Iowa Initiative for Artificial Intelligence

Final Report

Project title:	Text analytics from clinical notes to derive information for Mohs surgery in skin cancer
Principal Investigator:	Marta Van Beek, Andrew Poggemiller
Prepared by (IIAI):	Yanan Liu
Other investigators:	Nicky Marson, Thomas Tsilimigras
Date:	

Were specific aims fulfilled:	Y
Readiness for extramural proposal?	N
If yes … Planned submission date	
Funding agency	
Grant mechanism	
If no … Why not? What went wrong?	We finished the first goal and proved that Chatgpt can extract useful medical information from medical notes. The second and third goals were beyond the scope of this project.

Brief summary of accomplished results:

By analyzing clinical notes using generative AI and LLMs, we were able to extract tumor size, and answer 8 medical questions at high accuracy. The following are the question titles, the model accuracy, and the model's F1 score, for each of the 9 questions: Location: 91.78% Accuracy, 86.21% F1; Aggressive Tumor Pathology Type: 86.94% Accuracy, 77.91% F1; DFSP, 100% Acc, 0% F1; Recurrent Tumor, 96.91% Acc, 52.63% F1; Coordinated repair with ENT/oculoplastic: 97.26% Accuracy, 71.43% F1; Additional C&F: 97.94% Accuracy, 81.25% F1; Additional Excision: 58.62% Accuracy, 17.81% F1; Multiple Sites: 94.16% Accuracy, 85.47% F1; Size at Greatest Dimension: 98.18% Accuracy.

Research report:

Aims (provided by PI):

The primary goal of the project is to develop an AI powered system that can assist with triaging Mohs patients given high demand and limited manpower to complete the triaging system. It was noted that this was to be done ideally with the use of LLM to read patient notes and automatically triage them based upon their note characteristics.

There were second and third goals that were essentially stretch goals and were considered as a follow-up project. The second goal was to train the AI to identify and anticipate how many layers the skin cancer would go using standardized photos. The third goal was to attempt to implement it into the healthcare system at large within UIHC to change surgical triaging throughout the

university. These goals were all clearly outlined by Andrew Poggemiller at the very beginning of the semester and slightly changed. We completed the first goal of the project, the generation or extraction of answers to the questions.

Data:

An Excel sheet with 294 deidentified clinical notes from Mohs surgery patients were given. The Excel sheet includes columns for each of the 9 questions with answers to each patient's notes for checking the AI's accuracy. The notes contained 3 columns of patient notes: Physical Exam (PE), Assessment and Plan (A&P), and Addendum (A). The answer columns also had guesses on which of the 3 patient columns would help the most.

For tumor size column, 53 of the 294 notes had missing data issues for the Size at Greatest Dimension question. This was the only question that ran into any noticeable issues, and we could only use the PE notes for this question. Many of these notes either had a ground truth value but no size at all in the PE notes, or they didn't have a ground truth value to compare against at all. The latter was fixable but ignored in favor of time, the former was not fixable.

To curate the data and find these issues, we created a script to run through the ground truth column and search for null and non-numerical values and add those values to an array to write down. It then used regex to look for notes without numerical values and add those to an array. It also looked for notes with numerical values that were not in 'number x number' dimensional format or 'number cm/mm' format. Once all the note indexes of potentially bad notes were obtained, we manually looked at each PE note compared to its ground truth to ensure that the note was not valid to run on. We then kept a final array with a flag to turn on and off the ability to skip all these notes

AI/ML Approach:

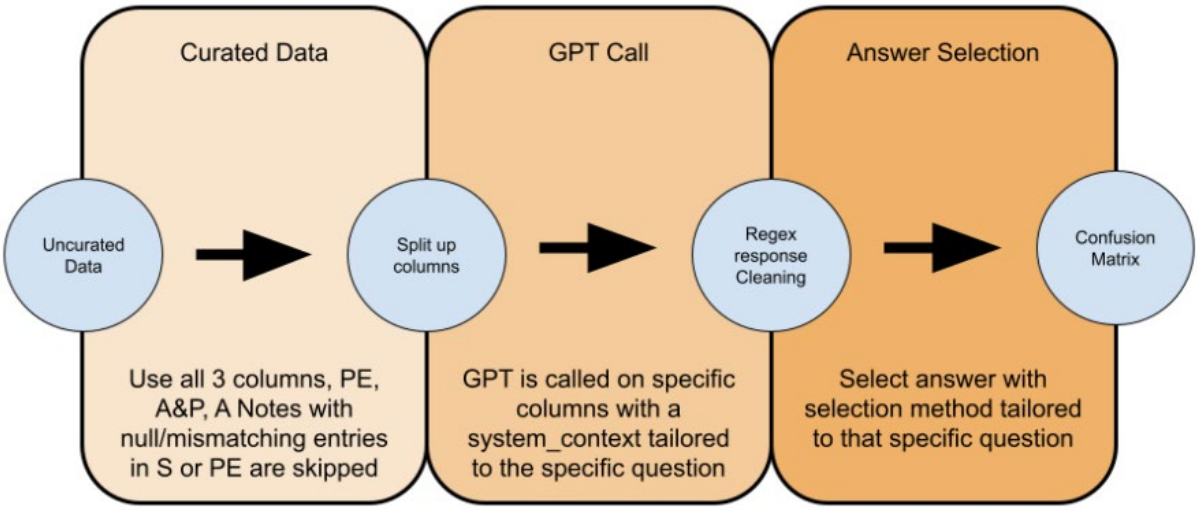
The aim of the first goal was to not train a model but instead have an LLM simply be called with a question and a whole patient note and see if it could accurately extract or predict information. We chose to use OpenAI's ChatGPT 4 version, ran through MS Azure to protect any data passed into the model from being trained on. We started with the tumor Size at Greatest Dimension question. It yielded ~60% accuracy when asking for the size of every tumor in the note and using regex to convert any tumors in mm to cm and select the largest dimension from the array. Then, we split it into 3 GPT calls, one for each patient column, and took an answer if two of the GPT responses agreed, else default to the PE answer. This improved the accuracy to ~75%. This is when we realized that many of the notes had issues, and we curated the data which gave us the true accuracy of ~92%. We made one more small change to fix the regex, bringing the accuracy to 95.7%.

We then split up to each work on the remaining 8 columns, all of which were yes or no questions. Even though we worked separately and on different questions, we applied the same 3 patient column split method to get the base accuracy of each of the tests. Then we focused on whichever needed the most help. Thomas began by focusing on fine tuning the question passed in with the patient note. The length and specific wording of the passed in question turned out to be extremely important across all questions.

Meanwhile, we explored the system_context parameter, which is essentially a message sent before a user asks a question to give the model context on what to expect next. This was by far the most important thing in affecting accuracy. The outline set in place for system_context was: task for the model -> GPT response format -> fallback behavior if the model is unsure about its answer -> criteria to determine yes or no. The specific wording in how strict the model needs to be to guess yes or no and if something is uncommon or not, drastically changed how often the model responded to the same question with the same patient input.

After this improvement, both Nicky and Thomas adopted the system_context parameter refining method. This is also when we introduced confusion matrices since we noticed that one of the columns had a very large number of false positives when examining results. Then, Nicky created different methods of selecting a GPT response such as majority vote. Another method is to prioritize a patient column, if its response is yes then call the other two columns to see if either agrees, otherwise select no. The latter turned out more useful in lowering false positives while usually not increasing false negatives by even 1. The multiple voting methods to find the right one for each individual question in combination with system_context and input question refining increased some accuracies and F1s by as much as 30%.

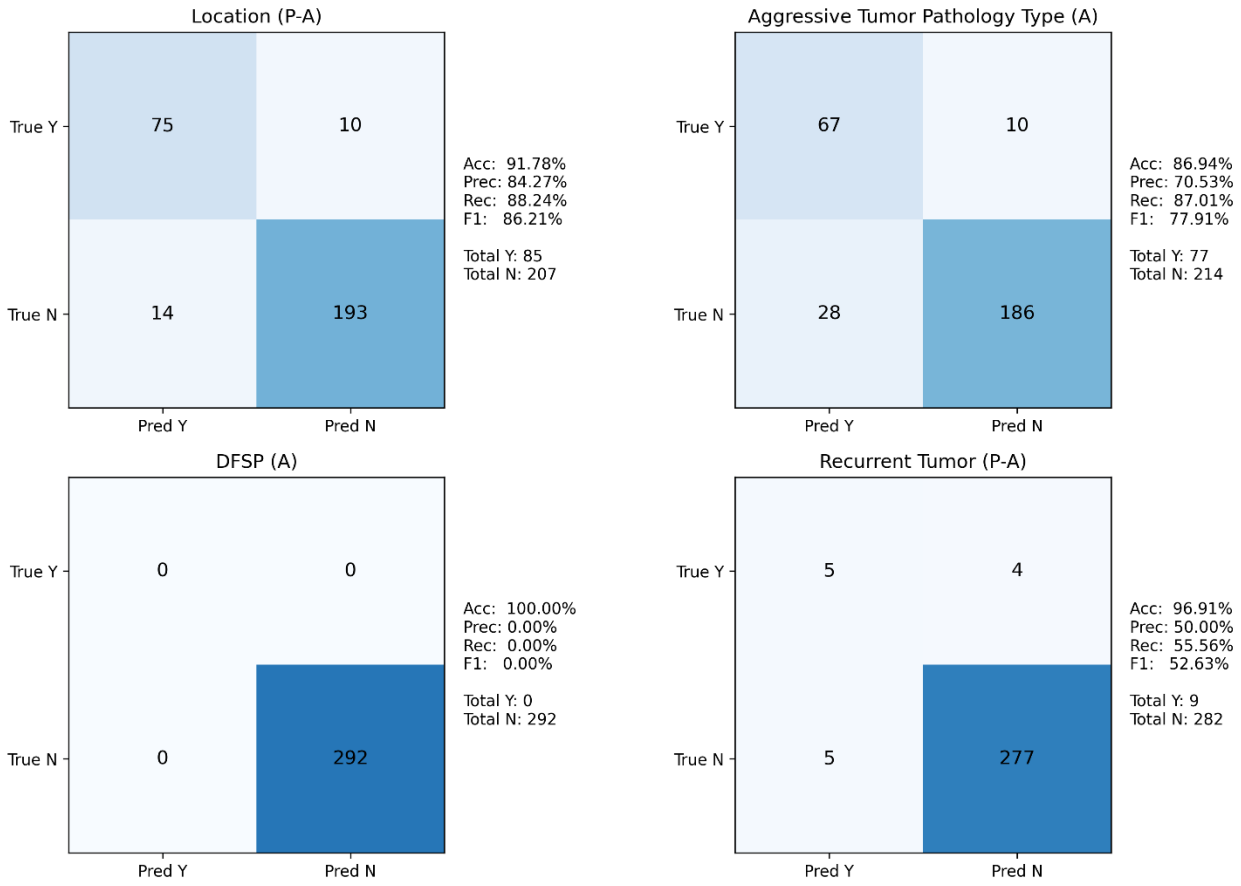
Experimental methods, validation approach:



Unlike typical ML models, we didn't use a train-validation-test split or k-fold cross validation since the goal was to use an LLM in its untrained state. The input is the system_context, which is like instructions to guide the model to the correct way of "thinking", and the user message, which is the question followed by the PE, A&P, and A columns. The model outputs are either a numerical answer for the size question, or yes or no for all other questions. Parameters used are max_tokens, temperature, and top_p. max_tokens is set low to keep the model from giving unnecessary information. The same is true for temperature, which gives the model less creativity in its answer, less like a detailed story and closer to a statistic. top_p controls the probability mass that the model samples from, we set this high, at 0.95, to ensure only the most confident responses are chosen, while keeping a small amount of variability.

Instead of formal validation metrics, we evaluated model performance by comparing its outputs against ground truth answers. This led to the creation of confusion matrices which guided the fine tuning of the system_context to be either more or less strict in borderline cases. In the future when the model is trained, it would likely start with train-validation-test methodology, after the data imbalances mentioned in the future project section get addressed.

Results:



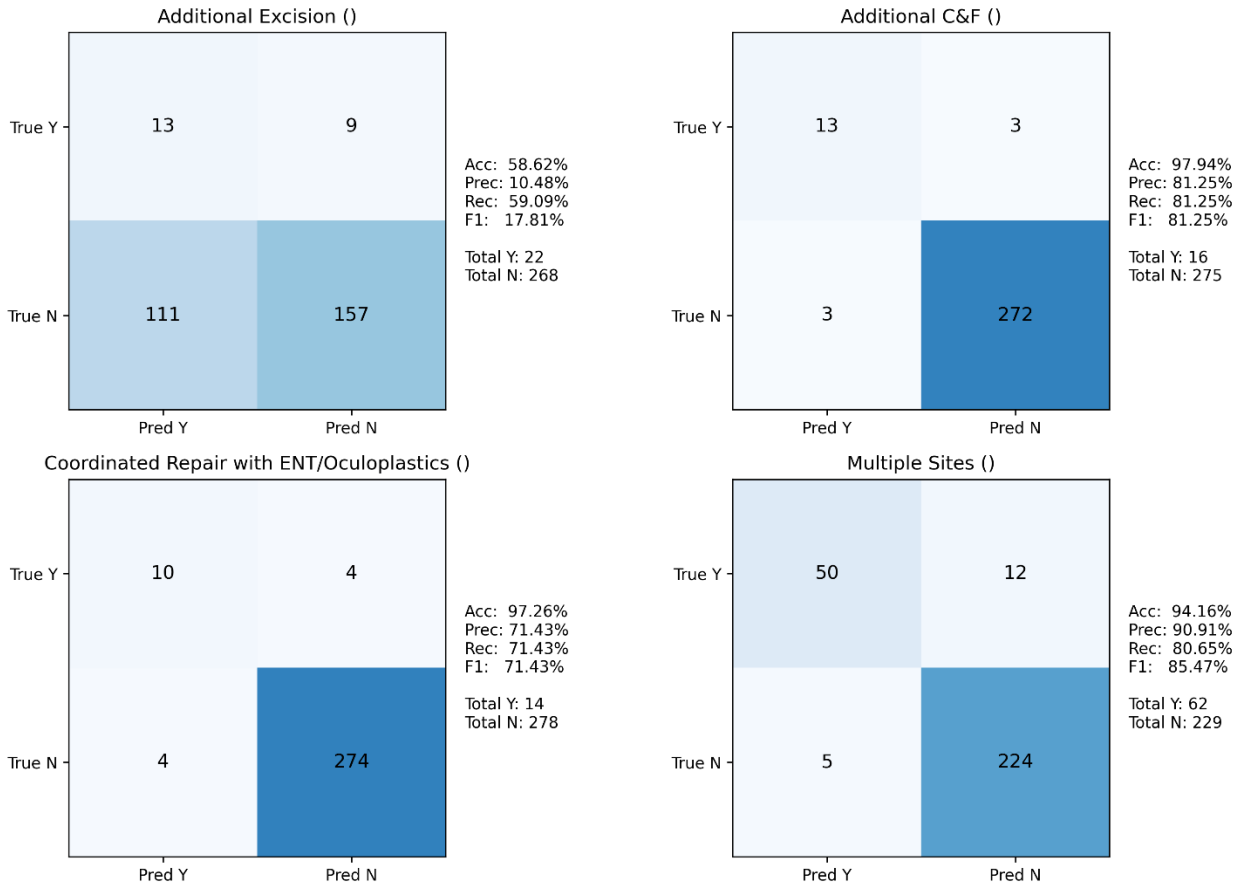


Figure 1. F1 score for 8 outputs

All accuracies but Additional Excision is >=87% and most F1 scores around 75%~85%. Detail information is shown in Figure 1.

Discussion:

The next step in improving the model further would be to train it train-validation-test splits. Guiding the LLM to the correct topic then giving it a clear outline in its context is essential. This is why training the model on patient notes and letting it know what constitutes a correct response to a certain question would help greatly. Although, this would require 9 separate models to be trained, 1 for each question, to ensure the highest accuracy. It is also important to know that giving it too much text in either the context or the user message can cause the model to get lost and hallucinate information. This is why we split each column of a single patient's note into 3 separate GPT calls, otherwise it would be too much for the model. The specification of how strict the model should be is essential in keeping the model from saying no every time in a subject such as cancerous tumors.

Several questions had issues with data imbalance. 4 of the 9 had 16 or less yes answers in the ground truth, which made it difficult to accurately judge how good the model truly is for those questions. As previously noted, although ground truth labels were provided, some of the tumor size data appeared to have been entered despite the absence of corresponding information in the patient notes. This discrepancy resulted in significant time lost during data validation and may have occurred due to the accidental removal of size details during the de-identified process.

Ideas/aims for future extramural project:

As indicated in Andrew's second and third objectives from the presentation, this work could serve as a foundation for a follow-up project. One natural extension would involve developing a user interface (UI), with the option to tailor the system either to individual patients or to multiple patients. Given that the core functionality has already been modularized into reusable functions, future development and maintenance are expected to be significantly more efficient. Another recommendation for future work involves obtaining the missing data for the four columns that currently lack ground truth values (i.e., the "Y" labels). While this task primarily involves data collection, it would provide a low-barrier entry point for new contributors to become familiar with the pipeline before engaging with more complex tasks such as Goal 2. While there remains some uncertainty regarding tum

Publications resulting from project:

Manuscript is being prepared.