order to meet the project requirements given to the team by the sponsors, our team is choosing an Agile methodology to approach the project requirements. By setting specific deliverables at the end of each sprint, we will ensure continuous progress and flexibility in adapting to changes. This iterative approach allows for regular feedback from the project sponsor and faculty advisor, enabling us to refine the solution incrementally while minimizing risks. It also helps maintain clear communication, keep the project on track, and ensure that all objectives are completed on time and within budget.

* The synthetic note generator must be capable of handling various types of clinical notes. Currently, the system is limited to radiation oncology prostate consult notes, but it will need to expand to include other clinical note types used in cancer treatment, such as on-treatment visits, treatment summaries, and follow-ups.
* The synthetic note generation system will be expanded to support additional disease sites beyond the initial focus on prostate cancer. The system will incorporate lung cancer, requiring the creation of sub-templates tailored to lung cancer-specific clinical notes. Furthermore, the design will be flexible to accommodate more cancer types in the future, ensuring scalability and adaptability as the system evolves.
* To address the limitation of the static template system's lack of variability, we will incorporate pre-trained models like Groq to rephrase sections of the synthetic notes. This rephrasing will maintain the original meaning and preserve all discrete feature values, while adding variability to mimic the writing styles of specific medical providers or institutions. Users will also have control over the degree of rephrasing, enabling them to adjust the variability from minor tweaks to significant changes, ensuring consistency and flexibility across different healthcare settings.
* To ensure that the synthetic notes generated are truly indistinguishable from real clinical notes, we will conduct a Turing Test-style experiment. In this test, both real and synthetic notes will be presented to physicians, who will attempt to identify which notes are synthetic. If they correctly identify synthetic notes, we will gather feedback from them on any notes they have about what allowed them to figure out the notes were synthetic, using this insight to analyze and improve the note generation process.
* To make the synthetic note generator available more to the public, we will develop a web-based configuration tool that allows users to define the data ranges they wish to include and select specific disease sites to target. This will make the system more extendable and user-friendly, giving users greater control over the customization of generated notes.
* The system should also be able to generate a series of sequential notes following a cancer patient’s care, from initial consultation through treatment and recovery. Starting with demographics, diagnosis, and care plans, it will create on-treatment notes documenting vitals, side effects, and responses to therapy. After treatment, follow-up notes will summarize recovery and long-term care. Randomized values and rephrasing will ensure the notes form a realistic, coherent narrative while maintaining clinical accuracy.

Almost all of the deliverables identified above can be done remotely, as all of them require research on the team’s side, while being able to effectively collaborate with the sponsors to ensure all the specific requirements are achieved. For specifically the Turing Test-style experiment, this will require interactions with different physicians, either virtually or in-person, in order to gather notes about their view on the synthetic or clinical notes shown to them.

#### **C.2 Milestones**

|  |  |  |
| --- | --- | --- |
| **Milestones** | **Expected Completion Date** | **Summar of milestone** |
| Create the Web Based Configuration Tool | November 2, 2024 | This task involves creating the web interface that users will interact with in order to create synthetic notes where they can specify the data ranges they want to include in their notes and what kind of disease sites to target. |
| Incorporate LLM to Rephrase Text | October 26, 2024 | Coming up with the proper prompt for an LLM, such as Groq, such that the model can rephrase yet maintain the original meaning of a clinical note and the different sections. |
| Support an Additional Disease Site | October 26, 2024 | Currently, the system only covers prostate cancer and will require to include at least lung cancer. This will require additional templates for lung cancer and should be designed in a way to handle more cancer types in the future. |
| Prepare data to create a Clinical Turing-test tool | October 26, 2024 | In order to validate the synthetic clinical notes that are being generated, we want to be able to see if physicians, while being shown a real and synthetic note, are able to tell the difference between the two. We want to determine what method will be most effective in asking physicians these questions, and using the feedback in order to update the current clinical note templates. |

#### **C.3 Resources**

In order to successfully complete the Synthetic Note Generation and LLM-Based Discrete Data Extraction project, several resources will be required. The Groq platform will be used for paraphrasing synthetic clinical notes, and will involve future costs for integrating it into the final web interface. To develop the web tool, frameworks such as Flask or Django will be necessary for the backend, along with front-end libraries like React to ensure a user-friendly interface. Cloud hosting services, such as AWS, Google Cloud, or Azure, may be needed in the future in order to deploy the interface for public access. A robust database system, either a relational or non-relational database, will be required to store generated notes and user inputs. Additionally, APIs for Groq integration and machine learning libraries for model fine-tuning will be critical. Version control systems like Git and GitHub will also be essential to manage the codebase and facilitate team collaboration. These resources will ensure the project meets its goals efficiently.

### **Appendix 1: Project Timeline**

Gantt Chart of project timeline:

<https://view.monday.com/embed/7620149158-8f2e43d20a3c65545058a163c0c870d5?r=use1>

<iframe src="https://view.monday.com/embed/7620149158-8f2e43d20a3c65545058a163c0c870d5?r=use1" width=770 height=500 style="border: 0; box-shadow: 5px 5px 56px 0px rgba(0,0,0,0.25);"></iframe>

### **Appendix 2: Team Contract (i.e. Team Organization)**

# Step 1: Get to Know One Another. Gather Basic Information.

**Task:** This initial time together is important to form a strong team dynamic and get to know each other more as people outside of class time. Consider ways to develop positive working relationships with others, while remaining open and personal. Learn each other’s strengths and discuss good/bad team experiences. This is also a good opportunity to start to better understand each other’s communication and working styles.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Team Member Name*** | ***Strengths each member bring to the group*** | ***Other Info*** | ***Contact Info*** |
| Sawiya Aidarus | Good with organization, good with communication, and planning things strategically. | *I’m looking forward to learning new skills and creating a project with my peers.* | [Aidarussa@vcu.edu](mailto:Aidarussa@vcu.edu)  612-478-2257 |
| Connor Holden | Good with organization, quick learner, and acquainted with professional environment. | *I’m excited to learn more about LLMs and apply my experiences.* | [Holdencj@vcu.edu](mailto:Holdencj@vcu.edu)  571-287-3963 |
| Shashank Sinha | Very flexible, can manage time very well, and previous experience. | *I’m looking forward to learning new skills and working on a real-world project.* | [Sinhas6@vcu.edu](mailto:Sinhas6@vcu.edu)  757-271-2877 |
| August Moses | Previous experience with topic, being able to outline project deliverables and meet them. | *I’m looking forward to learning how to meet the needs of the stakeholder.* | [Mosesa3@vcu.edu](mailto:Mosesa3@vcu.edu)  540-645-8564 |

|  |  |  |
| --- | --- | --- |
| ***Other Stakeholders*** | ***Notes*** | ***Contact Info*** |
| *Faculty Advisor:*  Preetam Ghosh |  | [Pghosh@vcu.edu](mailto:Pghosh@vcu.edu) |
| *Sponsor:*  Rishabh Kapoor |  | [Rishabh.kapoor@vcuhealth.org](mailto:Rishabh.kapoor@vcuhealth.org) |

# Step 2: Team Culture. Clarify the Group’s Purpose and Culture Goals.

**Task:** Discuss how each team member wants to be treated to encourage them to make valuable contributions to the group and how each team member would like to feel recognized for their efforts. Discuss how the team will foster an environment where each team member feels they are accountable for their actions and the way they contribute to the project. These are your Culture Goals (left column). How do the students demonstrate these culture goals? These are your Actions (middle column). Finally, how do students deviate from the team’s culture goals? What are ways that other team members can notice when that culture goal is no longer being honored in team dynamics? These are your Warning Signs (right column).

**Resources:** More information and an example Team Culture can be found in the Biodesign Student Guide “Intentional Teamwork” page ([webpage](https://biodesignguide.stanford.edu/toolkit/intentional-teamwork/) | [PDF](https://biodesignguide.stanford.edu/wp-content/uploads/2022/07/Intentional-Teamwork-v2.pdf))

|  |  |  |
| --- | --- | --- |
| ***Culture Goals*** | ***Actions*** | ***Warning Signs*** |
| *Being on time to every meeting* | * *Set up meetings reminders in discord* * *Send reminder e-mail the day before meeting* | * *Student misses first meeting without notice, warning is granted* * *Student misses' multiple meetings afterwards – issue is brought up with faculty advisor* |
| *Informing the group of any delays in completing assignments* | * *Weekly progress check during student meetings* * *Set reasonable deadlines and note when an extension is needed* | * *Student shows up for weekly meeting with no considerable work done, then a discussion would need to be had* * *Student shows up for weekly meeting with no considerable work done on multiple occasions – issue is brought up with faculty advisor* |
| *Have a good balance of work between members* | * *Check weekly at meetings that every has a fair workload, make sure everyone agrees* | * *Issues should be brought up during meetings and we can disperse work accordingly* * *Whoever feels like they aren’t being heard after communicating, they can bring up issues during meeting with faculty advisor* |

# Step 3: Time Commitments, Meeting Structure, and Communication

**Task:** Discuss the anticipated time commitments for the group project. Consider the following questions (don’t answer these questions in the box below):

* What are reasonable time commitments for everyone to invest in this project?
* What other activities and commitments do group members have in their lives?
* How will we communicate with each other?
* When will we meet as a team? Where will we meet? How Often?
* Who will run the meetings? Will there be an assigned team leader or scribe? Does that position rotate or will same person take on that role for the duration of the project?

**Required:** How often you will meet with your faculty advisor, where you will meet, and how the meetings will be conducted. Who arranges these meetings?

See examples below.

|  |  |  |
| --- | --- | --- |
| ***Meeting Participants*** | ***Frequency***  ***Dates and Times / Locations*** | ***Meeting Goals***  ***Responsible Party*** |
| *Students Only* | *Every other week on Thursday (6pm-7pm), On Discord Voice Channel* | *Update group on day-to-day challenges and accomplishments*  *(Connor will record these for the weekly progress reports and meetings with advisor)* |
| *Students Only* | *Every other week on Thursday (6pm-7pm), in ENGR West 0101* | *Actively work on project*  *(August will document these meetings by taking photos of whiteboards, physical prototypes, etc, then post on Discord and update Capstone Report )* |
| *Students + Faculty advisor* | *Every Thursday at 5 pm in Advisor’s office (conference room if they can reserve)* | *Update faculty advisor and get answers to our questions*  *(Connor will scribe; Sawiya will create meeting agenda and lead meeting)* |
| *Project Sponsor* | *First Friday at noon of every month*  *If sponsor is available, we’ll figure out Zoom or in person details*  *If not, then we’ll update the sponsor via email.* | *Update project sponsor and make sure we are on the right track (Connor will scribe; Sawiya will create meeting agenda and lead meeting; Shashank will present prototype so far)* |

# Step 4: Determine Individual Roles and Responsibilities

**Task:** As part of the Capstone Team experience, each member will take on a leadership role, ***in addition to*** contributing to the overall weekly action items for the project. Some common leadership roles for Capstone projects are listed below. Other roles may be assigned with approval of your faculty advisor as deemed fit for the project. For the entirety of the project, you should communicate progress to your advisor specifically with regard to your role.

* **Before meeting with your team**, take some time to ask yourself: what is my “natural” role in this group (strengths)? How can I use this experience to help me grow and develop more?
* **As a group,** discuss the various tasks needed for the project and role preferences. Then assign roles in the table on the next page. Try to create a team dynamic that is fair and equitable, while promoting the strengths of each member.

**Communication Leaders**

**Suggested:** Assign a team member to be the primary contact for the client/sponsor. This person will schedule meetings, send updates, and ensure deliverables are met.

**Suggested:** Assign a team member to be the primary contact for faculty advisor. This person will schedule meetings, send updates, and ensure deliverables are met.

**Common Leadership Roles for Capstone**

1. **Project Manager:** Manages all tasks; develops overall schedule for project; writes agendas and runs meetings; reviews and monitors individual action items; creates an environment where team members are respected, take risks and feel safe expressing their ideas.

**Required:** On Edusourced, under the Team tab, make sure that this student is assigned the Project Manager role. This is required so that Capstone program staff can easily identify a single contact person, especially for items like Purchasing and Receiving project supplies.

1. **Logistics Manager:** coordinates all internal and external interactions; lead in establishing contact within and outside of organization, following up on communication of commitments, obtaining information for the team; documents meeting minutes; manages facility and resource usage.
2. **Financial Manager:** researches/benchmarks technical purchases and acquisitions; conducts pricing analysis and budget justifications on proposed purchases; carries out team purchase requests; monitors team budget.
3. **Systems Engineer:** analyzes Client initial design specification and leads establishment of product specifications; monitors, coordinates and manages integration of sub-systems in the prototype; develops and recommends system architecture and manages product interfaces.
4. **Test Engineer:** oversees experimental design, test plan, procedures and data analysis; acquires data acquisition equipment and any necessary software; establishes test protocols and schedules; oversees statistical analysis of results; leads presentation of experimental finding and resulting recommendations.
5. **Manufacturing Engineer:** coordinates all fabrication required to meet final prototype requirements; oversees that all engineering drawings meet the requirements of machine shop or vendor; reviews designs to ensure design for manufacturing; determines realistic timing for fabrication and quality; develops schedule for all manufacturing.

|  |  |  |
| --- | --- | --- |
| ***Team Member*** | ***Role(s)*** | ***Responsibilities*** |
| Sawiya Aidarus | *Project Manager* | * *Send out weekly emails and other correspondence* * *Create schedule and meetings* * *Make sure everyone understands what is going on* |
| Connor Holden | Logistics Manager | * *Keep a detailed record of meeting notes and share with group* * *Coordinate meeting times* * *Follow up with communication of commitments* * *Obtaining information for the team* * *Manages facility and resource usage* |
| Shashank Sinha | Financial Manager | * *Keeps track of team budget* * *Sends in request for any additional money* * *Conducts pricing analysis* |
| August Moses | Test Engineer | * Oversees testing and implementation * Acquires equipment and any necessary software * Oversees statistical analysis of results |

# Step 5: Agree to the above team contract

*Team Member: Sawiya Aidarus* *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Team Member: Connor Holden* *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Team Member: Shashank Sinha* *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

*Team Member: August Moses* *Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_*

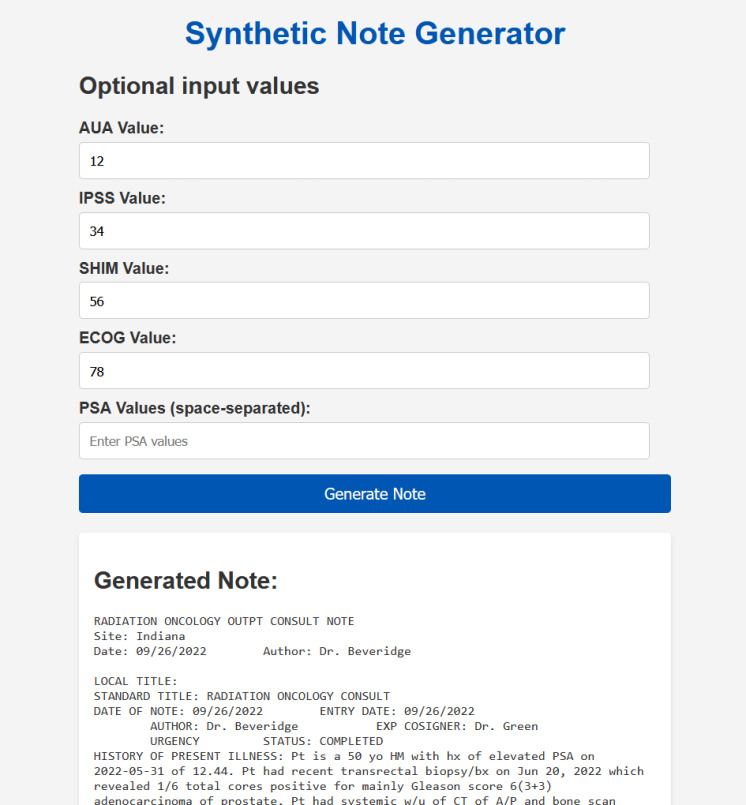
### Section D. Concept Generation

Our overarching design was built off a pre-existing note generation system titled synthetic-note-generator on Github. The initial system had no user interface and had to be manually ran, so we experimented with a series of user interfaces throughout our project.

Design concept 1:

Our first concept was a simple implementation of our note generator, with minimal options. It could generate one note at a time and the notes could not be downloaded. This was before our team started editing the note generator and making it more complex.

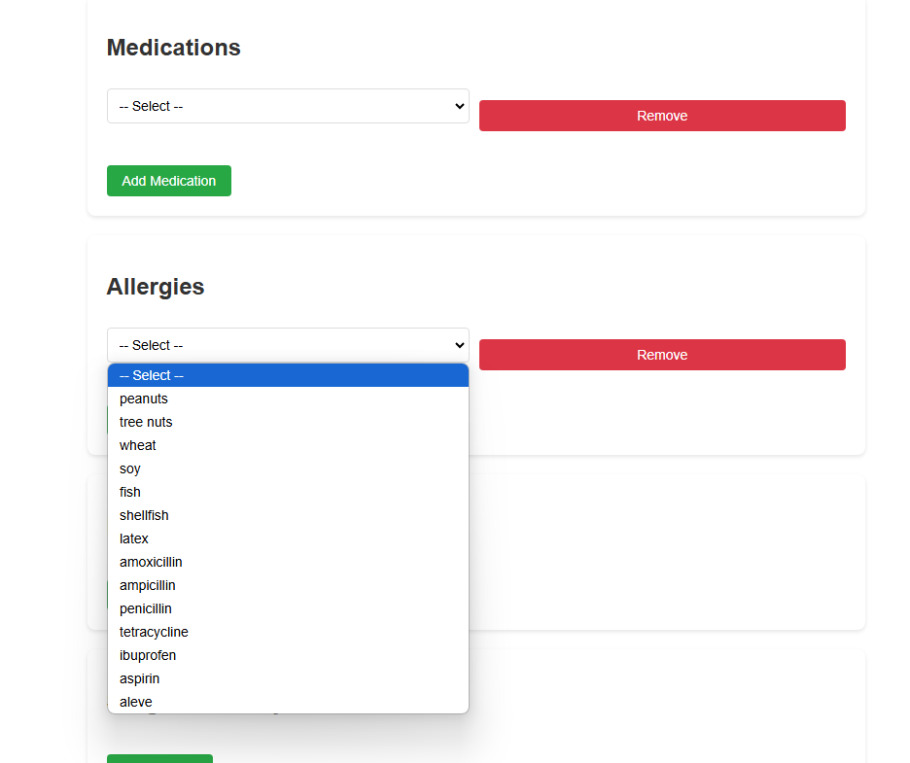
* Pros:
  + Has some input text boxes
  + Input is checked. Input values cannot be too high or too low
  + Aesthetically coherent
* Cons:
  + Too simple, not enough input boxes
  + Can’t download notes
  + Only generates one note at once
  + Didn’t integrate LLM text rephrasing



Design concept 2:

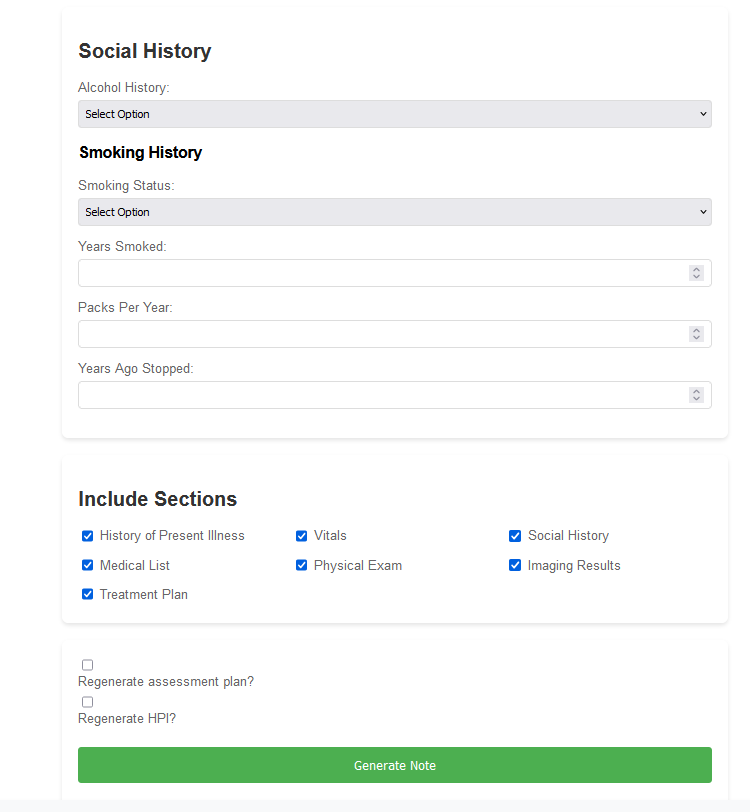
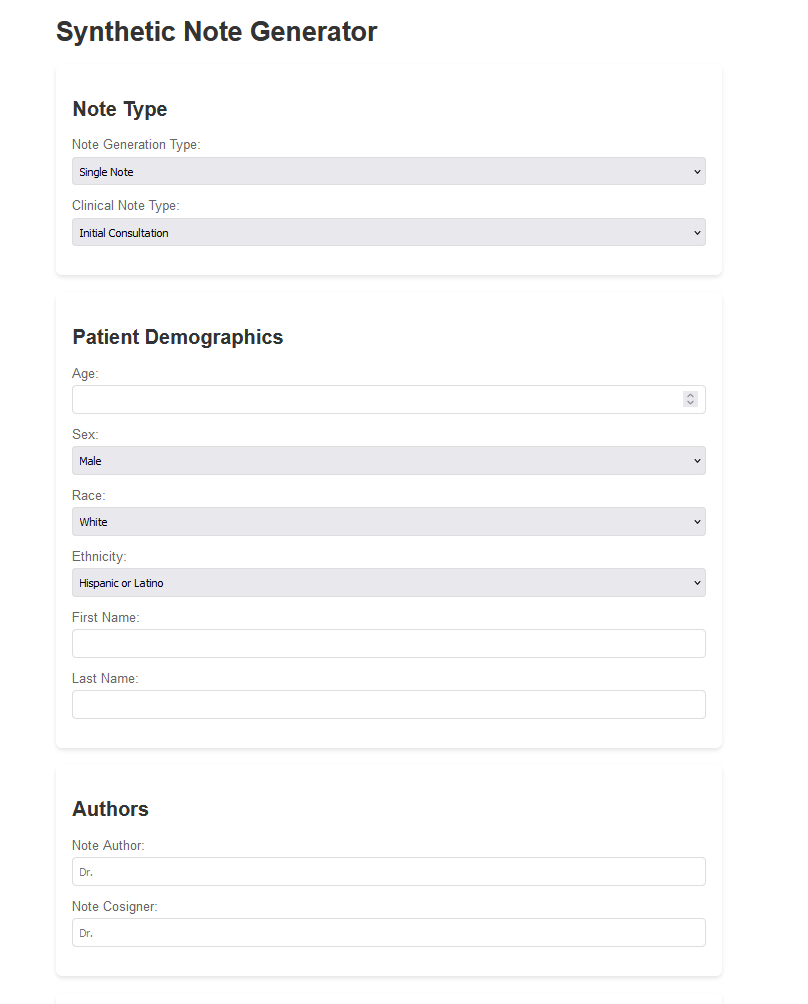
Our second concept was built off the previous concept.

* Pros:
  + Has more input text boxes and drop-down boxes with available options
  + Integrates LLM text-generation where users can choose to regenerate sections
  + Backend code was updated, leading to a more complete frontend design
* Cons:
  + Still can’t download notes
  + Still only generates one note at once
  + Slightly aesthetically disorganized



Design Concept 3:

* Pros:
  + Has input text boxes and drop-down boxes for ALL necessary variables now
  + Notes can now be downloaded
  + Sleek and easy to navigate user-interface
* Cons:
  + Still only generates one note at once



### Section E. Concept Evaluation and Selection

Based off the criteria presented from our sponsors we’ve generated 4 weighted criteria, with 1 being the least important category to 3 being the most important. Our least important category is the visualization of the web-tool. While good visualization promotes usability, visualization is often an aspect that can be changed at any point and doesn’t need to be too heavily focused on. Our ‘downloadable notes’ aspect is weighted at a 2 because it wasn’t discussed until later on in the project, however, it’s still important that we eventually implement that aspect. It’s important yet ‘skippable’ for the time being. The same applies to our ‘multiple note generation’ category. It’s important to integrate it in our final product but not a key concept at the moment. Our most important category is ‘LLM integration’. Since our whole project revolves around LLMs it’s vital that it remains a key focus no matter what design we choose. Despite our concept options being chronological implementations throughout the semester, reflection is important to determine if our current design should maintain any aspects from previous designs.

**Table 1 Decision Matrix -**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Weight | Design concept A | Design Concept B | Design Concept C |
| LLM Integration | 3 | No | Yes | Yes |
| Downloadable Notes | 2 | No | No | Yes |
| Multiple Note Generation | 2 | No | No | No\* |
| Organized and complete design | 1 | No | No | Yes |
|  |  |  |  |  |
| Total Score |  | 0 | 3 | 6 |

‘\*’ = To-do list

### Section F. Design Methodology

To evaluate, improve, and evolve the synthetic clinical note generation system, an iterative engineering design process will be utilized to ensure that the system meets its objectives and satisfies user needs. The team is utilizing an agile style methodology to constantly show results to the sponsors and advisors. This is done through having two weekly meetings, where the first meeting is used to show for the work that has been done throughout the week. Any necessary demonstrations are shown to the stakeholders, along with asking questions about the next steps for the project. After this meeting has concluded, the team has a separate meeting, where we discuss what we will each work on throughout the week and consistently keep in touch to ensure progress is being made by everyone. This process involves both verification—ensuring the project meets its technical specifications—and validation—confirming the system’s functionality aligns with the intended purpose. Stakeholder feedback is incorporated iteratively to refine the design.

Verification of the synthetic note generation system involves leveraging a Turing Test style experimentation while utilizing the web interface tool that has been developed. For the web interface tool, the group collectively discussed the best way to go about building the web application, ensuring it would meet both functional requirements and user expectations. The design discussions focused on creating an intuitive interface that allows users to easily customize note generation parameters, such as selecting disease types and specifying data ranges for placeholders. Regular testing of the interface ensured that it aligned with the goals of usability, scalability, and reliability. The interface was built using Flask, a python-based web framework that allows you to integrate the frontend and backend components together. Due to its lightweight and scalable nature, it was the perfect framework to utilize for the project.

In parallel, the Turing Test-style experimentation was planned to validate the synthetic notes. This will involve showing both real and synthetic notes to healthcare professionals and asking them to identify which ones were machine-generated. The results of these tests will provide critical feedback on the realism and coherence of the generated notes, highlighting areas where the system required improvement. Key metrics include the ability to generate notes indistinguishable from real ones and capturing feedback on features that made synthetic notes recognizable. This information guided iterative improvements to enhance the linguistic and contextual quality of the notes. By combining these verification and validation strategies, the team iteratively refined the system to produce high-quality synthetic clinical notes that are both functional and user-centered.

#### F.1 Computational Methods (e.g. FEA or CFD Modeling, example sub-section)

#### F.2 Experimental Methods (example subsection)

#### F.5 Validation Procedure

As part of our validation process, we plan to conduct a Turing Test-style experiment to ensure that the generated synthetic notes are truly representative of real clinical notes. This experiment will involve inviting physicians to participate in a study where they are presented with sections from both real and synthetic notes. For each section shown, they will be asked to identify whether it is real or synthetic and provide feedback explaining the reasoning behind their decision.

To facilitate this process, we will develop a web interface tool capable of randomly selecting sections from both real and synthetic notes and displaying them to the physician. The tool will include an input box for physicians to record their feedback, particularly focusing on the features that led them to classify the section as real or synthetic.

This feedback will be analyzed to identify patterns and areas where the synthetic notes can be improved. Based on these insights, we will refine the synthetic note generation process to ensure the notes closely mimic real ones in structure, style, and content.

The Turing Test-style experiment is scheduled to begin in January, coinciding with the completion of the web interface tool for generating and displaying synthetic notes. This marks the next phase of our project, moving from development to rigorous testing and validation.

### Section G. Results and Design Details

This section outlines the major results and design specifications of the Synthetic Note Generator web tool project, showcasing the implementation progress and key features developed to meet the project objectives.

**G.1 Architecture and System Design**

### **Frontend Design**

The web interface has been implemented using modern web technologies with the following key components:

1. **User Interface Layout**
   1. Modular section design for different input categories
   2. Responsive form elements for data entry
   3. Dynamic field generation for lists (medications, allergies, problems, surgical history)
   4. Real-time output display with split view for note text and JSON data
   5. Export functionality for both note text and structured data
2. **Input Categories**
   1. Note Type Selection
   2. Patient Demographics
   3. Medical Values
   4. Important Dates
   5. Dynamic Lists (Medications, Allergies, Problems, Surgical History)
   6. Vitals
   7. Staging Information
   8. Treatment Information
   9. Social History
   10. Section Toggles for Note Generation

### **Backend Architecture**

The system utilizes a Flask-based backend with the following components:

1. **Core Components**
   1. Flask web server handling HTTP requests
   2. Custom JSON encoder for handling complex data types
   3. Data validation and processing middleware
   4. Integration with existing Python-based note generator
   5. LLM integration for note regeneration (using Groq)
2. **Data Flow**

User Input → Form Validation → Data Processing → Note Generation → LLM Processing (if requested) → Response Formatting → Client Display

**G.2 Implementation Progress**

### **Completed Features**

1. **Frontend Development**
   1. Implemented comprehensive form interface
   2. Added dynamic field management for lists
   3. Created responsive design with CSS styling
   4. Integrated export functionality for generated content
   5. Added section toggles for note customization
2. **Backend Development**
   1. Established Flask server endpoints
   2. Implemented data validation and processing
   3. Integrated existing note generator code
   4. Added LLM regeneration capability
   5. Created JSON response formatting
3. **Data Management**
   1. Implemented dropdown population from constants
   2. Added dynamic list management
   3. Created data validation functions
   4. Established proper data type handling

### **Current Functionality**

The system currently supports:

1. **Input Processing**
   1. Patient demographic information
   2. Medical values (AUA, IPSS, SHIM, ECOG, PSA scores)
   3. Multiple list types (medications, allergies, problems, surgical history)
   4. Vital signs
   5. Staging information
   6. Treatment details
   7. Social history
2. **Note Generation**
   1. Basic note structure generation
   2. Section inclusion/exclusion
   3. LLM-based regeneration options
   4. Export capabilities for generated content

**G.3 Testing Results**

Initial testing has validated:

1. **Form Functionality**
   1. Proper data collection from all input fields
   2. Accurate dropdown population
   3. Dynamic field addition/removal
   4. Form validation
2. **Note Generation**
   1. Successful note creation with provided inputs
   2. Proper section toggling
   3. Accurate data representation in output
   4. Correct formatting of generated notes
3. **Data Processing**
   1. Accurate JSON conversion
   2. Proper handling of null/empty values
   3. Correct date formatting
   4. Appropriate type conversion

**G.4. Final Design Specifications**

### **Frontend Specifications**

* Interface supports all required input fields
* Responsive design works on standard desktop browsers
* Dynamic field management handles variable-length lists
* Export functionality supports both text and JSON formats

### **Backend Specifications**

* REST API endpoints handle all required operations
* Data validation ensures proper input formatting
* Note generation produces consistent output
* LLM integration provides note regeneration capability

### **Data Management**

* Supports all required medical data types
* Handles dynamic list management
* Provides proper data validation
* Ensures consistent data formatting

### **Performance Metrics**

* Form responsiveness: < 100ms
* Note generation time: < 2s
* LLM regeneration time: < 5s
* Export functionality: < 1s

### **Integration Requirements**

* Compatible with existing Python note generator
* Supports Groq LLM integration
* Maintains consistent data structure
* Provides proper error handling

Current implementation meets primary design objectives while providing a foundation for future enhancements. The system successfully generates synthetic medical notes with user-specified parameters and supports data export functionality.

### Section H. Societal Impacts of Design

In designing the Synthetic Note Generation and LLM-Based Discrete Data Extraction Project, it is critical to consider the wider societal impacts of the technology. The project aims to improve the accuracy and efficiency of clinical documentation in radiation oncology, and it is essential to assess how the design affects public health, safety, welfare, economics, and ethical considerations, among others.

#### H.1 Public Health, Safety, and Welfare

The design’s primary goal is to improve the efficiency and accuracy of clinical documentation in healthcare, particularly in radiation oncology, which directly affects patient care. By reducing errors in clinical notes and speeding up data extraction, this project enhances the quality of patient care and reduces the likelihood of medical errors.

* **Safety Features**: The use of synthetic data ensures no Protected Health Information (PHI) is exposed, protecting patient privacy. The system also implements robust security protocols (ISO 27001, NIST SP 800-53) to ensure data protection and prevent unauthorized access.
* **Public Health Impact**: Accurate and timely clinical documentation can improve treatment outcomes by reducing delays caused by incomplete or erroneous data. This contributes to overall public health by enabling better-informed clinical decisions.
* **Welfare**: The tool helps reduce the administrative burden on healthcare providers, allowing them to focus more on patient care, thus improving the welfare of healthcare workers and potentially leading to better job satisfaction.

#### H.2 Societal Impacts

The adoption of this technology could lead to widespread changes in how clinical notes are generated and utilized, with long-term implications for healthcare systems globally.

* **Improved Data Sharing**: The use of synthetic clinical notes ensures that healthcare institutions can share data without compromising patient privacy, potentially leading to better collaboration in research and treatment across institutions.
* **Reduction in Errors**: By automating the extraction of discrete data, this technology could significantly reduce human error, thus improving patient care on a large scale.
* **Potential for Increased Efficiency**: The reduction in manual note-taking and data entry could lead to more efficient use of healthcare resources, lowering operational costs and improving access to care.

#### H.3 Political/Regulatory Impacts

The project’s emphasis on data privacy and the use of synthetic data positions it well within the current regulatory environment, where the privacy of patient information is of paramount concern. Several political and regulatory factors must be considered:

* **Compliance with Regulations**: The design adheres to privacy regulations like HIPAA in the U.S. and GDPR in Europe by ensuring that no real patient data is used in the generation of synthetic notes. This makes the technology compliant with key privacy regulations that govern healthcare data.
* **Potential for Policy Change**: As healthcare systems move towards digital solutions, there may be political shifts towards increased automation in documentation and data management. This could encourage the adoption of similar technologies to enhance data integrity and healthcare delivery.

#### H.4. Economic Impacts

The project can have significant economic implications for healthcare institutions, patients, and broader industries involved in healthcare technology.

* **Cost Savings**: By reducing manual data entry and extraction, healthcare providers can decrease labor costs, which is especially important in an environment where healthcare costs are rising.
* **Job Creation**: The development of synthetic data generation tools and web-based applications could create jobs in software development, cybersecurity, and healthcare data management.
* **Market Disruption**: The technology could disrupt traditional healthcare data management systems, encouraging competition and innovation among companies developing EMR and clinical documentation tools.

#### H.5 Environmental Impacts

While the environmental impact of this project may not be as direct as in other industries, it is essential to consider potential impacts:

* **Energy Use**: The use of cloud-based servers for synthetic note generation and machine learning model fine-tuning could lead to increased energy consumption. Therefore, it is important to optimize the design for energy efficiency and to consider using sustainable cloud service providers.
* **Reduction in Paper Usage**: By enabling more efficient digital clinical documentation, the system could contribute to a reduction in paper usage, supporting environmental sustainability goals in healthcare.

#### H.6 Global Impacts

This project could have significant global impacts, particularly in the context of healthcare systems in both developed and developing countries.

* **Global Data Sharing**: The use of synthetic data facilitates safer cross-border healthcare data sharing, promoting international collaborations in research and patient care.
* **Scalability in Low-Resource Settings**: With an affordable and easy-to-deploy system, the project has the potential to be scaled to low-resource healthcare settings where clinical documentation is often a challenge. This could improve healthcare delivery in underserved areas globally.
* **Global Health Research**: By providing accurate data more quickly, the project could accelerate medical research and global health studies, particularly in oncology, where timely data is critical.

#### H.7. Ethical Considerations

The ethical implications of generating and using synthetic clinical data are paramount in this project, as the technology must maintain the integrity of patient privacy and the accuracy of healthcare documentation.

* **Data Privacy**: Since synthetic notes are being generated to ensure the safety of patient data, ethical considerations around data privacy are central. The system ensures that no real patient information is used, mitigating the risk of unintended PHI exposure.
* **Bias in Data Generation**: There is a risk that the synthetic data could introduce biases if the underlying data used to train the models is not representative of diverse patient populations. To mitigate this, the team will ensure that the synthetic data generation process uses diverse datasets.
* **Transparency and Trust**: Ethical considerations also involve ensuring transparency in how synthetic data is generated and used, as well as fostering trust in the technology by clinicians and patients. The system must be auditable and ensure that synthetic data generation is performed in an ethical and transparent manner.
* **Accountability**: Given the role of automated systems in clinical decision-making, there must be accountability for the outcomes of using synthetic data, ensuring that healthcare providers do not overly rely on AI-generated notes without human oversight.

### Section I. Cost Analysis

Our group has not used any funding or made any purchases for the project thus far. All work has been conducted using internal resources, provided tools, and free software.

### Section J. Conclusions and Recommendations

Our group has made significant progress in developing a preliminary design for the synthetic note generation tool, with several key features implemented to meet the project's goals. While we do not yet have a final design, we have successfully worked on expanding the tool's capabilities to support new types of medical notes. This evolution was guided by the engineering design process, and we’ve learned valuable lessons along the way.

Some of the major achievements in the preliminary design include:

* **Web Tool Development:** We designed and refined a user-friendly web-based interface, ensuring smooth data entry and usability for clinicians and other users.
* **Groq & LLM Integration:** Our team focused on integrating Groq’s LLM with the tool, updating Python code to ensure that responses from the language model met our requirements for generating synthetic notes. This integration was essential for the functionality of the tool.
* **Turing Test Implementation:** We have also implemented various updates and conducted testing cycles to enhance the tool’s ability to pass a clinical Turing Test. This ensures that the synthetic notes generated are realistic and coherent.
* **Python Code Updates:** Throughout the project, we made necessary updates to the Python code to accommodate new types of medical notes and incorporated internal feedback to ensure seamless functionality of the tool.

While these developments are promising, we recognize that further work is needed to achieve a final design. In the future, the tool could be improved by adding more advanced features, optimizing the user interface, and expanding its compatibility with additional disease sites and medical contexts. Additionally, as we continue to refine the integration with Groq’s LLM, we plan to explore further ways to improve the accuracy and realism of the synthetic notes generated.

Looking ahead, this project will continue as a senior design project, where key milestones will include further testing, optimization, and potentially expanding the scope of the tool to meet additional healthcare needs. We recommend that future work focuses on refining the integration with Groq's LLM, conducting rigorous testing in clinical environments, and enhancing the overall robustness of the web tool.

**References**

Provide a numbered list of all references in order of appearance using APA citation format. The reference page should begin on a new page as shown here.

[1] VCU Writing Center. (2021, September 8). *APA Citation: A guide to formatting in APA style.* Retrieved September 2, 2024. <https://writing.vcu.edu/student-resources/apa-citations/>

[2] Teach Engineering. *Engineering Design Process*. TeachEngineering.org. Retreived September 2, 2024. <https://www.teachengineering.org/populartopics/designprocess>

[3] Lee, J., et al. (2020). Errors in clinical notes in radiation oncology: A systematic review. *International Journal of Radiation Oncology, Biology, Physics, 106*(3), 531-538.

[4] American Medical Association. (2020). *Ethical issues in the use of artificial intelligence in healthcare*. AMA Journal of Ethics, 22(12), 1157–1165. https://doi.org/10.1001/amajethics.2020.1157

[5] World Health Organization. (2016). *Global strategy on human resources for health: Workforce 2030*. World Health Organization. <https://www.who.int/publications/i/item/9789241511131>

[6] Zink, T., & Berman, D. (2019). *The ethical use of artificial intelligence in healthcare*. The Journal of Law, Medicine & Ethics, 47(2), 334-343. <https://doi.org/10.1177/1073110519851169>