AI in Healthcare Documentation: A Tool for Validating Specificity and Accuracy in Medical Notes

Valeria Nicole Fontánez Ortiz, Mentor: Joanne Brenes Catinchi, PhD
Polytechnic University of Puerto Rico
Undergraduate Research Program for Honor and Outstanding Students (URP-HOS) 2024-2025

HSI STEM TITLE III
POLYTECHNIC UNIVERSITY OF PUERTO RICO
A MULTIFACETED APPROACH TO STUDENT CENTERED STEM EDUCATION
P031C210139

Abstract

Accurate clinical documentation is essential for effective care and diagnosis, as vague or incomplete notes can contribute to serious medical errors.This research proposes an AI-powered validation tool that uses ChatGPT and Google Gemini AI to evaluate the accuracy and specificity of SOAP-formatted clinical notes. The tool integrates a Forest-Based Validation Algorithm to classify each section as Valid, Vague, Not Defined, or Missing. Built using Agile methodology, the tool was tested through both quantitative metrics (accuracy, precision, error rate) and human review. Results show the tool enhances documentation quality while supporting clinician workflows and ethical AI use in healthcare.

Introduction

SOAP notes, structured into the Subjective, Objective, Assessment, and Plan sections, are the most widely used method for documenting clinical information across medical specialties (Podder et al., 2023; Sudarsan et al., 2021). With the rise of artificial intelligence in healthcare, models such as ChatGPT and Google Vertex AI are now used to generate these notes automatically (Leong et al., 2024). While they offer documentation speed and efficiency (Bongurala et al., 2024), they also risk producing vague, incomplete, or fabricated content, known as AI hallucinations (Liao et al., 2023; Ghaith et al., 2022). These concerns highlight the need for robust validation methods to ensure AI-generated documentation aligns with established medical standards and regulatory requirements, including HIPAA and World Health Organization guidelines. This study addresses this need by evaluating approaches to detect and classify deficiencies in AI-generated SOAP notes.

Objectives

- Develop an AI-powered validation tool to flag vague or inaccurate clinical notes.
- Ensure validated notes align with SOAP standards.
- Support clinical workflows while protecting patient privacy.
- Evaluate performance against human review and reach ≥60% accuracy.

Methodology

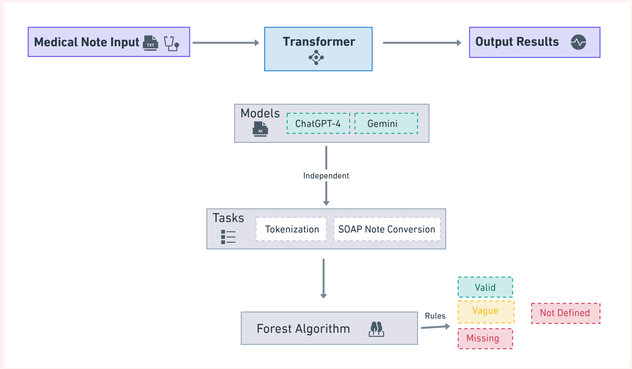

Figure 1: Medical Note Validation Workflow

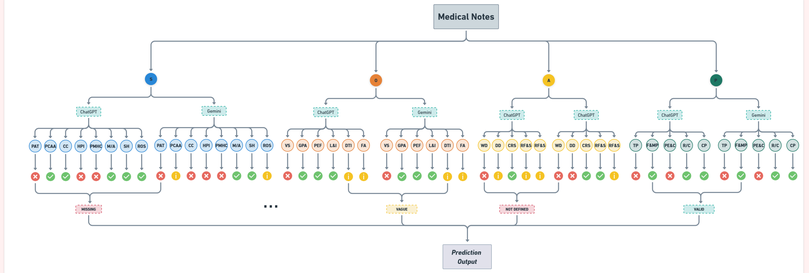

Figure 2: Forest Validation Algorithm

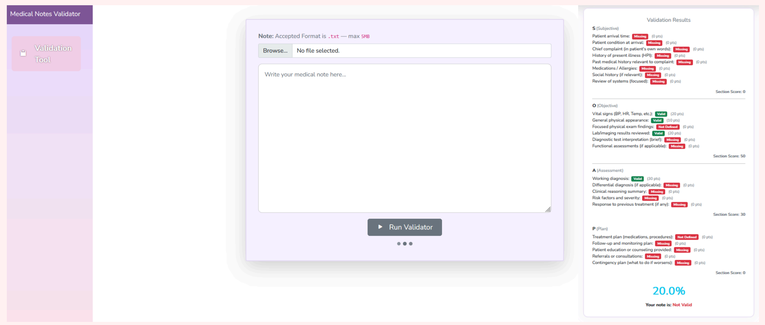

Figure 3: Medical Notes Validation Tool Interface

Results

The evaluation metrics indicate that the system achieved an accuracy of 33%, meaning that only one-third of the classifications aligned with human judgment. The precision score of 100% suggests that when the tool labeled a section as valid, it was always correct according to human evaluation, with no false positives recorded. However, the error rate was high at 67%, reflecting a substantial proportion of incorrect classifications in comparison to human review.

The disagreement rate was calculated at 23%, derived from 319 "Not Defined" outputs across 57 runs and 24 SOAP subsections per run. This elevated hallucination rate highlights the frequency with which the tool produced undefined or absent content, signaling a key area for model refinement.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + FN + TN)} = \frac{(1 + 5)}{1 + 0 + 12 + 5} \approx 0.333$$

$$Precision = \frac{TP}{(TP + FP)} = \frac{(1)}{(1 + 0)} = 1$$

$$Error = \frac{(FP + FN)}{(TP + FP + FN + TN)} = \frac{(0 + 12)}{1 + 0 + 12 + 5} \approx 0.67$$

$$Disagreement\ Rate = \frac{Total\ Not\ Defined}{Total\ Runs * Subsections\ per\ Run} = \frac{319}{57\ x\ 24} \approx 0.23$$

Figure 4: Equation Results

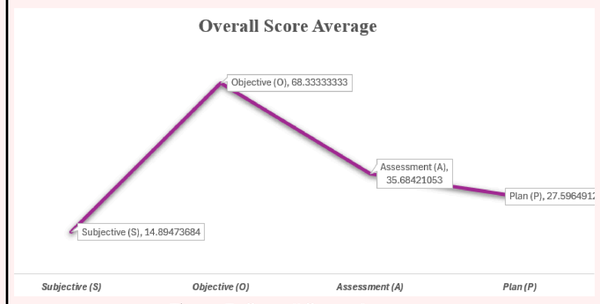

Figure 5: Overall Score Average

The Objective section achieved the highest average score (68.33), indicating stronger alignment with expected content. Assessment (35.68) and Plan (27.60) scored lower, while Subjective had the lowest average (14.89), suggesting it was the most challenging section for the tool to validate accurately.

Conclusions and Future Work

The validation tool identified vague and incomplete sections with some agreement to human review. ChatGPT and Gemini showed partial alignment and the observed discrepancies are likely due to differences in training data. The tool maintained adherence to SOAP formatting and documentation standards. It operated without storing identifiable data and complied with HIPAA and WHO guidelines. The performance approached the 60% for one note, supporting its potential in real-world workflows. However, the presence of errors and occasional hallucinations indicates that the prompt design may require refinement to improve output reliability. Future work for this research could include extending compatibility to additional file formats, improving validation methods for specialized use cases, and implementing structured expert review processes to safeguard accuracy and reliability.

References

Centers for Disease Control and Prevention. (2024, July 10). Health Insurance Portability and Accountability Act of 1996 (HIPAA). https://www.cdc.gov/phlp/php/resources/health-insurance-portability-and-accountability-act-of-1996-hipaa.html

Podder, V., Lew, V., & Ghassemzadeh, S. (2023). SOAP notes. In StatPearls. StatPearls Publishing. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK482263/