Visualizing LLM Decision-Making: A Transparent Approach to Extracting Renal Tumor Characteristics from Pathology Reports

Chen Chen (cc4865), Anna Dominic (amd9200)

Abstract—The proliferation of Electronic Health Records (EHRs) presents a unique opportunity to leverage Large Language Models (LLMs) in medical diagnostics, particularly for extracting and interpreting complex data from pathology reports. Our project aims to address the challenge of accurately identifying renal tumor characteristics, namely tumor margin status, lymph node staging and presence of multiple tumors. We start by exploring ways to generate more insights from open-sourced models using BERTViz to visualize the attention mechanism of the open-sourced models, providing insights into their decision-making process. Since close-sourced models do not allow us to visualize attention weights, we also developed a chain-of-thought (CoT) prompt pipeline understand the LLM's reasoning at each step of refining the lengthy input text to the portion relevant to the labels being predicted. This approach aims to not only reduce hallucination errors, but also make the decision-making process of LLMs transparent to clinicians and foster trust. We hope our effort sets a precedent for applying LLMs more effectively in EHR analysis. We will also implement an interactive confusion matrix with various performance metrics to compare CoT against traditional prompt engineering, as well compare performance across races. With our initiatives, we hope to establish a scalable framework for accurate, transparent, and equitable EHR data extraction, improving collaboration and clinical outcomes.

1 Introduction

1.1 Electronic Health Records (EHR)

Electronic health records (EHRs) are digital versions of patients' medical histories maintained by healthcare providers in a centralized system. They generally include comprehensive health information like medical history, diagnoses, medications and treatment plans. EHRs aims to facilitate coordinated care across different departments and specialists by enabling healthcare professionals to access data more efficiently. Insightful analytics can also be aggregated from large amounts of EHRs to help clinicians make informed and customized treatment decisions. By centralizing and digitizing patient information, EHRs help streamline healthcare workflows and enhance the overall quality of care.

1.2 Challenges of extracting information from EHRs using machine learning

Extracting information from electronic health records (EHRs) is challenging due to data inconsistency and unstructured data formats. Different healthcare professionals often document patient data in diverse ways, using varied terminologies. For instance, one provider might use "hypertension," while another might label it as "high blood pressure." Such variations complicate data standardization. Moreover, inconsistencies in data formats and coding systems across different EHR platforms (e.g. hospitals) make it difficult to standardize data preprocessing for machine learning models. Additionally, much of the critical clinical information is embedded in lengthy free-text clinical notes, requiring natural language processing tools to decipher and categorize the data correctly. This process often requires frequent verification by domain experts, making it manual and tedious.

1.3 Using Large Language Models (LLM) for Feature Extraction from EHRs

Significant advancements have been made in employing Large Language Models (LLMs) to extract information from Electronic Health Records (EHR). For our project, we aim to enhance the capability of LLMs in extracting pivotal features for predicting renal functions post-nephrectomies via visualization techniques. We intend to focus on classifying tumor margin status, tumor lymph node staging and indicator of multiple tumors from pathology reports, since accurate extractions are crucial for determining the appropriate treatment plan and prognosis. This can be a challenging task for LLMs as pathology reports are extensive, intricate, and laden with medical terminology pertaining to various organs. The complexity can lead LLMs to make naive mistakes or produce hallucinations, drawing from irrelevant phrases and syntax. Additionally, LLMs are also known as black boxes, making it challenging for clinicians to understand the rationale behind their predictions.

Our project seeks to mitigate these challenges and improve the efficacy of LLMs in processing pathology reports for more accurate feature extractions. Our main contributions will be three fold:

1. Use BERTViz for validating the models' predictions and ensuring their alignment with clinical relevance. By understanding the attention patterns, we can ensure that the models are not only accurate but also focusing on clinically pertinent information, thereby enhancing the reliability and trustworthiness of the automated classification process.
2. Utilize Chain-of-Thought prompt engineering to iteratively refine input text, identifying the most relevant information to generate accurate responses and minimize hallucination errors.
3. Visualize changes in classification performance between Chain-of-Thought and naive approaches using an interactive confusion matrix that breaks down performance by patient race, feature of interest, and prompt engineering approach.

In summary, our project aims to leverage techniques like BertViz and Chain-of-Thought prompt engineering to improve the accuracy and explainability of large language models in extracting crucial features from pathology reports. By incorporating an interactive confusion matrix that analyzes classification performance across various patient demographics, features of interest, and prompt engineering approaches, we will offer clinicians a tool to compare the performance of chain-of-thought prompting across different patient groups. By reducing the potential for hallucination through Chain-of-Thought prompting, we aim to enable clinicians to harness the feature extraction capabilities of LLMs for pathology reports more reliably. This will streamline the process compared to manual labeling, fostering efficient human-AI collaboration.

2 Related Works

Our research is situated within a rapidly evolving landscape, where the integration of Large Language Models (LLMs) into healthcare analytics, particularly through the analysis of Electronic Health Records (EHR), represents a cutting-edge frontier. The CPLLM model, outlined in "Clinical Prediction with Large Language Models" (Ben Shoham and

Rappoport, 2023) [1], underscores a significant leap forward. By fine-tuning pre-trained LLMs with clinical prompts, CPLLM has demonstrated satisfactory performance in predicting patient diagnoses over conventional models and even state-of-the-art methods like Med-BERT. This approach not only leverages the deep contextual understanding inherent in LLMs but also sidesteps the need for domain-specific pre-training, thus opening new avenues for EHR-based predictive analytics.
Further elaborating on the utility of LLMs in medical contexts, the study "Integrating Retrieval-Augmented Generation for Medical Question Answering" [2] expands upon the foundational work of CPLLM by proposing an ensemble learning framework to enhance the accuracy and reliability of medical QA systems. This innovative approach employs both boosting and clustering methods to refine the selection and integration of relevant information, demonstrating the potential of LLMs when augmented with strategic information retrieval techniques to address complex biomedical inquiries.
Moreover, the application of Chain of Thought (CoT) prompting in Nephrology, as explored in "Chain of Thought Utilization in Large Language Models and Application in Nephrology," [3] provides a compelling example of how LLMs can be guided to perform more nuanced and accurate clinical analyses. Through a series of interconnected queries, this methodology deepens the model's reasoning capabilities, highlighting the benefits of CoT in enhancing diagnostic precision and offering a richer contextual understanding of patient data.
These studies collectively point to a significant gap in current research — the need for more sophisticated methods to interpret and visualize how LLMs process and prioritize EHR data to make predictions. While CPLLM offers a robust framework for disease prediction using LLMs, challenges remain in understanding the "why" behind its predictions, particularly in addressing naive errors and hallucinations.
Our project aims to build upon the groundwork laid by CPLLM and related methodologies by introducing advanced visualization techniques to uncover the inner workings of LLMs in processing EHR data. By doing so, we intend to refine the application of LLMs in healthcare, enhancing their interpretability, reliability, and overall utility in clinical predictions. This endeavor will use the CPLLM model as a baseline for our investigation, providing a foundation for our explorations into improving LLM-based predictions of renal function post-nephrectomies.

3 Background

The project will utilize Electronic Health Records (EHRs) from NYU Langone Health, specifically pathology reports for nephrectomies from 2000 to 2024, to extract margin status, lymph node staging and presence of multiple masses for kidney tumors.

3.1 Data Privacy

Our mentor, Dr. Madhur Nayan, has facilitated our inclusion in the Institutional Review Board (IRB) and sponsored our accounts, granting us access to the NYU Langone Health's data lake. This pivotal support enables us to leverage the extensive resources and data available at NYU Langone Health for our research endeavors. Mindful of the sensitive nature of Patient Health Information (PHI) and the paramount importance of data privacy, we plan to employ all Large Language Models (LLMs) within NYU Langone's secured virtual desktop infrastructure and via NYUTron, ensuring that all interactions with these models are conducted in a secure and encrypted environment. Furthermore, we are dedicated to using synthetic data for demonstrations and sharing only aggregate performance statistics with the teaching staff. This approach ensures the ethical handling of data while upholding the integrity and confidentiality of patient information.

3.2 Features of Interest and Clinical Importance

From pathology reports, we will extract the following three features for patients that have had nephrectomies (partial or complete kidney removal) as treatment for kidney tumors:

1. **Margin status:** The margin status indicates whether cancer cells are present at the edge of the removed tissue, providing insight into whether the tumor has been entirely excised or tumor cells have been left behind. Correct identification of margin status from pathology reports is crucial for nephrectomy patients because it directly impacts their treatment and prognosis. If the margins are positive (cancer cells are present), there's a higher risk that residual cancer cells remain, necessitating further treatment like additional surgery or adjuvant therapy. Negative margins (no cancer cells) suggest a more favorable outcome and reduce the likelihood of recurrence. Thus, accurate margin status assessment guides clinicians in developing appropriate follow-up strategies, improving patient outcomes, and reducing unnecessary procedures.
2. **Lymph Node Staging**: Lymph node staging (pN) determines whether cancer has metastasized to regional lymph nodes, which materially influences the patient's prognosis and treatment plan. Higher pN stages (i.e. pN1) often correlate with more advanced disease and require more aggressive treatments, including additional surgeries, systemic therapies, or radiation. Conversely, lower pN stages (i.e. pN0) indicate less aggressive cancer, potentially allowing for more conservative management. Thus, correctly staging lymph nodes helps clinicians tailor treatment strategies, predict patient outcomes, and optimize follow-up care.
3. **Indicator for Multiple Masses**: Identifying multiple kidney tumors correctly in pathology reports is critical for nephrectomy patients as it can alter treatment plan and prognosis. The presence of multiple tumors suggests a higher likelihood of underlying genetic conditions like hereditary syndromes, which may require further screening and surveillance. It can also indicates a more aggressive disease course, often necessitating more extensive surgical approaches and more diligent post-operative monitoring. Recognizing multiple tumors early on allows clinicians to plan appropriately for potential metastatic risks and recurrence, ensuring that treatment strategies are more customized for patients with less frequent conditions such as multiple tumors.

Accurate extraction of these features from pathology reports is essential for optimizing the management of nephrectomy patients. By obtaining precise information on margin status, lymph node staging, and the presence of multiple tumors, clinicians can better understand each patient's condition and tailor treatment plans accordingly. This allows for more efficient, personalized care, reduces unnecessary procedures, and ensures that follow-up strategies are effectively aligned with individual patient needs. Additionally, efficient extraction of these important features using machine learning facilitates more accurate prediction of renal function post-surgery. This predictive capability allows clinicians to anticipate potential complications and further refine treatment plans to optimize recovery and long-term health outcomes.

3.3 Pathology report structure

Pathology reports are detailed medical documents that contain information on a patient's diagnostic findings, often on specimen collected during surgery. They are often lengthy and mention multiple organs, making it challenging when the features being extracted pertain to only one organ. The following example skeleton of a pathology report illustrates this concept:

1.VENA CAVA, VENTRAL MARGIN: ...(5.1 CM) SURGICAL MARGINS ARE FREE OF TUMOR 2. URETER Biopsy ... 3. KIDNEY, RIGHT: PARTIAL NEPHRECTOMY - ONCOCYTOMA (2.4 CM)... margins Involved by invasive carcinoma ... 4. KIDNEY, LEFT: PARTIAL NEPHRECTOMY - ONCOCYTOMA 1.6 cm ... MARGINS ARE NEGATIVE FOR CARCINOMA

In this example above, the true label is positive margin for the kidney tumor. However, LLM can potentially misclassify negative margin or identify the wrong tumor size due to the vena cava tumor. Hallucination errors like this highlight why we're using Chain-of-Thought prompting to iteratively narrow down the input text, ensuring that only relevant parts are considered for the final labels. By refining the input in this manner, we aim to minimize misclassification

risks and improve the accuracy of identifying critical features like margin status, lymph node staging and presence of multiple tumors.

3.4 Objectives

To mitigate these types of errors, our project aims at providing users with more insights on how the model identifies and prioritizes information pertinent to the classification task. We expect to achieve our goal through the following aspects:

- Analyze and classify Electronic Health Records (EHRs) using GPT-2, BERT, and BioBERT to identify nephrectomy details, classify nephrectomy type and surgical margin status, and visualize attention mechanisms with BERTViz for deeper insights into the decision-making process.
- Streamlining data preprocessing to exclude irrelevant words and sections, thereby enhancing the model's focus on the most informative elements
- Visualize a chain-of-thought prompt pipeline that details the predicted labels along with the rationale behind LLM's prediction
- Assess whether a prompt engineering system, trained to minimize errors at each step above, will yield higher accuracy than a simplistic approach that merely requests classification on renal tumor margin status, lymph node staging and presence of multiple tumors. This will be done via an interactive confusion matrix that breaks down performance by patient race, feature of interest, and prompt engineering approach.

With these approaches, we hope to enhance human-AI collaboration and foster trust among physicians by making the LLM's multi-step reasoning transparent. Physicians can more easily understand the rationale behind each result and locate the pertinent text for verification, mitigating the opaque "black-box" nature of LLMs.

4 METHODS

4.1 Open-sourced Models

In this study, we leveraged three open-source models—GPT-2, BERT, and BioBERT—to analyze and classify Electronic Health Records (EHRs). Additionally, BERTViz was employed to visualize the attention mechanisms within these models, offering insights into the decision-making processes during text classification. A small labelled sample of the dataset was used for evaluation and comparison, since EHR data is largely unlabelled and unstructured.

4.1.1 GPT-2

GPT-2, a generative language model, was employed for extracting specific information from EHRs through a series of guided prompts. This model is particularly adept at generating coherent and contextually relevant text, making it suitable for the extraction tasks outlined below. The following tasks were formulated as prompts and processed sequentially by GPT-2 to derive the necessary information:

Tasks:

- Confirm if the text describes a nephrectomy.
- Determine if it is a partial or radical nephrectomy.
- Specify which kidney was operated on.

These tasks were designed to utilize GPT-2's strength in understanding and generating detailed textual responses, thereby facilitating the extraction of precise medical information from the EHRs.

4.1.2 BERT and BioBERT

BERT (Bidirectional Encoder Representations from Transformers) and its domain-specific variant, BioBERT, were utilized to classify the EHR reports. These models were chosen for their robust performance in natural language understanding tasks, particularly in the medical domain for BioBERT. The classification tasks involved determining whether the reports pertained to a partial or radical nephrectomy and whether the surgical margin was positive or negative.
The preprocessing pipeline involved the following steps:

- **Tokenization**: The EHR texts were tokenized to convert them into a format suitable for input into the models. This step included padding and truncation to handle varying text lengths, ensuring consistency across inputs.
- **Model Prediction**: The tokenized texts were fed into BERT and BioBERT to generate predictions. The models output logits, which were then converted to probabilities to determine the most likely class.
- **Mapping Predictions**: The predicted logits were mapped to human-readable labels (e.g., 'Partial' vs. 'Radical' and 'Positive' vs. 'Negative').

To enhance the model's comprehension, a sentence-sequence pair approach was employed:

- **Sentence A**: Task prompt (e.g., "Determine whether the surgical margin is positive or negative.")
- **Sentence B**: EHR text.

This methodology provided context to the model, enabling it to focus on specific tasks while considering the relevant sections of the EHR text.

4.1.3 BERTViz for Visualization

BERTViz was utilized to visualize the attention weights of the BERT and BioBERT models [4]. Snapshots of this interactive tool can be seen in Figure 7 and Figure 8. This tool aids in elucidating the internal workings of the models by highlighting the parts of the text that the models focus on during prediction.

The following are insights from visualization we aim to gain from using BERTViz:

- **Understanding Model Focus**: By visualizing attention weights, we could identify which sections of the EHR text were considered most important by the models during the classification tasks.
- **Token Influence Analysis**: The visualizations facilitated the identification of tokens that had a significant impact on the model's decision-making process, thereby highlighting influential and non-influential segments of the text.
- **Model Interpretation**: These insights contributed to a better understanding of the model's interpretability, providing transparency in how the models arrived at their predictions.

4.2 Close-sourced Models: Chain-of-Thought Prompting

4.2.1 "Brute Force" Prompt

Using GPT4, Our plan involves constructing a Chain-of-Thought prompt pipeline that iteratively narrows down the relevant text for feature extraction. Before designing these Chain-of-Thought prompts, we will establish a "brute force" prompt that serves as a baseline for comparison. See "brute force" prompt in Figure 1 that requests all three features of interest simultaneously, providing clear instructions on identifying the signs that can pinpoint the values of each feature. Note that the prompt begins with the instruction to GPT-4 to "only answer based on the part of [pathology report] that discuss nephrectomies or concern renal tumors." However, as we will later observe in the results, this instruction is not consistently followed.

4.2.2 "Chain of Thought" Prompt

Contrarily, the Chain-of-Thought approach breaks down the long prompt into following four sequential steps that iteratively narrow down the section of relevant text for extraction. A skeleton of this setup is demonstrated in Figure 2.

- **Step 1.** Identifying the section discussing renal tumors; this can be more challenging in reports where sections on different organs are not clearly numbered

Prompt: Return the text in path_report_text that discuss renal tumors and call it text_renal. If the sections of path_report_text are numbered or lettered, keep the number or letter of the section in text_renal. You can find this typically by looking for a section that starts with "(left right) kidney" or "renal tumor" (both case irrelevant) and followed by "nephrectomy" shortly after, but this is not the only rule. Text_renal should include sections that talk about lymph nodes that relate to kidney or nephrectomy too. There may be two sections of text not adjacent to each other that both discuss renal tumors, so parse the entire path_report_text even if you have found text related to renal tumors already. Connect pieces of texts you find to fall under text_renal with "...". Finally, output "reason_text_renal" that describe how you created text_renal.

- **Step 2.** Decide if there are multiple tumors mentioned
  Prompt: Decide whether text_renal mentions multiple tumor. If yes, set mult_mass_ind to 1, otherwise 0. The first sign for multiple tumor is mentioning more than one tumor sizes in formats such as (...cm). An example is you may see "(...cm)" or "...cm" twice, or (..cm and ..cm). Note that if text_renal mentions greatest dimension and additional dimension, those together describe one tumor only. Another sign for multiple tumor is text_renal mentioning two distinct types of renal tumor. Additional signs also include "two (multiple) foci" and seeing "greatest dimension" twice. Also output "reason_mult_mass_ind" that describes the reason behind the value for mult_mass_ind.

- **Step 3.** 3. For every renal tumor, pinpointing phrase(s) pertaining to tumor margin status and lymph node status
  Prompt: From text_renal, create text_renal_marg_pn that only contain text that describes tumor surgical margin status and lymph node status. If previously mult_mass=1, do this for every renal tumor. Sentences that discuss surgical margin may contain keywords 'margin', "surgical margin", "edge", "surface", "tumor is ...present", "tumorcarcinoma...invades...at the inked...". This is not an exhaustive list, also use your own judgment. For lymph node status, these sentences should contain "pN" followed by "0", "1", "2", "3" or "X" or a pattern like lymph nodes ([number]/[number])". Connect pieces of texts you find to fall under text_renal_marg_pn with "...". . Also output "reason_text_renal_margin_pn" that describes the reason behind the value for text_renal_marg_pn.

- **Step 4.** Generating the appropriate response, i.e. classifying positive margins if present on any renal tumor, extracting the maximum dimension of the largest tumor, and indicating presence for multiple tumors)
  Prompt:
  ***
  Within text_renal_marg_pn, identify whether the surgical/resection margin is positive (neph_margin). Any patterns like "tumor is presenttumor (carcinoma) invades at the inked surface(or typos for edge)/surgical (resection) margin/capsule" is a positive margin (neph_margin=1). If you see this pattern, ignore all other language on margin and assign neph_margin=1. If you do not see this pattern, look for language that indicates negative margin such as "free of carcinoma", "negative for carcinoma", etc. If you see pattern "tumor is present/tumor (carcinoma) invades fat/sinus/vascular space/renal vein/artery margin", this does not mean neph_margin=1, look for other language to indicate whether margin is positive or negative. If you cannot find anything that describes surgical/resection margin, neph_margin = "unknown".
  ***
  Next, identify whether cancer has spread to lymph nodes (neph_stageN). Typically, you should see this denoted as pN followed by "0", "1", "2", "3" or "X". If lymph node is not mentioned, neph_stageN = pNX. If you see "lymph nodes (03)", output neph_stageN as pN0. You can also use "[number] lymph nodes negative for tumor" to decide on neph_stageN. If mult_mass_ind = 1, which means multiple tumors, only produce neph_stageN with pN of the highest level you've found in text_renal_marg_pn, and output neph_margin=1 if any tumor has positive margin.
  ***
  The final output I need are neph_margin, neph_stageN and mult_mass_ind. Summarize your reasoning ('reason_output') for how you came up with neph_margin, neph_stageN and mult_mass_ind. In reason_output, make sure you mention how you decided on mult_mass_ind from text_renal.

The Chain-of-Thought approach divides the 'brute force' prompt into four steps, with each prompt progressively narrowing down the input text. This iterative process helps GPT-4 focus on relevant information, reducing hallucination errors that might arise from processing text unrelated to kidneys, our domain of interest. We have shown a video demonstration [5] using Microsoft's Azure AI Studio to illustrate this approach on the following synthetically generated pathology report.

John Doe, Male, 55 y.o., underwent the following procedures on 5/24/2020: TERMINAL ILEUM, APPENDIX AND RIGHT COLON: HEMICOLECTOMY: - severe CHRONIC ACTIVE COLITIS, involving right colon and appendix, consistent with crohns disease. - NEGATIVE FOR DYSPLASIA AND MALIGNANCY. - TERMINAL ILEUM WITH MILD REACTIVE CHANGES. - FOURTEEN LYMPH NODES, NEGATIVE FOR CARCINOMA (0/14). RENAL TUMOR, LEFT: PARTIAL NEPHRECTOMY - CLEAR CELL RENAL CELL CARCINOMA (2.3 CM), SEE SUMMARY CASE SUMMARY: KIDNEY MACROSCOPIC SUMMARY: Procedure:Partial nephrectomy Specimen Laterality:Left Tumor Site:Lower pole Tumor Size:Tumor greatest dimension:2.3cm Tumor Focality:Unifocal Macroscopic Extent of Tumor:Tumor limited to kidney MICROSCOPIC SUMMARY: Histologic Type:Clear cell renal cell carcinoma Sarcomatoid Features:Not identified Tumor Necrosis:Not identified Histologic Grade:G2: Nuclei slightly irregular, approximately 15 um; nucleoli evident Microscopic Tumor Extension:Tumor limited to kidney Margins:Uninvolved by invasive carcinoma Lymph-Vascular Invasion:Not identified Pathologic Staging (pTNM): Primary Tumor (pT):pT1a: Tumor 4 cm or less in greatest dimension, limited to the kidney Regional Lymph Nodes (pN):No nodes submitted or found Number of Lymph Nodes Examined:Not applicable, see Comment Number of Lymph Nodes Involved:Not applicable, see Comment Distant Metastasis (pM):Not applicable Pathologic Findings in Non-Neoplastic Kidney:Significant pathologic alterations None identified

The input text of 185 words was iteratively refined to 'CLEAR CELL RENAL CELL CARCINOMA (2.3 CM), ...Margins:Uninvolved...Regional Lymph Nodes (pN):No nodes submitted or found' which formed the basis for producing three labels: negative tumor margin, no lymph node submitted (pNX), and no multiple tumors detected.

4.3 Performance Comparison: Interactive Confusion Matrix

Using D3.js and TensorFlow in Observable, we created an interactive confusion matrix [6] (Figure 3) that enables dynamic comparisons between prompt engineering approaches (brute force vs. chain-of-thought), race, features of interest, and any combination of the three for GPT4. Note that race data was synthetically generated to protect patient anonymity. This customizable matrix allows users to normalize the confusion matrix according to predictions (where rows sum to 100%) or ground truth (where columns sum to 100%), change the color scale, and show or hide values on the matrix. At the bottom of this interactive confusion matrix dashboard, accuracy, weighted average precision, weighted average recall, and weighted average F1 scores are displayed for any combination of prompt engineering approach, race ("all", "White", "Asian", "Black or African American" and "Hispanic"), and feature of interest. Weighted average was selected to weight each statistic by the number of occurrences of a class to take into

account class imbalances. Figure 4-6 demonstrates the interface of the interactive confusion matrix.

This approach enables:

1. **Comparing Performance:** We can compare the overall performance of brute force versus chain-of-thought prompt engineering approaches, helping to evaluate the added value of prompts that reduce input text to the relevant portions.
2. **Assessing Fairness Across Races:** Within each approach, we can assess fairness across different races. If certain races have notably high error rates, it may indicate underlying characteristics in these patients' pathology reports that affect the suitability of the prompts, and thus warrant further understanding of the nature of the errors.
3. **Parameter Toggles:**The tool's flexibility allows users to set one parameter constant and toggle others, leading to effective comparisons. For instance, setting "Feature of Interest" to margin allows us to compare performance between races and prompt engineering approaches or fix a race while toggling between features of interest.

The high level of interactivity and personalization offered by this tool makes it flexible enough to adapt to various analysis needs and supports a comprehensive evaluation of the different combinations.

5 EVALUATION

5.1 Open-sourced Models

5.1.1 GPT-2 Performance

While GPT-2 demonstrated some capability in generating relevant information, it also produced nonsensical outputs, often referred to as hallucinations. This issue is likely due to the following factors:

- **Lengthy Texts**: The extensive length of the EHR texts diluted the model's focus, leading to less coherent and relevant outputs.
- **Text Truncation and Preprocessing**: Performance improved when the texts were truncated or preprocessed. Techniques such as removing irrelevant sections and simplifying the text helped the model focus on the essential information.
- **Lack of Pre-training on Domain-specific Data**: GPT-2 might have performed better if it had been pre-trained on domain-specific data. However, due to the limited availability of labeled EHR data, extensive pre-training was not feasible.

Despite these challenges, GPT-2's ability to generate detailed text highlights its potential for tasks requiring narrative responses, provided it is adequately pre-trained and the input text is appropriately managed.

5.1.2 BERT and BioBERT Performance

BERT and BioBERT were employed for classification tasks, specifically to determine the type of nephrectomy and the surgical margin status. These models exhibited significantly better performance compared to GPT-2, attributed to their robust architecture designed for understanding and classifying text. Key observations include:

- **Small Sample Size**: The evaluation was constrained by the limited availability of labeled data, with only five reports being annotated. As a result, reporting detailed performance metrics was deemed impractical.
- **Impact of Preprocessing**: The application of natural language processing (NLP) techniques, such as tokenization, normalization, and truncation, enhanced model performance. Preprocessing helped in reducing the noise and focusing the model's attention on relevant parts of the text.
- **Performance and Sentence Length**: The performance of both BERT and BioBERT deteriorated with increasing sentence length, indicating a challenge in handling very lengthy texts.

Overall, BERT and BioBERT demonstrated robust classification capabilities, especially when the input text was well-preprocessed and concise.

5.1.3 Attention Visualization with BERTViz

BERTViz was employed to visualize the attention mechanisms of the models, providing valuable insights into their decision-making processes. The key findings from the attention visualizations are as follows:

- **Influential Tokens**: Most tokens were found to be non-influential, with the models' attention predominantly focused on specific tokens such as '[SEP]' (the separator token). This focus on the separator token suggests that the models heavily relied on the structure of the input text.
- **Keywords Attention**: Words directly related to the classification tasks, such as 'partial', 'radical', 'left', 'right', and 'nephrectomy', received significant attention. This indicates that the models correctly identified and concentrated on the critical terms relevant to the tasks.
- **Visualization Insights**: The visualization helped in understanding the distribution of attention across the text, highlighting which parts of the input were deemed important by the models. This insight is crucial for validating the model's focus and ensuring that the predictions are based on clinically relevant information.

5.1.4 Summary of Findings

In summary, while GPT-2 showed potential for generating narrative responses, its performance was hindered by lengthy texts and a lack of domain-specific pre-training. On the other hand, BERT and BioBERT excelled in classification tasks, particularly when aided by effective preprocessing techniques. The attention visualizations provided by BERTViz were instrumental in validating the focus and reliability of the models, ensuring their outputs were grounded in clinically relevant information.

5.2 Close-sourced Models: GPT4

5.2.1 Brute Force vs. Chain-of-Thought Prompt Engineering

On 100 pathology reports, we can compare the overall performance of the brute force and chain-of-thought prompt engineering approaches by setting race to "All" and choosing "Feature of Interest" in the interactive confusion matrix. Figure 4-6 aggregate this comparison across all three features in terms of accuracy, false negative rate, and false positive rate. We observe that, in all three metrics the chain-of-thought approach consistently outperforms the brute force approach due to the input text being shortened to the portion relevant to the question. The most significant improvements are the reduction in false positive rate in lymph node staging (neph_stageN) from 38% to 4%, and the reduction in false negative rate for margin status from 36% to 7%.

These reductions are crucial because lower false positive and false negative rates mean more reliable clinical decision-making. In particular, reducing the false positive rate in predicting lymph node metastasis minimizes unnecessary treatments. A false positive prediction means that the model incorrectly identifies lymph node metastasis when it is not actually present. This misdiagnosis can lead to patients undergoing invasive and often stressful treatments that they do not need, such as additional surgeries, chemotherapy, or radiation therapy. These treatments can have significant side effects, causing unnecessary pain, fatigue, and other health complications, all while increasing healthcare costs. On the other hand, reducing the false negative rate ensures that fewer cases of lymph node metastasis are missed. A false negative occurs when the model fails to identify the condition even though it's present. Missing this diagnosis may result in patients not receiving the treatment they need, allowing the cancer to progress and potentially spread further. This can lead to more advanced stages of disease that are harder to treat and have

worse outcomes. By improving accuracy and reducing false negatives, patients receive timely and appropriate treatment, leading to better overall health outcomes and a higher likelihood of successful recovery. The improvement in accuracy and reduction of false positive and false negative rates lead to greater precision in extracting important features for better predictions of post-nephrectomy kidney function. Although PR-AUC was not directly calculated, the accuracy, precision, and recall all exceed 90% using the chain-of-thought method. From this, we can infer that the chain-of-thought approach significantly improves upon the 35% PR-AUC achieved by the CPLLM method in one of our baseline papers.

5.2.2 Performance across races

Below are several observations from comparing Brute Force and Chain-of-Thought Performances Across Races:

1. **General Comparison**: The brute force approach exhibits more variation in error rates among races compared to the chain-of-thought approach. The chain-of-thought method shows higher accuracy, weighted-average precision, recall, and F1 scores across all racial groups.
2. **Margin Status**: The White population shows the poorest performance in margin status, followed by Black, Hispanic, and Asian groups.
3. **Lymph Node Staging**: As 65 samples lack lymph node assessment (pNX), only 35 pathology reports are considered in this analysis. Consequently, variations among racial groups are not significant, and all groups achieve very high accuracy (>90%).
4. **Multiple Tumors**: Under the chain-of-thought approach, the multiple tumor indicator shows more variation among racial groups than other features. Black and African American groups have significantly lower weighted-average performance.

With more labeled data, the fairness of model predictions across races can be assessed more thoroughly for both the brute force and chain-of-thought approaches through this interactive confusion matrix.

6 Conclusion

6.1 Open-sourced models

Looking ahead, several avenues for future research and development are suggested by our findings. Enhancing the performance of models like GPT-2 could involve pre-training them on larger, domain-specific datasets to improve their ability to generate accurate and relevant outputs in the medical domain. Additionally, addressing the challenge of handling longer texts is crucial. This could be achieved through segmenting texts into manageable chunks or utilizing models specifically designed for long-form text processing.
Finally, building on the success of BERTViz, there is a clear opportunity to develop enhanced visualization tools. Such tools would not only aid in understanding model behavior but also increase trust and transparency in automated classification systems, ultimately supporting more effective decision-making in clinical settings.

6.2 Close-sourced models and Interactive Confusion Matrix

If we had more time, we would have labeled 1,000 reports to understand aggregate performance through the interactive confusion matrix in a more representative way. With additional time, we would also conduct subgroup analysis, incorporating more patient demographics such as gender and age group. Nevertheless, the chain-of-thought interactive confusion matrix establishes a framework for a pipeline using LLMs to extract clinical features from pathology reports.
In the future, leveraging this framework with a more comprehensive dataset and an expanded demographic analysis will allow for a deeper understanding of the nuanced differences between racial, gender, and age groups. Such insights will guide the development of more equitable and accurate models for clinical feature extraction, ultimately leading to improved patient care and outcomes.

7 FIGURES

Fig. 1: Brute Force Prompt for Extracting Features of Kidney Tumor from Pathology Reports

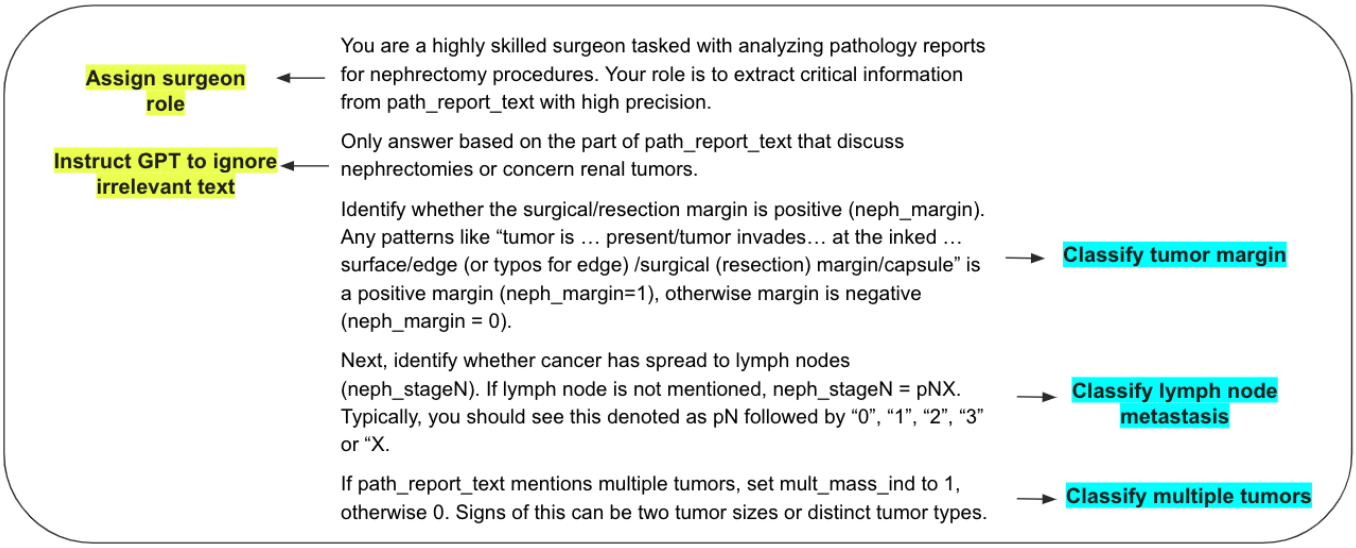


Fig. 2: Structure of Chain-of-Thought Prompts for Extracting Features of Kidney Tumor from Pathology Reports

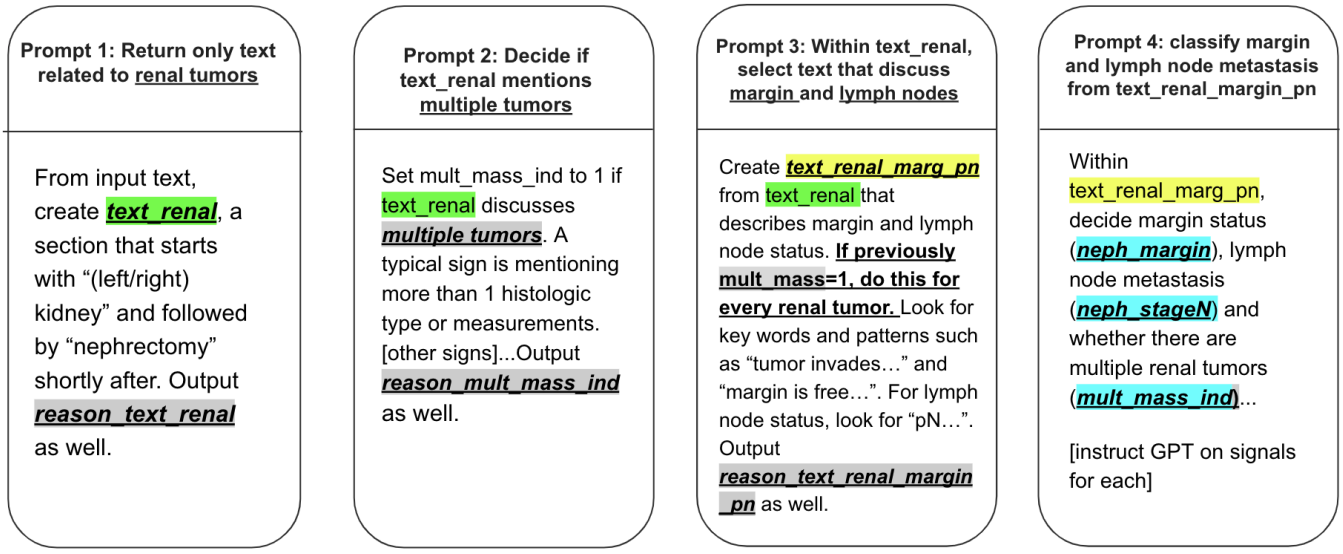
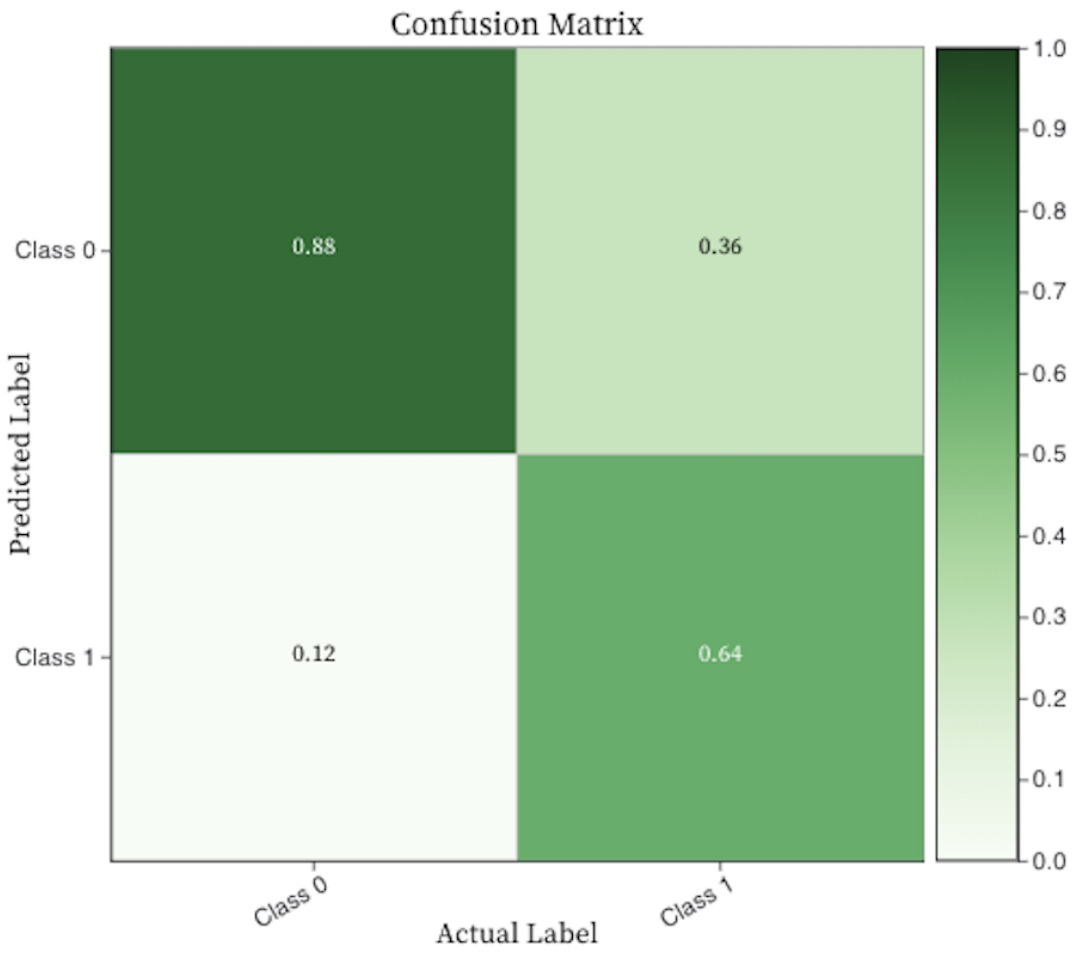


Fig. 3: Interactive Confusion Matrix for Renal Tumor Extraction



Accuracy: 0.7684
Weighted Average Precision: 0.7684
Weighted Average Recall: 0.7622
Weighted Average F1: 0.7646

Select Prompt Engineering Approach

Brute Force

Select Race

All

Select Feature of Interest

Margin

Change the color scale of the confusion matrix

green

Display cell values on confusion matrix

On Off

Normalize confusion matrix

Prediction Ground Truth All No

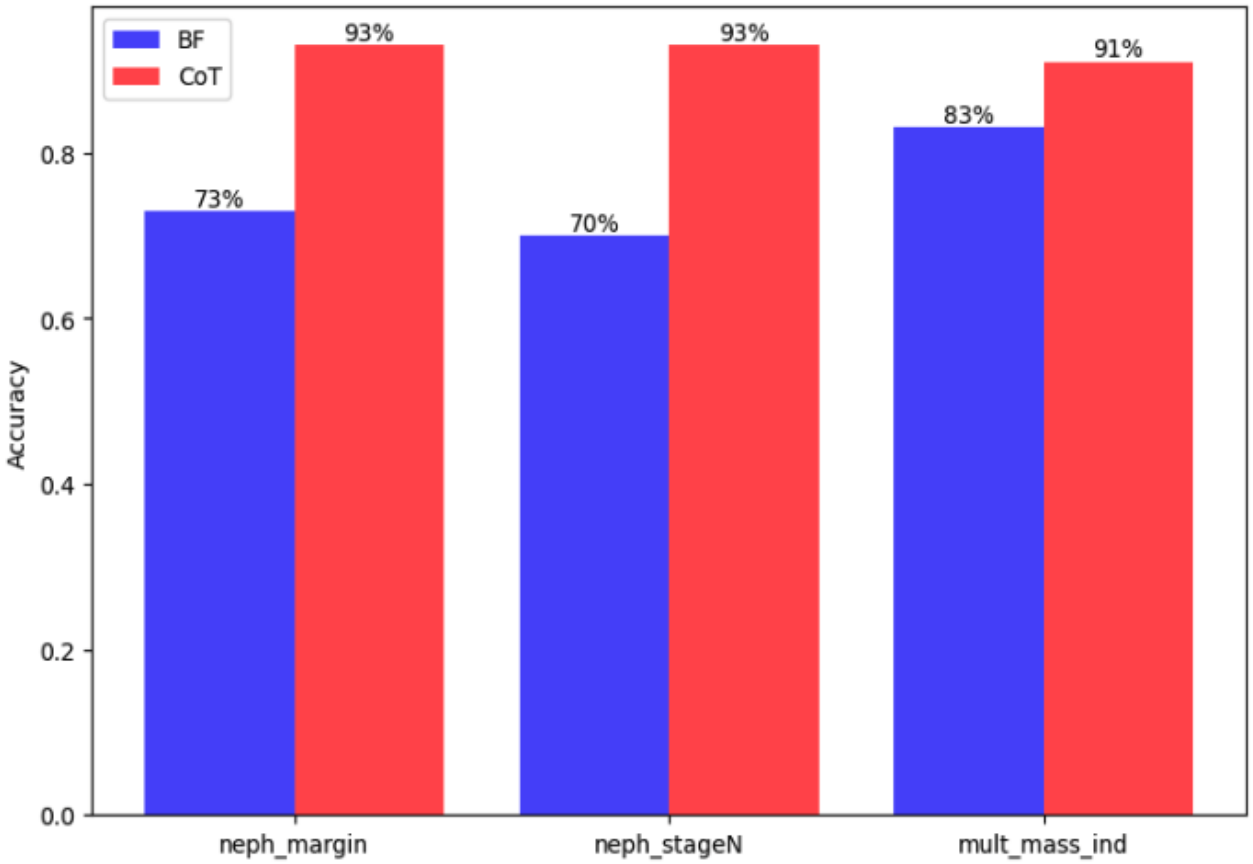


Fig. 4: Accuracy - Brute Force vs CoT

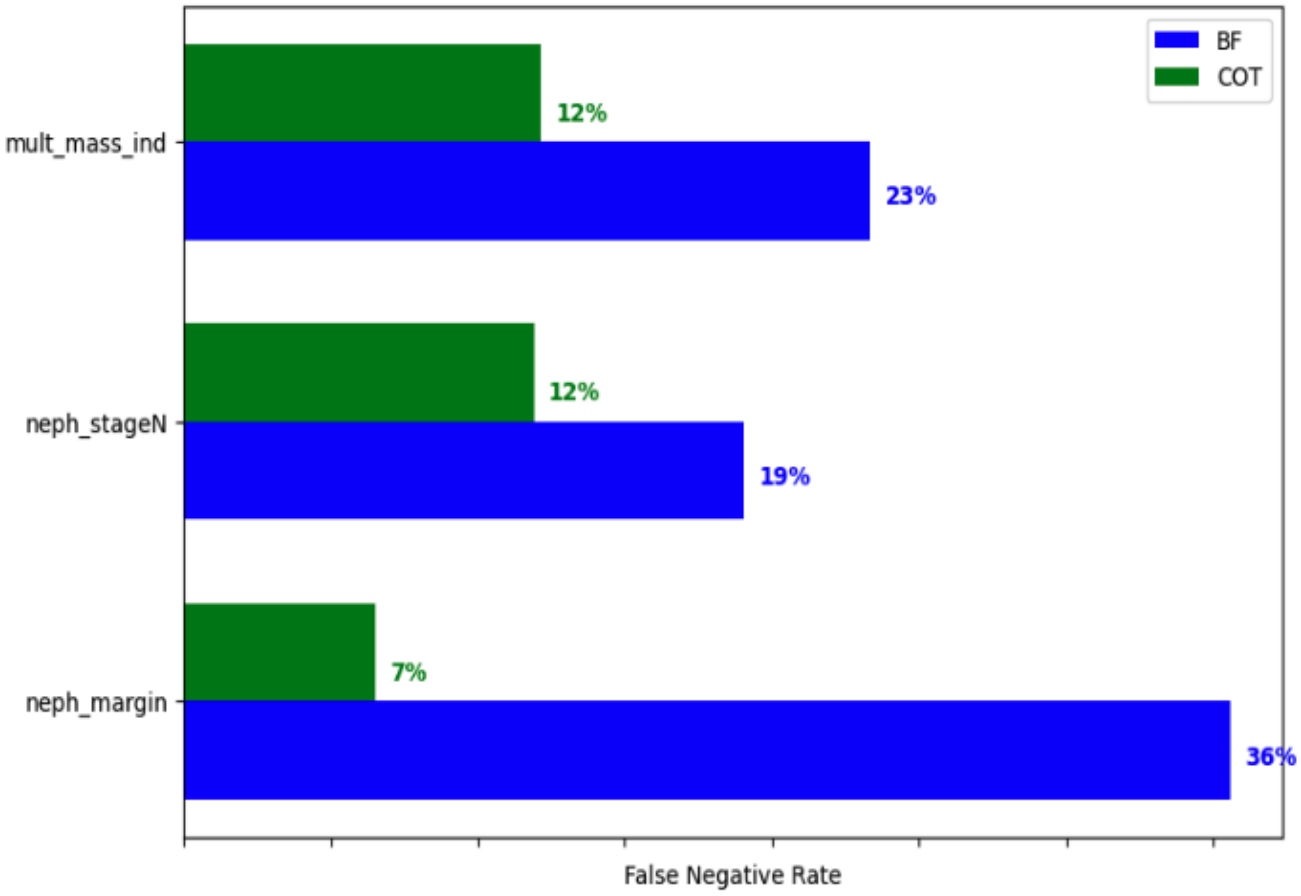


Fig. 5: False Negative Rate - Brute Force vs CoT

Fig. 6: False Positive Rate - Brute Force vs CoT

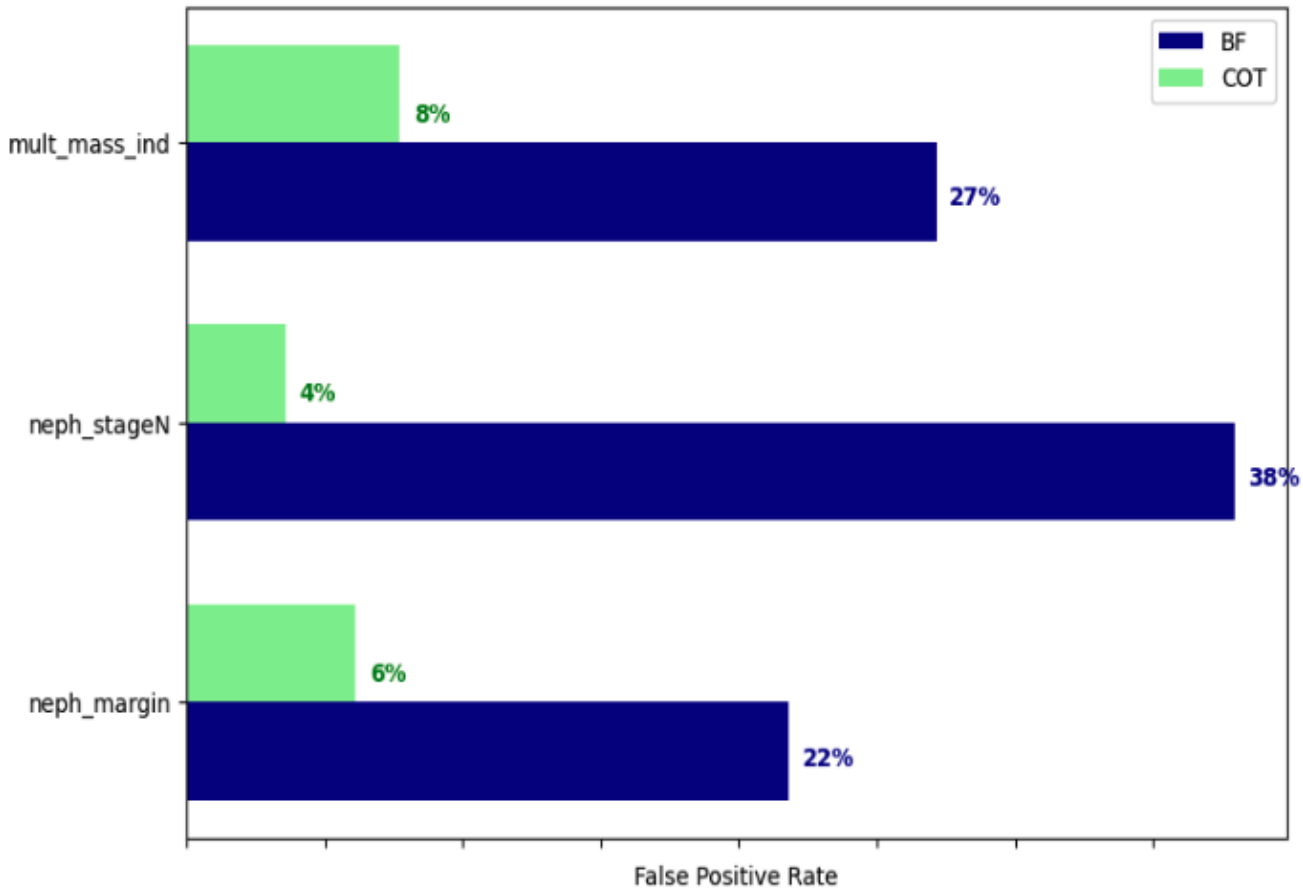


Fig. 7: BERTViz Model View: Visualizing Attention Weights

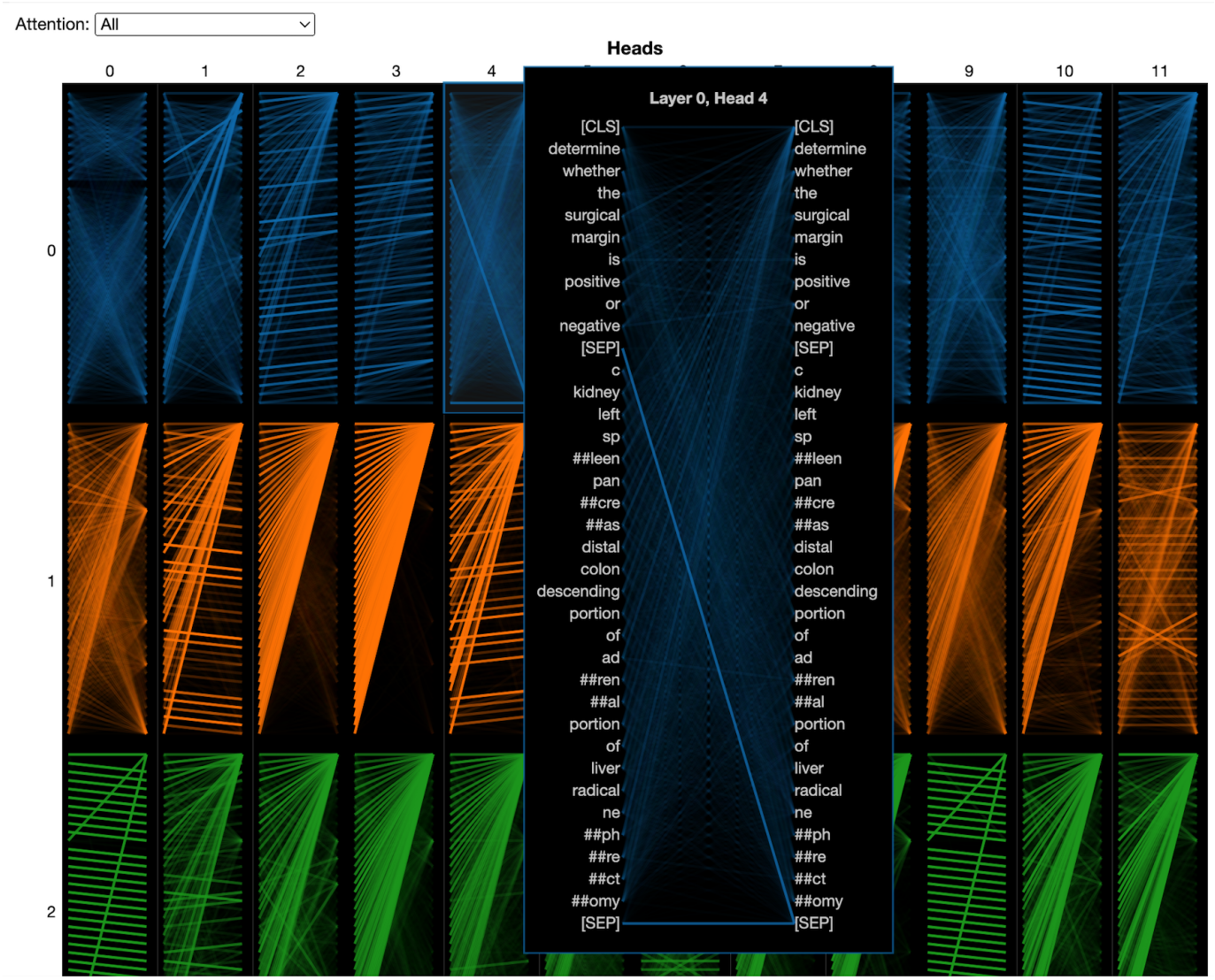
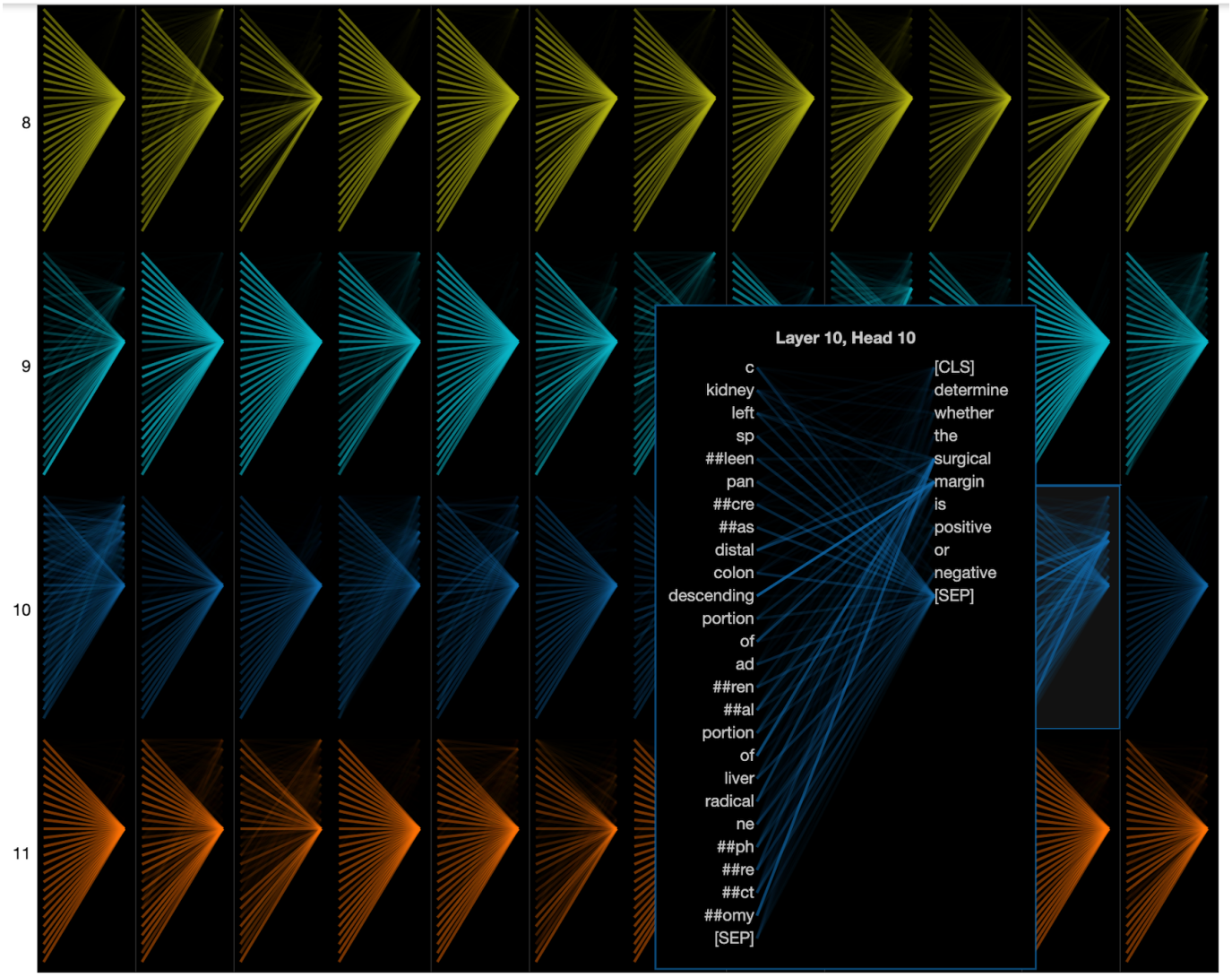


Fig. 8: BERTViz Model View: Visualizing Attention Weights



References

[1] O. B. Shoham and N. Rappoport, "Cpllm: Clinical prediction with large language models," *arXiv preprint arXiv:2309.11295*, 2024, v2. [Online]. Available: https://doi.org/10.48550/arXiv.2309.11295 2

[2] J. Miao, C. Thongprayoon, S. Suppadungsuk, O. A. Garcia Valencia, and W. Cheungpasitporn, "Integrating retrieval-augmented generation with large language models in nephrology: Advancing practical applications," *Medicina*, vol. 60, no. 3, p. 445, 2024. [Online]. Available: https://doi.org/10.3390/medicina60030445 2

[3] J. Miao, C. Thongprayoon, S. Suppadungsuk, P. Krisanapan, Y. Radhakrishnan, and W. Cheungpasitporn, "Chain of thought utilization in large language models and application in nephrology," *Medicina*, vol. 60, no. 1, p. 148, 2024. [Online]. Available: https://doi.org/10.3390/medicina60010148 2

[4] A. Dominic. (n.d.) Bertviz visualization for feature extraction from pathology report. Google Colab. [Online]. Available: https://colab.research.google.com/drive/14gOEHHH_a4gHhHrHtzr0hQhM7J6Arvnu?usp=sharing 3

[5] C. Chen, "Chain without prompt: Extracting relevant text from pathology reports," *Loom*, 2024, accessed May 2. [Online]. Available: https://www.loom.com/share/baf52ecc0af84cbd99f23a48017eeb3c?sid=22781f6a-4d38-4d4d-a12c-a9542942a3eb 4

[6] ——, "Visualizing llm decision-making: Interactive confusion matrix for renal tumor extraction," *Observable*, 2024, accessed May 8. [Online]. Available: https://observablehq.com/d/d08122e535e5740f 4