

# 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:

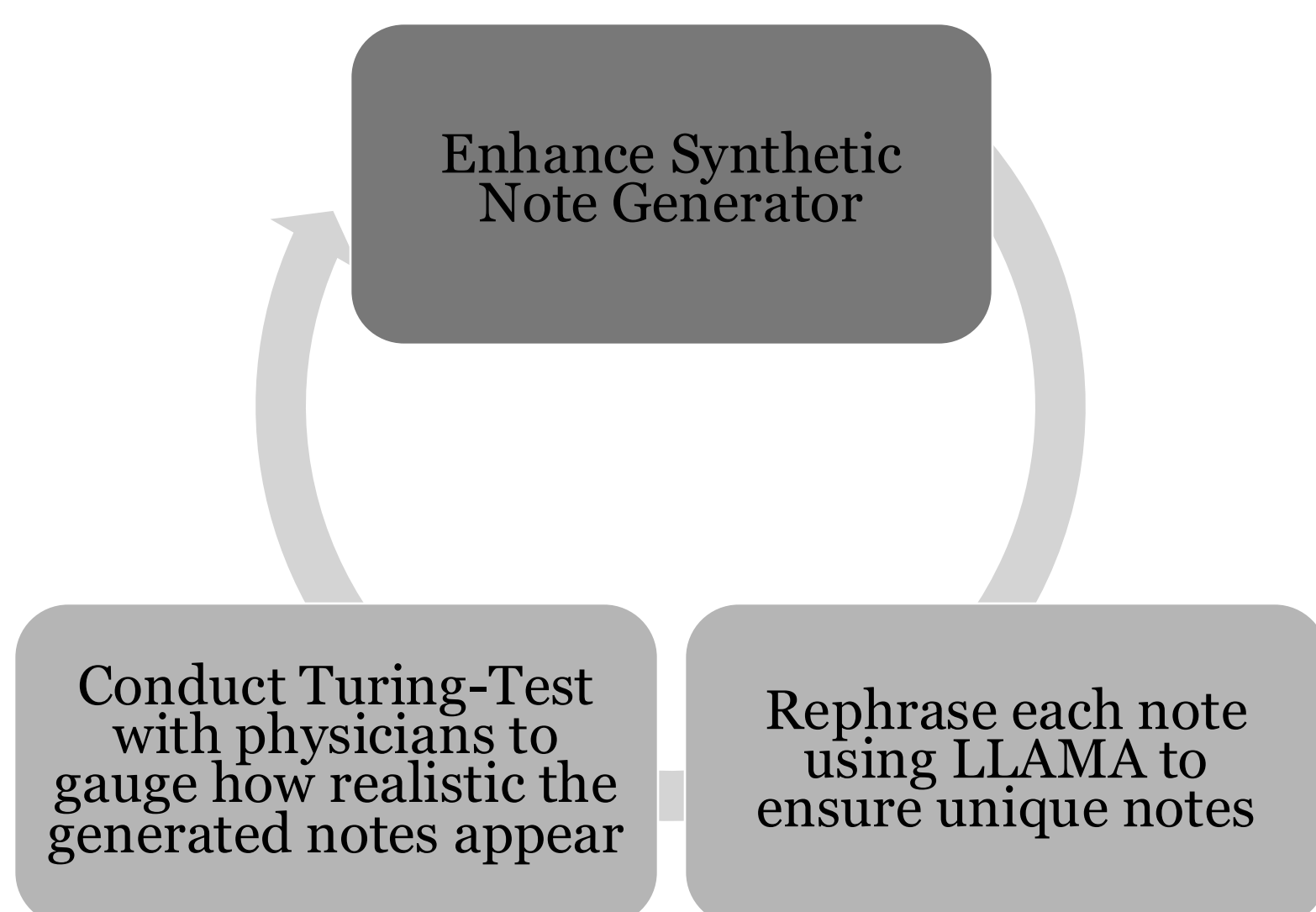
- Training of medical professionals without risking patient privacy
- Development and testing of healthcare software systems
- Medical research studies that require large datasets

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

Synthetic Note Generator

Choose Generation Type

Single Note Generation

Bulk Note Generation

Landing Page with Single or Bulk Note Generation

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

Patient Demographics

Age:

Sex:

Race:

Ethnicity:

First Name:

Last Name:

Generated Note:

RADIATION CONSULT RESULT

Sites: Maryland

Date: May 28, 2022 Author: Dr. Hurley

LOCAL TITLE:

STANDARD TITLE: RADIATION ONCOLOGY CONSULT

DATE OF NOTE: May 28, 2022 ENTRY DATE: May 28, 2022

AUTHOR: Dr. Hurley EOP COORDINER: Dr. Hosley

URGENCY: STATUS: COMPLETED

Here is the reconstructed note:

CHIEF COMPLAINT: Newly diagnosed low prostate cancer.

HISTORY OF PRESENT ILLNESS:

Kid is a 76 year old white male with a clinical T2cN0M0 Gleason Gleason score 7(4+3) prostate cancer, with a PSA of 5.72. Initial PSA on 2019-03-05 was 3.58. The most recent PSA from Oct 28, 2021 shows 5.72. A biopsy performed on 01/17/2022 confirmed the diagnosis. He had a colonoscopy 9 months ago, no polyps but has internal hemorrhoids, denies rectal pain/bleeding.

Physical Exam:

Blood Pressure: 164/108

Respirations: 15

Weight: 213 lbs

Pain: 5

Temperature: 99.75 F

Pulse: 115

Karotid: 70

ECOG: ECOG 2

Past Medical/Surgical History:

None

Active Outpatient Medications

- 1) UREA 20% CREAM APPLY A SUFFICIENT AMOUNT EXTERNALLY EVERY DAY
- 2) BIOSUNASTATIN CA 40MG TAB
- 3) BIOSUNASTATIN CA 40MG TAB
- 4) BIOSUNASTATIN CA 40MG TAB
- 5) BIOSUNASTATIN CA 40MG TAB
- 6) BIOSUNASTATIN CA 40MG TAB
- 7) BIOSUNASTATIN CA 40MG TAB
- 8) BIOSUNASTATIN CA 40MG TAB

Allergies: aspirin, tetracycline, amoxicillin

CT abdomen and pelvis on 2022-02-02: No evidence of metastatic disease in the abdomen or pelvis

MRI abdomen and pelvis on 2021-08-11: No evidence of metastatic disease in the abdomen or pelvis

Bone scan on 2022-01-27: No evidence for skeletal metastatic involvement is noted at this time.

SOCIAL HISTORY:

Tobacco Use: N/A 54 packs per year, has smoked for approximately 38 years.

Quit 17 years ago.

Alcohol Use: currently drinks 0-3 beers per week.

Family Hx:

father had rectum ca, sister had ovarian ca

CANCER TREATMENT HISTORY:

Note Data:

```

{
  "allergies": [
    "aspirin",
    "tetracycline",
    "amoxicillin"
  ],
  "age": 76,
  "base_data": "2022-05-28",
  "biopsy": {
    "biopsy_date": "2022-01-17",
    "biopsy_type": null,
    "gleason": {
      "primary": 4,
      "secondary": 3,
      "total": 7
    },
    "left_cores": 4,
    "right_cores": 3,
    "total_cores": 8
  },
  "bone_scan": "2022-01-27",
  "colonoscopy": true,
  "dose_data": null,
  "ecog": 2,
  "family_history": true,
  "ips": 5,
  "medications": [
    "UREA 20% CREAM APPLY A SUFFICIENT AMOUNT EXTERNALLY EVERY DAY",
    "BIOSUNASTATIN CA 40MG TAB",
    "BIOSUNASTATIN CA 40MG TAB",
    "BIOSUNASTATIN CA 40MG TAB",
    "BIOSUNASTATIN CA 40MG TAB",
    "BIOSUNASTATIN CA 40MG TAB",
    "BIOSUNASTATIN CA 40MG TAB",
    "BIOSUNASTATIN CA 40MG TAB"
  ],
  "note_date": "2021-11-01",
  "note_author": "Dr. Hurley",
  "note_coauthor": "Dr. Hosley",
  "note_type": "consult",
  "patient": {
    "age": 76,
    "date_of_birth": "1945-08-01",
    "ethnicity": "HISPANIC OR LATINO",
    "first_name": "Charles",
    "last_name": "Hosley",
    "race": "White",
    "sex": "male"
  },
  "pelvic_ct": "2022-02-02",
  "pelvic_mri": "2021-08-11",
  "performance_score": 70,
  "prior_treatment": {
    "chemotherapy_drug_prescribed": null,
    "chemotherapy_prescribed": null,
    "hormone_therapy_date": "2022-10-10",
    "hormone_therapy_prescribed": null,
    "prior_ct": null,
    "prior_rt_date": "2017-07-18"
  }
}

```

## 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

Clinical Note Section Quiz

**Physical Exam:**

Date Vital Measurement Qualifiers

03/22/2024 12:56 Temp F (C) 98 (36.7) Oral

\*\* Pulse 63

\*\* Respir 16 Spontaneous

\*\* BP 101/68 R Arm, Sitting, Adult

Cuff, CuffAutomated

\*\* Ht in (cm) 76 (193.04) Actual

\*\* Wt lbs (kg) 213.2 (96.71) Actual, Standing Weight

\*\* CIG in (cm) Unavailable

\*\* Pain 0

\*\* PCx (LMin) 97 Room Air

General: Well appearing Male, no acute distress

HEENT: No scleral icterus

Neck: No palpable lymphadenopathy

Cardiovascular: Warm well perfused extremities

Pulmonary: Respirations non-labored

Abdo: Soft, nontender, nondistended

Musculoskeletal: No lower extremity edema

Psych: Normal affect

Neuro: Alert, awake, oriented x 3. Moves all extremities spontaneously.

Is this section from a real or synthetic clinical note?

Real

Synthetic

Why do you think so?

The formatting looks off and also doesn't seem like it came from a real note due to the lack of variability.

Next

User sees the results of their quiz at the end

Clinical Note Section Quiz

**Imaging:**

personally reviewed and described in HPI

CT Chest 3/8/24

Findings: Previously noted right upper lobe nodule definitely increased...

Quiz Completed!

You got 3/5 correct.

**Review Your Answers:**

- Question 1: You guessed **real**, Actual: **real**. Reasoning: Looks like real data
- Question 2: You guessed **synthetic**, Actual: **synthetic**. Reasoning: Doesn't look real
- Question 3: You guessed **synthetic**, Actual: **real**. Reasoning: The formatting looks off and also doesn't seem like it came from a real note due to the lack of variability.
- Question 4: You guessed **real**, Actual: **real**. Reasoning: This looks like a real clinical note because it includes a specific comparison to prior imaging, a cautious interpretation ("concerning for very slow growing neoplasm"), and a typical radiology disclaimer directing to the full report. Synthetic notes often lack such detail and natural phrasing.
- Question 5: You guessed **synthetic**, Actual: **real**. Reasoning: Lack of information.

## 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.