A Comprehensive Exploration and Evaluation of Foundational and Subspecialized Large Language Models in Healthcare and Medicine

by

Felicia Liu

Supervisor: Dr. Farzad Khalvati
April 7th, 2025

B.A.Sc. Thesis

Division of Engineering Science
UNIVERSITY OF TORONTO

** Final Thesis Report **

A Comprehensive Exploration and Evaluation of Foundational and Subspecialized Large Language Models in Healthcare and Medicine

Submitted by

Felicia Liu

Supervised by

Dr. Farzad Khalvati

University of Toronto

Faculty of Applied Science and Engineering

April 7, 2025

Abstract

This thesis investigates the effectiveness of Large Language Models (LLMs) for medical imaging tasks, specifically glioma classification and segmentation, and compares their performance to that of traditional convolutional neural networks (CNNs). While LLMs have shown strong performance in text-based healthcare tasks, their utility in image-based applications remains underexplored. Using the BraTS 2020 dataset of multi-modal brain MRIs, we evaluated a general-purpose vision-language LLM (LLaMA 3.2 Instruct) both before and after fine-tuning, and benchmarked its performance against custom 3D CNNs. For glioma classification (Low-Grade vs. High-Grade), the CNN achieved 80% accuracy and balanced precision and recall. The general LLM reached 76% accuracy but suffered from a specificity of only 18%, often misclassifying Low-Grade tumors. Fine-tuning improved specificity to 55%, but overall performance declined (e.g., accuracy dropped to 72%). For segmentation, three methods - center point, bounding box, and polygon extraction, were implemented. CNNs accurately localized gliomas, though small tumors were sometimes missed. In contrast, LLMs consistently clustered predictions near the image center, with no distinction of glioma size, location, or placement. Fine-tuning improved output formatting but failed to meaningfully enhance spatial accuracy. The bounding polygon method yielded random, unstructured outputs. Overall, CNNs outperformed LLMs in both tasks. LLMs showed limited spatial understanding and minimal improvement from fine-tuning, indicating that, in their current form, they are not well-suited for image-based tasks. More rigorous fine-tuning or alternative training strategies may be needed for LLMs to achieve better performance, robustness, and utility in the medical space.

Acknowledgements

Firstly, I want to thank Jay Yoo, my PhD mentor, for his ongoing support and guidance. He has been an incredible help throughout this process, meeting with me weekly and offering valuable advice that has greatly influenced my work. I've learned so much from his insights and mentorship. I also want to thank Dr. Farzad Khalvati for fostering a collaborative and motivating environment. His thoughtful feedback and suggestions during our team meetings have played a critical role in shaping the direction of my undergraduate thesis. Lastly, I'd like to acknowledge the rest of the IMICs lab. While I haven't had the chance to work closely with everyone, I truly appreciate their questions and feedback during the monthly meetings. Their comments have been very helpful and have pushed my work forward. I'm truly grateful for everyone and all the support I've received along the way.

Contents

1 Introduction **1**

1.0.1 Context . . . 1

1.0.2 Research Gap and Significance of Proposed Work . . . 1

1.0.3 Goals and Objectives . . . 1

2 Background and Literature Review **2**

2.1 Artificial Intelligence and Deep Learning in Healthcare . . . 2

2.2 CNNs in Traditional Medical Imaging . . . 3

2.2.1 Overview of CNNs . . . 3

2.2.2 Applications of CNNs in Healthcare . . . 4

2.2.3 Fine-tuning CNNs for Performance . . . 4

2.2.4 Limitations of CNNs . . . 5

2.3 Large Language Models in Healthcare . . . 5

2.3.1 Overview of LLMs . . . 5

2.3.2 Strengths of LLMs . . . 6

2.3.3 Effective Applications of LLMs in Healthcare . . . 7

2.3.4 A Paradigm Shift in AI in Healthcare . . . 7

2.4 Research Framework and Rationale . . . 8

2.4.1 Motivation for Research . . . 8

2.4.2 Objectives, Hypothesis, and Approach . . . 8

2.4.3 Research Goals and Significance . . . 8

3 Design Decisions **9**

3.1 Dataset and Model Selection . . . 9

3.1.1 BraTS 2020 Dataset . . . 9

3.1.2 Llama-3.2-11B-Vision-Instruct . . . 10

3.1.3 Unsloth Vision Fine-tuning . . . 10

3.1.4 SciNet Computation Resources . . . 11

3.1.5 Google Colab Computation Resources . . . 11

3.2 Data Pre-Processing . . . 12

3.2.1 Image Data Processing . . . 12

3.2.2 Label Data Processing . . . 13

3.2.3 Segmentation Data Processing . . . 13

3.2.4 Creating Conversation Datasets for Fine-tuning . . . 14

3.3 Evaluation Approaches . . . 15

3.3.1 CNN Evaluation Approach . . . 15

3.3.2 LLM Evaluation Approach . . . 15

4 Image Classification Task **16**
4.1 Introduction . . . 16
4.1.1 Tumor Grade Classification Task . . . 16
4.2 Methodology and Implementation . . . 16
4.2.1 CNN Baseline Model Parameters and Training . . . 16
4.2.2 LLM General Model Setup and Prompting . . . 17
4.2.3 LLM Subspecialised Model Fine-tuning . . . 18
4.2.4 LLM Consistency and Robustness Test . . . 18
4.2.5 Evaluation Metrics . . . 19
4.3 Results . . . 19
4.3.1 CNN Baseline Performance . . . 19
4.3.2 LLM General Model Classification Performance . . . 21
4.3.3 LLM Subspecialised Model Classification Performance . . . 23
4.3.4 LLM Consistency Test Classification Performance . . . 26
4.4 Discussion . . . 26
4.4.1 Key Findings . . . 26
4.4.2 Limitations . . . 27
4.5 Conclusion . . . 29

5 Image Segmentation Task **29**
5.1 Introduction . . . 29
5.1.1 Tumor Voxel Segmentation Task . . . 30
5.1.2 Segment Anything Model for Segmentation . . . 30
5.1.3 Tumor Center Point Segmentation Task . . . 31
5.1.4 Tumor Bounding Box Segmentation Task . . . 31
5.1.5 Tumor Bounding Polygon Segmentation Task . . . 31
5.2 Methodology and Implementation . . . 31
5.2.1 CNN Baseline Model Parameters and Training . . . 31
5.2.2 LLM General Model Setup and Prompting . . . 32
5.2.3 LLM Subspecialised Models Fine-tuning . . . 33
5.2.4 Evaluation Metrics . . . 33
5.3 Results . . . 34
5.3.1 CNN Baseline Performance . . . 34
5.3.2 LLM General Center Point Model Performance . . . 38
5.3.3 LLM Subspecialised Center Point Model Performance . . . 40
5.3.4 LLM General Bounding Box Model Performance . . . 41
5.3.5 LLM Subspecialised Bounding Box Model Performance . . . 43
5.3.6 LLM General Bounding Polygon Model Performance . . . 45

5.3.7 LLM Subspecialised Bounding Polygon Model Performance . . . 46
5.4 Discussion . . . 48
5.4.1 Key Findings . . . 48
5.4.2 Limitations . . . 50
5.5 Conclusion . . . 51

6 Conclusion and Future Work **52**

A Classification Task: CNN Baseline Models **57**

B Segmentation Task: CNN Baseline Models **59**

C Baseline CNN Segmentation **61**

D General LLM Center Point Segmentation **64**

E Fine-Tuned LLM Center Point Segmentation **67**

F General LLM Bounding Box Segmentation **70**

G Fine-Tuned LLM Bounding Box Segmentation **73**

H General LLM Polygon Segmentation **76**

I Fine-Tuned LLM Polygon Segmentation **79**

List of Figures

1 Diagram of a standard CNN pipeline. 3
2 Example of results from brain tumor segmentation using C-CNN enhanced (blue), core (green), and edema (red) regions segmented [5]. . . 4
3 Diagram of a standard transformer architecture [10]. 6
4 The BraTS dataset provides multi-modal MRI scan [17]–[21]. 10
5 Full methodology pipeline from raw data to evaluation and comparison. 12
6 Scan modality visualization for an LGG patient [17]–[21]. 12
7 Scan modality visualization for an HGG patient [17]–[21]. 13
8 Sample segmentations using a) center points, b) bounding boxes, c) bounding polygons, and d) the given form of binary voxels. 14
9 Sample LLM conversation dataset conversion for the classification task. 14
10 Diagram illustrates how predictions for each patient were made by the LLM. 16
11 Sample response for one patient from the general LLM model. Most responses follow the correct format although there are some outliers. The tally is shown at the bottom and this scan was predicted as LGG. 17
12 LLM general model prompt sample for the classification task. 18
13 FLAIR scan slices of a) High Grade Glioma, and b) Low Grade Glioma. These images clearly illustrate the distinct features of each condition, and the LLM is expected to predict them accurately [17]–[21]. 19
14 Training and validation accuracy and loss curves for the CNN model (learning rate = 4e-7, epochs = 200). 20
15 Visual representation of the ratio of HGG to LGG labels across each test patient's axial scan slices. 21
16 Visual representation of the ratio of HGG to LGG labels across each test patient's coronal scan slices. 22
17 Visual representation of the ratio of HGG to LGG labels across each test patient's sagittal scan slices. 22
18 LLM model loss after fine-tuning for 100 steps with a learning rate of 2e-5. 23
19 LLM model loss after fine-tuning for one epoch with a learning rate of 5e-6. 23
20 Visual representation of the ratio of HGG to LGG labels across test patient axial scan slices with glioma evaluated using the small general LLM model. 24

21 Visual representation of the ratio of HGG to LGG labels across test patient axial scan slices with glioma evaluated using the fine-tuned LLM model of 100 steps. 25
22 Visual representation of the ratio of HGG to LGG labels across test patient axial scan slices with glioma evaluated using the fine-tuned LLM model of one epoch. 25
23 LLM model prompt sample for the center point segmentation task. . . 32
24 LLM model prompt sample for the bounding box segmentation task. . 33
25 LLM model prompt sample for the bounding polygon segmentation task. This prompt was adapted from the methodology used in a previous study that successfully applied prompt-based LLM segmentation. Source: [29] 33
26 Training and validation loss curves for the CNN model (learning rate = 5e-5, epochs = 100) showing stable training and no overfitting. 34
27 Evaluation metrics for the CNN baseline model predictions. 35
28 Baseline CNN segmentation visualization showed accurate glioma segmentation with correct size, location, and voxel boundaries. 36
29 Baseline CNN segmentation visualization showed accurate glioma segmentation, but with less precise tumor boundaries. 36
30 Baseline CNN segmentation visualization showed overestimation of glioma size, capturing broader brain regions or the entire brain. 36
31 Baseline CNN segmentation visualization showed challenges in detecting smaller, irregularly shaped gliomas. 37
32 Baseline CNN segmentation visualization showed accurate detection of smaller gliomas in potentially higher-quality scans. 37
33 Baseline CNN segmentation visualization showed accurate identification and segmentation of complex gliomas with unusual or multiple boundaries. 37
34 Baseline CNN segmentation visualization showed incomplete segmentation of complex gliomas with irregular shapes. 38
35 Evaluation metrics for the general LLM center point model predictions. 38
36 General LLM center point segmentation visualization showed, for all patients, the model predicted points near the center of the frame with some variance, regardless of the ground truth glioma location. 39
37 LLM model fine-tuning loss (learning rate = 1e-5, steps = 100). 40
38 Evaluation metrics for the fine-tuned LLM center point model predictions. 40

39 Fine-tuned LLM center point segmentation visualization showed, for all pa- tients, the model still predicted points near the center of the frame with more variance, regardless of the ground truth glioma location. . . 41
40 Evaluation metrics for the general LLM bounding box model predictions. 41
41 General LLM bounding box segmentation visualization shows that the model's predictions are notably sparse. 43
42 General LLM bounding box segmentation visualization shows that the model consistently predicts gliomas near the center of the image. . . . 43
43 General LLM bounding box segmentation visualization shows that the model's predictions lack clear differentiation between larger and smaller gliomas. 43
44 LLM model fine-tuning loss (learning rate = 8e-6, steps = 200). 44
45 Evaluation metrics for the fine-tuned LLM bounding box model predictions. 44
46 Fine-tuned LLM bounding box segmentation visualization shows no noticeable improvement in the placement of bounding boxes. 45
47 Fine-tuned LLM bounding box segmentation visualization shows continued insensitivity to glioma size. 45
48 Evaluation metrics for the general LLM bounding polygon model predictions. 45
49 General LLM bounding polygon segmentation visualization showed random predicted shapes, centered in the image with varying sizes. 46
50 LLM model fine-tuning loss (learning rate = 2e-5, steps = 100). 47
51 Evaluation metrics for the fine-tuned LLM bounding polygon predictions. 47
52 Fine-tuned LLM bounding polygon segmentation visualization shows that the segmentations remain largely unchanged. 48
53 Training and validation accuracy and loss curves for the CNN model (learning rate = 1e-4, epochs = 50). 57
54 Training and validation accuracy and loss curves for the CNN model (learning rate = 1e-6, epochs = 80). 57
55 Training and validation accuracy and loss curves for the CNN model (learning rate = 5e-7, epochs = 100). 58
56 Training and validation loss (learning rate = 1e-4, epochs = 70). . . . 59
57 Training and validation loss (learning rate = 1e-5, epochs = 150). . . . 59
58 Training and validation loss (learning rate = 1e-5, epochs = 200). . . . 60

List of Tables

1 Available GPU types and memory allocations in Google Colab Pro [25]. 11
2 Evaluation of the CNN baseline model (learning rate = 4e-7, epochs = 200). 20
3 Evaluation of the general LLM model across axial, coronal, and sagittal imaging orientations during testing. 21
4 Evaluation of the fine-tuned LLM models in comparison to the general LLM. 24
5 Evaluation of the CNN baseline model (learning rate = 5e-5, epochs = 100). 35
6 Evaluation of the general LLM center point model. 39
7 Evaluation of the fine-tuned LLM center point model. 40
8 Evaluation of the general LLM bounding box model. 42
9 Evaluation of the fine-tuned LLM bounding box model. 43
10 Evaluation of the general LLM bounding polygon model. 46
11 Evaluation of the fine-tuned LLM bounding polygon model. 47
12 Metrics for the CNN model (learning rate = 1e-4, epochs = 50). 57
13 Metrics for the CNN model (learning rate = 1e-6, epochs = 80). 58
14 Metrics for the CNN model (learning rate = 5e-7, epochs = 100). . . . 58
15 Metrics for the CNN model (learning rate = 1e-4, epochs = 70). 59
16 Metrics for the CNN model (learning rate = 1e-5, epochs = 150). . . . 60
17 Metrics for the CNN model (learning rate = 1e-5, epochs = 200). . . . 60

1 Introduction

1.0.1 Context

In recent years, there has been a significant shift in the field of medical artificial intelligence (AI) from traditional convolutional neural networks (CNNs) to large-scale deep learning models, particularly large language models (LLMs). CNNs have demonstrated strong performance in various healthcare applications, including disease diagnosis, tumor classification, and medical image segmentation. However, CNNs have inherent limitations, such as their dependence on large annotated datasets and inability to capture long-range, multimodal relationships in medical data. Conversely, LLMs have rapidly evolved and their ability to process and understand large volumes of text has enabled wide adoption in medical natural language processing (NLP) tasks. LLMs have been successfully applied to a variety of healthcare text-based tasks trained using large-scale datasets, transforming healthcare by enhancing clinical decision-making, patient triaging, and automating clinical note generation.

1.0.2 Research Gap and Significance of Proposed Work

The current applications of LLMs in healthcare primarily assume the availability of large-scale datasets and focus on text-based tasks. This focus overlooks two critical gaps: 1) the performance of LLMs when fine-tuned with limited, domain-specific data, and 2) the applicability of LLMs beyond text, particularly in vision-based tasks such as medical image classification and segmentation. While general-purpose LLMs excel in text processing, there remains a lack of comprehensive overviews evaluating accuracy, robustness, and the overall utility of medical LLMs. When available data is limited, or when applied to non-text-based problems, the effectiveness of medical LLMs is unclear and addressing this gap is central for advancing healthcare AI. This thesis will provide insight into when fine-tuning LLMs is necessary and how they can be adapted for non-textual medical tasks. Understanding how general out-of-the-box LLMs compare with subspecialized models can guide model selection and training strategies for many new medical applications. By broadening their utility, medical LLMs have the potential to revolutionize fields like radiology, leading to more accurate and personalized healthcare.

1.0.3 Goals and Objectives

The primary goal of this thesis was to evaluate the effectiveness of general-purpose LLMs compared to subspecialized models across various healthcare tasks, with a focus on limited data and non-textual inputs. The first objective was to compare general and subspecialized (fine-tuned) models. We assessed whether general LLMs, such as Large

Language Model Meta AI (Llama) Models, could perform healthcare tasks effectively without fine-tuning, or if specialized training on niche datasets significantly improved performance. The second objective was to evaluate LLMs for vision-based tasks by implementing and testing medical tasks such as brain tumor classification and tumor segmentation. Through these objectives, this research contributed to a more informed understanding of when general models sufficed and when subspecialized models were necessary in healthcare, offering a framework for selecting optimal models and training needs for various medical applications to achieve robust performance.

2 Background and Literature Review

2.1 Artificial Intelligence and Deep Learning in Healthcare

Artificial intelligence (AI) has rapidly emerged as a transformative force in healthcare, offering innovative solutions to some of the field's most complex challenges. AI was initially developed as simple rule-based systems, such as "if ... then...", but has since evolved into sophisticated algorithms capable of processing vast, multidimensional datasets to support medical decision-making [1]. One of AI's greatest strengths lies in its ability to learn and recognize intricate and complex patterns from large diverse data sources, enabling these computer algorithms to perform complex tasks, for example, translating a patient's entire medical record with reports, notes, and scans, into a single, predictive value that can be used for diagnosis [2].

In medical imaging, deep learning models, such as Convolutional Neural Networks (CNN), have played a significant role in enhancing the detection and classification of tumors and lesions, which are traditionally labor-intensive tasks managed by radiologists [3]. Not only may these AI systems assist physicians in identifying suspicious findings (outlier detection), but they can also contribute to personalized patient care by recommending tailored protocols, monitoring radiation exposure, and minimizing diagnostic errors [3]. Beyond imaging, AI models have successfully shown to be able to integrate structured data (e.g., vital signs, demographics) and unstructured data (e.g., clinical notes) in predicting patient hospitalization needs, allowing for real-time triaging in emergency situations and optimized resource allocation in hospitals [4].

Current AI systems are highly dynamic and continuously improve as they access and process more data, becoming increasingly autonomous, generalisable, and adaptive in clinical environments [2]. This combination of pattern recognition, adaptability, and predictive capability positions AI as a vital tool in modern medicine, enhancing diagnostic accuracy, improving efficiency, and enabling more personalized, data-driven patient care.

2.2 CNNs in Traditional Medical Imaging

CNNs have become central to many image-based healthcare applications due to their exceptional performance in medical imaging tasks. This section will detail the state of CNNs in the medical imaging space.

2.2.1 Overview of CNNs

CNNs are a type of deep learning model designed for processing grid-like data, such as images. They use convolutional layers to apply filters over the input, extracting and learning local features like edges and textures from the images. These features are progressively combined through multiple layers in depth, allowing the network to learn abstract representations as a whole. CNNs also leverage weight sharing, where filters are reused across the image, reducing parameters that need to be adjusted to achieve good predictive results and improving model efficiency. These characteristics enable CNNs to be highly effective for tasks like Image Classification, where the goal is to assign input data to one of several predefined categories, or classes, as CNNs can automatically extract relevant patterns and are robust to shifts in the input image. A standard medical CNN pipeline is given in Figure 1 illustrating the flow from data collection and preprocessing, through convolutional and pooling layers for feature extraction, to a fully connected layer that generates a binary classification output.

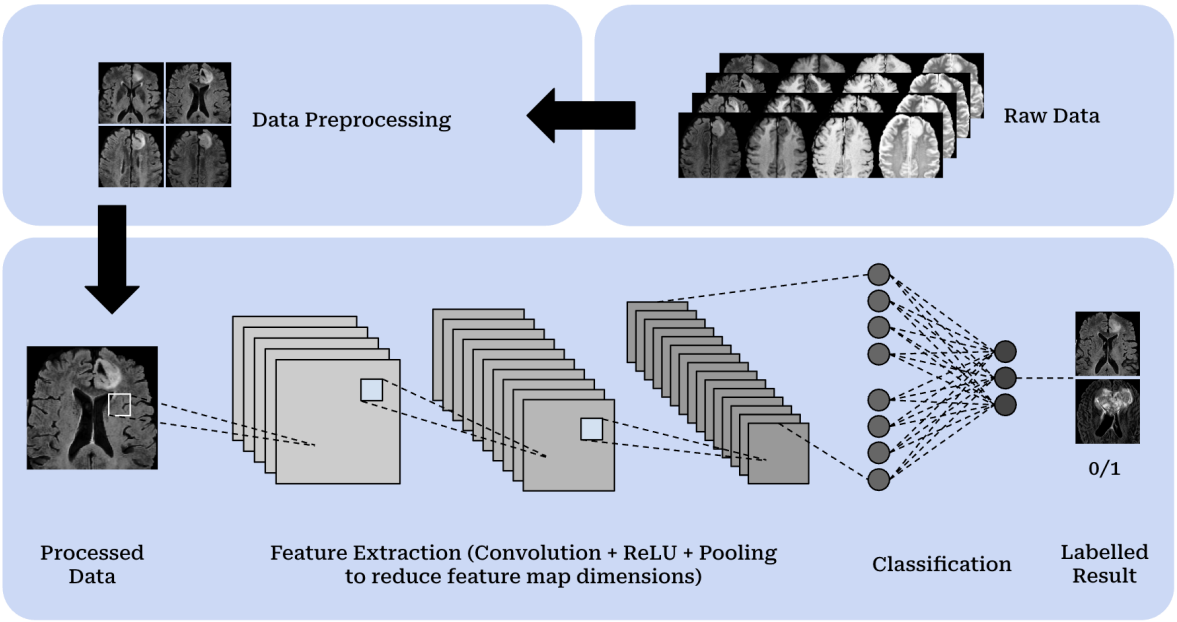


Figure 1: Diagram of a standard CNN pipeline.

2.2.2 Applications of CNNs in Healthcare

CNNs have demonstrated very high performance in a variety of medical imaging tasks, particularly in classification and segmentation. A Cascade CNN with a Distance-Wise Attention mechanism developed for brain tumor segmentation using MRI scans, achieved competitive Dice scores of 0.9203 for whole tumor and 0.8726 for tumor core, demonstrating its effectiveness for accurate and efficient tumor localization [5].

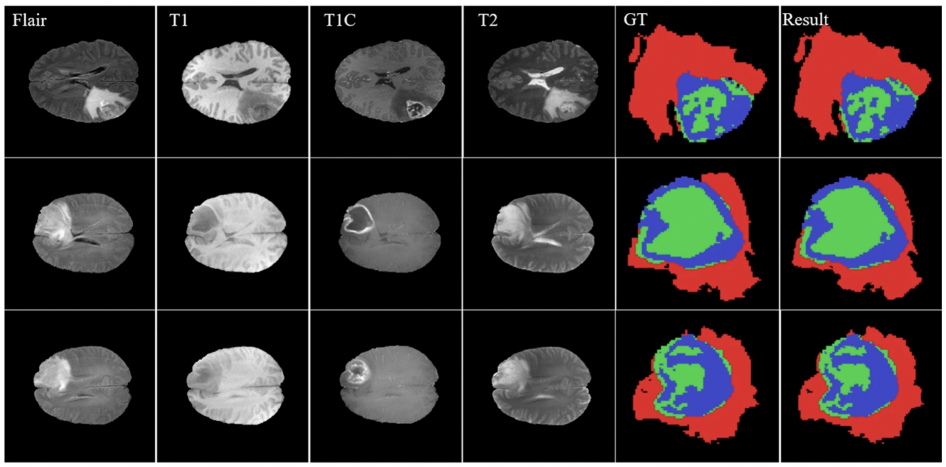


Figure 2: Example of results from brain tumor segmentation using C-CNN enhanced (blue), core (green), and edema (red) regions segmented [5].

For lung disease classification, [6] created a model using lung spectrogram features that achieved 97% accuracy, outperforming previous methods and contributing to high-performance diagnostics for conditions like pneumonia, tuberculosis, and COVID-19 [6]. In retinal image analysis, a ResNet101 based model achieved 94.17% accuracy in classifying eye diseases, such as diabetic retinopathy, while Swin-Unet demonstrated 86.19% mean pixel accuracy in segmenting retinal blood vessels, providing valuable insights for disease detection and treatment planning [7]. Additionally, a CNN designed for classifying normal eyes, cataracts, diabetic retinopathy, and glaucoma achieved 84% accuracy, excelling in diabetic retinopathy detection but highlighting the need for further improvements in glaucoma classification [8].

These applications demonstrate the strength of CNNs in image classification and segmentation tasks, offering reliable, high-accuracy tools for clinicians and radiologists.

2.2.3 Fine-tuning CNNs for Performance

Fine-tuning CNNs on domain-specific medical datasets is a common practice to improve performance by freezing the lower-layer weights and adapting higher layers. With fine-tuning, models are able to leverage learned features from a large dataset while specializing in the new task where data is limited, for example, disease classification

and anomaly detection. This adaptability makes CNNs a powerful tool for medical image analysis, though their limitations underscore the need for more versatile models in healthcare applications.

2.2.4 Limitations of CNNs

CNNs are highly effective for image-based tasks, such as detecting and classifying tumors in medical imaging, but they have limitations when working with other sources of data.

CNNs are not as well suited for developing a semantic understanding of images, even as they can be trained to perform specific image processing tasks [9]. This semantic understanding is often crucial for accurate diagnosis, especially in the medical field. For example, when diagnosing brain tumors, key details from a patient's medical history or symptoms are essential, but CNNs cannot capture or leverage such information directly in their classification models. Additionally, CNNs require large labeled datasets for supervised training, which may not always be available, particularly in specialized fields with rare conditions. They may also have difficulty generalizing to new imaging modalities or diverse patient populations, especially if the quality or diversity of training data is limited, which can result in suboptimal performance. Furthermore, CNNs typically require fixed input sizes, making it challenging to handle variable-sized or very large scans, a common issue in medical imaging.

While CNNs remain a powerful tool for image classification and segmentation tasks, addressing these gaps requires exploring other AI and deep learning approaches capable of processing multimodal data, adapting to a variety of conditions, and working with smaller, more specialized datasets. These advancements are essential for providing more robust, flexible, and accurate diagnostic support for radiologists and clinicians.

2.3 Large Language Models in Healthcare

In recent years, LLMs have gained significant attention, opening up new possibilities for healthcare applications beyond imaging, including enhancing clinical decision-making and automating documentation processes. This section will introduce LLMs and the impact they have had on medical imaging.

2.3.1 Overview of LLMs

LLMs are deep learning models designed to understand and generate natural language text. Unlike CNNs, which specialize in images, LLMs process textual data, enabling tasks like translation, summarization, and question answering. They are trained

on vast datasets of text, learning patterns, syntax, and contextual meaning. LLMs leverage tokenizers, tools used to convert raw text into a suitable representation that the LLM can later process by breaking into smaller units. Once the text is tokenized, LLMs primarily use transformers, an architecture that leverages self-attention to process text, learning from the context and relationships between tokens. This mechanism allows the model to weigh the importance of each word in a sentence, regardless of its position, making it adept at understanding context and handling very long-range dependencies in text. Figure 3 visually represents a standard transformer architecture, showcasing key components such as multi-head self-attention, positional encoding, and skip connections that enable efficient processing of sequential data.

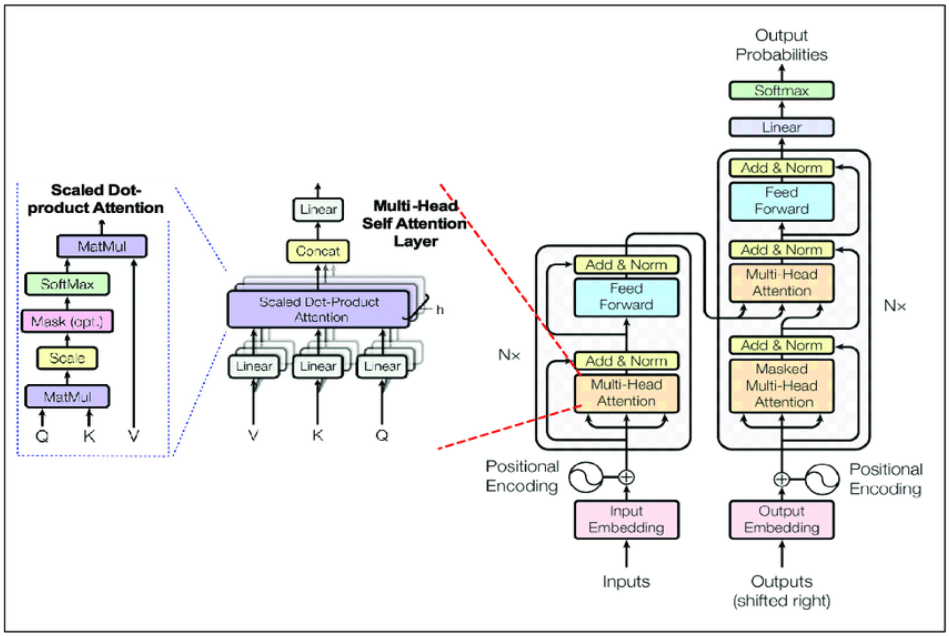


Figure 3: Diagram of a standard transformer architecture [10].

2.3.2 Strengths of LLMs

LLMs excel at processing unstructured textual data, making them highly valuable in healthcare settings where clinical notes, patient histories, and other textual information are critical for diagnosis. LLMs can extract meaningful insights from these textual sources, identifying patterns and nuances that may be overlooked in image-based analysis. Unlike CNNs, LLMs are equipped to handle more than local spatial data, and can integrate multimodal data, processing text with image or sensor data into tokens to create a more comprehensive understanding of a patient's condition. This ability to process and contextualize text opens up possibilities for enhancing diagnostic accuracy, improving decision-making, and supporting clinicians with valuable insights derived from a variety of data sources and modes.

2.3.3 Effective Applications of LLMs in Healthcare

Key and high performing LLMs include GPT (Generative Pre-trained Transformer) and Llama. GPT models, like ChatGPT, excel in text generation for tasks such as conversational agents and content creation. Llama models prioritize efficiency and scalability, handling tasks like summarization and question answering across multiple languages and domains. Fine-tuning these models on domain-specific data has become a common practice, enabling them to successfully specialize in various applications, including healthcare.

For example, LLMs have been successfully used for clinical note generation, with [11] developing HEAL, a 13B Llama2-based model, is designed for medical conversations and automated scribing, achieving 78.4% accuracy on PubMedQA, a dataset with 1,000 question-answer pairs derived from PubMed abstracts [11]. Another model, Med-PaLM, trained on the MultiMedQA benchmark (which includes MedQA with 3,000 USMLE-style questions, MedMCQA with 1,000 clinical questions, PubMedQA, and HealthSearchQA, a dataset of medical questions searched online), excelled in medical concept identification and surpasses human scribes in correctness and completeness [11], [12]. PaLM, a 540-billion parameter model, and its instruction-tuned variant, Flan-PaLM, both trained on extensive medical data, achieved impressive accuracy across the MultiMedQA datasets, including 67.6% on MedQA, surpassing prior results by more than 17% [12]. LLMs have also found success in clinical decision support systems [13] and patient triaging [14], addressing primary care concerns [15], and summarizing key findings [12]. These models demonstrate the capacity of LLMs to process large volumes of medical information, enhancing clinical workflows, decision-making, and overall healthcare efficiency and accuracy.

2.3.4 A Paradigm Shift in AI in Healthcare

Current research in medical LLMs predominantly focuses on text-based tasks using large-scale datasets, driving advancements in natural language processing. However, this approach assumes an abundance of textual data and largely ignores scenarios with limited data or non-textual inputs, such as image-based medical tasks. CNNs excel in image analysis, but they cannot process textual information, limiting their integration into comprehensive diagnostic workflows.

While LLMs are widely applied to text-based tasks through fine-tuning on large datasets, their potential in multimodal and image-based medical applications remains largely untapped. This highlights the need to assess their accuracy, robustness, and utility in data-limited, non-textual tasks, motivating further exploration in these areas.

2.4 Research Framework and Rationale

2.4.1 Motivation for Research

This research aims to deepen our understanding of LLMs and their effectiveness in healthcare, extending beyond the commonly explored areas of large-scale, text-based datasets. While LLMs have demonstrated considerable success in text-heavy tasks, there remains a significant gap in understanding their potential in scenarios with limited data or tasks involving non-text inputs, such as image-based medical tasks. Furthermore, this research seeks to evaluate when subspecialized models are necessary versus when general-purpose foundational models are sufficient, addressing a key gap in current LLM research.

2.4.2 Objectives, Hypothesis, and Approach

The core objective of this research is to assess the performance of both general-purpose LLMs and subspecialized models in medical imaging tasks, evaluating their accuracy, robustness, and utility, especially in limited data use cases. The research will explore the integration of image analysis, examining the potential for LLMs to handle non-text based tasks. By comparing out-of-the-box LLMs with fine-tuned versions, the study aims to determine when fine-tuning is beneficial and how it impacts performance, especially in specialized healthcare tasks with limited size datasets. We hypothesize that fine-tuned models will yield better accuracy and robustness in scenarios where the data closely matches the task. We anticipate that LLMs, primarily optimized for text-based tasks, may struggle with image-based applications unless adapted or fine-tuned for visual data. Fine-tuning is expected to enhance their performance, especially with niche or limited datasets.

2.4.3 Research Goals and Significance

This research aims to provide a clearer understanding of when and where LLMs are most effective in healthcare. By comparing general and subspecialized models, we will establish a framework for selecting the optimal models and training approaches for various medical tasks, highlighting when fine-tuning is most beneficial. Extending LLMs to image classification and understanding when they perform well could revolutionize fields such as radiology, leading to more accurate and personalized healthcare.

3 Design Decisions

3.1 Dataset and Model Selection

To thoroughly investigate the capabilities of medical LLMs, we aimed to evaluate their performance across a variety of medical tasks. Image classification and segmentation were selected as they play a foundational role in enabling more advanced applications such as object detection, scene understanding, and medical diagnosis [16]. This study first focused on classifying brain gliomas using the Multimodal Brain Tumor Segmentation Challenge 2020 (BraTS 2020) dataset. For this evaluation, we compared the performance of the Llama 3.2 Vision Instruct LLM model, both in its out-of-the-box form (referred to as the "general LLM") and its subspecialized form fine-tuned with Unsloth (referred to as the "fine-tuned LLM"), against a custom-built, small-scale 3D CNN, which served as the baseline due to this deep learning model's proven effectiveness in medical image classification, providing a comparable benchmark. The second task focused on image segmentation using the same models and dataset, aiming to assess their ability to accurately detect tumor boundaries. Comparing classification and segmentation performance across these CNN and LLM models provided valuable insights into their strengths and limitations when applied to critical medical imaging tasks. The source code and experimental details for this thesis are publicly available at https://github.com/liufeli2/Healthcare-LLM-Thesis.

3.1.1 BraTS 2020 Dataset

The BraTS 2020 (Brain Tumor Segmentation) dataset was part of an international competition aimed at advancing machine learning methods for brain tumor analysis, particularly for gliomas, which are the most common type of primary brain tumors. Gliomas arise from glial cells in the brain and are classified into two main categories: Low-Grade Glioma (LGG), which is less prominent and aggressive, and High-Grade Glioma (HGG), which is more defined and malignant. The dataset provided multimodal MRI scans for each patient, consisting of four different types of scans: T1-weighted (T1), T1-weighted with contrast enhancement (T1ce), T2-weighted (T2), and Fluid-Attenuated Inversion Recovery (FLAIR). These imaging modalities together give a comprehensive view of the tumor's structure and its interaction with surrounding tissues. Additionally, the dataset included expert-annotated segmentation masks and each patient was labeled as either an HGG or LGG patient. The BraTS 2020 dataset was chosen due to its relevance and quality for glioma classification, its annotations, and its relatively larger sample size of 365 patients make it an ideal dataset for training and testing models [17]–[21].

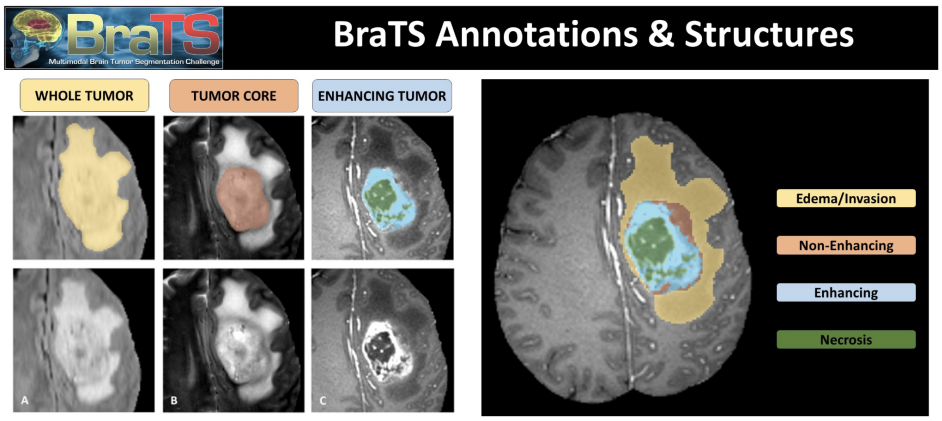


Figure 4: The BraTS dataset provides multi-modal MRI scan [17]–[21].

3.1.2 Llama-3.2-11B-Vision-Instruct

The Llama 3.2 Vision Instruct model was selected due to its high performance in NLP and multimodal tasks, and its ability to integrate both text and image inputs. It excels in visual recognition, image reasoning, image captioning, and visual question answering (VQA), making it a promising candidate for medical image classification and segmentation. The model benefits from large-scale pretraining, allowing it to process medical images effectively for both identifying tumor presence (classification) and accurately delineating tumor boundaries (segmentation). By testing its out-of-the-box performance, we can assess how well a general model handles these specialized medical tasks and understand the adaptability of general LLMs to the medical domain [22].

3.1.3 Unsloth Vision Fine-tuning

Unsloth was chosen as the fine-tuning tool due to its simplified user pipeline and optimized features that enable efficient training of large language models (LLMs). This lightweight framework accelerates training while minimizing memory usage, making it ideal for limited hardware environments such as Google Colab GPUs. Unsloth integrates seamlessly with popular packages like Hugging Face Transformers, bitsandbytes (for 8-bit and 4-bit quantization), and Parameter-Efficient Fine-Tuning (PEFT). It supports parameter-efficient techniques such as Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA), which allow for task-specific adaptations without updating all model parameters. The base model remains frozen while small trainable adapters are updated, significantly reducing memory demands. QLoRA further improves efficiency by applying 4-bit quantization to model weights, preserving performance while lowering memory usage. The bitsandbytes library facilitates low-bit precision operations, optimizing matrix multiplication and gradient updates for quantized models. Together, these features make Unsloth an effective and resource-efficient

solution for fine-tuning on specialized datasets, such as medical image classification tasks, while ensuring optimal performance even with constrained resources [23].

3.1.4 SciNet Computation Resources

SciNet at the University of Toronto was chosen for its high-performance computing resources, including powerful GPU nodes (NVIDIA A100s), which are ideal for large-scale deep learning tasks. SciNet uses a Slurm-based scheduling system and supports custom software environments through the modules system. Users access resources via SSH, with MIST being a primary platform for intensive computations. SciNet provides storage, along with documentation and support for environment setup and troubleshooting [24]. Note that, at the time of this research, bitsandbytes could not be built on SciNet's MIST system, with the issue unresolved despite recent bug fixes. This prevented Unsloth from being downloaded, built, or imported correctly for fine-tuning, requiring an alternative computational resource to proceed [24].

3.1.5 Google Colab Computation Resources

Google Colab is a cloud-based platform developed by Google Research, enabling users to write and execute Python code directly in the browser [25]. It offered access to powerful computational resources, including GPUs and TPUs, which are essential for deep learning applications. We utilized Google Colab Pro, which granted access to advanced GPUs such as the NVIDIA A100, L4, and T4. Note that bitsandbytes imported correctly on Google Colab, making it the platform of choice for all LLM experiments. The A100 is optimized for high-performance tasks, particularly large-scale LLM training, with substantial tensor core counts and memory bandwidth [26]. The L4 provides a cost-effective solution for inference, delivering 2 to 4 times the performance of the T4, which was employed for smaller tasks such as generating plots and calculating evaluation metrics [26]. This selection of GPUs allowed for optimized task execution, balancing both performance and cost-efficiency.

Table 1: Available GPU types and memory allocations in Google Colab Pro [25].

GPU Type	System RAM	GPU RAM	Disk
T4 GPU	12.7 GB	15.0 GB	235.7 GB
L4 GPU	53.0 GB	22.5 GB	235.7 GB
A100 GPU	83.5 GB	40.0 GB	235.7 GB

3.2 Data Pre-Processing

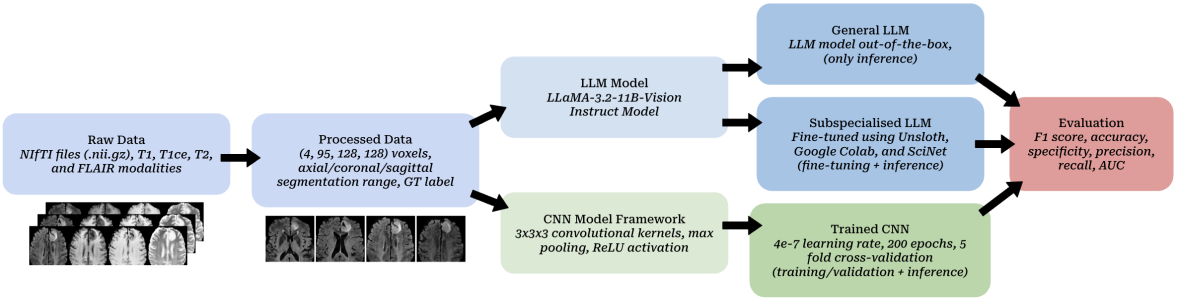


Figure 5: Full methodology pipeline from raw data to evaluation and comparison.

3.2.1 Image Data Processing

All experiments, including classification and segmentation tasks using CNNs and LLMs, were conducted with the same cleaned data to ensure a fair comparison. The BraTS 2020 dataset consists of 365 patients, each with four MRI scans (T1, T1ce, T2, and FLAIR) providing 95 axial slices per scan, resulting in 95x128x128 voxel arrays of intensity. The scans are processed as integer arrays (0-256 per pixel), normalized, and centered around the glioma using radiologist-annotated segmentations. All images are resized to a consistent shape of (4, 95, 128, 128). To focus on the capabilities of the CNN and LLM models, no additional image processing techniques, such as filtering or histogram equalization, were applied. The dataset is split into separate cohorts: 310 training, 62 validation (for each of 5-folds), and 55 test samples (15%). Care was taken to ensure that all scans and slices for a single patient were kept within the same cohort, preventing any overlap of information between the training, validation, and test sets. The dataset was also imbalanced in that there are far fewer LGG patients compared to the HGG patients. All training and validation data was balanced by oversampling the LGG class. Sample patient scans from the BraTS 2020 dataset are given in Figure 6 and Figure 7.

Figure 6: Scan modality visualization for an LGG patient [17]–[21].

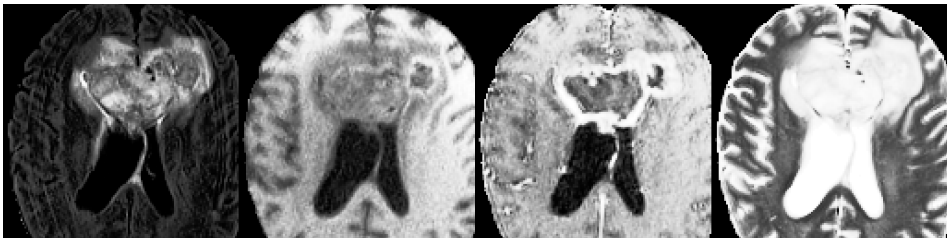


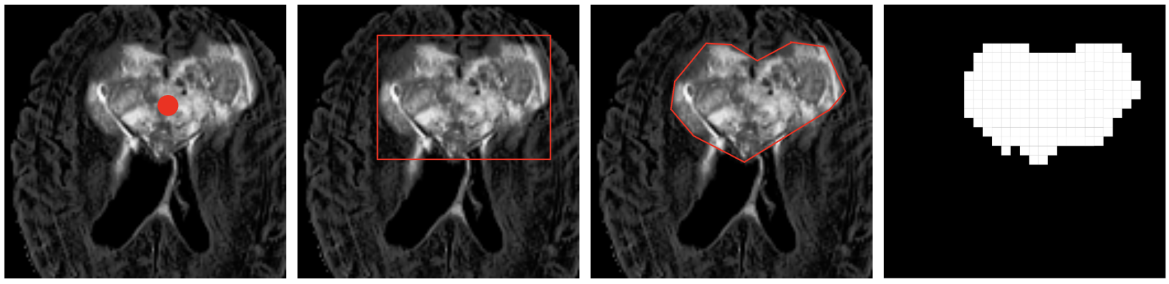
Figure 7: Scan modality visualization for an HGG patient [17]–[21].

3.2.2 Label Data Processing

The BraTS 2020 dataset labels each patient as either LGG or HGG, using 0 for LGG and 1 for HGG. These labels, stored alongside the MRI images, were extracted and translated into binary format to train models for distinguishing between low and high-grade gliomas used in the CNN classification tasks.

3.2.3 Segmentation Data Processing

For the segmentation tasks, each of the 365 patients had a 95x128x128 scan structured as a voxel occupancy grid, where zero values indicated regions without glioma and non-zero values represented glioma presence. To ensure viable segmentation, the first and last slices containing glioma were identified in the axial, coronal, and sagittal planes, allowing for the exclusion of non-glioma slices, ensuring that only relevant slices were used in training, thus minimizing potential confusion. Three types of segmentation data were extracted. The first method involved defining a bounding box for each axial slice containing glioma. This was done by extracting the minimum and maximum coordinates along both the x and y axes (xmin, xmax, ymin, ymax), creating a square bounding box. The bounding box was represented by four points: top-left, top-right, bottom-right, and bottom-left, stored in a (4, 2) array for each slice. The second approach focused on the center of the glioma, where the center of the bounding box was calculated and recorded as a single (1, 2) coordinate for each axial slice containing glioma. This representation was simpler but still provided the important information on glioma localization. The third, most complex approach involved representing the glioma with polygons. For each axial slice, the boundary of the glioma was traced using 10 to 15 points, following a clockwise direction starting from the north-most point. Smaller gliomas used fewer points, while larger or more irregularly shaped ones used more. The polygon data was stored as a list of coordinates with shape (10-15, 2), capturing the detailed contours of the glioma. This method offered the most precise and detailed segmentation, providing data for training segmentation models with a higher degree of accuracy in tumor shape and location.



a) Center Point b) Bounding Box c) Bounding Polygon d) Binary Voxels

Figure 8: Sample segmentations using a) center points, b) bounding boxes, c) bounding polygons, and d) the given form of binary voxels.

3.2.4 Creating Conversation Datasets for Fine-tuning

For all classification and segmentation tasks, the data is structured into a format compatible with Unsloth for LLM fine-tuning, converting raw information into a conversation dataset. This format is consistent across tasks, with slight variations depending on the specific instructions and ground truth data. The conversation structure includes the task instruction and relevant image data, ensuring the LLM has the necessary context. Each task begins with a clear instruction (e.g., classifying brain scans or segmenting glioma regions) embedded in the dataset. For classification tasks, the label (e.g., "Low Grade Glioma", "High Grade Glioma") serves as the ground truth, while segmentation tasks use center points, bounding boxes, or polygons. The conversation dataset is organized as exchanges between a "user" and an "assistant," with the assistant's response based on the ground truth data. This approach ensures proper preparation for fine-tuning the LLM, making it easy to feed into Unsloth and adapt the model for specific tasks. The process is standardized for this efficient training while preserving task-specific details.

```
def convert_to_conversation(sample):
   instruction = "Classify the brain scan as Low Grade Glioma (0), High Grade Glioma (1),
or No Glioma (2). Respond only in the following format: Choice: <0, 1, or 2> Reasoning:
<Provide concise reasoning using 10 keywords based on the scan's visual features>."

   conversation = [
       { "role": "user",
         "content" : [
           {"type" : "text",  "text"  : instruction},
           {"type" : "image", "image" : sample['image']} ]
       },
       { "role" : "assistant",
         "content" : [
           {"type": "text", "text": f"Choice: {sample['label']}"} ]
       },
   ]
   return { "messages" : conversation }
```

Figure 9: Sample LLM conversation dataset conversion for the classification task.

3.3 Evaluation Approaches

Due to differences in input processing, the CNN and LLM approaches differed in how the data was processed for training, inference, and evaluation.

3.3.1 CNN Evaluation Approach

The CNN model leveraged full 3D convolutions across all four imaging modalities (T1, T1ce, T2, FLAIR), enabling it to capture comprehensive spatial information from the entire volume of each scan. With the power of GPUs, which process the data quickly, the full 3D scans for each patient were fed into the model, rather than processing each 2D slice individually. This allowed the CNN to discern a single metric per patient, such as one LGG/HGG classification or one output voxel segmentation mask, which directly corresponds to the 3D ground truth mask for that patient. This approach benefits from the ability to process the entire scan, leveraging spatial relationships between slices, offering a more holistic view of the tumor.

3.3.2 LLM Evaluation Approach

The LLM approach, by nature, was limited to 2D inputs, as the LLM Vision Model only supports 2D inputs. This constraint, coupled with the slower inference times of the LLM, motivated several key design decisions. Instead of processing the full 3D scan, which would require handling each slice individually, we focused on the axial slices from the FLAIR modality. FLAIR is considered the superior imaging modality for detecting brain anomalies [27], and limiting the data to this modality helped reduce processing time while still providing high-quality information for the task. Furthermore, because each slice prediction had to be processed separately, we implemented a tallying technique to determine the final patient classification. Specifically, we used a majority vote (winner-takes-all) across all slice predictions for each patient, ensuring that the most consistent label (HGG or LGG) prevailed. This approach is visualized in Figure 10. For segmentation tasks, we computed evaluation metrics (such as Dice coefficient and Hausdorff distance, discussed later) for each slice independently, before averaging these metrics across the entire patient to obtain a final score. This slice-by-slice evaluation, while a limitation of the LLM approach, was a necessary adaptation given the model's inability to process the full 3D spatial context.

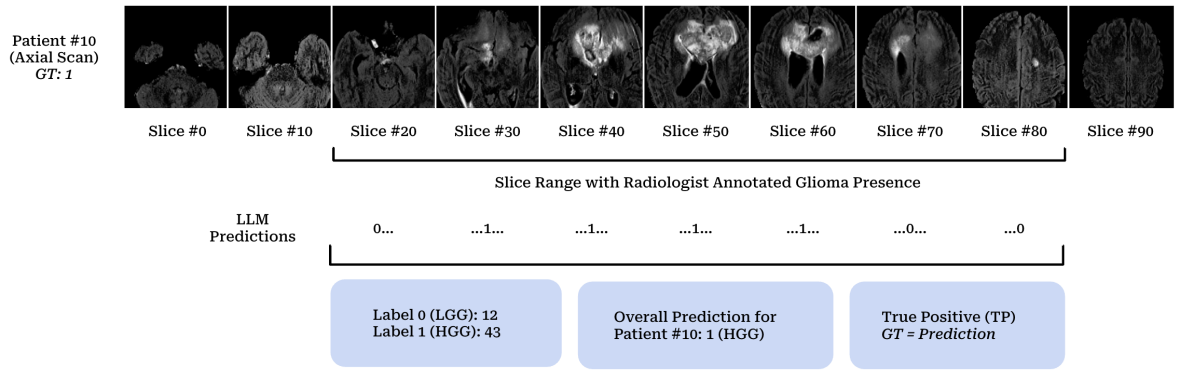


Figure 10: Diagram illustrates how predictions for each patient were made by the LLM.

These design choices highlight the fundamental differences between the CNN and LLM approaches. Ultimately, the goal of this exploration was to assess the performance of both general-purpose and specialized LLMs in medical imaging tasks, comparing them against highly effective image processing models like CNNs. Despite the differences in evaluation and training techniques, this comparison remains valuable for understanding the potential of LLMs in this domain. However, it is important to acknowledge these methodological differences, as they make a direct one-to-one comparison between the two approaches challenging.

4 Image Classification Task

4.1 Introduction

4.1.1 Tumor Grade Classification Task

The goal of this task is to classify each patient as either an HGG or LGG patient based on their MRI scans, utilizing all four imaging modalities as a single 3D input (a 95x128x128 image with a depth of four modalities). The corresponding binary ground truth labels are 0 for LGG and 1 for HGG [17]–[21].

4.2 Methodology and Implementation

4.2.1 CNN Baseline Model Parameters and Training

The 3D CNN was built with an input shape of (N, 4, 95, 128, 128), where N=16 was the batch size. The first convolutional layer applied 8 filters with a kernel size of 3x3x3, stride of 1, and padding of 1. The second convolutional layer used 16 filters, also with a 3x3x3 kernel, stride of 1, and padding of 1. A max pooling operation with a 2x2x2 kernel and stride of 2 followed each convolutional layer. The fully connected

layers included 128 hidden units, with a final output layer for binary classification, distinguishing between LGG and HGG. Rectified Linear Unit activation was used for the hidden layers, and a Sigmoid function was applied to the logits. Learning rates were varied between 1e-4 and 5e-7 to train a suitable and generalising model, and model checkpoints were saved and reloaded during training to ensure robustness. After training and validation, the best model was selected and tested using the separate cohort of 55 patients.

4.2.2 LLM General Model Setup and Prompting

The general out-of-the-box LLM was tested using the same test cohort of 55 patients as the CNN. The segmentation labels were used to select the appropriate axial slice depth range for each patient. Each slice from the FLAIR scans was fed separately to the LLM, which classified the slice as either Low Grade Glioma (LGG-0), High Grade Glioma (HGG-1), or No Glioma (NG-2) based on visual features. The model also responded with concise reasoning using 10 keywords related to the scan's visual characteristics to ensure the model wasn't making a random decision. Predictions for each slice were aggregated and the final patient classification was determined using a majority vote (winner-takes-all) across all slice predictions. A sample output response from the general LLM is shown in Figure 11, and the corresponding prompt is provided in Figure 12 for the classification task.

```
Scan #: 77/220
Slice #2 => Choice: 2 Reasoning: No mass, no contrast
enhancement, normal brain structure.
Slice #3 => Choice: 1 Reasoning: Heterogeneous mass, irregular
borders, necrosis, edema
Slice #4 => ERROR: Unknown >>||  || Choice: 1 ||  || Reasoning:
Heterogeneous mass, irregular margins, necrosis
Slice #5 => Choice: 2 Reasoning: No visible tumor, no mass
effect, normal brain structure, no
Slice #6 => ERROR: Unknown >>||  || Step 1: Identify the visual
features of the brain scan. ||  || The brain scan shows a
…
Slice #74 => Choice: 2 Reasoning: No visible tumor, normal brain
structure, no abnormal enhancement.
Slice #75 => Choice: 0 Reasoning: Small, round, non-enhancing
lesion with minimal edema.
Slice #76 => Choice: 2 Reasoning: No tumor visible, normal brain
structure, no irregularities, no
Slice #77 => Choice: 0 Reasoning: Uniform gray matter, no
contrast enhancement, no mass effect, no
Slice #78 => ERROR: Unknown >>||  || Classification: ||  ||
Choice: 2 ||  Reasoning: No mass, no abnormal enhancement
Slice #79 => Choice: 1 Reasoning: Enhancing mass, irregular
margins, heterogeneous signal intensity.
Guesses: [2, 1, 3, 2, 3, 2, 1, 1, 0, 2, 0, 0, 1, 3, 2, 1, 0, 3,
2, 1, 2, 3, 2, 0, 2, 2, 2, 0, 0, 1, 0, 2, 1, 2, 0, 0, 0, 0, 0, 1,
2, 2, 1, 0, 1, 1, 1, 1, 2, 0, 2, 2, 2, 1, 0, 0, 0, 0, 1, 1, 0, 1,
0, 0, 2, 1, 3, 2, 2, 3, 3, 3, 2, 0, 2, 0, 3, 1]
Tally: [24, 20, 24, 10]
True Label: LGG || Prediction Label: LGG
```

Figure 11: Sample response for one patient from the general LLM model. Most responses follow the correct format although there are some outliers. The tally is shown at the bottom and this scan was predicted as LGG.

Prompt: Classify the brain scan as Low Grade Glioma (0), High Grade Glioma (1), or No Glioma (2). Respond only in the following format: Choice: <0, 1, or 2> Reasoning: <Provide concise reasoning using 10 keywords based on the scan's visual features>.

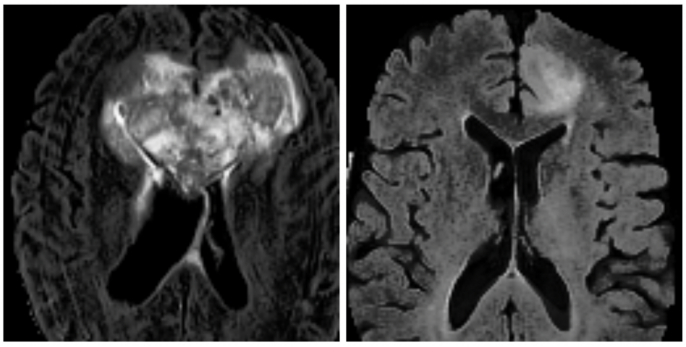
Figure 12: LLM general model prompt sample for the classification task.

4.2.3 LLM Subspecialised Model Fine-tuning

For Unsloth fine-tuning, the brain scan dataset is first balanced to address class imbalances, and each image is formatted into a structured prompt for classification. The pre-trained general vision model is then loaded from Unsloth using the FastVisionModel class, which consists of vision, language, attention, and MLP layers adapted to the task. LoRA is applied with rank 16 and alpha 16 to update only a small set of parameters, reducing computational overhead while maintaining generalization. Gradient checkpointing is also used to optimize memory usage, allowing for larger batches and more complex architectures without exceeding memory limits. Fine-tuning is conducted using the SFTTrainer class, designed for supervised fine-tuning. A batch size of 16 is used with gradient accumulation over 4 steps to balance memory constraints and model performance. The learning rate is set to 2e-6, selected through experimentation for stable and fast convergence. The AdamW optimizer, with 8-bit precision, accelerates training, which is carried out for a maximum of 100 steps, with logging at each step. Upon completion, the model is saved as a LoRA-adapted version to Hugging Face for future testing. The same prompt-based testing procedure used for the general LLM is applied to the specialized model by selecting the fine-tuned version.

4.2.4 LLM Consistency and Robustness Test

In addition to the primary objectives comparing the subspecialized LLM, the general LLM, and the baseline CNN in performance, we also conducted a consistency test to evaluate the robustness of the out-of-the-box LLM in this classification task. This involved inputting the same image and prompt into the LLM 95 times, simulating the maximum number of axial slices (95) if a glioma were present throughout all slices, to observe the prediction distribution. A robust model should consistently produce the same prediction for identical inputs. This test was performed on both a HGG and LGG image, as shown in Figure 13, using the general LLM and will be repeated with the subspecialized LLM once fine-tuning is complete. If the general LLM shows inconsistency, we anticipate that fine-tuning will improve its reliability. We also plan to expand this test with additional examples to further assess prediction consistency.



a) High Grade Glioma b) Low Grade Glioma

Figure 13: FLAIR scan slices of a) High Grade Glioma, and b) Low Grade Glioma. These images clearly illustrate the distinct features of each condition, and the LLM is expected to predict them accurately [17]–[21].

4.2.5 Evaluation Metrics

Accuracy measures the overall correctness of a model by calculating the proportion of correct predictions (true positives and true negatives) out of all predictions made. However, it can be misleading in imbalanced datasets where one class dominates. Precision focuses on the quality of positive predictions by measuring the proportion of true positives among all instances predicted as positive, indicating how many of the positive predictions are actually correct. Recall or sensitivity measures the model's ability to correctly identify all actual positive cases, highlighting its effectiveness at detecting positive instances. Specificity assesses the model's capacity to correctly identify negative cases, reflecting the model's ability to avoid false positives. The F1 score, defined as the harmonic mean of precision and recall, provides a balanced measure, which is especially useful in situations with imbalanced classes. The Area Under the Curve (AUC) evaluates the model's ability to distinguish between classes across different thresholds, with higher values indicating better performance. AUC cannot be accurately calculated for the LLM, as the model does not generate probabilities. Instead, tally scores are used as a proxy for probability, but this approach lacks credibility and doesn't offer meaningful insight into model confidence. Together, these metrics offer a comprehensive view of a model's strengths and weaknesses in classification.

4.3 Results

4.3.1 CNN Baseline Performance

The CNN training results are summarized below, where different learning rates and epoch counts were evaluated to determine the most stable configuration and optimal test performance. The best-performing model, trained with a learning rate of 4e-7 over

200 epochs, is shown in Figure 14 and was used as the baseline. The training curve showed consistent improvement and the validation curve remained steady without a rise, indicating no overfitting of the model.

Table 2: Evaluation of the CNN baseline model (learning rate = 4e-7, epochs = 200).

Metric	Training	Validation	Testing
Accuracy	0.9677	0.7581	0.8000
F1 Score	0.9792	0.8485	0.8706
Precision	1.0000	0.8400	0.9024
Recall	0.9592	0.8571	0.8409
AUC	0.9975	0.7221	0.8202
Specificity	1.0000	0.3846	0.6364

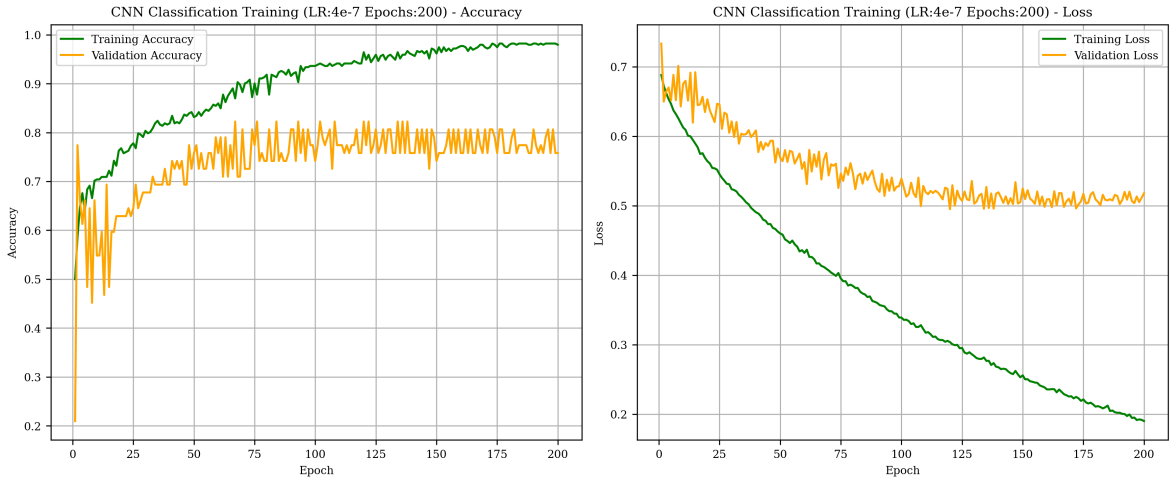


Figure 14: Training and validation accuracy and loss curves for the CNN model (learning rate = 4e-7, epochs = 200).

After 200 epochs, the model achieved an accuracy of 0.9677, F1 score of 0.9792, precision of 1.0000, and recall of 0.9592 on the training set. The AUC reached 0.9975, indicating excellent classification capability. In the validation set, accuracy dropped to 0.7581, with a slight decrease in F1 score (0.8485) and precision (0.8400), although recall remained strong at 0.8571. The AUC for validation was 0.7221, suggesting moderate classification ability. The testing set yielded an accuracy of 0.8000, F1 score of 0.8706, precision of 0.9024, and recall of 0.8409, with an AUC of 0.8202, showing balanced performance. Specificity was perfect (1.0000) in the training set, but dropped to 0.3846 in validation and increased slightly to 0.6364 in testing, indicating challenges in avoiding false positives in validation and test cohorts.

Several alternative models were also trained but were not selected as the baseline model due to more unstable convergence, as their validation accuracy fluctuated. Nev-

ertheless, these models still demonstrated strong performance, with AUC scores ranging from 69% to 75% and F1 scores between 81% and 88%. Detailed results for these models are provided in Appendix A.

4.3.2 LLM General Model Classification Performance

The general LLM model's performance was assessed across three different imaging orientations: axial, coronal, and sagittal using the evaluation outlined in the Methodology. The results are presented below in Figure 15, Figure 16, and Figure 17 respectively. In the figures, the vertical axis is modeled using a Glioma Label Distribution Ratio (GLDR) to highlight the distribution of glioma types across scan slices for each patient for visualization. For a scan with 100 slices, if 25 slices are labeled as LGG and 75 as HGG, the patient would have a ratio of 0.75, indicating the proportion of HGG-labeled slices compared to the total number of slices.

Table 3: Evaluation of the general LLM model across axial, coronal, and sagittal imaging orientations during testing.

Metric	Axial	Coronal	Sagittal
Accuracy	0.7818	0.8000	0.7963
F1 Score	0.8776	0.8889	0.8842
Precision	0.7963	0.8000	0.8077
Recall	0.9773	1.0000	0.9767
Specificity	0.0000	0.0000	0.0909

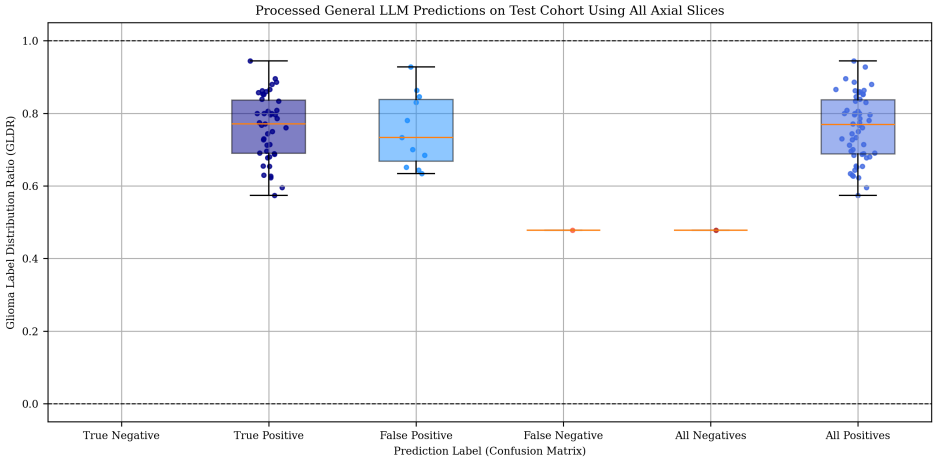


Figure 15: Visual representation of the ratio of HGG to LGG labels across each test patient's axial scan slices.

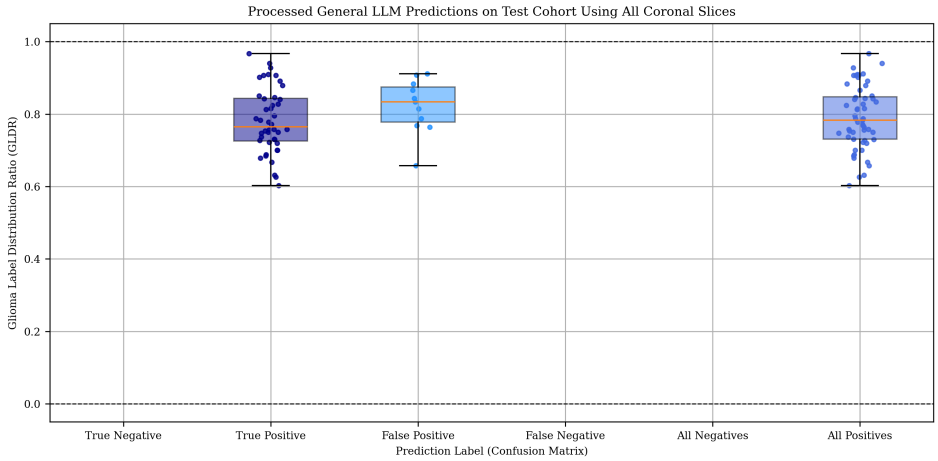


Figure 16: Visual representation of the ratio of HGG to LGG labels across each test patient's coronal scan slices.

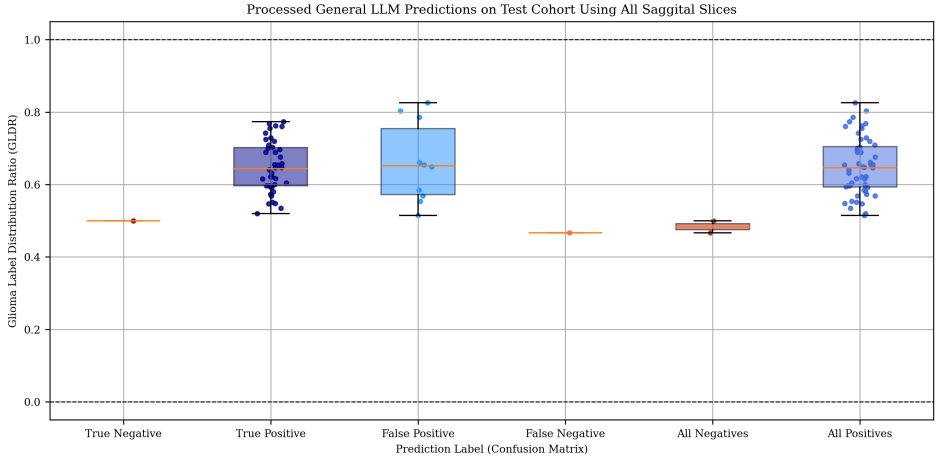


Figure 17: Visual representation of the ratio of HGG to LGG labels across each test patient's sagittal scan slices.

The model demonstrates strong recall and accuracy across all orientations, with the coronal achieving the highest accuracy (0.8000) and perfect recall (1.0000), while the axial and sagittal orientations also exhibit strong recall values (0.9773 and 0.9767, respectively). At first glance, these metrics suggest good performance. However, the performance is misleading due to the class imbalance in the dataset, where the majority of slices are HGG, and far fewer are LGG. When examining the data in the figures, it becomes evident that the model predominantly predicts HGG across all cases. This is in line with the imbalance, as the model appears to simply and consistently predict the majority class (HGG) rather than differentiating between HGG and LGG as shown by the zero specificity across all slice orientations. I had expected to potentially see LGG ground truths with ratios closer to 0.5, where the model might label some slices as LGG and others as HGG, however, this was not the case. Certain LGG scans demonstrated

ratios in the 90% range, indicating failure of the general LLM model to differentiate between the classes.

Overall, the results indicate that while the model can be said to be accurate in detecting HGG, it is not effectively performing classification and identifying LGG patients, and is heavily biased in favour of the imbalance in the dataset, hence the high accuracy, precision, recall, and F1 score. These results reflect the performance of the general LLM model. With a fine-tuned model, we plan to compare these results with those from the fine-tuned model through similar tests.

4.3.3 LLM Subspecialised Model Classification Performance

The training loss curves for fine-tuning the LLM for classification are shown in Figure 18 and Figure 19 below. The first model was trained for 100 steps in Unsloth with a batch size of 4, while the second, more robust model, was fine-tuned for one full epoch using a larger batch size of 16.

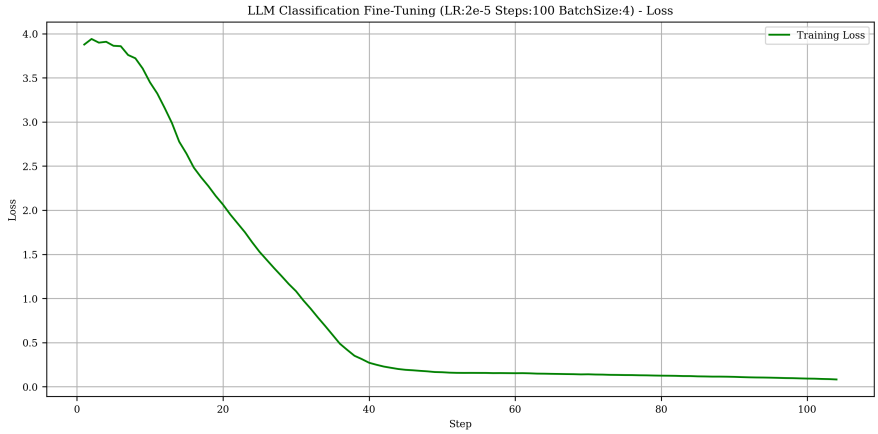


Figure 18: LLM model loss after fine-tuning for 100 steps with a learning rate of 2e-5.

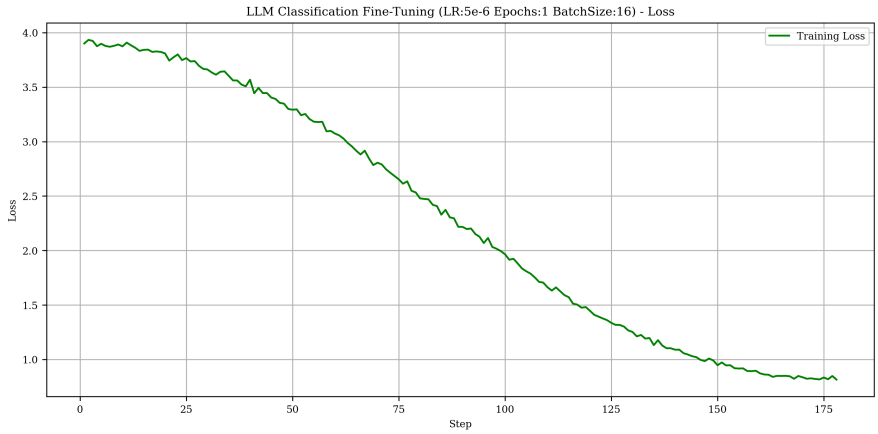


Figure 19: LLM model loss after fine-tuning for one epoch with a learning rate of 5e-6.

The fine-tuned LLM model was evaluated solely on the axial imaging orientation, using the same methodology as the general LLM to maintain consistency while reducing computational demands. Because the fine-tuning process was performed exclusively on slices containing glioma, the fine-tuned models were also evaluated only on this subset. To ensure a fair comparison, the general LLM was assessed on the same glioma containing slices (referred to as the "General LLM (small)") to ensure that performance is compared on equivalent data across models. The results are once again, visualized using the Glioma Label Distribution Ratio (GLDR), which represents the proportion of glioma types averaged across slices per patient. These are shown in Figure 20 for the small general LLM, Figure 21 for the fine-tuned model trained with 100 steps, and Figure 22 for the model fine-tuned for one full epoch.

Table 4: Evaluation of the fine-tuned LLM models in comparison to the general LLM.

Model	Accuracy	F1 Score	Precision	Recall	Specificity
General LLM (small)	0.7636	0.8602	0.8163	0.9091	0.1818
Fine-tuned LLM (100 Steps)	0.76	0.8537	0.8333	0.875	0.3
Fine-tuned LLM (Full Epoch)	0.6667	0.7733	0.8529	0.7073	0.5

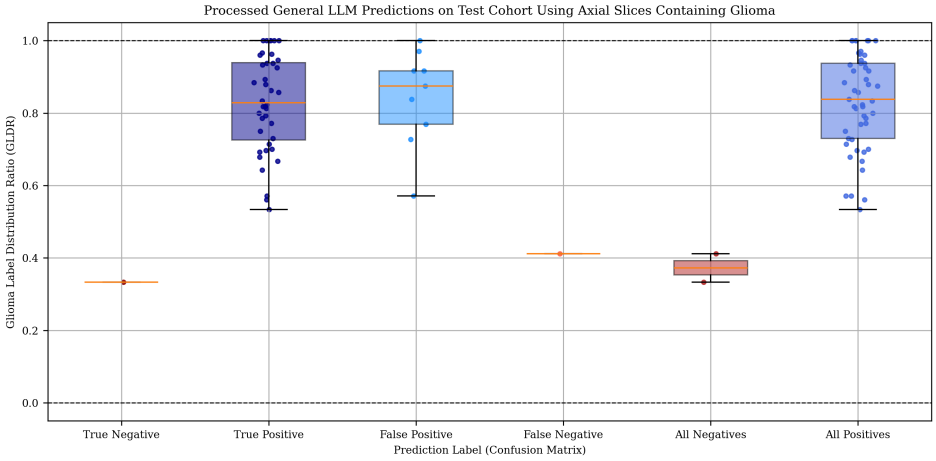


Figure 20: Visual representation of the ratio of HGG to LGG labels across test patient axial scan slices with glioma evaluated using the small general LLM model.

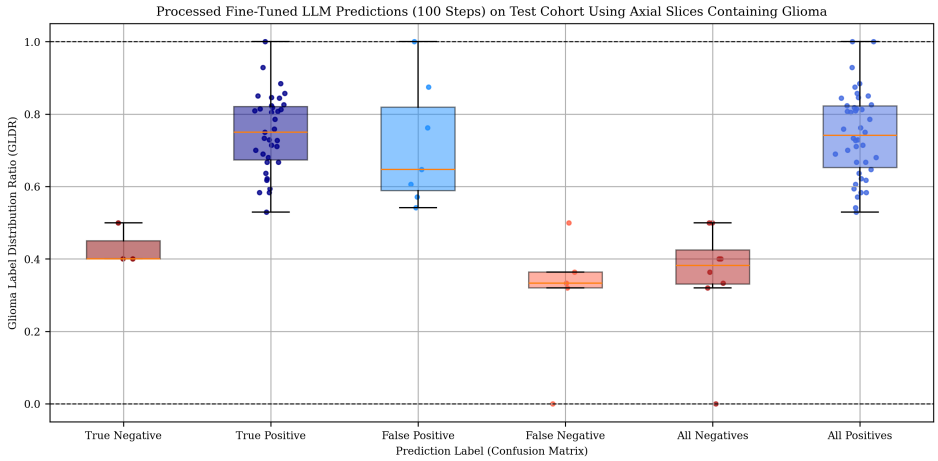


Figure 21: Visual representation of the ratio of HGG to LGG labels across test patient axial scan slices with glioma evaluated using the fine-tuned LLM model of 100 steps.

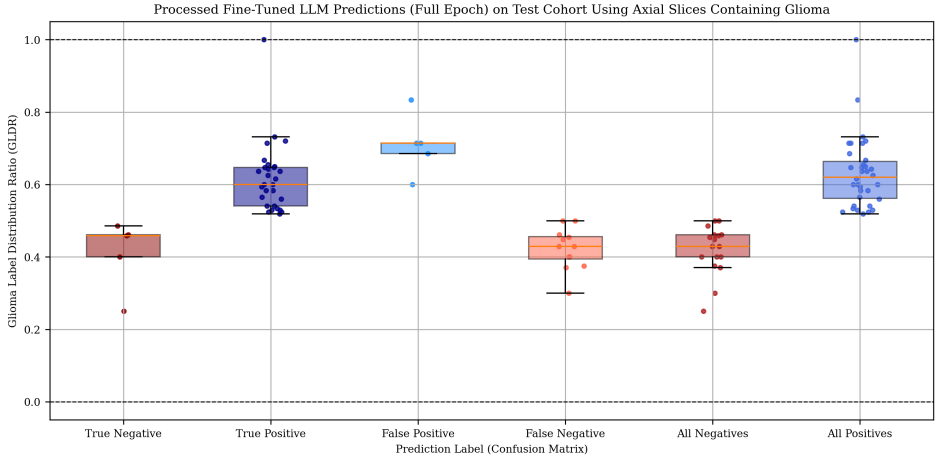


Figure 22: Visual representation of the ratio of HGG to LGG labels across test patient axial scan slices with glioma evaluated using the fine-tuned LLM model of one epoch.

The results for the full-epoch and 100-step fine-tuned models show some mixed outcomes. For the full-epoch model, there was a noticeable decline in performance, with an accuracy of 0.6667 and an F1 score of 0.7733. The recall dropped to 0.7073, indicating a reduction in the model's ability to identify true positives compared to the 100-step model, which achieved a higher recall of 0.875. The specificity for the full-epoch model was 0.5, suggesting that it still struggled with distinguishing between HGG and LGG despite the fine-tuning effort. On the other hand, the 100-step fine-tuned model showed relatively stable performance, with accuracy at 0.76 and an F1 score of 0.8537, but the precision remained low at 0.8333. Both models demonstrated only marginal improvements over the general model. The full-epoch model, in particular, did not perform as expected, as it failed to improve its ability to differentiate between the two glioma types. Ideally, we would have expected the LGG predictions to cluster closer

to a ratio of 0-0.1 and the HGG predictions to be near 0.9-1, but this was not the case. Instead, the models produced predictions that were roughly split between HGG and LGG, suggesting that neither model has learned a clear distinction between the two classes and seems to treat the two types as interchangeable, rather than accurately distinguishing between them. While both fine-tuned models showed some marginal improvement over the general LLM, the results were not as promising as anticipated.

4.3.4 LLM Consistency Test Classification Performance

For the consistency test, the general LLM model consistently predicted the HGG image as HGG across all 95 trials with the same prompt, which demonstrates good consistency. However, 77 of the 95 trials with the LGG image were incorrectly predicted as HGG, while only 18 were correctly identified as LGG. This led to an overall prediction of HGG, which is incorrect. The split in predictions, especially with the same input, raises concerns about the model's consistency, as it flipped between predictions despite identical input. This behavior is unusual and indicates a lack of stability in the general model's response. The fine-tuned LLM results show that for the HGG image, the model incorrectly predicted 23 out of 95 trials as LGG and 72 as HGG, which is worse compared to the general LLM's performance. For the LGG image, 42 trials were predicted as LGG and 53 as HGG, resulting in a near 50-50 split. This outcome is somewhat expected because as the results above show, the model has not learned to distinguish between LGG and HGG effectively. The inconsistent predictions, especially for the LGG case, suggest that fine-tuning did not enhance the model's ability to differentiate the classes and there is still a need for alternative and improved fine-tuning strategies.

4.4 Discussion

4.4.1 Key Findings

The core objective of this research was to assess and compare the performance of general-purpose large language models (LLMs) and subspecialized models in medical imaging tasks, particularly focusing on their accuracy, robustness, and practical utility in data-limited image-based environments. The evaluation centers on determining when fine-tuning general LLM models provides measurable benefits in performance over simply using them for inference out-of-the-box.

The CNN baseline demonstrated competitive performance with a training accuracy of 0.9677, validation accuracy of 0.7581, and testing accuracy of 0.8000. Precision and recall were particularly strong in the training phase, achieving 1.0000 and 0.9592, respectively, while the testing phase maintained a balanced F1 score of 0.8706. The AUC

remained high in testing (0.8202), although specificity dropped to 0.6364, indicating some difficulty in correctly identifying negative cases (LGG). The baseline as a whole shows it is accurate and robust.

In contrast, the general LLM model showed promising accuracy and recall across axial, coronal, and sagittal orientations, with the coronal view achieving the highest accuracy (0.8000) and perfect recall (1.0000). However, these metrics were misleading due to the dataset's class imbalance, where high-grade glioma (HGG) slices vastly outnumbered low-grade glioma (LGG) slices. The model predominantly predicted HGG for nearly all samples, resulting in near-zero specificity across orientations. The GLDR visualizations further emphasized this issue, as most patient scans had HGG favouring ratios, even in cases where LGG was the ground truth. The expectation was to observe more balanced predictions in borderline cases due to the "winner-takes-all" voting system, but this was not reflected in the results. In its current out-of-the-box form, the general LLM model's inability to distinguish between HGG and LGG limits its clinical utility and classification reliability.

The results for the full-epoch and 100-step fine-tuned models show limited improvement over the general model. The full-epoch model showed a decrease in performance, with lower accuracy (0.6667) and recall (0.7073) compared to the 100-step model, which maintained a higher recall of 0.875. Neither model demonstrated significant improvements in distinguishing between HGG and LGG, as the predictions remained split between the two types, contrary to expectations of more distinct clustering. This suggests that the fine-tuning process has not yet effectively enabled the model to differentiate between glioma types, particularly LGG gliomas, with both models showing only marginal differences. The results indicate that further refinement is necessary to achieve more accurate classification.

These findings align with the study's hypothesis that general-purpose LLMs, while versatile and successfully applicable into a variety of medical tasks, are not inherently well-suited for specialized tasks like medical image classification even with fine-tuning. The CNN's stronger performance in differentiating glioma types shows the advantage of domain-specific models in handling class imbalance and nuanced medical data.

4.4.2 Limitations

A major limitation of this study is the difference in input data between the CNN and the LLM models. The CNN used full 3D convolutions and all four MRI modalities (FLAIR, T1, T1ce, and T2), giving it access to rich spatial and multi-modal information. This comprehensive input may contribute to its stronger performance. In contrast, the LLM was restricted to 2D axial slices from only the FLAIR modality

given its necessary input structure for prompting. Each slice was classified individually, and patient-level predictions were determined by majority vote across slices. This approach limited the LLM's access to complete spatial and modality information. Although it's important to acknowledge this difference when evaluating the methods and understanding how the comparison was conducted, it ultimately has minimal impact on the study's goals. The primary focus is on comparing the best-performing versions of the CNN and LLM to assess their overall performance, regardless of input disparities.

Another key limitation is the inconsistency in the general-purpose LLM's predictions. In consistency testing, the LLM correctly identified an HGG in all 95 trials. However, for an LGG, it misclassified the case as HGG 77 times and correctly identified it only 18 times. This inconsistency, despite identical inputs, highlights the model's instability and unreliability, which also translate into uncertainty in the investigation of the general LLM. We saw further in this consistency test that fine-tuning failed to achieve better results. Instead of improving the model's ability to distinguish between HGG and LGG, the fine-tuning process appears to have introduced additional instability. The fact that the model's predictions for LGG images are nearly evenly split between LGG and HGG classifications suggests that the model has not developed a clear decision boundary between the two classes.

The next significant limitation in this study is the restricted scope of fine-tuning performed on the LLM. The fine-tuning process could have been far more robust, but was heavily limited by resource constraints. The batch size of just 16 2D images per iteration was quite small, especially when compared to the CNN, which used 16 3D patients with four MRI modalities per sample. This was due to the memory constraints of Colab GPUs. The batch size can heavily impact the model's ability to generalize, potentially hindering the LLM's capacity to capture the complex patterns needed to differentiate between glioma subtypes with fine-tuning. In addition to the batch size, the fine-tuning process was constrained by the relatively short training duration. With only 100 steps and one full epoch completed in these fine-tuned models, they may not have had enough time to converge to an optimal solution. While the training ran for over 10 hours, the limited number of iterations meant the LLM was not exposed to enough data to learn robust features and exploit the underlying patterns in the data, which could have been better captured with additional epochs. The extended runtime was costly on Colab, and further training would have incurred significant expenses, limiting the ability to run additional epochs. Another challenge during fine-tuning was the absence of a validation curve. Unlike the CNN training, which included regular validation checks to monitor for overfitting, the fine-tuning of the LLM did not include such checks due to the extensive time constraints. Without this important feedback, it was difficult to assess whether the model was overfitting to the training data or if the

fine-tuning was progressing effectively. To mitigate this, very small learning rates were used in an attempt to make incremental updates, but without the ability to validate the model's performance consistently, there was uncertainty regarding how well the model was learning. The combination of these resource limitations, small batch size, limited training duration, and lack of validation, likely contributed to suboptimal fine-tuning results for both subspecialised LLM models.

4.5 Conclusion

The CNN model outperformed the general-purpose LLM in terms of accuracy, robustness, and reliability in classifying glioma types, particularly in handling class imbalances and distinguishing between LGG and HGG gliomas. While the LLM showed some promise, its initial performance was hindered by low specificity and inconsistent predictions, especially for LGG cases. Although fine-tuning was attempted, the results were not as effective as anticipated due to limitations in training duration, batch size, and resource constraints. The LLM's potential remains, but further refinement and more robust fine-tuning are necessary for it to effectively tackle specialized medical imaging tasks like glioma classification.

5 Image Segmentation Task

5.1 Introduction

In this glioma segmentation task, we explore four different methods to represent and compare ground truth segmentation with the voxel grid ground truth from the dataset. The objective is to assess the effectiveness of these various segmentation strategies, particularly when transitioning from the CNN baseline models to both the general LLM and fine-tuned LLM approaches.

This next stage will allow us to evaluate the model's performance across different medical domains, as initially intended, providing valuable insights into its adaptability and generalization. Similar to the glioma classification task, we will process the new dataset and fine-tune the model to create a subspecialized version. We will then perform a performance comparison between the CNN, the general LLM, and the fine-tuned LLM, focusing on key factors such as consistency, utility, and robustness. This will help us gain a comprehensive understanding of the applicability and usefulness of LLMs in various image-based healthcare tasks, particularly those with limited data.

5.1.1 Tumor Voxel Segmentation Task

Traditional CNN-based segmentation methods generate a 3D volumetric prediction by downsampling input data and subsequently upsampling it to reconstruct a segmentation mask. This allows for a structured, voxel-wise representation of the tumor, which can be directly compared to the ground truth for evaluation using standard metrics such as Dice similarity coefficient and Hausdorff distance.

However, when applying LLMs to segmentation, directly outputting a full 3D segmentation mask in numerical form presents significant challenges. Token length limitations make it infeasible to represent the entire mask explicitly, and even if possible, post-processing such a long sequence would be computationally expensive and prone to inconsistencies. Hence for LLMs, we considered using a different and existing general-purpose segmentation model.

5.1.2 Segment Anything Model for Segmentation

The Segment Anything Model (SAM), developed by Meta, is a general-purpose segmentation model that can segment objects based on user prompts such as points, bounding boxes, or text descriptions of colour and location [28]. It leverages a Vision Transformer (ViT) architecture to generate high-dimensional image embeddings, enabling rapid and flexible segmentation. Despite its proven strong performance across various segmentation tasks, we opted not to use SAM for our segmentation task for several key reasons. Firstly, SAM was trained on the BraTS 2020 dataset, providing it with a domain-specific advantage that could introduce bias into our comparisons. Secondly, SAM relies on explicit user prompts, such as seed points or bounding boxes, which may provide external guidance not typically available to CNN-based models. This explicit prompting introduces additional context that makes direct comparison with CNN-based approaches less valid. Furthermore, the ViT architecture incorporates self-attention mechanisms, allowing the SAM model to capture long-range dependencies and global context within the image. This feature makes it harder to separate the contributions of spatial image understanding from the impact of prompt-based enhancements in LLMs.

Given that our primary objective is to evaluate the capabilities of a general-purpose vision LLM and compare it to a full CNN baseline model, the use of SAM also would complicate this comparison. Thus, we chose to explore segmentation approaches that provide more efficient and structured representations for the LLM, allowing for better alignment with token limitations while still capturing similar accuracy information, but in a more condensed form compared to a full segmentation mask. We selected 3 distinct segmentation prompting strategies to structure the input.

5.1.3 Tumor Center Point Segmentation Task

The goal of this method is to predict the center of the glioma in each MRI scan by identifying the center of the glioma for each axial slice. The ground truth for each slice consists of the center of glioma (extracted from the bounding box), represented as a single (x, y) coordinate, indicating the x and y coordinates of the glioma's center. This approach simplifies the task by focusing on the central location of the glioma, while still providing essential information for glioma localization. The model is expected to predict the center coordinates in the same format for each slice.

5.1.4 Tumor Bounding Box Segmentation Task

The goal of this method is to segment the glioma present in each MRI scan by predicting a bounding box for each axial slice containing the glioma. The ground truth for each slice consists of a bounding box, represented by the minimum and maximum coordinates along the x and y axes (xmin, xmax, ymin, ymax), forming a square around the glioma. This bounding box is defined by four points: top-left, top-right, bottom-right, and bottom-left, stored in an array of shape (4, 2) for each slice. The model is expected to predict the bounding boxes in the same format for each slice, with the goal of accurately identifying the glioma's location within the image.

5.1.5 Tumor Bounding Polygon Segmentation Task

The goal of this method is to predict the detailed boundaries of the glioma in each MRI scan by representing the glioma as a polygon. For each axial slice containing glioma, the boundary was traced using 10 to 15 points, starting from the north-most point and following a clockwise direction. Smaller gliomas were represented with fewer points, while larger or irregularly shaped gliomas used more. The ground truth for each slice was stored as a list of coordinates with shape (10-15, 2), capturing the precise and tight contours of the glioma. The model is expected to predict the polygon representation of the glioma in the same format, providing a detailed and accurate segmentation of the tumor shape and location.

5.2 Methodology and Implementation

5.2.1 CNN Baseline Model Parameters and Training

The 3D CNN model was designed for the segmentation of brain tumor images, outputting an upsampled binary mask with a shape of (95, 128, 128). The model took inputs with a shape of (N, 4, 95, 128, 128), where N=24 represented the batch size of patient scans. The first convolutional layer applied 16 filters with a 3x3x3 kernel,

a stride, and padding of 1. The second convolutional layer utilized 32 filters with the same kernel size, stride, and padding. Each convolution was followed by a max pooling operation with a 2x2x2 kernel and a stride of 2. The bottleneck consisted of a third convolutional layer with 64 filters. For the decoding component, two transposed convolution layers were used for upsampling, followed by a final convolution layer that output a single-channel segmentation mask. ReLU activation was applied after each convolution and deconvolution, while the output was passed through a sigmoid activation function for binary segmentation. The model was trained for 70 epochs with a learning rate of 0.0001, and model checkpoints were saved regularly to avoid overfitting. The model's performance was evaluated using Dice loss, and after training, the best-performing model was saved for further testing on the same separate cohort of 55 patients used in the classification task. During inference and testing, the predicted masks were thresholded using a tuned parameter of 0.7 to create a binary output before being evaluated.

5.2.2 LLM General Model Setup and Prompting

The general out-of-the-box LLM was evaluated on the same test cohort of 55 patients used for the CNN baseline. Segmentation labels were utilized to identify the relevant axial slice depth range for each patient. For each case, individual slices from the FLAIR scans were processed sequentially, with the LLM tasked to predict either the tumor's center point, bounding box (defined by its four corner coordinates), or bounding polygon, depending on the segmentation task. Evaluation metrics were computed independently for each slice and subsequently aggregated across all slices for a given patient to produce a final performance score. While the slice-by-slice evaluation limited the model's ability to capture the full 3D spatial context, this approach was a necessary compromise due to the LLM's architectural constraints. The specific prompt used for each segmentation method is illustrated in Figure 23, Figure 24, and Figure 25.

Prompt:
You are an expert medical AI assistant specializing in glioma segmentation on FLAIR-mode brain scans. Given a 128x128 grayscale brain scan, output the bounding box around the tumor using the four corner vertices. Ensure the bounding box tightly encloses the entire tumor without extending into non-tumor regions. The bounding box output must be formatted strictly as: [(row_min, col_min), (row_min, col_max), (row_max, col_max), (row_max, col_min)] where (row, col) are integers between 0 and 127, with (0,0) at the top-left and row increasing downward, and col increasing rightward. Do not output any other text or explanation, only the coordinate list in the exact format above.

Figure 23: LLM model prompt sample for the center point segmentation task.

Prompt:
You are an expert medical AI assistant specializing in glioma segmentation on FLAIR-mode brain scans. Given a 128x128 grayscale brain scan, output the center point of the tumor as a single coordinate (row, col). The tumor region is the brightest, high-intensity abnormality distinct from normal brain structures. Ensure the predicted center point accurately represents the geometric center of the tumor, which is generally round in shape. The output must be formatted strictly as: (row, col) where row and col are integers between 0 and 127, with (0,0) at the top-left, row increasing downward, and col increasing rightward. Do not output any other text or explanation, only the coordinate in the exact format above.

Figure 24: LLM model prompt sample for the bounding box segmentation task.

Prompt:
For each glioma in a 128×128 grayscale FLAIR-mode brain scan, output coordinates of a 10 to 15 point polygon that encloses the tumor region. These points should be arranged in a clockwise direction and should accurately trace the tumor boundary. The output should be a tuple in the format of ((row1, col1), (row2, col2), ..., (rowN, colN)), where N is between 10 and 15 points, and row and col are integers between 0 and 127. The tumor region is the brightest, high-intensity abnormality distinct from normal brain structures. For image [query image], what is the output? Output only the coordinates in the exact format specified, without any additional text or explanation.

Figure 25: LLM model prompt sample for the bounding polygon segmentation task. This prompt was adapted from the methodology used in a previous study that successfully applied prompt-based LLM segmentation. Source: [29]

5.2.3 LLM Subspecialised Models Fine-tuning

The subspecialized models were fine-tuned using the same general procedure as in the classification task, but adapted for the new segmentation prompts and ground truths. Each image was first formatted into a task-specific prompt, and the dataset was balanced where needed. The LLM was loaded using the Unsloth FastVisionModel, and fine-tuning was applied using LoRA with rank and alpha set to 16, targeting only a subset of parameters to improve efficiency. To support the higher memory demands of the segmentation tasks, gradient checkpointing and gradient accumulation (over four steps) were used, allowing batch sizes between 8 and 24 depending on the available GPU memory on Google Colab A100s. The learning rate was selected per task to ensure stable convergence, and training was performed for up to 200 steps using the 8-bit AdamW optimizer. After training, the adapted models were saved to Hugging Face and evaluated using the same prompt-based testing strategy as the general LLM.

5.2.4 Evaluation Metrics

For segmentation tasks, the primary evaluation metrics used are the Dice coefficient and the 95% Hausdorff Distance, which provide complementary perspectives on model performance. The Dice coefficient measures the spatial overlap between the predicted segmentation and the ground truth, and is defined as twice the intersection over the sum

of the predicted and true volumes. It is particularly useful in medical imaging where high overlap is essential for clinical utility. In contrast, the 95% Hausdorff Distance captures the spatial agreement at the boundaries by measuring the 95th percentile of distances between the closest points on the predicted and true segmentation surfaces. This metric is more sensitive to boundary errors, making it a valuable complement to Dice, especially in irregular or complex tumor shapes. Additionally, precision, recall, and specificity are computed at the voxel level to provide insight into the model's ability to correctly identify tumor versus non-tumor regions. While these metrics are also used in classification, their voxel-wise application in segmentation tasks reflects the granularity and spatial precision required in medical image analysis. Together, these measures form a robust evaluation framework for assessing segmentation quality in terms of both overlap and boundary accuracy.

5.3 Results

5.3.1 CNN Baseline Performance

The CNN segmentation baseline was trained with various configurations, ultimately achieving the best results with 100 epochs and a batch size of 16. The training results for the best-performing model are summarized below. Figure 26 shows the training and validation loss curves along with the Dice coefficient, while Figure 27 visualizes the evaluation metrics across test patients.

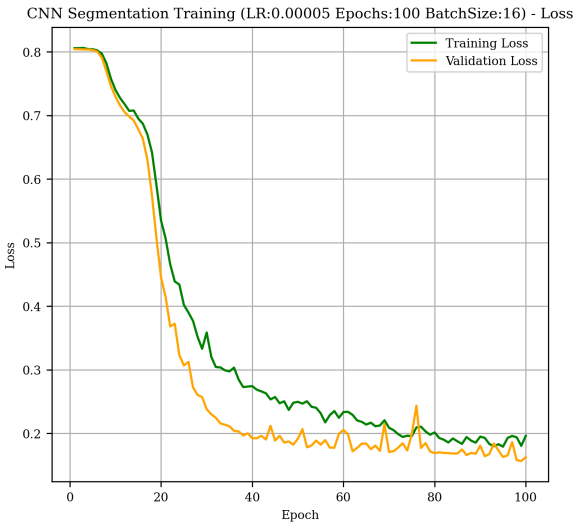


Figure 26: Training and validation loss curves for the CNN model (learning rate = 5e-5, epochs = 100) showing stable training and no overfitting.

Table 5: Evaluation of the CNN baseline model (learning rate = 5e-5, epochs = 100).

Metric	Training	Validation	Testing
Dice Coefficient	0.6020	0.6158	0.5942
95% Hausdorff Distance	53.9291	57.9943	50.5580
Precision	0.7011	0.7225	0.6765
Recall	0.7220	0.7109	0.7163
Specificity	0.9251	0.9506	0.9445

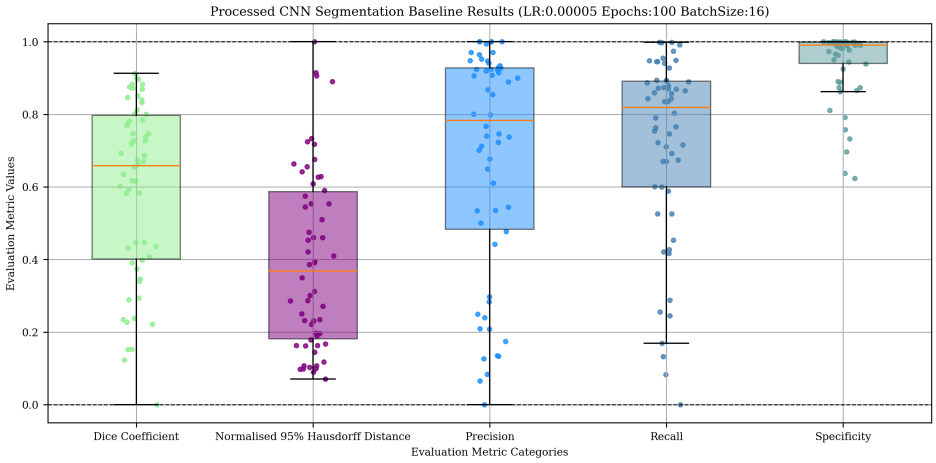


Figure 27: Evaluation metrics for the CNN baseline model predictions.

After 100 epochs, the model achieved a Dice coefficient of 0.6020, a 95% Hausdorff Distance of 53.9291, precision of 0.7011, recall of 0.7220, and specificity of 0.9251 on the training set. The validation set yielded a Dice coefficient of 0.6158, a 95% Hausdorff Distance of 57.9943, precision of 0.7225, recall of 0.7109, and specificity of 0.9506. On the testing set, the model achieved a Dice coefficient of 0.5942, a 95% Hausdorff Distance of 50.5580, precision of 0.6765, recall of 0.7163, and specificity of 0.9445. These results show a consistent performance across training, validation, and testing sets, with the model demonstrating reliable segmentation accuracy, good tumor boundary alignment, and effective detection of tumor regions while maintaining a strong ability to identify non-tumor areas.

Other baseline CNN models were also trained but not selected due to suboptimal convergence, indicated by slight plateaus during training, and had Dice testing scores ranging from 55% to 57%. Detailed results for these models are provided in Appendix B.

To qualitatively assess the performance of the CNN baseline model, we visualized its segmentations on the 55 patients from the testing cohort. The results, displayed in the figures below, highlight both successes and challenges. In each figure, the ground truth segmentation is shown in blue, while the CNN model's predictions are represented in red. This visualization offers a clear overview of the model's performance, showcasing

areas where it performs well and identifying regions where it struggles.

Figure 28 illustrates that the model accurately segmented gliomas, with segmentations closely aligning with the ground truth. The model effectively captured the tumor's size, location, and voxel boundaries, reflecting strong performance in these instances by maintaining the general shape and spatial distribution of the glioma. In contrast, Figure 29 demonstrates that while the segmentations for other gliomas were still largely accurate, some limitations emerged. Although the model successfully identified the correct regions and preserved the general 3D shape of the brain, it struggled with the precision of tumor boundaries. In these cases, the predicted pixels tended to remain around the correct regions, suggesting that while the overall segmentation was generally accurate, finer tumor details were missed. There were also instances where the model overestimated the glioma size, as seen in Figure 30, where it captured broader regions of the brain, sometimes even encompassing the entire brain. This over-segmentation could be attributed to poorer contrast in the images, which made the glioma less distinct, or the presence of other brain anomalies that the model mistakenly identified as part of the glioma. In such cases, the segmentation was less focused on the tumor itself and more on surrounding brain structures, resulting in inaccurate tumor boundaries.

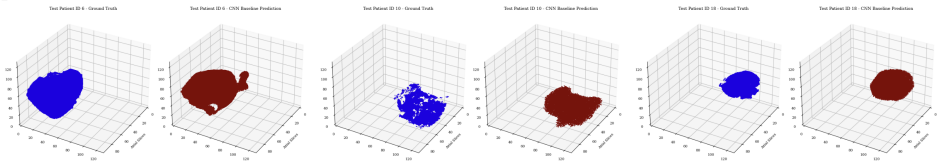


Figure 28: Baseline CNN segmentation visualization showed accurate glioma segmentation with correct size, location, and voxel boundaries.

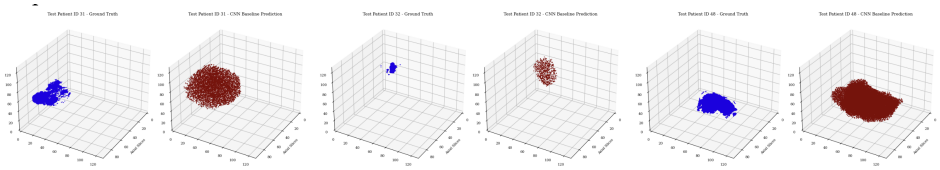


Figure 29: Baseline CNN segmentation visualization showed accurate glioma segmentation, but with less precise tumor boundaries.

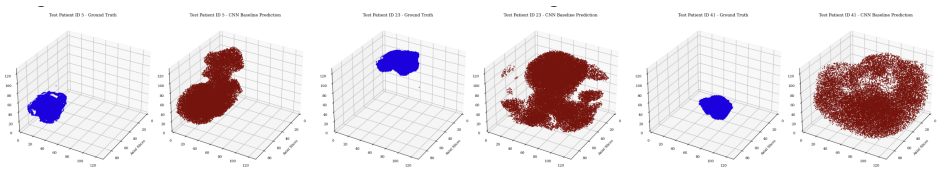


Figure 30: Baseline CNN segmentation visualization showed overestimation of glioma size, capturing broader brain regions or the entire brain.

Smaller gliomas, particularly those with irregular shapes, presented a significant challenge for the model, as shown in Figure 31. In many cases, the model struggled to detect these tumors, often predicting a scattered cloud of points rather than a well-defined glioma. This issue may have stemmed from the model either over-segmenting the entire brain or failing to distinguish small gliomas from surrounding tissues. As a result, these instances resulted in lower accuracy, with the model having difficulty precisely localizing the tumor. However, some smaller gliomas were accurately detected, as seen in Figure 32, indicating that the model's performance was potentially influenced by the quality of the MRI scan.

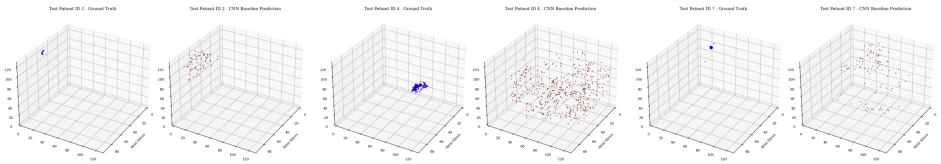


Figure 31: Baseline CNN segmentation visualization showed challenges in detecting smaller, irregularly shaped gliomas.

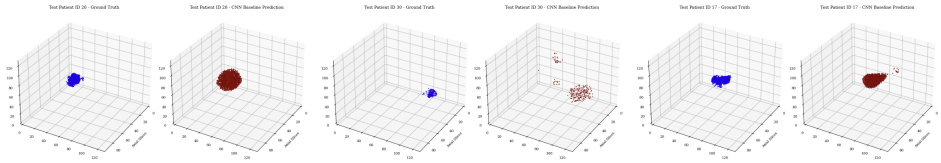


Figure 32: Baseline CNN segmentation visualization showed accurate detection of smaller gliomas in potentially higher-quality scans.

In some cases where the glioma had unusual boundaries or where multiple gliomas were present, as shown in Figure 33, the CNN baseline model still demonstrated strong performance in accurately identifying and segmenting these complex scenarios. The model effectively predicted the boundaries of each glioma, even in the presence of multiple or irregularly shaped tumors. However, in other cases, such as in Figure 34, the model struggled to capture the full extent of complex gliomas, especially those with irregular shapes. These segmentations often resulted in smaller areas being captured than the actual size of the glioma, indicating that the model still faced difficulties with accurately delineating complex tumor boundaries.

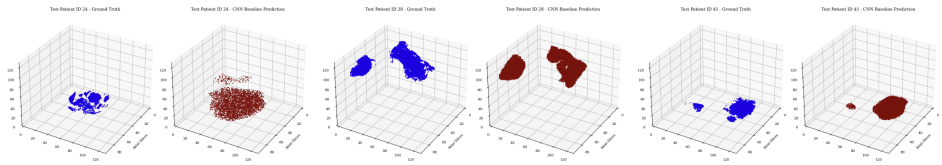


Figure 33: Baseline CNN segmentation visualization showed accurate identification and segmentation of complex gliomas with unusual or multiple boundaries.

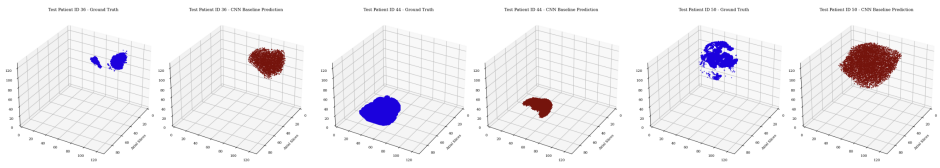


Figure 34: Baseline CNN segmentation visualization showed incomplete segmentation of complex gliomas with irregular shapes.

In summary, while the CNN baseline model demonstrated strong performance in segmenting gliomas across a range of cases, challenges remain, particularly with smaller gliomas and those with irregular shapes or boundaries. Further baseline CNN segmentation visualizations can be found in Appendix C.

5.3.2 LLM General Center Point Model Performance

For the center point segmentation task, the performance of the general LLM model was evaluated using only the axial imaging orientation. We also used different evaluation metrics as substitutes for the Dice coefficient and 95% Hausdorff distance. Instead of the Dice coefficient, we calculated the percentage of predicted points within the ground truth bounding box for each slice, as an area-related proxy. For the 95% Hausdorff distance, we used the Euclidean distance between the predicted and ground truth center points and took the 95th percentile of all Euclidean distances for each patient. We also measured the 95th percentile of the closest distances from the predicted center point to the ground truth bounding box, providing a metric that accounts for tumor size. All distance-based metrics were normalized to the 128x128 image size. Ideally, we would expect to see low distances and a high percentage of points within the bounding box. The results are presented below in Figure 35.

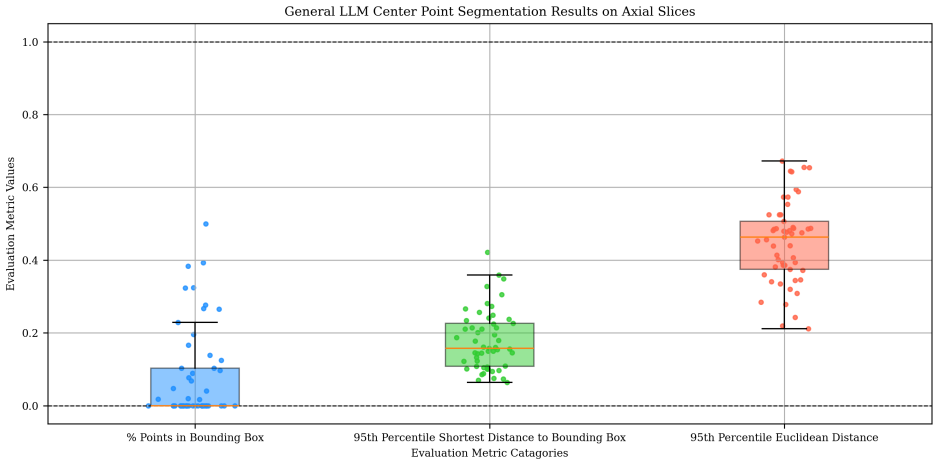


Figure 35: Evaluation metrics for the general LLM center point model predictions.

Table 6: Evaluation of the general LLM center point model.

Metric	Value
Average % Points in Bounding Box	0.0805
Average 95th Percentile Shortest Distance to Bounding Box	22.8566
Average 95th Percentile Euclidean Distance	57.4112

The results of the general LLM model on center point segmentation were not promising. Ideally, a high percentage of predicted points should fall within the ground truth bounding box, but the median value was close to zero, indicating that most predicted points were not aligned with the true tumor location. This suggests that the accuracy of the center point segmentation was largely random, as seen in Figure 36. The model tended to predict points near the center of the frame with some variance, contributing to the low percentage of points within the bounding box. For smaller or more distant gliomas, predicted points were often outside the box, resulting in a median percentage near zero. In terms of distance metrics, the model's predictions were, on average, about half the frame's distance away from the true tumor location. The 95th percentile shortest distance to the bounding box averaged 22.8566, while the 95th percentile Euclidean distance was 57.4112. These distances reflect the model's tendency to predict points near the image center, rather than accurately localizing tumor centers. This pattern suggests that the model is not effectively differentiating between different gliomas, regardless of their size or location, which is again reflected in the visualizations in Figure 36. The predictions are clustered near the center, as the model is failing to identify relevant features within the image to guide tumor localization. Further general LLM center point segmentation visualizations can be found in Appendix D.

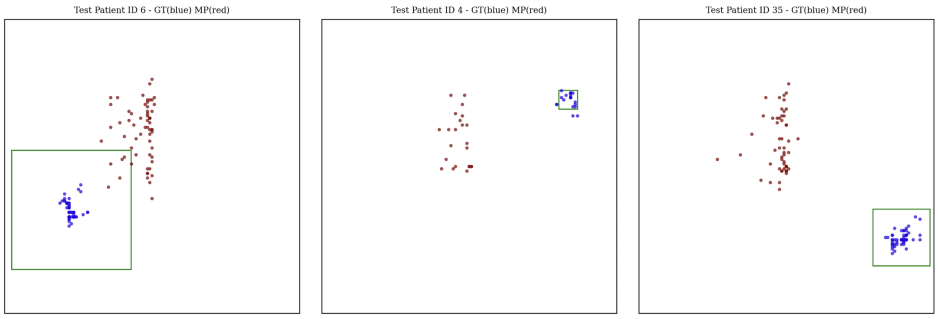


Figure 36: General LLM center point segmentation visualization showed, for all patients, the model predicted points near the center of the frame with some variance, regardless of the ground truth glioma location.

5.3.3 LLM Subspecialised Center Point Model Performance

The general center point model was fine-tuned for 100 steps with a batch size of 16, and the plot is shown below in Figure 37. With fine-tuning, we would expect an improvement in the model's performance, leading to an increase in the percentage of points within the ground truth bounding box and a decrease in the distances and we would anticipate the points becoming more clustered around the correct center points. Figure 38 presents the evaluation results.

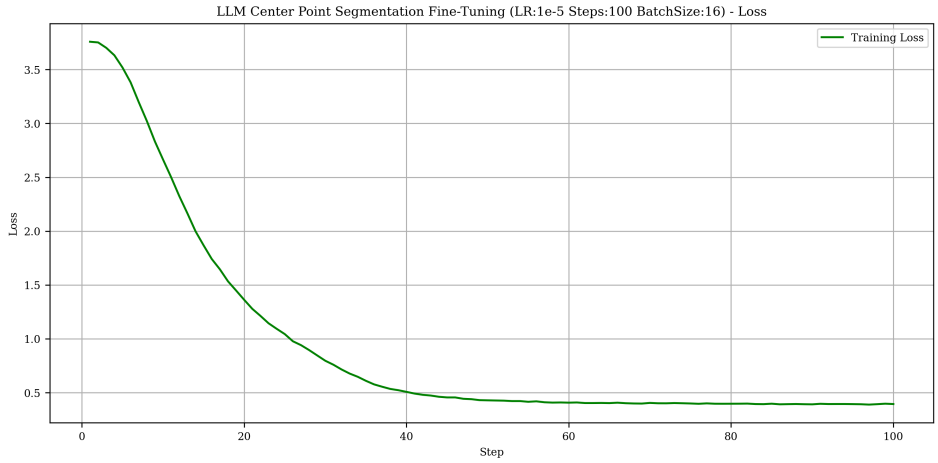


Figure 37: LLM model fine-tuning loss (learning rate = 1e-5, steps = 100).

Table 7: Evaluation of the fine-tuned LLM center point model.

Metric	Value
Average % Points in Bounding Box	0.0800
Average 95th Percentile Shortest Distance to Bounding Box	26.2855
Average 95th Percentile Euclidean Distance	68.4812

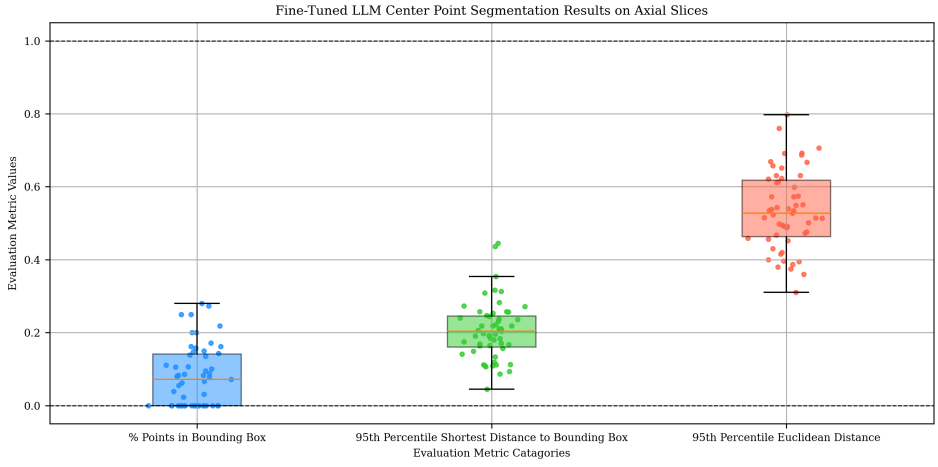


Figure 38: Evaluation metrics for the fine-tuned LLM center point model predictions.

After fine-tuning, we observed that the model's center point predictions became more spread out compared to the previous results, as shown in the visualizations in Figure 39. However, the points still remain predominantly centered in the frame. In terms of distance metrics, the 95th percentile shortest distance to the bounding box (26.2855) and the 95th percentile Euclidean distance (68.4812) reflect the increased variance in predictions. As seen in Figure 39, some points are closer to the ground truth, while others are farther away, resulting in average distances that show no significant improvement compared to the initial model. Further fine-tuned LLM center point segmentation visualizations can be found in Appendix E.

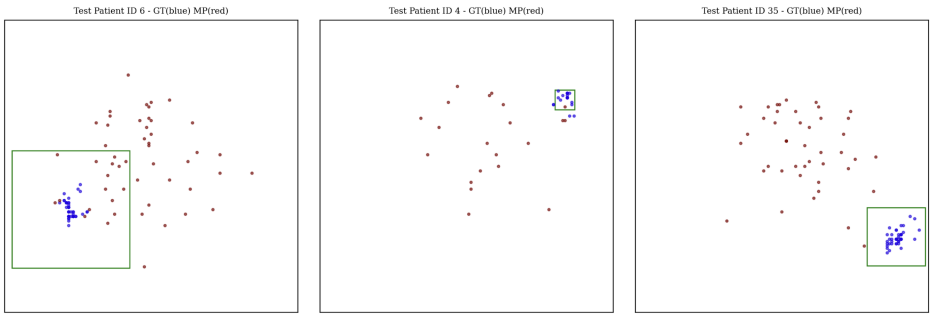


Figure 39: Fine-tuned LLM center point segmentation visualization showed, for all patients, the model still predicted points near the center of the frame with more variance, regardless of the ground truth glioma location.

5.3.4 LLM General Bounding Box Model Performance

In this segmentation task, the general LLM model's performance was assessed using the axial imaging orientation. The results are summarized in the table and Figure 40 below, providing an overview of the model's performance.

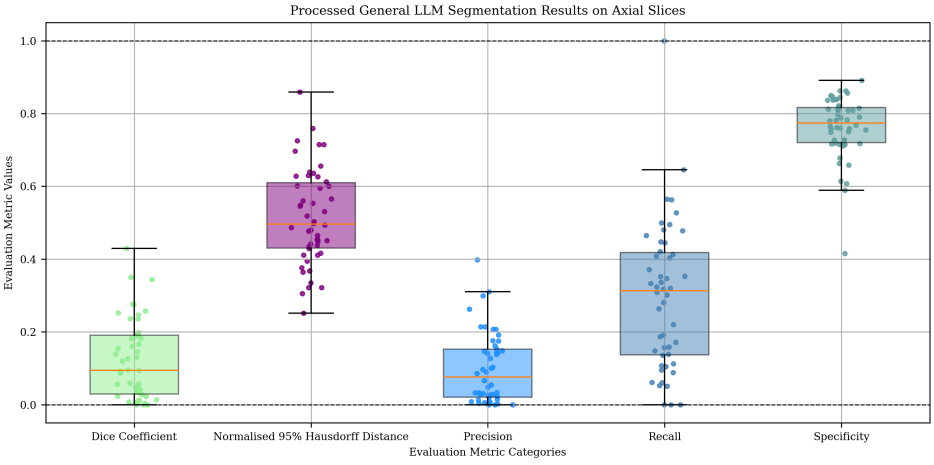


Figure 40: Evaluation metrics for the general LLM bounding box model predictions.

Table 8: Evaluation of the general LLM bounding box model.

Metric	Value
Dice Coefficient	0.1219
95^{th} Hausdorff Distance	65.9103
Precision	0.0989
Recall	0.2941
Specificity	0.7623

The general LLM model's performance on the bounding box test shows a Dice coefficient of 0.1219, indicating poor overlap between the predicted and actual tumor areas. The 95th percentile Hausdorff distance of 65.9103 suggests that the predicted segmentation is consistently far from the true tumor boundaries, reflecting significant localization errors. With a precision of 0.0989, only 9.89% of the model's tumor predictions are correct, indicating a high rate of false positives and poor tumor identification. These results highlight the model's challenges in both accurately identifying and localizing gliomas.

The quantitative results presented above are further supported by visualizations of the ground truth and model-predicted bounding boxes. Examining these visualizations in the figures below shows several key observations regarding the general model's prediction behavior.

The model's predictions are notably sparse, as seen in Figure 41. In several instances, the LLM failed to respond in the expected format, often generating redundant explanations despite being provided with sufficient tokens. This resulted in fewer meaningful predictions, limiting the data available for analysis and making it challenging to obtain reliable results. We aimed to evaluate the general model's ability to determine both the location and size of gliomas. In terms of location, as seen in Figure 42, the model's predictions consistently fall in the middle of the image, regardless of the actual glioma positions. The predicted bounding boxes are frequently centered, showing little regard for the intensity or other relevant image features. This indicates that the model is not effectively utilizing the spatial information needed to accurately localize gliomas. When it comes to glioma size, as shown in Figure 43, the model's predictions do not exhibit any clear bias between larger and smaller tumors. Regardless of the glioma's actual size, the predicted bounding boxes remain similar in shape, with minimal variation and often concentrated in the middle of the frame. This suggests that the model struggles to capture differences in tumor size or spatial characteristics, further highlighting the limitations of the general LLM model in accurately segmenting gliomas. Further general LLM bounding box segmentation visualizations can be found in Appendix F.

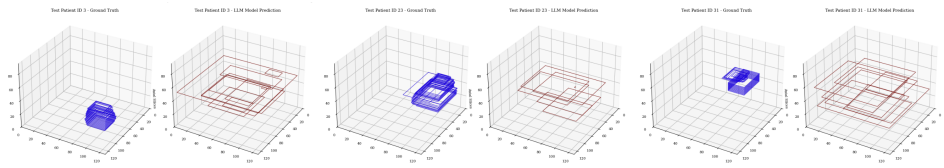


Figure 41: General LLM bounding box segmentation visualization shows that the model's predictions are notably sparse.

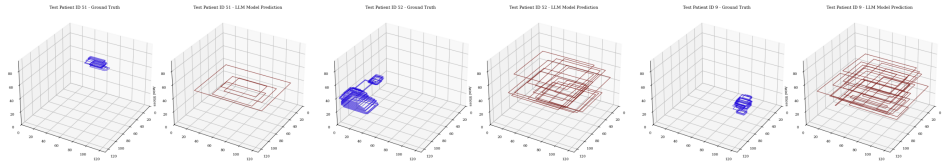


Figure 42: General LLM bounding box segmentation visualization shows that the model consistently predicts gliomas near the center of the image.

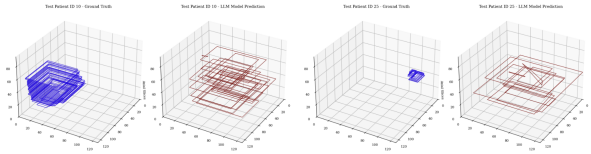


Figure 43: General LLM bounding box segmentation visualization shows that the model's predictions lack clear differentiation between larger and smaller gliomas.

5.3.5 LLM Subspecialised Bounding Box Model Performance

The general bounding box model was fine-tuned for 200 steps with a batch size of 16, and the plot is shown below in Figure 44. With fine-tuning, we would expect the model to improve its ability to differentiate the location and size of gliomas, leading to an increase in Dice coefficient overlap, precision, recall, and specificity, and a decrease in the 95% Hausdorff distance, indicating the model's improved accuracy in predicting the tumor's bounding box. Figure 45 presents the evaluation results.

Table 9: Evaluation of the fine-tuned LLM bounding box model.

Metric	Value
Dice Coefficient	0.1085
95^{th} Hausdorff Distance	67.3812
Precision	0.0882
Recall	0.2413
Specificity	0.8137

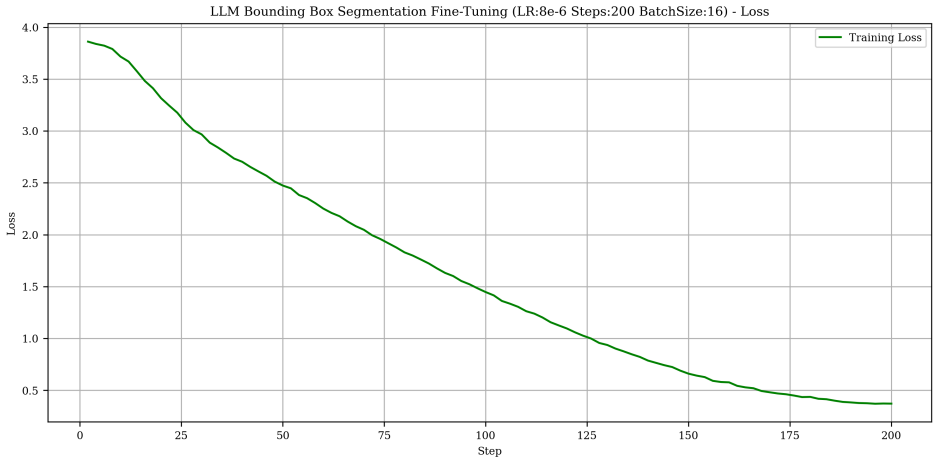


Figure 44: LLM model fine-tuning loss (learning rate = 8e-6, steps = 200).

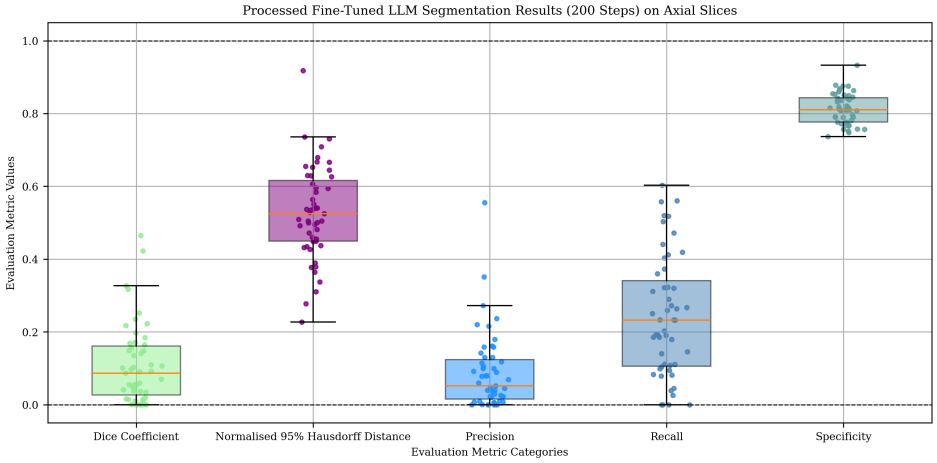


Figure 45: Evaluation metrics for the fine-tuned LLM bounding box model predictions.

The Dice coefficient decreased to 0.1085, indicating that there was no significant improvement in the model's ability to segment the gliomas accurately. The 95% Hausdorff distance slightly increased to 67.3812, suggesting that the model continues to struggle with aligning predicted center points with the true tumor locations. Precision decreased to 0.0882, while recall dropped to 0.2413, reflecting a decline in the model's ability to correctly identify true positive centers. However, specificity improved to 0.8137, indicating that the model became more effective at minimizing false positives. Overall, the fine-tuning did not result in substantial improvements in the model's performance.

The qualitative visualizations of the bounding box segmentations show some improvement in the model's ability to generate responses, as seen in Figure 46 and Figure 47. Previously, the model struggled to provide consistent answers, but now it regularly produces more bounding box predictions across slices. However, despite the increase in responses, there is no noticeable improvement in the placement of the bound-

ing boxes after fine-tuning. The bounding boxes remain in similar locations as before, as shown in Figure 46, and the model continues to show no sensitivity to the size of the gliomas, as seen in Figure 47. This suggests that the model is not adjusting its predictions based on tumor size. It still tends to place bounding boxes near the center of the image, often on the larger side. While this may minimize Dice loss, it indicates that the model has not learned to segment gliomas effectively. Further fine-tuned LLM bounding box segmentation visualizations can be found in Appendix G.

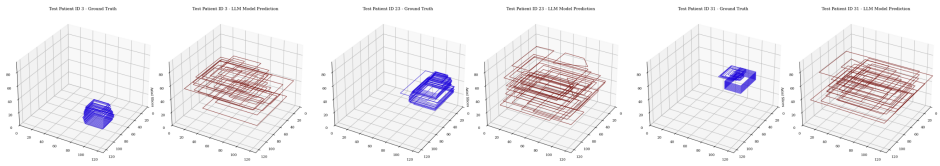


Figure 46: Fine-tuned LLM bounding box segmentation visualization shows no noticeable improvement in the placement of bounding boxes.

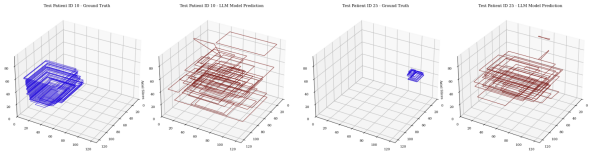


Figure 47: Fine-tuned LLM bounding box segmentation visualization shows continued insensitivity to glioma size.

5.3.6 LLM General Bounding Polygon Model Performance

In this last segmentation task, the general LLM model's performance was evaluated using the bounding polygon test again with just the axial imaging orientation. The results, as summarized in the table and Figure 48 below, providing an overview of the model's performance.

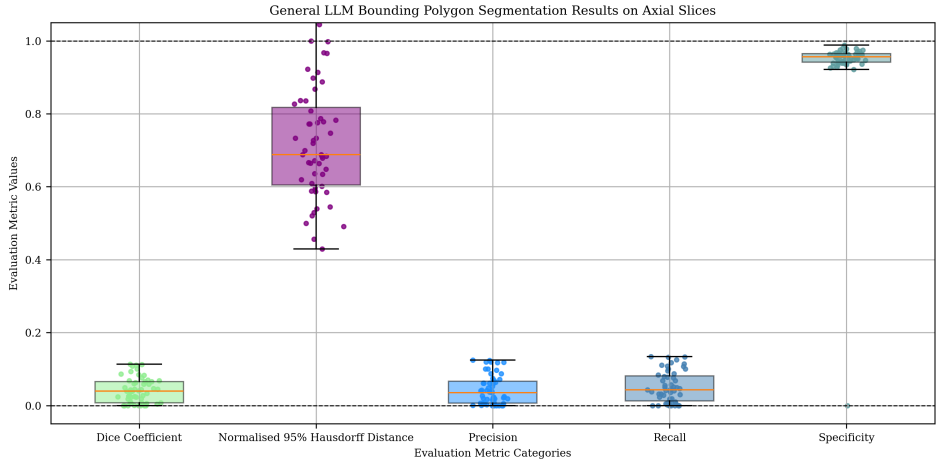


Figure 48: Evaluation metrics for the general LLM bounding polygon model predictions.

Table 10: Evaluation of the general LLM bounding polygon model.

Metric	Value
Average Dice Coefficient	0.0412
Average 95% Hausdorff Distance	92.4296
Average Precision	0.0432
Average Recall	0.0502
Average Specificity	0.9371

The general LLM model's performance on the bounding polygon test shows a Dice coefficient of 0.0412, indicating a very low overlap between the predicted and actual tumor areas. The 95th percentile Hausdorff distance of 92.4296 suggests that the predicted segmentation is significantly distant from the true tumor boundaries, reflecting substantial localization errors. With a precision of 0.0432, only 4.32% of the model's tumor predictions are correct, suggesting a high rate of false positives and poor tumor identification. Recall is also low at 0.0502, indicating that the model is not effectively identifying a significant proportion of the actual tumors. However, the specificity of 0.9371 shows that the model is good at avoiding false positives, correctly identifying regions of the image where tumors are not present.

The visualizations of the bounding polygon segmentations did not reveal any discernible trends or improvements. As seen in Figure 49, the predicted shapes appeared random, varying in size and consistently placed in the center of the image. This shows no clear ability to localize tumors to specific ground truth sizes or locations. For adjacent slices, which are often very similar in scans and would logically require similar segmentations, the model made drastically different predictions for each slice. As a result, the polygon segmentations resembled irregular, star-like shapes, emphasizing the model's inability to localize and segment gliomas across slices. Further general LLM bounding polygon segmentation visualizations can be found in Appendix H.

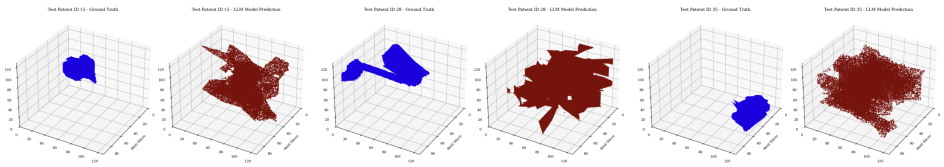


Figure 49: General LLM bounding polygon segmentation visualization showed random predicted shapes, centered in the image with varying sizes.

5.3.7 LLM Subspecialised Bounding Polygon Model Performance

The general bounding polygon model was fine-tuned for 150 steps with a batch size of 4, and the plot is shown below in Figure 50. With fine-tuning, we would expect the

model to improve its ability to differentiate the location and size of gliomas, leading to an increase in Dice coefficient overlap, precision, recall, and specificity, and a decrease in the 95% Hausdorff distance. Figure 51 presents the evaluation results.

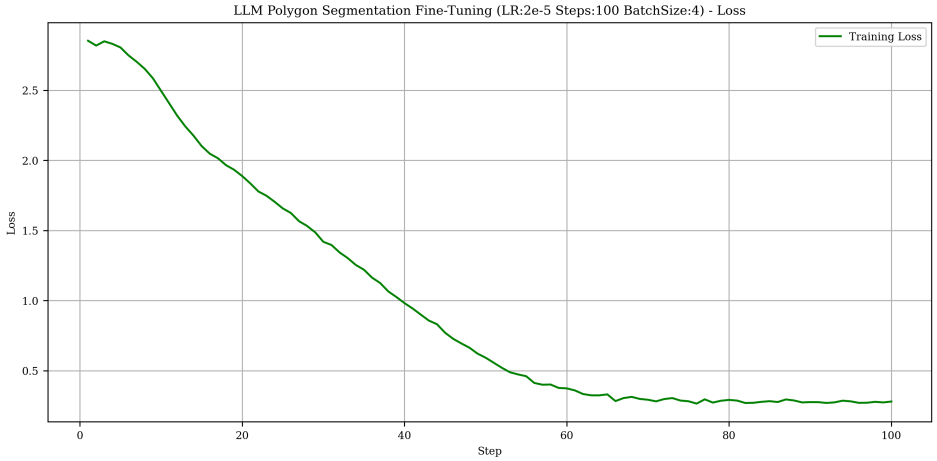


Figure 50: LLM model fine-tuning loss (learning rate = 2e-5, steps = 100).

Table 11: Evaluation of the fine-tuned LLM bounding polygon model.

Metric	Value
Average Dice Coefficient	0.0335
Average 95% Hausdorff Distance	86.9709
Average Precision	0.0487
Average Recall	0.0296
Average Specificity	0.9568

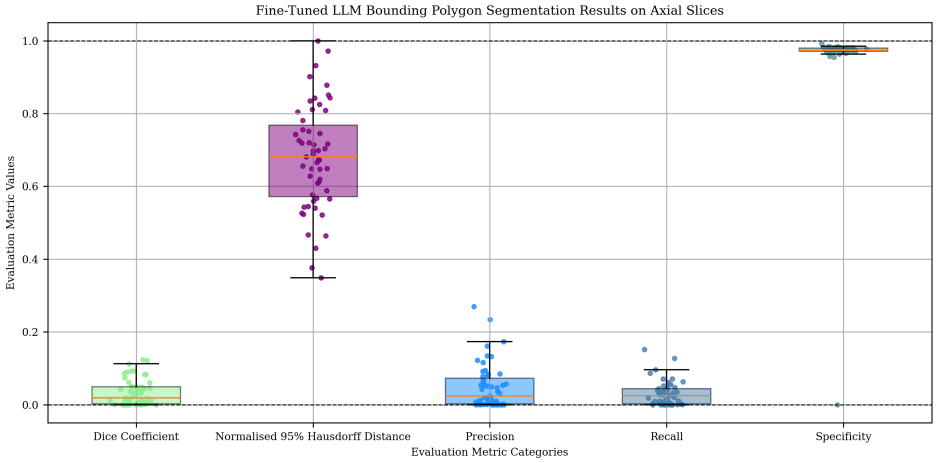


Figure 51: Evaluation metrics for the fine-tuned LLM bounding polygon predictions.

The Dice coefficient decreased to 0.0335, indicating that the fine-tuning did not significantly enhance the model's ability to segment the gliomas accurately. The 95%

Hausdorff distance slightly improved to 86.9709, suggesting that while the model is still struggling with the precise alignment of predicted tumor boundaries, it has not shown substantial progress. Precision increased to 0.0487, though still low, indicating a slight improvement in the model's ability to correctly identify true tumor boundaries. However, recall dropped to 0.0296, showing that the model is still failing to capture a significant portion of the actual tumors. Specificity improved to 0.9568, demonstrating that the model has become more effective at avoiding false positives. Overall, the fine-tuning resulted in marginal improvements but did not lead to significant performance gains in accurately segmenting gliomas.

After fine-tuning, the bounding polygon predictions appear slightly more concentrated around the center of the image, as shown in Figure 52, but overall, they remain largely unchanged from the previous results. This suggests that fine-tuning did not lead to significant improvements in the model's ability to localize or accurately segment gliomas. The pattern of random, inconsistent shapes persists, and the segmentations still do not align with the actual tumor locations or sizes. This raises the possibility that the current evaluation method may not have been the most effective in capturing meaningful improvements in the model's performance. Further fine-tuned LLM bounding polygon segmentation visualizations can be found in Appendix I.

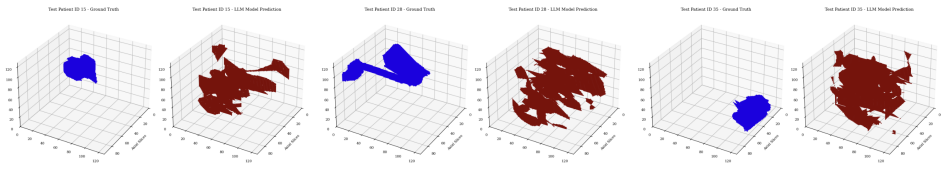


Figure 52: Fine-tuned LLM bounding polygon segmentation visualization shows that the segmentations remain largely unchanged.

5.4 Discussion

5.4.1 Key Findings

The core objective of this research was to evaluate and compare the performance of general-purpose large language models (LLMs) and subspecialized models in medical image segmentation tasks, with a focus on their accuracy, robustness, and practical utility in environments with limited labeled data. The evaluation aimed to assess when fine-tuning general LLMs provides measurable benefits over simply using them for inference out-of-the-box in the context of segmentation.

The CNN baseline model showed solid performance in glioma segmentation, achieving a Dice coefficient of 0.5942 on the testing set. It successfully segmented larger gliomas, accurately capturing tumor boundaries and general shapes. In these cases, the

model aligned well with the ground truth, demonstrating reliable tumor detection and boundary delineation. However, challenges emerged with smaller gliomas and those with irregular shapes. For these tumors, the model struggled to accurately localize and define boundaries, sometimes predicting scattered points instead of a well-defined tumor shape. This was particularly noticeable with gliomas of unusual sizes or strange shapes, where the model tended to either overestimate or under-represent the tumor extent. Additionally, the model occasionally over-segmented the brain, capturing broader regions beyond the glioma, such as other brain structures or anomalies. This over-segmentation could have been due to low contrast in the images or the presence of other brain features that the model misidentified as part of the tumor. Despite this, the model still demonstrated a strong ability to identify and segment gliomas in more straightforward cases, with fewer issues in detecting larger or more distinct tumors.

The general LLM model's performance in center point segmentation was notably poor, with the predicted points mostly scattered around the center of the image rather than aligning with the true tumor location. Only 8.05% of the predicted points fell within the ground truth bounding box, reflecting a random approach to identifying tumor centers. The high 95th percentile Euclidean and shortest distances (57.4112 and 22.8566, respectively) further demonstrated that the model struggled to accurately localize gliomas, often predicting points far from the actual tumor locations. After fine-tuning, there was a slight increase in the variance of the predictions, indicating the model was attempting to explore different regions of the image. However, the improvement was minimal, with the percentage of points within the bounding box remaining nearly unchanged at 8.00%. The distance metrics slightly worsened (95th percentile Euclidean distance: 68.4812, shortest distance: 26.2855), suggesting that the model had not learned to focus on identifying the correct tumor centers. Instead, the predictions remained centered in the image. These results highlight that fine-tuning did not yield meaningful improvements in the model's ability to localize gliomas, and the predictions remained largely random.

The general LLM model's performance on the bounding box segmentation task also revealed significant shortcomings. With a Dice coefficient of 0.1219, the model showed poor overlap between the predicted and actual tumor regions. The 95th percentile Hausdorff distance of 65.9103 indicated that the predicted tumor boundaries were consistently far from the true tumor locations, and with a precision of 0.0989 and recall of 0.2941, the model struggled to accurately identify and localize gliomas, with a high rate of false positives and missed tumor regions. Visualizations revealed that the model's bounding boxes were often centered in the image, irrespective of the tumor's actual location or size, suggesting that the model did not effectively utilize the spatial information or tumor features, highlighting its limitations in precise tumor segmentation.

After fine-tuning for 200 steps, the model's performance showed little improvement. The Dice coefficient slightly decreased to 0.1085, and the 95th percentile Hausdorff distance rose to 67.3812, reflecting the continued difficulty in aligning predicted bounding boxes with actual tumor boundaries. Precision and recall further declined to 0.0882 and 0.2413, respectively, indicating a deterioration in the model's ability to detect true positive centers. However, specificity improved to 0.8137, suggesting better handling of false positives. Despite these marginal changes, qualitative analysis of the bounding box segmentations revealed that while the model was now consistently producing more bounding box predictions, the placement of these boxes remained largely unchanged, and were still placed near the center of the image. The model also showed no sensitivity to the gliomas' size, indicating it had not learned to make the critical distinctions necessary for accurate segmentation. Instead of focusing on relevant tumor features, the model's predictions remained centered and arbitrary, demonstrating that fine-tuning did not lead to meaningful improvements in performance.

The general LLM model's performance on the bounding polygon segmentation task also yielded poor results. The Dice coefficient of 0.0412 indicates minimal overlap between predicted and true tumor areas, while a 95th percentile Hausdorff distance of 92.4296 highlights significant localization errors. Precision (0.0432) and recall (0.0502) were very low, reflecting a high false-positive rate and poor tumor identification. However, specificity was high (0.9371), indicating the model avoided false positives but failed to accurately segment tumors. Visualizations showed random, irregular shapes centered in the image, with no correlation to actual tumor locations or sizes. Adjacent slices, which should have similar segmentations, showed drastically different predictions. After fine-tuning for 150 steps, the model showed no significant improvements, with the Dice coefficient decreasing to 0.0335, the Hausdorff distance improving slightly to 86.9709, precision rising to 0.0487, and specificity increasing to 0.9568. However, recall dropped to 0.0296, and the bounding polygons still failed to align with true tumor locations, suggesting limited impact from fine-tuning and ineffective evaluation.

The findings of this study align with the hypothesis that general-purpose LLMs, despite their versatility in a range of medical tasks, are not inherently suited for specialized tasks like medical image segmentation, even with fine-tuning.

5.4.2 Limitations

A major limitation of this study is the difference in input data between the CNN and the LLM models, which mirrors the limitation observed in the classification task. The CNN utilized full 3D convolutions and all four MRI modalities (FLAIR, T1, T1ce, and T2), providing rich spatial and multi-modal information, which likely contributed to its

stronger performance in segmentation. In contrast, the LLM was restricted to 2D axial slices from only the FLAIR modality. Each slice was processed independently, with patient-level predictions determined by majority voting across slices. This restricted input, while necessary for the LLM's prompting structure, limited its access to full spatial and modality information, affecting its segmentation performance.

The restricted scope of fine-tuning, was another limitation shared with the classification task, also hindered the LLM's performance in segmentation. The small batch size of 16 2D images per iteration, limited by Colab GPU memory constraints, and the short training duration of just 100 to 200 steps, significantly impacted the model's ability to learn robust features. This was compounded by the lack of validation checks, which would have allowed for better monitoring of the model's learning process. While small learning rates were used to mitigate overfitting, the absence of a validation curve introduced uncertainty regarding the model's true performance. These resource constraints likely contributed to the suboptimal fine-tuning results, which were consistent with the issues encountered during classification.

Another limitation was the different prompting strategies required to enable the LLM to perform segmentation. Unlike the CNN, which directly learns spatial relationships from voxel-based 3D data, the LLM requires text-based input for its segmentation task. This conversion of voxel locations into tokens for prompting could have introduced complexity in learning spatial relationships, as the model was tasked with interpreting segmentation as a sequence of tokens rather than directly processing physical voxel locations. Converting these locations back into segmentation predictions likely added an additional layer of abstraction, making it more difficult for the model to learn accurate tumor boundaries. This approach could have impeded the model's ability to effectively capture spatial details and localize gliomas, which are typically better represented by the direct manipulation of voxel-based data in models like CNNs.

5.5 Conclusion

The CNN model outperformed the general-purpose LLM in terms of accuracy, robustness, and reliability in segmenting gliomas, particularly in accurately identifying tumor boundaries and handling varying tumor sizes. While the LLM showed some potential, its initial performance was limited by poor spatial localization and high variability in segmentation results. Fine-tuning the LLM did not yield significant improvements due to constraints in training duration, batch size, and available resources. Although the LLM has potential for medical image segmentation, further refinement, more robust fine-tuning, and improvements in input processing are needed for it to effectively handle specialized tasks like glioma segmentation.

6 Conclusion and Future Work

In conclusion, this research provided valuable insights into the effectiveness of general-purpose large language models (LLMs) versus specialized models in medical imaging tasks, particularly in glioma classification and segmentation. Our findings demonstrated that while general-purpose LLMs showed some promise, they were significantly outperformed by specialized models, such as Convolutional Neural Networks (CNNs), in terms of accuracy, robustness, and handling class imbalances. The CNN baseline excelled in both glioma classification and segmentation, achieving strong performance metrics and effectively addressing challenges like class imbalance and localization errors. In contrast, the LLM struggled, with issues such as low specificity, inconsistent predictions, and difficulties in learning spatial relationships due to its text-based nature. Fine-tuning the LLM resulted in only marginal improvements, hindered by constraints like limited data, small batch sizes, and insufficient training duration.

These findings suggest that while general-purpose LLMs have potential, they are not yet well-suited for specialized medical tasks like glioma classification and image segmentation. Specialized models, particularly CNNs, proved to be more accurate and robust, underscoring the importance of domain-specific models in handling complex medical data. This research emphasizes the need for tailored approaches in healthcare AI, offering a clearer understanding of when and where LLMs are most effective and establishing a framework for selecting optimal models and training strategies for various medical tasks. Ultimately, fine-tuning was found to have limited impact unless substantial adaptations were made to align the models more closely with the specific requirements of the medical tasks at hand.

There is still more work to be done in this area. First, more rigorous fine-tuning with larger batch sizes, longer training durations, and proper validation will be necessary. This will allow for a more robust comparison of the fine-tuned results, ensuring that the improvements seen are genuine and not influenced by resource constraints or training limitations. Additionally, it will be crucial to determine whether these findings are replicable across different datasets. While this study primarily focused on the BraTS dataset, it would be valuable to explore whether similar trends emerge when applied to other datasets. This could help assess the generalizability of our results and identify any dataset-specific differences or challenges. Evaluating the value of text-based LLMs in image-based tasks raises the immediate question of how these models compare to foundational models dedicated to imaging tasks.

A further comparative study between text-based LLMs and specialized foundational imaging models would also provide insights into their relative performance and highlight strengths and weaknesses in different contexts.

References

[1] M. A. Rahman, E. Victoros, J. Ernest, *et al.*, "Impact of Artificial Intelligence (AI) Technology in Healthcare Sector: A Critical Evaluation of Both Sides of the Coin," *Clinical Pathology*, vol. 17, p. 2632010X241226887, Jan. 2024, ISSN: 2632-010X. DOI: 10.1177/2632010X241226887. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10804900/ (visited on 01/19/2025).

[2] J. Bajwa, U. Munir, A. Nori, *et al.*, "Artificial intelligence in healthcare: Transforming the practice of medicine," *Future Healthcare Journal*, vol. 8, no. 2, e188–e194, Jul. 2021, ISSN: 2514-6645. DOI: 10.7861/fhj.2021-0095. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8285156/ (visited on 01/19/2025).

[3] A. Derevianko, S. F. M. Pizzoli, F. Pesapane, *et al.*, "The Use of Artificial Intelligence (AI) in the Radiology Field: What Is the State of Doctor–Patient Communication in Cancer Diagnosis?" *Cancers*, vol. 15, no. 2, p. 470, Jan. 2023, ISSN: 2072-6694. DOI: 10.3390/cancers15020470. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9856827/ (visited on 01/19/2025).

[4] L.-H. Yao, K.-C. Leung, C.-L. Tsai, *et al.*, "A Novel Deep Learning–Based System for Triage in the Emergency Department Using Electronic Medical Records: Retrospective Cohort Study," *Journal of Medical Internet Research*, vol. 23, no. 12, e27008, Dec. 2021, ISSN: 1439-4456. DOI: 10.2196/27008. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8749584/ (visited on 01/19/2025).

[5] R. Ranjbarzadeh, A. Bagherian Kasgari, S. Jafarzadeh Ghoushchi, *et al.*, "Brain tumor segmentation based on deep learning and an attention mechanism using MRI multi-modalities brain images," en, *Scientific Reports*, vol. 11, no. 1, p. 10 930, May 2021, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-021-90428-8. [Online]. Available: https://www.nature.com/articles/s41598-021-90428-8 (visited on 01/19/2025).

[6] Z. Tariq, S. K. Shah, and Y. Lee, "Lung Disease Classification using Deep Convolutional Neural Network," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2019, pp. 732–735. DOI: 10.1109/BIBM47256.2019.8983071. [Online]. Available: https://ieeexplore.ieee.org/document/8983071/?arnumber=8983071 (visited on 01/19/2025).

[7] F. Faria, M. B. Moin, P. Debnath, *et al.*, "Explainable convolutional neural networks for retinal fundus classification and cutting-edge segmentation models for

retinal blood vessels from fundus images," May 2024. DOI: 10.48550/arXiv.2405.07338.

[8] G. Verma, "Retinal Image Analysis for Disease Classification using Convolutional Neural Networks," in *2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, ISSN: 2768-0673, Oct. 2024, pp. 1284–1288. DOI: 10.1109/I-SMAC61858.2024.10714588. [Online]. Available: https://ieeexplore.ieee.org/document/10714588 (visited on 01/19/2025).

[9] T. Ersavas, M. A. Smith, and J. S. Mattick, "Novel applications of Convolutional Neural Networks in the age of Transformers," en, *Scientific Reports*, vol. 14, no. 1, p. 10 000, May 2024, Publisher: Nature Publishing Group, ISSN: 2045-2322. DOI: 10.1038/s41598-024-60709-z. [Online]. Available: https://www.nature.com/articles/s41598-024-60709-z (visited on 01/19/2025).

[10] *Exploring Architectures and Capabilities of Foundational LLMs*, en-US. [Online]. Available: https://www.aporia.com/learn/exploring-architectures-and-capabilities-of-foundational-llms/ (visited on 01/19/2025).

[11] D. Yuan, E. Rastogi, G. Naik, *et al.*, "A Continued Pretrained LLM Approach for Automatic Medical Note Generation," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds., Mexico City, Mexico: Association for Computational Linguistics, Jun. 2024, pp. 565–571. DOI: 10.18653/v1/2024.naacl-short.47. [Online]. Available: https://aclanthology.org/2024.naacl-short.47 (visited on 10/05/2024).

[12] K. Singhal, S. Azizi, T. Tu, *et al.*, "Large language models encode clinical knowledge," en, *Nature*, vol. 620, no. 7972, pp. 172–180, Aug. 2023, Publisher: Nature Publishing Group, ISSN: 1476-4687. DOI: 10.1038/s41586-023-06291-2. [Online]. Available: https://www.nature.com/articles/s41586-023-06291-2 (visited on 09/29/2024).

[13] K. J. Prabhod, "Integrating Large Language Models for Enhanced Clinical Decision Support Systems in Modern Healthcare," en, *Journal of Machine Learning for Healthcare Decision Support*, vol. 3, no. 1, pp. 18–62, Jun. 2023, Number: 1, ISSN: 2347-9817. [Online]. Available: https://medlines.uk/index.php/JMLHDS/article/view/23 (visited on 10/05/2024).

[14] L. Masanneck, L. Schmidt, A. Seifert, *et al.*, "Triage Performance Across Large Language Models, ChatGPT, and Untrained Doctors in Emergency Medicine: Comparative Study," EN, *Journal of Medical Internet Research*, vol. 26, no. 1,

e53297, Jun. 2024, Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. DOI: 10.2196/53297. [Online]. Available: https://www.jmir.org/2024/1/e53297 (visited on 10/05/2024).

[15] H. Mondal, R. De, S. Mondal, *et al.*, "A large language model in solving primary healthcare issues: A potential implication for remote healthcare and medical education," en-US, *Journal of Education and Health Promotion*, vol. 13, no. 1, p. 362, Sep. 2024, ISSN: 2277-9531. DOI: 10.4103/jehp.jehp_688_23. [Online]. Available: https://journals.lww.com/jehp/fulltext/2024/09280/a_large_language_model_in_solving_primary.362.aspx (visited on 10/05/2024).

[16] R. Archana and P. S. E. Jeevaraj, "Deep learning models for digital image processing: A review," en, *Artificial Intelligence Review*, vol. 57, no. 1, p. 11, Jan. 2024, ISSN: 1573-7462. DOI: 10.1007/s10462-023-10631-z. [Online]. Available: https://doi.org/10.1007/s10462-023-10631-z (visited on 04/06/2025).

[17] S. Bakas, M. Reyes, A. Jakab, *et al.*, *Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge*, arXiv:1811.02629 [cs, stat], Apr. 2019. DOI: 10.48550/arXiv.1811.02629. [Online]. Available: http://arxiv.org/abs/1811.02629 (visited on 10/05/2024).

[18] S. Bakas, H. Akbari, A. Sotiras, *et al.*, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," eng, *Scientific Data*, vol. 4, p. 170 117, Sep. 2017, ISSN: 2052-4463. DOI: 10.1038/sdata.2017.117.

[19] B. H. Menze, A. Jakab, S. Bauer, *et al.*, "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)," eng, *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015, ISSN: 1558-254X. DOI: 10.1109/TMI.2014.2377694.

[20] S. Bakas, H. Akbari, A. Sotiras, *et al.*, *BRATS-TCGA-GBM*, en-US. [Online]. Available: https://www.cancerimagingarchive.net/analysis-result/brats-tcga-gbm/ (visited on 10/06/2024).

[21] *BRATS-TCGA-LGG*, en-US. [Online]. Available: https://www.cancerimagingarchive.net/analysis-result/brats-tcga-lgg/ (visited on 10/06/2024).

[22] *Meta-llama/Llama-3.2-11B-Vision-Instruct · Hugging Face*, Dec. 2024. [Online]. Available: https://huggingface.co/meta-llama/Llama-3.2-11B-Vision-Instruct (visited on 01/19/2025).

[23] *Llama 3.2 Vision Fine-tuning with Unsloth*, en. [Online]. Available: https://unsloth.ai/blog/vision (visited on 01/19/2025).

[24] *Mist - SciNet Users Documentation*. [Online]. Available: https://docs.scinet.utoronto.ca/index.php/Mist (visited on 01/19/2025).

[25] *Google Colab*. [Online]. Available: https://research.google.com/colaboratory/faq.html (visited on 04/06/2025).

[26] *NVIDIA L4 vs. A100 GPUs: Choosing the Right Option for Your AI Needs*, en. [Online]. Available: https://www.e2enetworks.com/blog/nvidia-l4-vs-a100-gpus-choosing-the-right-option-for-your-ai-needs (visited on 04/06/2025).

[27] R. Bakshi, S. Ariyaratana, R. H. B. Benedict, *et al.*, "Fluid-Attenuated Inversion Recovery Magnetic Resonance Imaging Detects Cortical and Juxtacortical Multiple Sclerosis Lesions," *Archives of Neurology*, vol. 58, no. 5, pp. 742–748, May 2001, ISSN: 0003-9942. DOI: 10.1001/archneur.58.5.742. [Online]. Available: https://doi.org/10.1001/archneur.58.5.742 (visited on 04/06/2025).

[28] A. Kirillov, E. Mintun, N. Ravi, *et al.*, "Segment Anything," en, in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 3992–4003, ISBN: 9798350307184. DOI: 10.1109/ICCV51070.2023.00371. [Online]. Available: https://ieeexplore.ieee.org/document/10378323/ (visited on 04/06/2025).

[29] L. Zhu, T. Chen, D. Ji, *et al.*, "LLaFS: When Large Language Models Meet Few-Shot Segmentation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, ISSN: 2575-7075, Jun. 2024, pp. 3065–3075. DOI: 10.1109/CVPR52733.2024.00296. [Online]. Available: https://ieeexplore.ieee.org/document/10657904/ (visited on 04/06/2025).

A Classification Task: CNN Baseline Models

This appendix provides results for alternative CNN baseline models trained for the classification task. Although these models were not chosen due to less stable convergence, they still achieved strong performance. For each model, training and validation plots are included to show learning behavior, along with a summary table reporting evaluation metrics across the training, validation, and test sets.

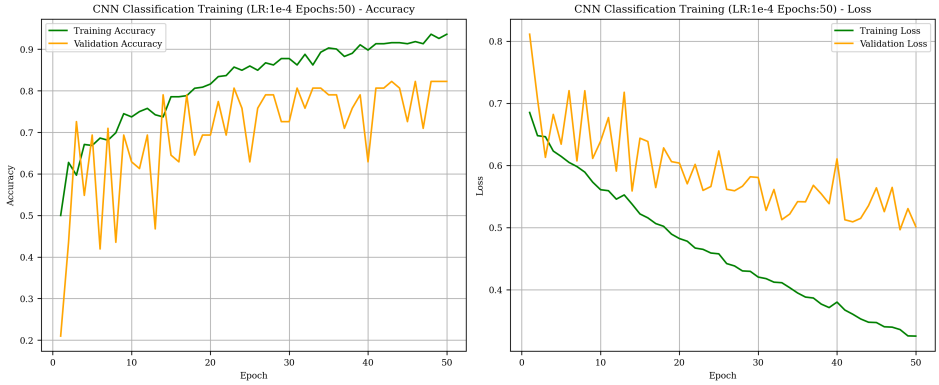


Figure 53: Training and validation accuracy and loss curves for the CNN model (learning rate = 1e-4, epochs = 50).

Table 12: Metrics for the CNN model (learning rate = 1e-4, epochs = 50).

Metric	Training	Validation	Testing
Accuracy	0.8266	0.5968	0.7091
F1 Score	0.8841	0.7191	0.8140
Precision	0.9371	0.8000	0.8333
Recall	0.8367	0.6531	0.7955
AUC	0.9128	0.6295	0.7025
Specificity	0.9375	0.3846	0.3636

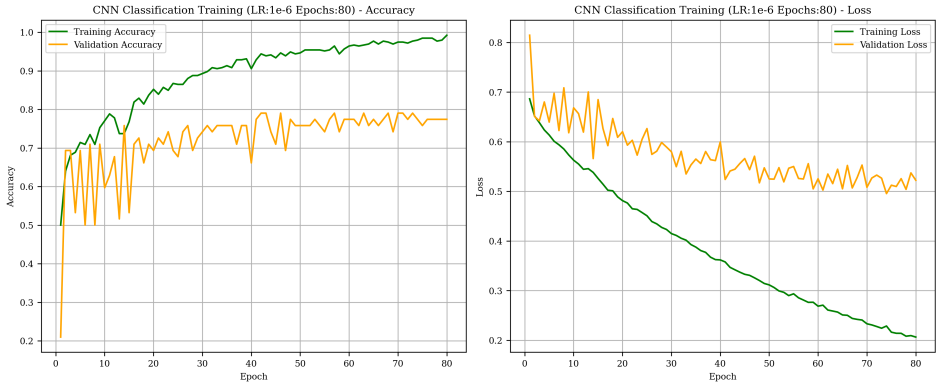


Figure 54: Training and validation accuracy and loss curves for the CNN model (learning rate = 1e-6, epochs = 80).

Table 13: Metrics for the CNN model (learning rate = 1e-6, epochs = 80).

Metric	Training	Validation	Testing
Accuracy	0.9879	0.7742	0.8182
F1 Score	0.9923	0.8627	0.8864
Precision	1.0000	0.8302	0.8864
Recall	0.9847	0.8980	0.8864
AUC	0.9994	0.6766	0.7562
Specificity	1.0000	0.3077	0.5455

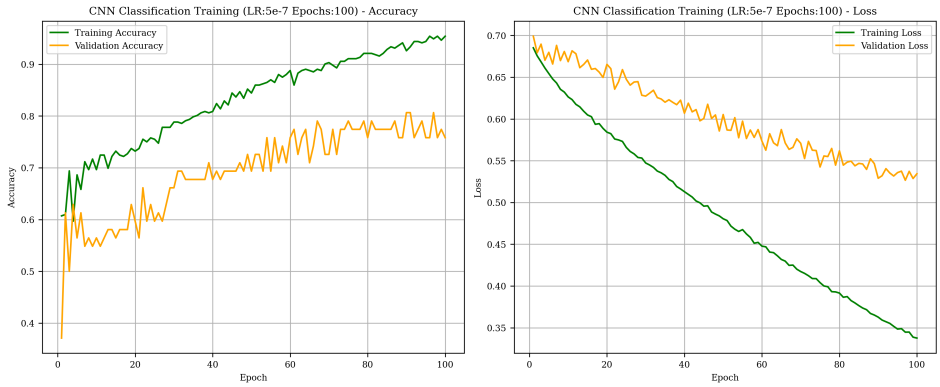


Figure 55: Training and validation accuracy and loss curves for the CNN model (learning rate = 5e-7, epochs = 100).

Table 14: Metrics for the CNN model (learning rate = 5e-7, epochs = 100).

Metric	Training	Validation	Testing
Accuracy	0.9435	0.7581	0.7636
F1 Score	0.9634	0.8515	0.8506
Precision	0.9892	0.8269	0.8605
Recall	0.9388	0.8776	0.8409
AUC	0.9878	0.6829	0.6942
Specificity	0.9615	0.3077	0.4545

B Segmentation Task: CNN Baseline Models

This appendix provides results for alternative CNN baseline models trained for the segmentation task. Although these models were not selected due to suboptimal convergence, they still produced reasonable performance. For each model, training and validation plots are included to show learning behavior, along with a summary table reporting evaluation metrics across the training, validation, and test sets.

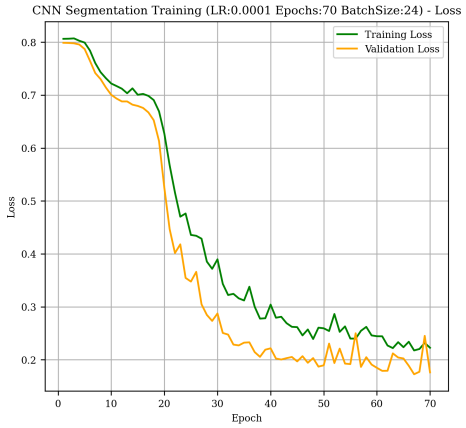


Figure 56: Training and validation loss (learning rate = 1e-4, epochs = 70).

Table 15: Metrics for the CNN model (learning rate = 1e-4, epochs = 70).

Metric	Training	Validation	Testing
Dice Coefficient	0.5820	0.5992	0.5751
95% Hausdorff Distance	62.3937	64.6147	59.5499
Precision	0.6180	0.6359	0.6025
Recall	0.7771	0.7506	0.7692
Specificity	0.9025	0.9315	0.9243

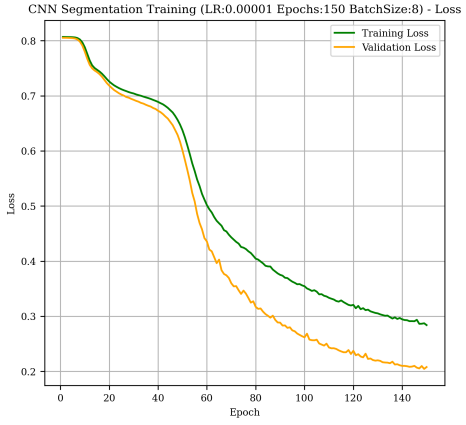


Figure 57: Training and validation loss (learning rate = 1e-5, epochs = 150).

Table 16: Metrics for the CNN model (learning rate = 1e-5, epochs = 150).

Metric	Training	Validation	Testing
Dice Coefficient	0.5477	0.6096	0.5456
95% Hausdorff Distance	64.8179	61.8089	66.8668
Precision	0.6062	0.6505	0.5427
Recall	0.7719	0.7931	0.7606
Specificity	0.8790	0.9085	0.9082

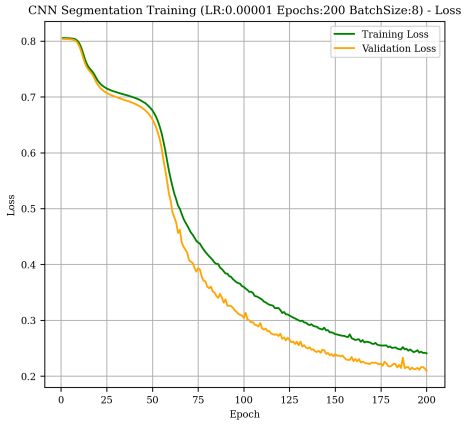


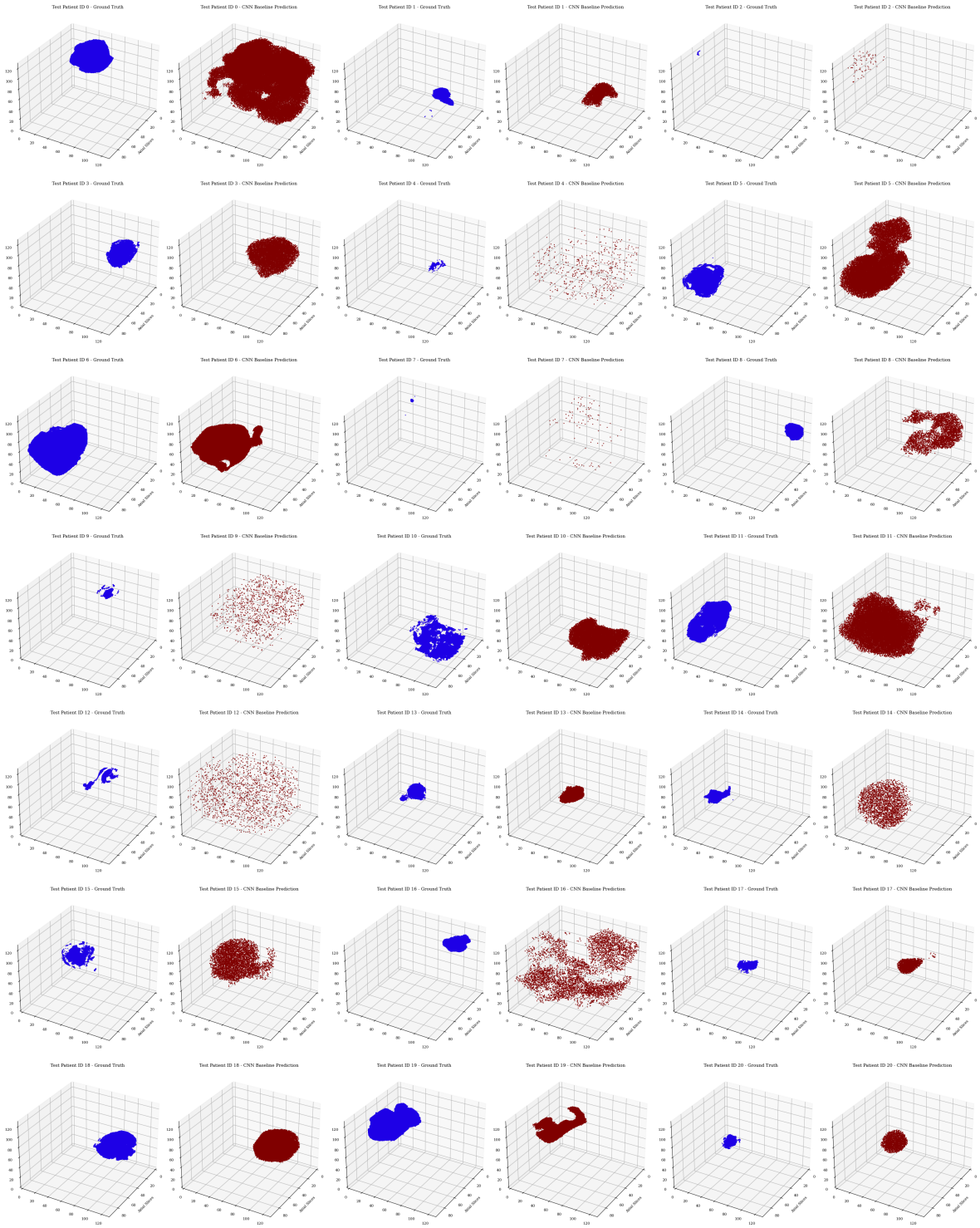
Figure 58: Training and validation loss (learning rate = 1e-5, epochs = 200).

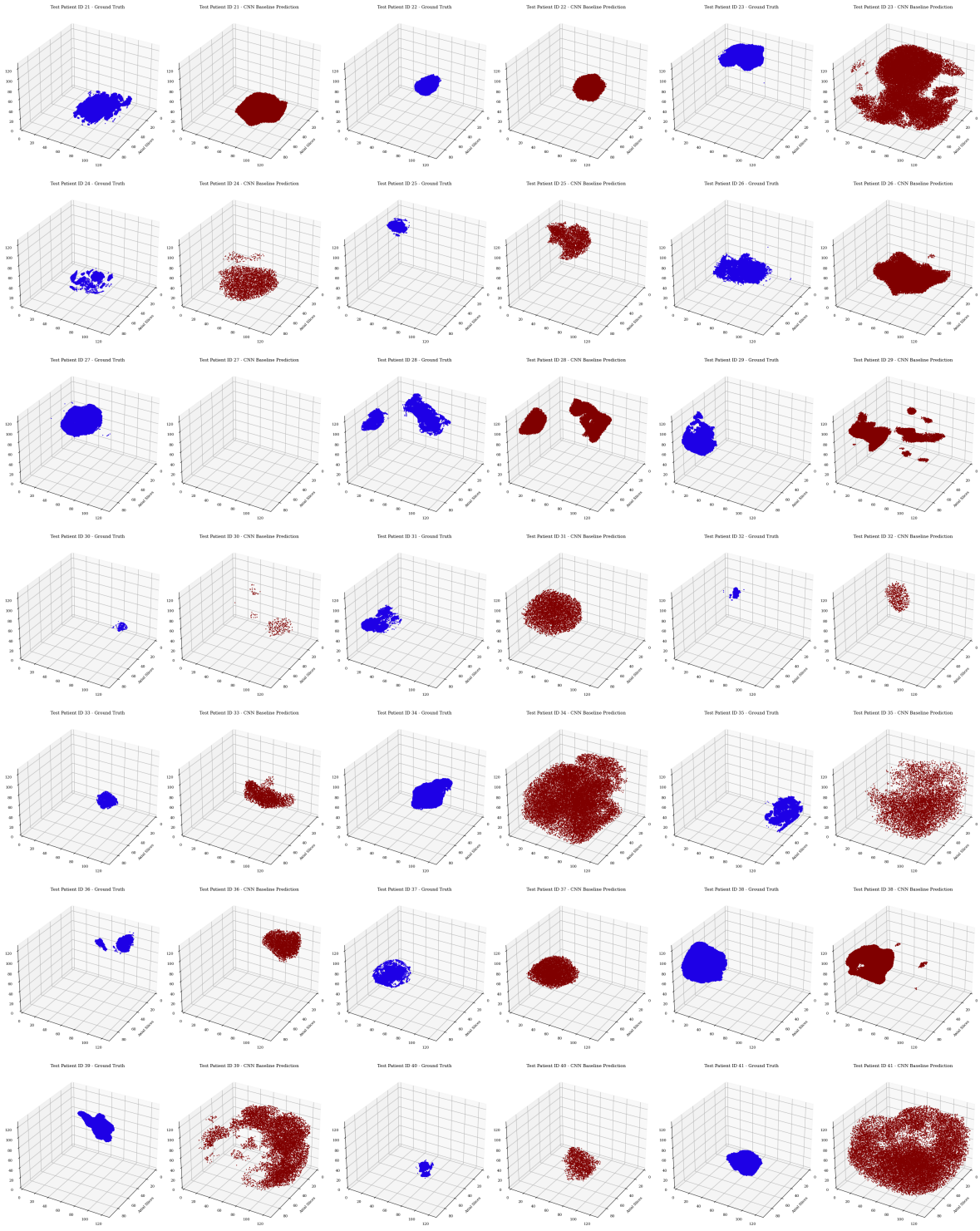
Table 17: Metrics for the CNN model (learning rate = 1e-5, epochs = 200).

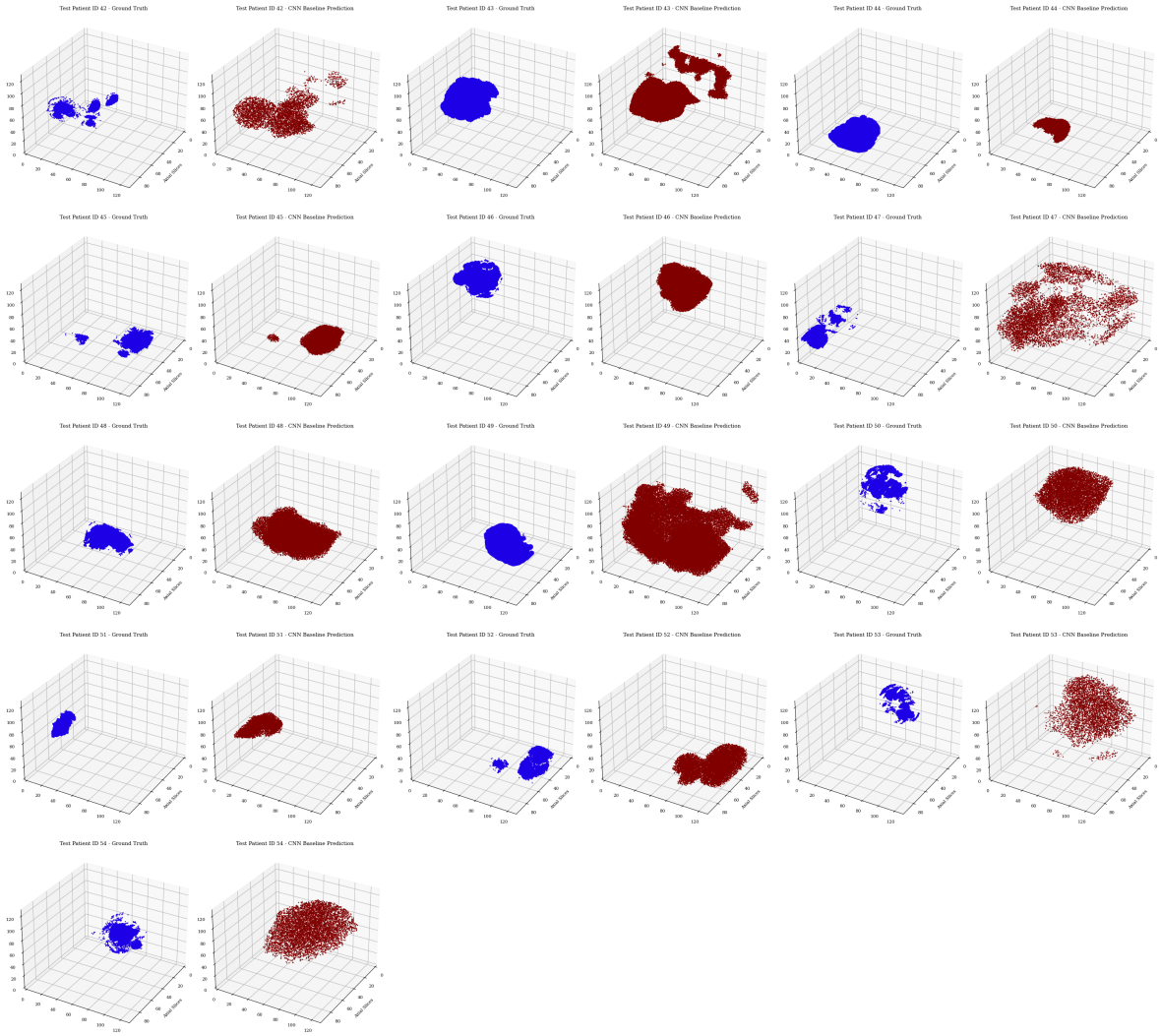
Metric	Training	Validation	Testing
Dice Coefficient	0.5739	0.5990	0.5596
95% Hausdorff Distance	62.1015	62.8311	58.2910
Precision	0.6244	0.6381	0.6064
Recall	0.7721	0.7613	0.7643
Specificity	0.8949	0.9203	0.9172

C Baseline CNN Segmentation

This section presents the segmentation visualizations generated by the baseline CNN model for all 55 test patients.

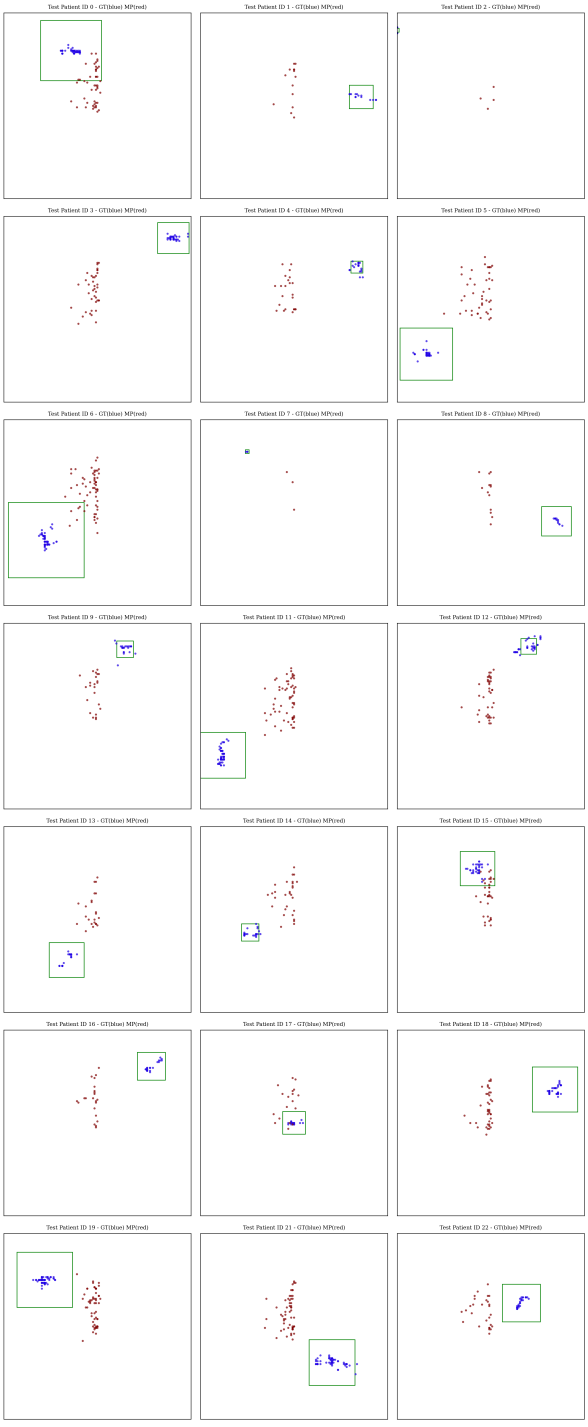


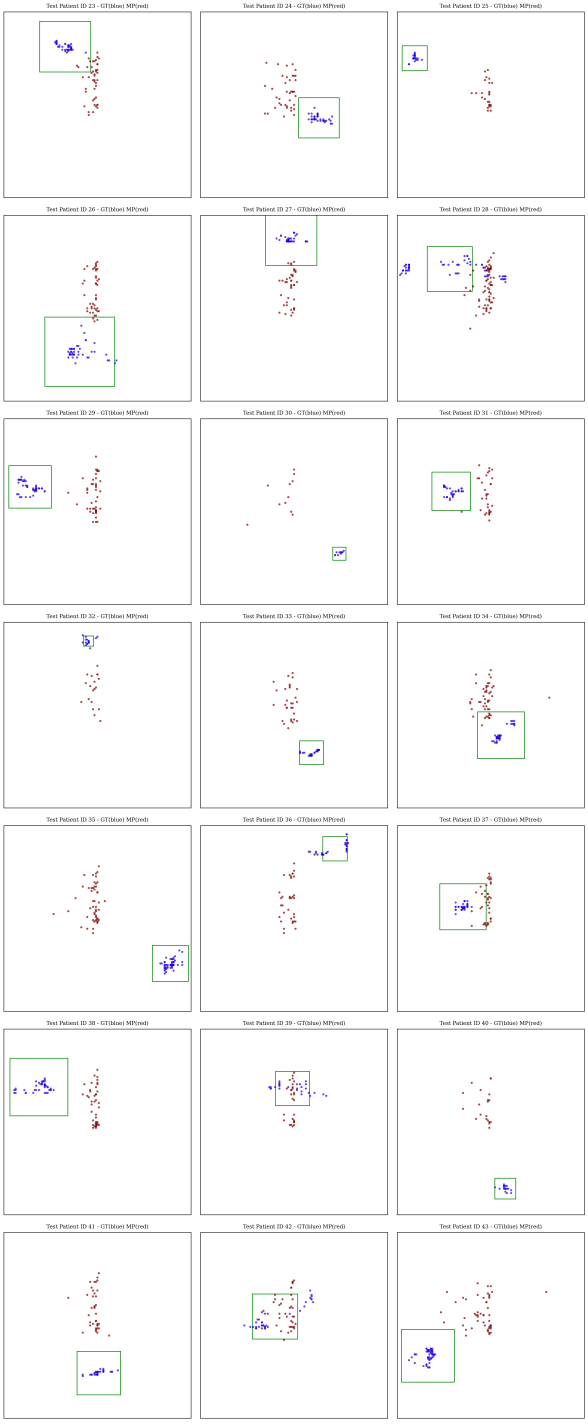

Test Patient ID 21 - Ground Truth
Test Patient ID 21 - CNN Baseline Prediction
Test Patient ID 22 - Ground Truth
Test Patient ID 22 - CNN Baseline Prediction
Test Patient ID 23 - Ground Truth
Test Patient ID 23 - CNN Baseline Prediction
Test Patient ID 24 - Ground Truth
Test Patient ID 24 - CNN Baseline Prediction
Test Patient ID 25 - Ground Truth
Test Patient ID 25 - CNN Baseline Prediction
Test Patient ID 26 - Ground Truth
Test Patient ID 26 - CNN Baseline Prediction
Test Patient ID 27 - Ground Truth
Test Patient ID 27 - CNN Baseline Prediction
Test Patient ID 28 - Ground Truth
Test Patient ID 28 - CNN Baseline Prediction
Test Patient ID 29 - Ground Truth
Test Patient ID 29 - CNN Baseline Prediction
Test Patient ID 30 - Ground Truth
Test Patient ID 30 - CNN Baseline Prediction
Test Patient ID 31 - Ground Truth
Test Patient ID 31 - CNN Baseline Prediction
Test Patient ID 32 - Ground Truth
Test Patient ID 32 - CNN Baseline Prediction
Test Patient ID 33 - Ground Truth
Test Patient ID 33 - CNN Baseline Prediction
Test Patient ID 34 - Ground Truth
Test Patient ID 34 - CNN Baseline Prediction
Test Patient ID 35 - Ground Truth
Test Patient ID 35 - CNN Baseline Prediction
Test Patient ID 36 - Ground Truth
Test Patient ID 36 - CNN Baseline Prediction
Test Patient ID 37 - Ground Truth
Test Patient ID 37 - CNN Baseline Prediction
Test Patient ID 38 - Ground Truth
Test Patient ID 38 - CNN Baseline Prediction
Test Patient ID 39 - Ground Truth
Test Patient ID 39 - CNN Baseline Prediction
Test Patient ID 40 - Ground Truth
Test Patient ID 40 - CNN Baseline Prediction
Test Patient ID 41 - Ground Truth
Test Patient ID 41 - CNN Baseline Prediction

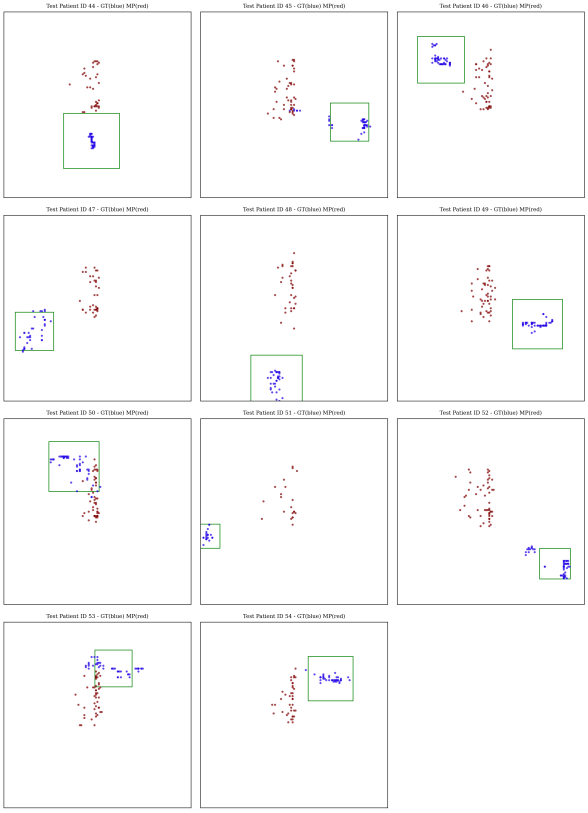

Test Patient ID 42 - Ground Truth
Test Patient ID 42 - CNN Baseline Prediction
Test Patient ID 43 - Ground Truth
Test Patient ID 43 - CNN Baseline Prediction
Test Patient ID 44 - Ground Truth
Test Patient ID 44 - CNN Baseline Prediction
Test Patient ID 45 - Ground Truth
Test Patient ID 45 - CNN Baseline Prediction
Test Patient ID 46 - Ground Truth
Test Patient ID 46 - CNN Baseline Prediction
Test Patient ID 47 - Ground Truth
Test Patient ID 47 - CNN Baseline Prediction
Test Patient ID 48 - Ground Truth
Test Patient ID 48 - CNN Baseline Prediction
Test Patient ID 49 - Ground Truth
Test Patient ID 49 - CNN Baseline Prediction
Test Patient ID 50 - Ground Truth
Test Patient ID 50 - CNN Baseline Prediction
Test Patient ID 51 - Ground Truth
Test Patient ID 51 - CNN Baseline Prediction
Test Patient ID 52 - Ground Truth
Test Patient ID 52 - CNN Baseline Prediction
Test Patient ID 53 - Ground Truth
Test Patient ID 53 - CNN Baseline Prediction
Test Patient ID 54 - Ground Truth
Test Patient ID 54 - CNN Baseline Prediction

D General LLM Center Point Segmentation

This section presents the center point segmentation visualizations generated by the general LLM model for all 55 test patients.

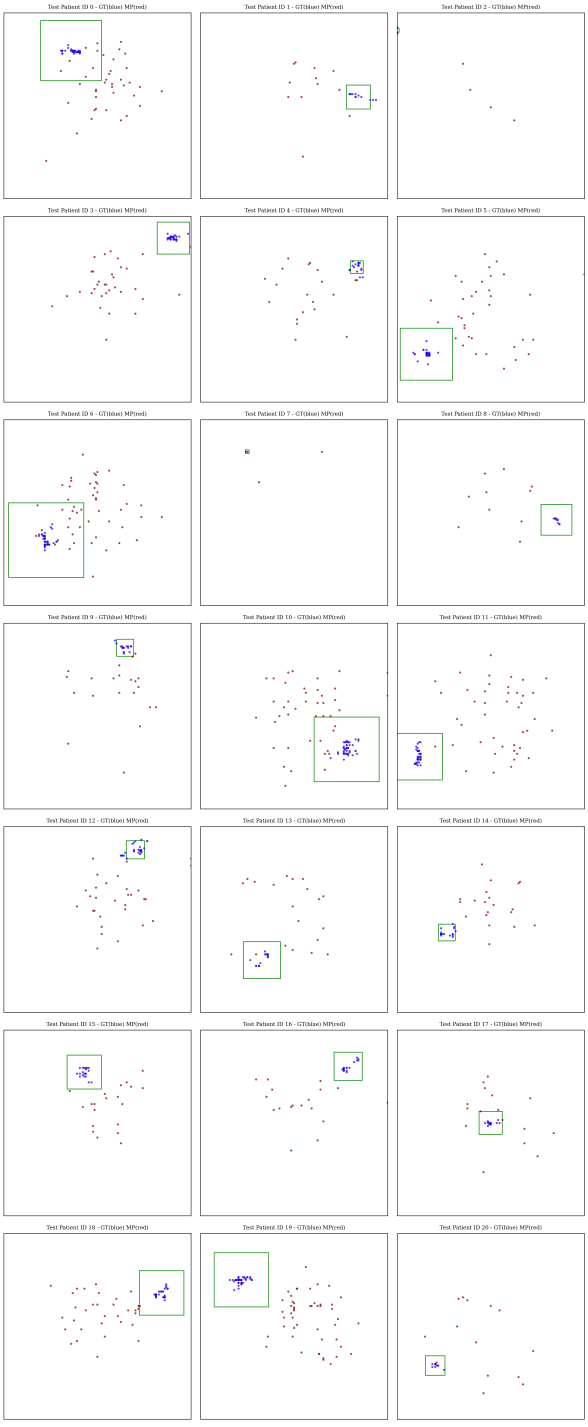



Test Patient ID 23 - GT(blue) MP(red)
Test Patient ID 24 - GT(blue) MP(red)
Test Patient ID 25 - GT(blue) MP(red)
Test Patient ID 26 - GT(blue) MP(red)
Test Patient ID 27 - GT(blue) MP(red)
Test Patient ID 28 - GT(blue) MP(red)
Test Patient ID 29 - GT(blue) MP(red)
Test Patient ID 30 - GT(blue) MP(red)
Test Patient ID 31 - GT(blue) MP(red)
Test Patient ID 32 - GT(blue) MP(red)
Test Patient ID 33 - GT(blue) MP(red)
Test Patient ID 34 - GT(blue) MP(red)
Test Patient ID 35 - GT(blue) MP(red)
Test Patient ID 36 - GT(blue) MP(red)
Test Patient ID 37 - GT(blue) MP(red)
Test Patient ID 38 - GT(blue) MP(red)
Test Patient ID 39 - GT(blue) MP(red)
Test Patient ID 40 - GT(blue) MP(red)
Test Patient ID 41 - GT(blue) MP(red)
Test Patient ID 42 - GT(blue) MP(red)
Test Patient ID 43 - GT(blue) MP(red)

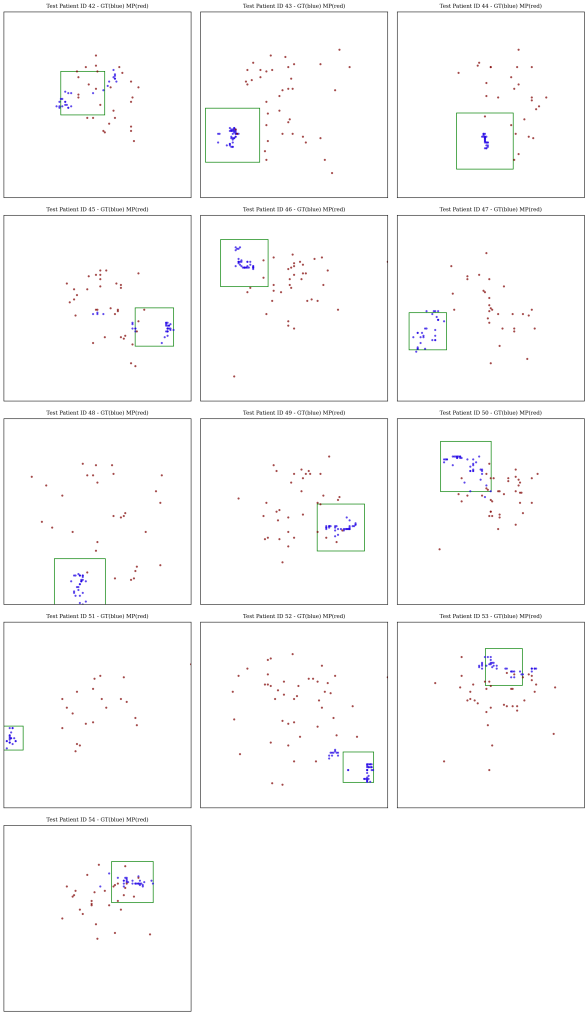

Test Patient ID 44 - GT(blue) MP(red)
Test Patient ID 45 - GT(blue) MP(red)
Test Patient ID 46 - GT(blue) MP(red)
Test Patient ID 47 - GT(blue) MP(red)
Test Patient ID 48 - GT(blue) MP(red)
Test Patient ID 49 - GT(blue) MP(red)
Test Patient ID 50 - GT(blue) MP(red)
Test Patient ID 51 - GT(blue) MP(red)
Test Patient ID 52 - GT(blue) MP(red)
Test Patient ID 53 - GT(blue) MP(red)
Test Patient ID 54 - GT(blue) MP(red)

E Fine-Tuned LLM Center Point Segmentation

This section presents the center point segmentation visualizations generated by the fine-tuned LLM model for all 55 test patients.

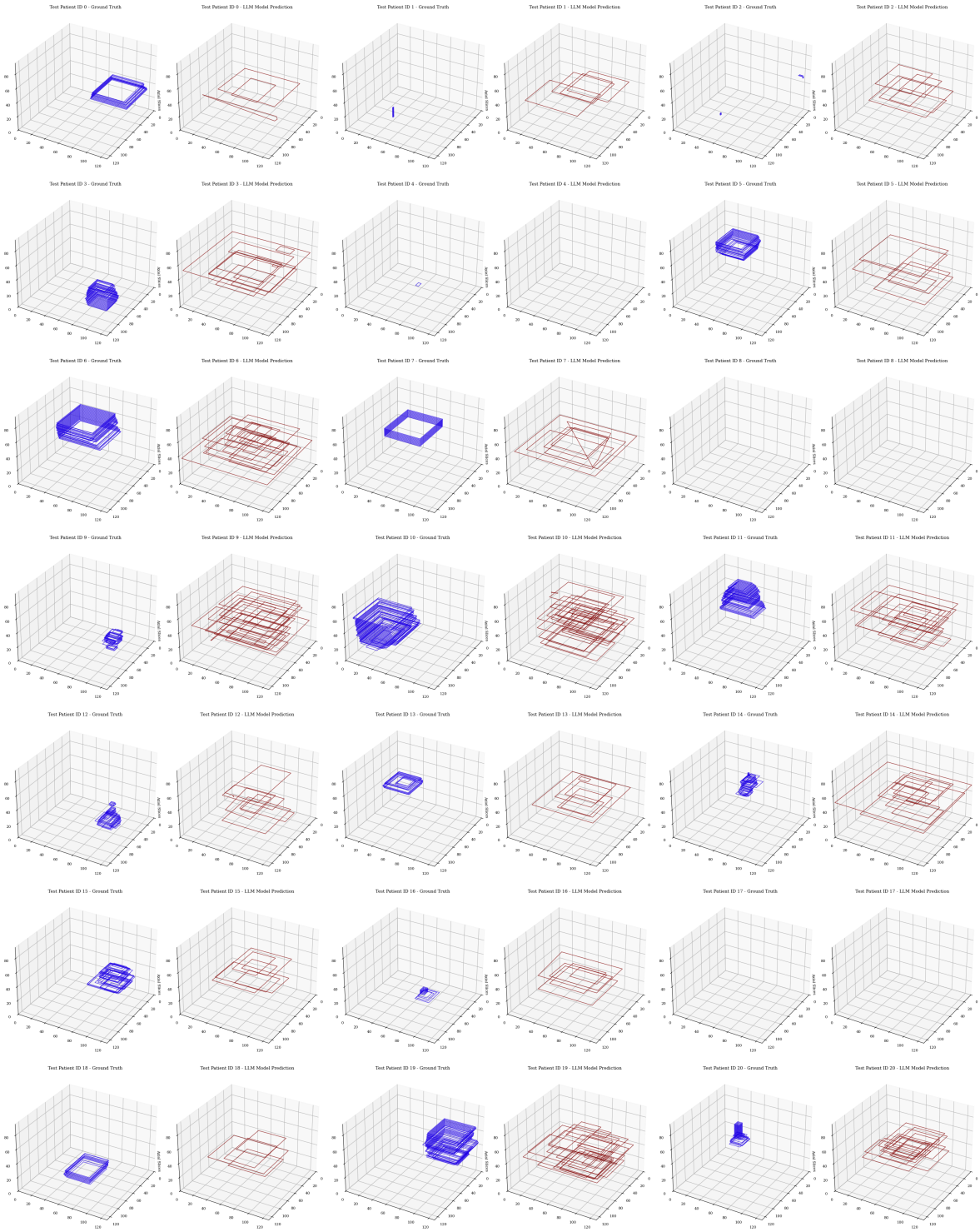


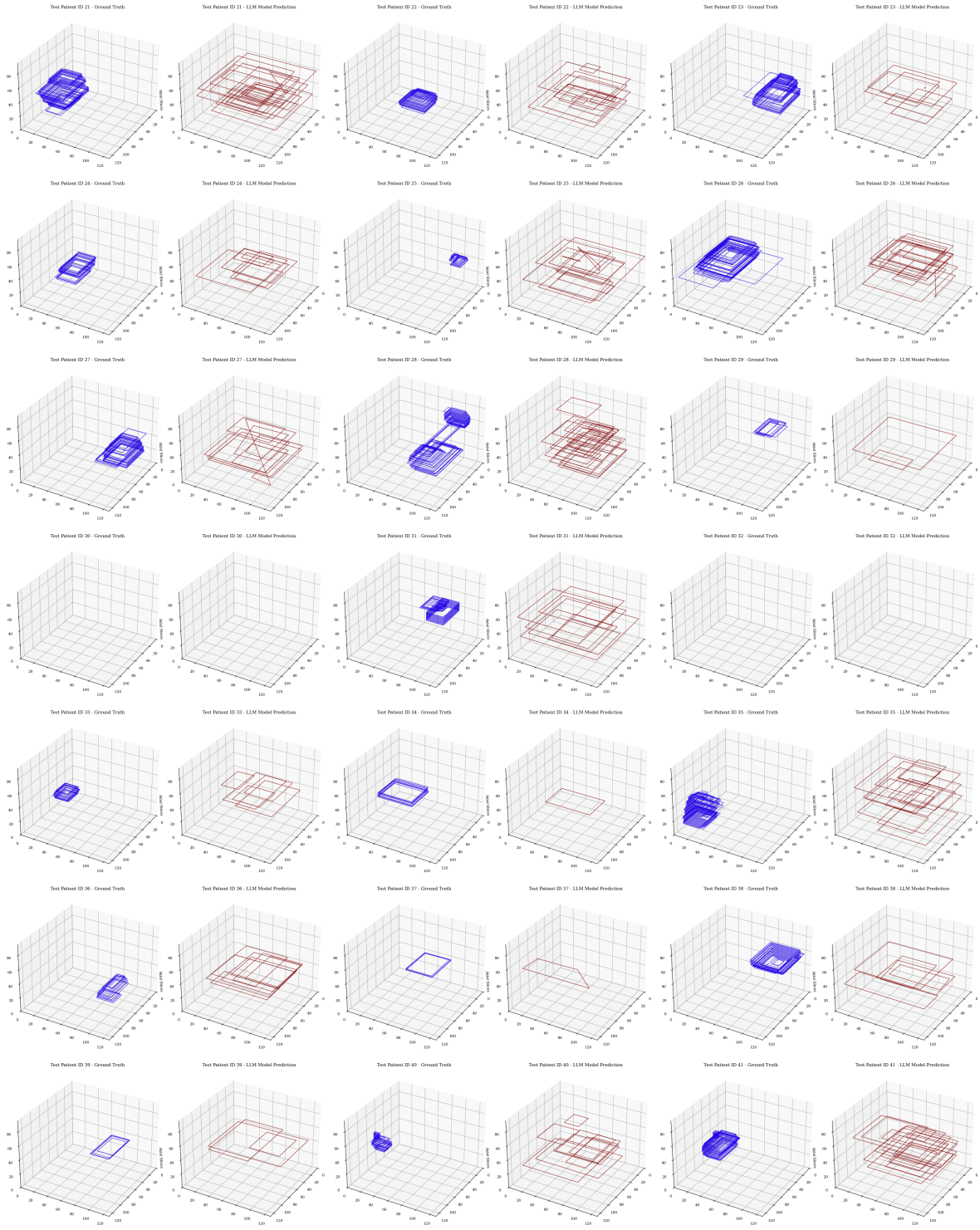

Test Patient ID 21 - GT(blue) MP(red)
Test Patient ID 22 - GT(blue) MP(red)
Test Patient ID 23 - GT(blue) MP(red)
Test Patient ID 24 - GT(blue) MP(red)
Test Patient ID 25 - GT(blue) MP(red)
Test Patient ID 26 - GT(blue) MP(red)
Test Patient ID 27 - GT(blue) MP(red)
Test Patient ID 28 - GT(blue) MP(red)
Test Patient ID 29 - GT(blue) MP(red)
Test Patient ID 30 - GT(blue) MP(red)
Test Patient ID 31 - GT(blue) MP(red)
Test Patient ID 32 - GT(blue) MP(red)
Test Patient ID 33 - GT(blue) MP(red)
Test Patient ID 34 - GT(blue) MP(red)
Test Patient ID 35 - GT(blue) MP(red)
Test Patient ID 36 - GT(blue) MP(red)
Test Patient ID 37 - GT(blue) MP(red)
Test Patient ID 38 - GT(blue) MP(red)
Test Patient ID 39 - GT(blue) MP(red)
Test Patient ID 40 - GT(blue) MP(red)
Test Patient ID 41 - GT(blue) MP(red)

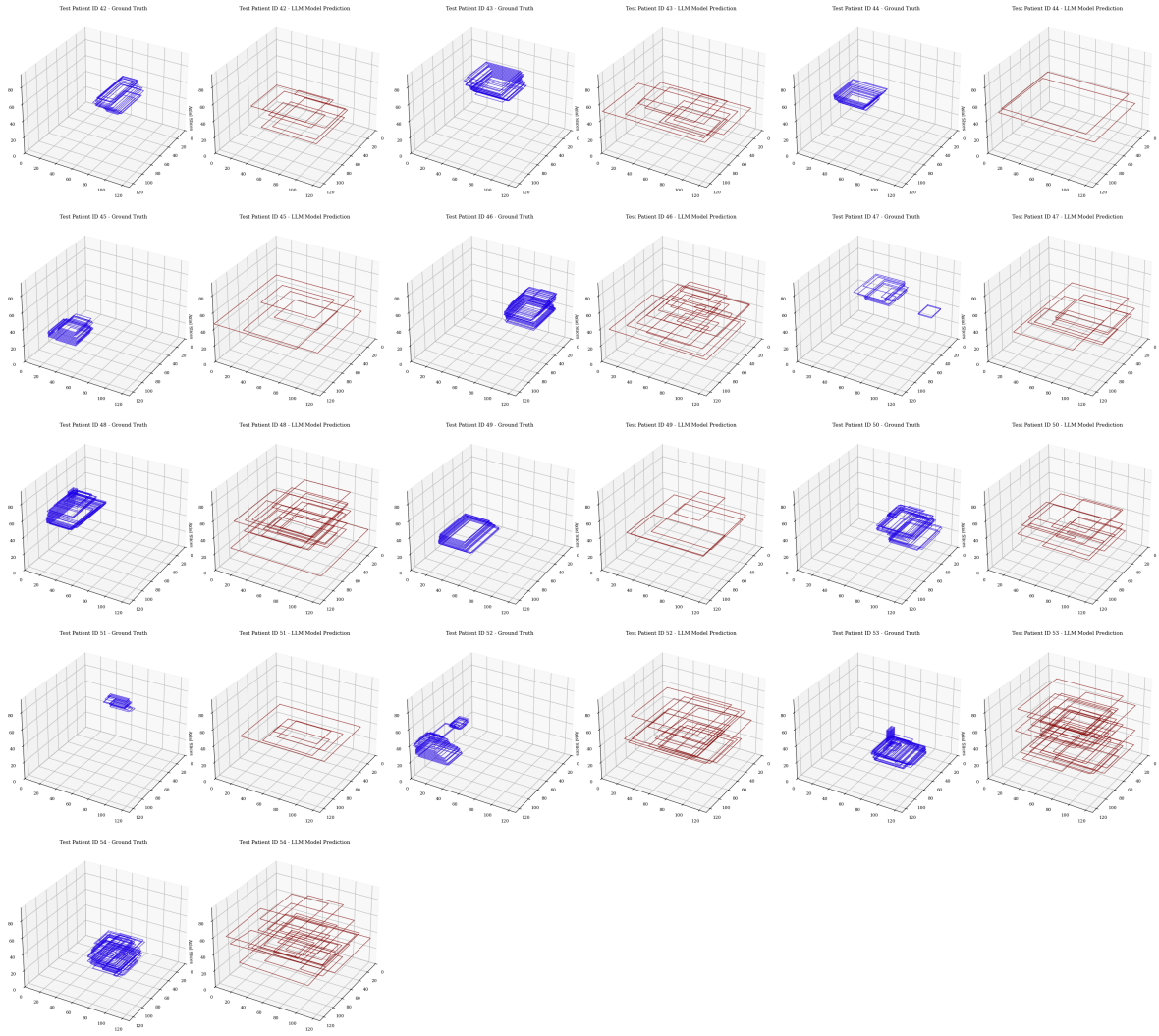

Test Patient ID 42 - GT(blue) MP(red)
Test Patient ID 43 - GT(blue) MP(red)
Test Patient ID 44 - GT(blue) MP(red)
Test Patient ID 45 - GT(blue) MP(red)
Test Patient ID 46 - GT(blue) MP(red)
Test Patient ID 47 - GT(blue) MP(red)
Test Patient ID 48 - GT(blue) MP(red)
Test Patient ID 49 - GT(blue) MP(red)
Test Patient ID 50 - GT(blue) MP(red)
Test Patient ID 51 - GT(blue) MP(red)
Test Patient ID 52 - GT(blue) MP(red)
Test Patient ID 53 - GT(blue) MP(red)
Test Patient ID 54 - GT(blue) MP(red)

F General LLM Bounding Box Segmentation

This section presents the bounding box segmentation visualizations generated by the general LLM model for all 55 test patients.

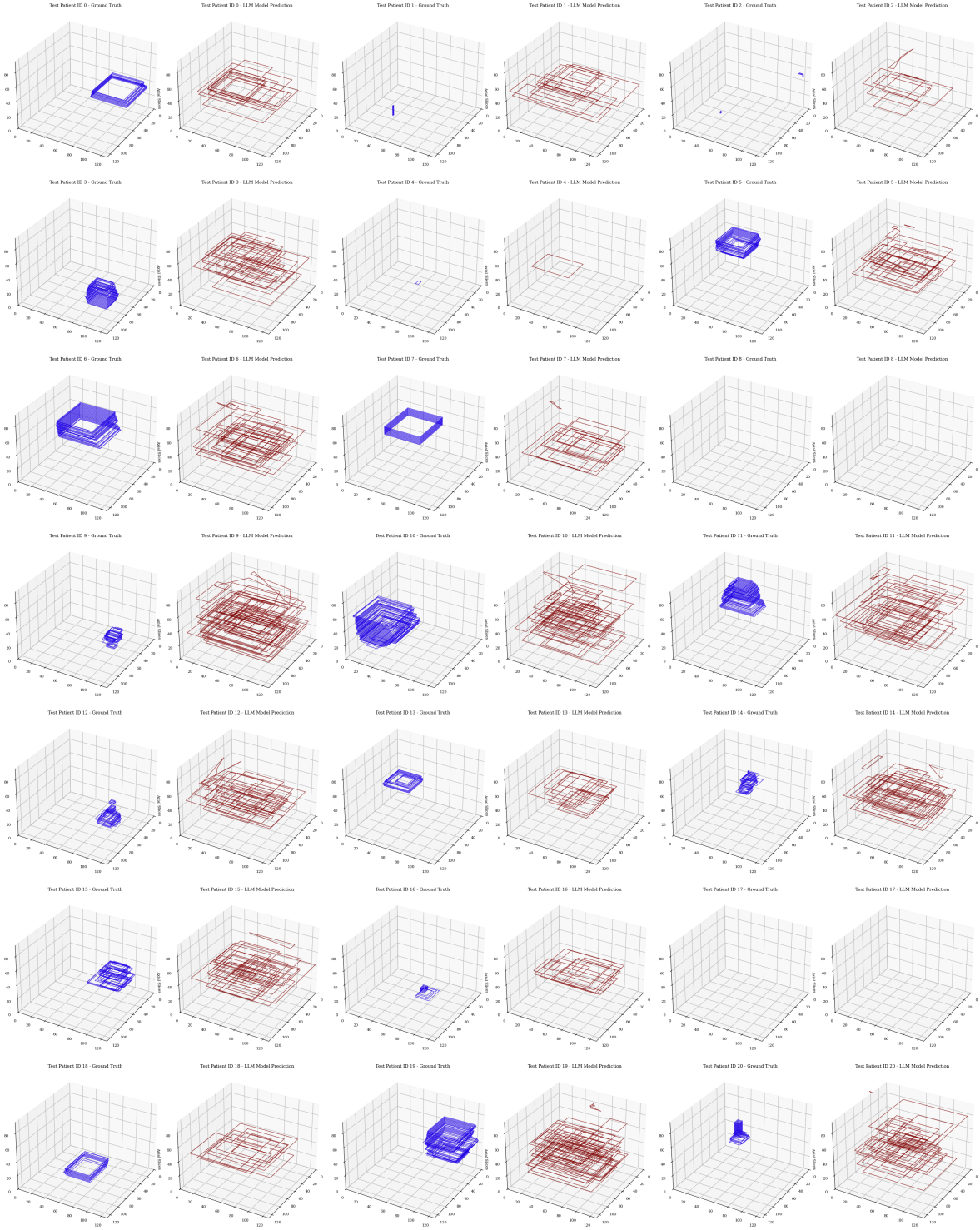


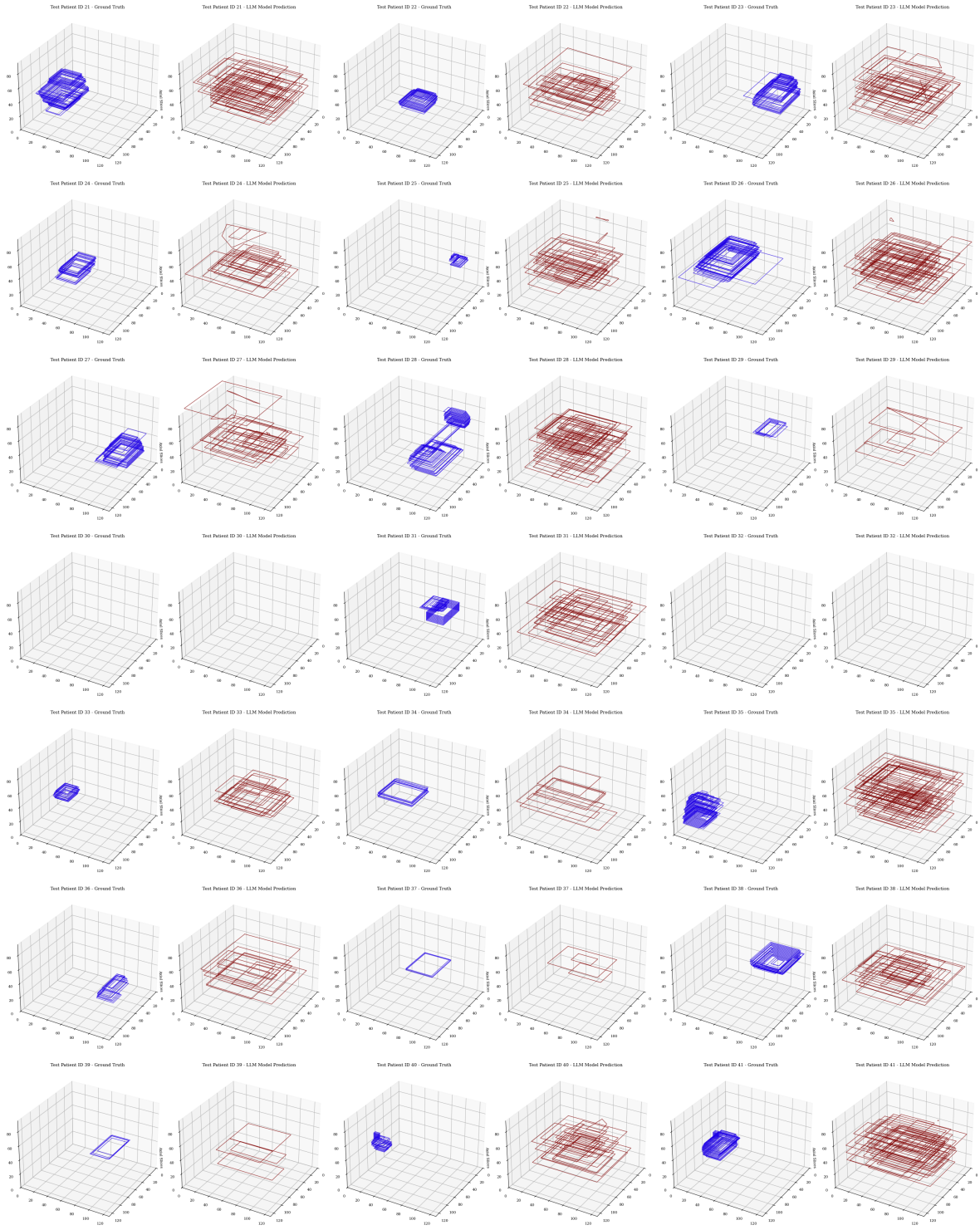

Test Patient ID 21 - Ground Truth
Test Patient ID 21 - LLM Model Prediction
Test Patient ID 22 - Ground Truth
Test Patient ID 22 - LLM Model Prediction
Test Patient ID 23 - Ground Truth
Test Patient ID 23 - LLM Model Prediction
Test Patient ID 24 - Ground Truth
Test Patient ID 24 - LLM Model Prediction
Test Patient ID 25 - Ground Truth
Test Patient ID 25 - LLM Model Prediction
Test Patient ID 26 - Ground Truth
Test Patient ID 26 - LLM Model Prediction
Test Patient ID 27 - Ground Truth
Test Patient ID 27 - LLM Model Prediction
Test Patient ID 28 - Ground Truth
Test Patient ID 28 - LLM Model Prediction
Test Patient ID 29 - Ground Truth
Test Patient ID 29 - LLM Model Prediction
Test Patient ID 30 - Ground Truth
Test Patient ID 30 - LLM Model Prediction
Test Patient ID 31 - Ground Truth
Test Patient ID 31 - LLM Model Prediction
Test Patient ID 32 - Ground Truth
Test Patient ID 32 - LLM Model Prediction
Test Patient ID 33 - Ground Truth
Test Patient ID 33 - LLM Model Prediction
Test Patient ID 34 - Ground Truth
Test Patient ID 34 - LLM Model Prediction
Test Patient ID 35 - Ground Truth
Test Patient ID 35 - LLM Model Prediction
Test Patient ID 36 - Ground Truth
Test Patient ID 36 - LLM Model Prediction
Test Patient ID 37 - Ground Truth
Test Patient ID 37 - LLM Model Prediction
Test Patient ID 38 - Ground Truth
Test Patient ID 38 - LLM Model Prediction
Test Patient ID 39 - Ground Truth
Test Patient ID 39 - LLM Model Prediction
Test Patient ID 40 - Ground Truth
Test Patient ID 40 - LLM Model Prediction
Test Patient ID 41 - Ground Truth
Test Patient ID 41 - LLM Model Prediction

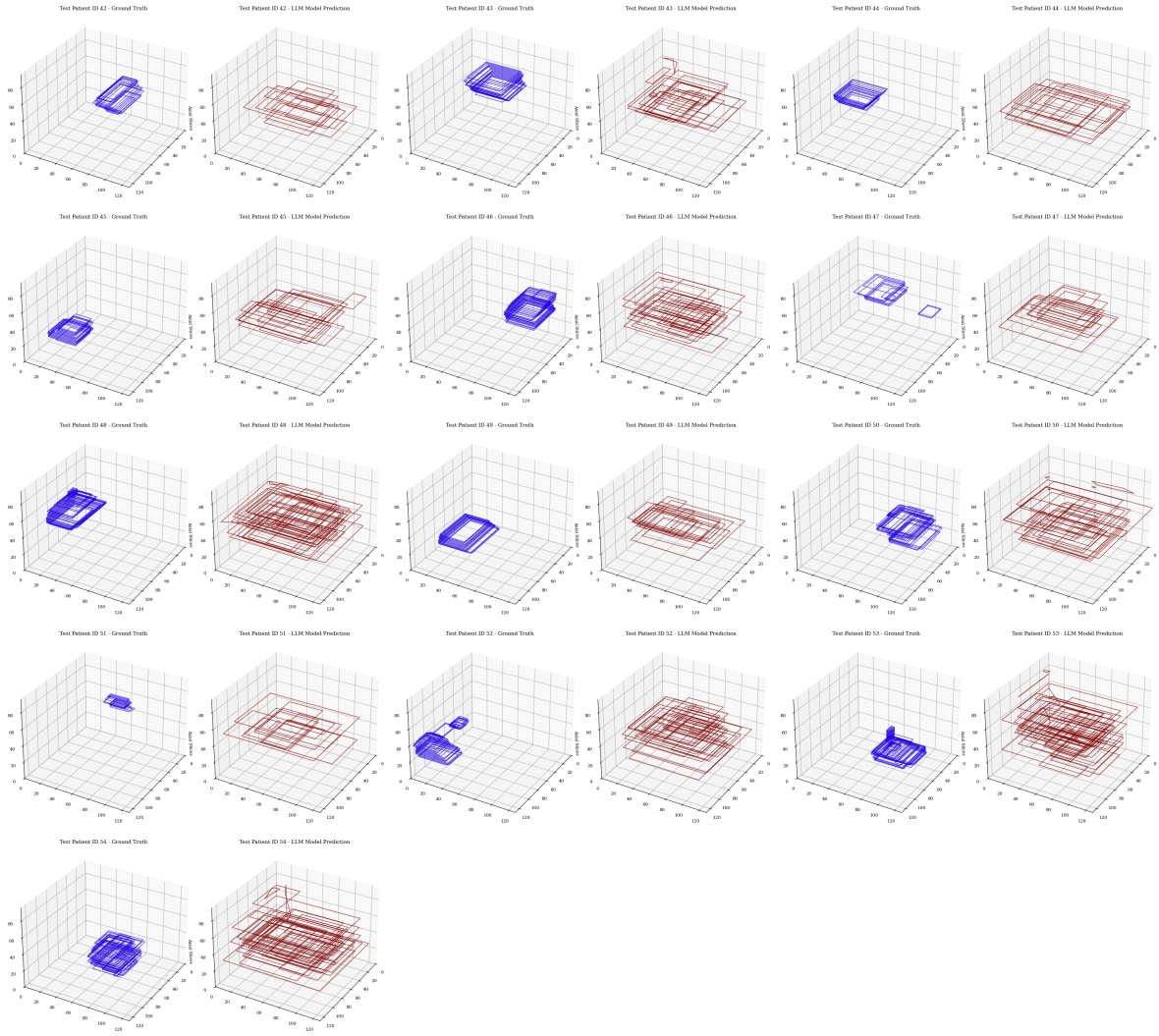

Test Patient ID 42 - Ground Truth
Test Patient ID 42 - LLM Model Prediction
Test Patient ID 43 - Ground Truth
Test Patient ID 43 - LLM Model Prediction
Test Patient ID 44 - Ground Truth
Test Patient ID 44 - LLM Model Prediction
Test Patient ID 45 - Ground Truth
Test Patient ID 45 - LLM Model Prediction
Test Patient ID 46 - Ground Truth
Test Patient ID 46 - LLM Model Prediction
Test Patient ID 47 - Ground Truth
Test Patient ID 47 - LLM Model Prediction
Test Patient ID 48 - Ground Truth
Test Patient ID 48 - LLM Model Prediction
Test Patient ID 49 - Ground Truth
Test Patient ID 49 - LLM Model Prediction
Test Patient ID 50 - Ground Truth
Test Patient ID 50 - LLM Model Prediction
Test Patient ID 51 - Ground Truth
Test Patient ID 51 - LLM Model Prediction
Test Patient ID 52 - Ground Truth
Test Patient ID 52 - LLM Model Prediction
Test Patient ID 53 - Ground Truth
Test Patient ID 53 - LLM Model Prediction
Test Patient ID 54 - Ground Truth
Test Patient ID 54 - LLM Model Prediction

G Fine-Tuned LLM Bounding Box Segmentation

This section presents the bounding box segmentation visualizations generated by the fine-tuned LLM model for all 55 test patients.

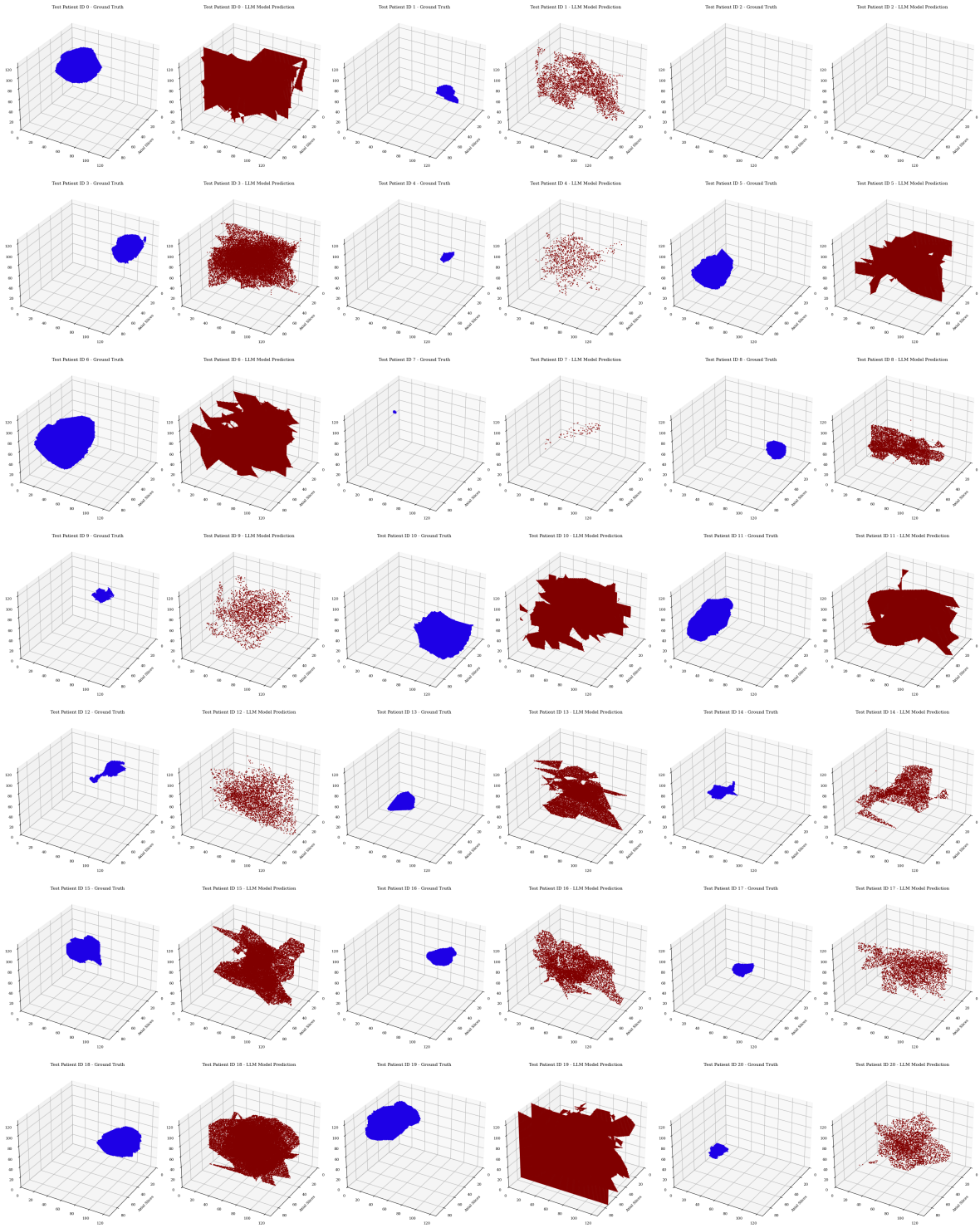


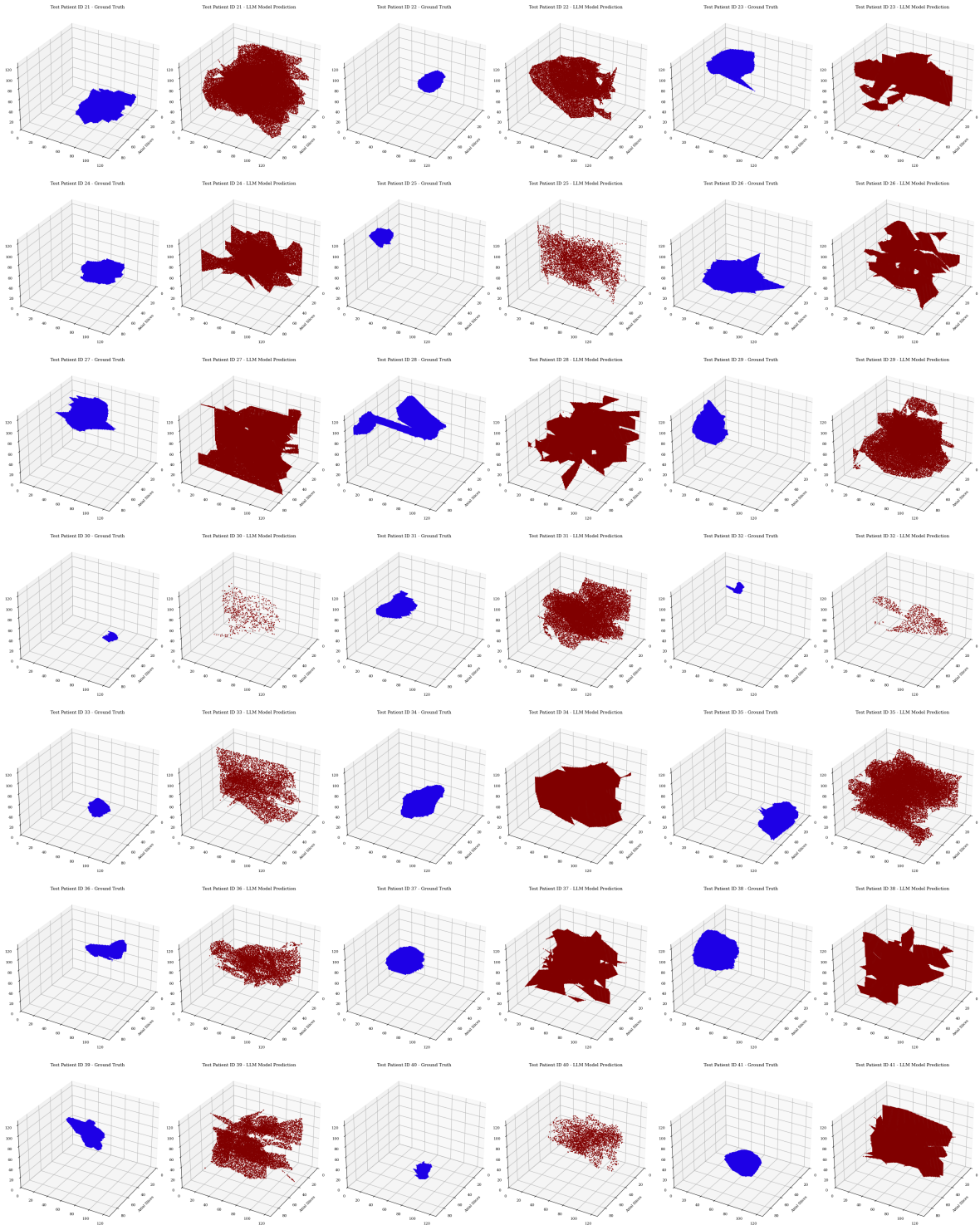

Test Patient ID 21 - Ground Truth
Test Patient ID 21 - LLM Model Prediction
Test Patient ID 22 - Ground Truth
Test Patient ID 22 - LLM Model Prediction
Test Patient ID 23 - Ground Truth
Test Patient ID 23 - LLM Model Prediction
Test Patient ID 24 - Ground Truth
Test Patient ID 24 - LLM Model Prediction
Test Patient ID 25 - Ground Truth
Test Patient ID 25 - LLM Model Prediction
Test Patient ID 26 - Ground Truth
Test Patient ID 26 - LLM Model Prediction
Test Patient ID 27 - Ground Truth
Test Patient ID 27 - LLM Model Prediction
Test Patient ID 28 - Ground Truth
Test Patient ID 28 - LLM Model Prediction
Test Patient ID 29 - Ground Truth
Test Patient ID 29 - LLM Model Prediction
Test Patient ID 30 - Ground Truth
Test Patient ID 30 - LLM Model Prediction
Test Patient ID 31 - Ground Truth
Test Patient ID 31 - LLM Model Prediction
Test Patient ID 32 - Ground Truth
Test Patient ID 32 - LLM Model Prediction
Test Patient ID 33 - Ground Truth
Test Patient ID 33 - LLM Model Prediction
Test Patient ID 34 - Ground Truth
Test Patient ID 34 - LLM Model Prediction
Test Patient ID 35 - Ground Truth
Test Patient ID 35 - LLM Model Prediction
Test Patient ID 36 - Ground Truth
Test Patient ID 36 - LLM Model Prediction
Test Patient ID 37 - Ground Truth
Test Patient ID 37 - LLM Model Prediction
Test Patient ID 38 - Ground Truth
Test Patient ID 38 - LLM Model Prediction
Test Patient ID 39 - Ground Truth
Test Patient ID 39 - LLM Model Prediction
Test Patient ID 40 - Ground Truth
Test Patient ID 40 - LLM Model Prediction
Test Patient ID 41 - Ground Truth
Test Patient ID 41 - LLM Model Prediction

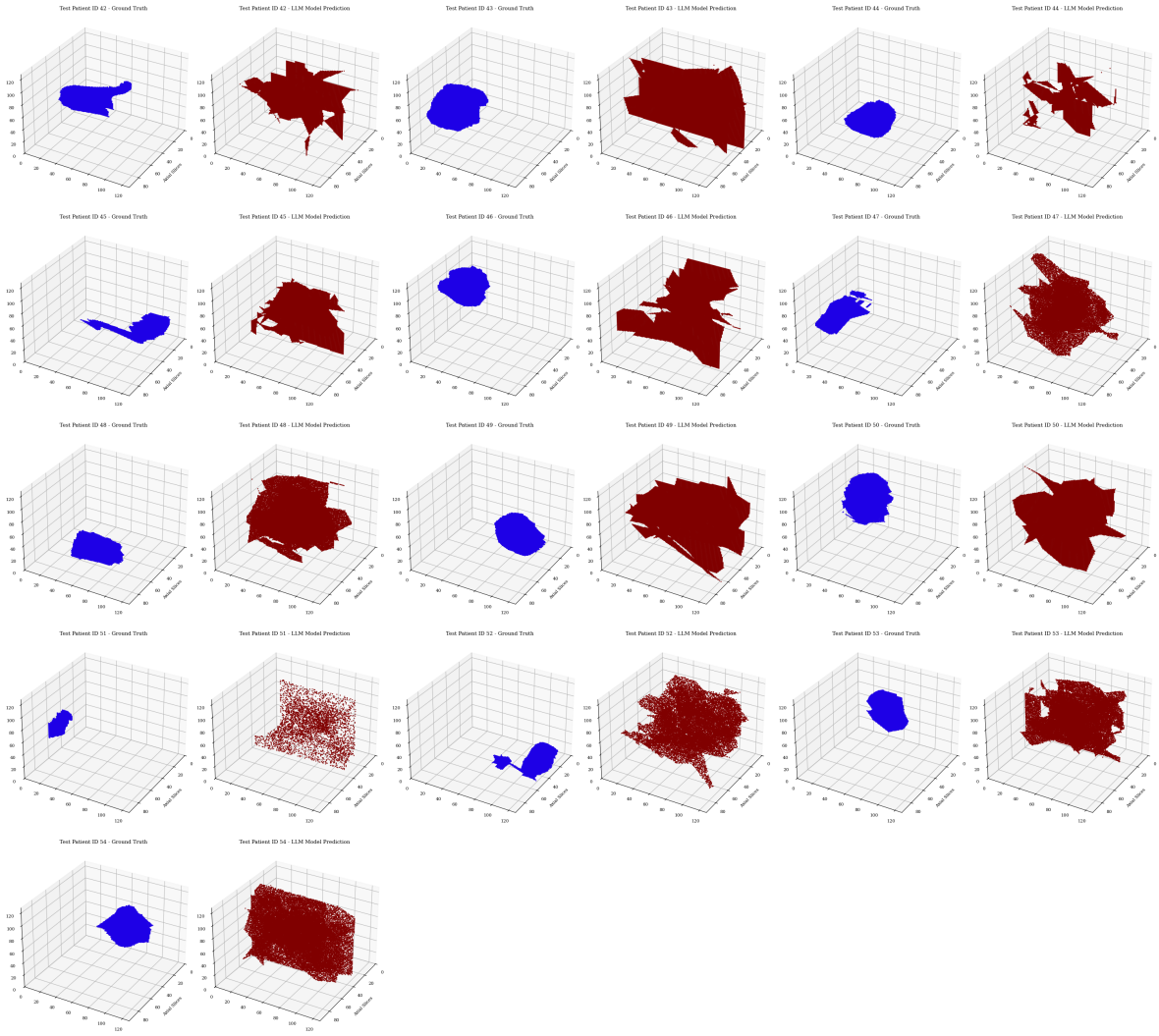

Test Patient ID 42 - Ground Truth
Test Patient ID 42 - LLM Model Prediction
Test Patient ID 43 - Ground Truth
Test Patient ID 43 - LLM Model Prediction
Test Patient ID 44 - Ground Truth
Test Patient ID 44 - LLM Model Prediction
Test Patient ID 45 - Ground Truth
Test Patient ID 45 - LLM Model Prediction
Test Patient ID 46 - Ground Truth
Test Patient ID 46 - LLM Model Prediction
Test Patient ID 47 - Ground Truth
Test Patient ID 47 - LLM Model Prediction
Test Patient ID 48 - Ground Truth
Test Patient ID 48 - LLM Model Prediction
Test Patient ID 49 - Ground Truth
Test Patient ID 49 - LLM Model Prediction
Test Patient ID 50 - Ground Truth
Test Patient ID 50 - LLM Model Prediction
Test Patient ID 51 - Ground Truth
Test Patient ID 51 - LLM Model Prediction
Test Patient ID 52 - Ground Truth
Test Patient ID 52 - LLM Model Prediction
Test Patient ID 53 - Ground Truth
Test Patient ID 53 - LLM Model Prediction
Test Patient ID 54 - Ground Truth
Test Patient ID 54 - LLM Model Prediction

H General LLM Polygon Segmentation

This section presents the bounding polygon segmentation visualizations generated by the general LLM model for all 55 test patients.

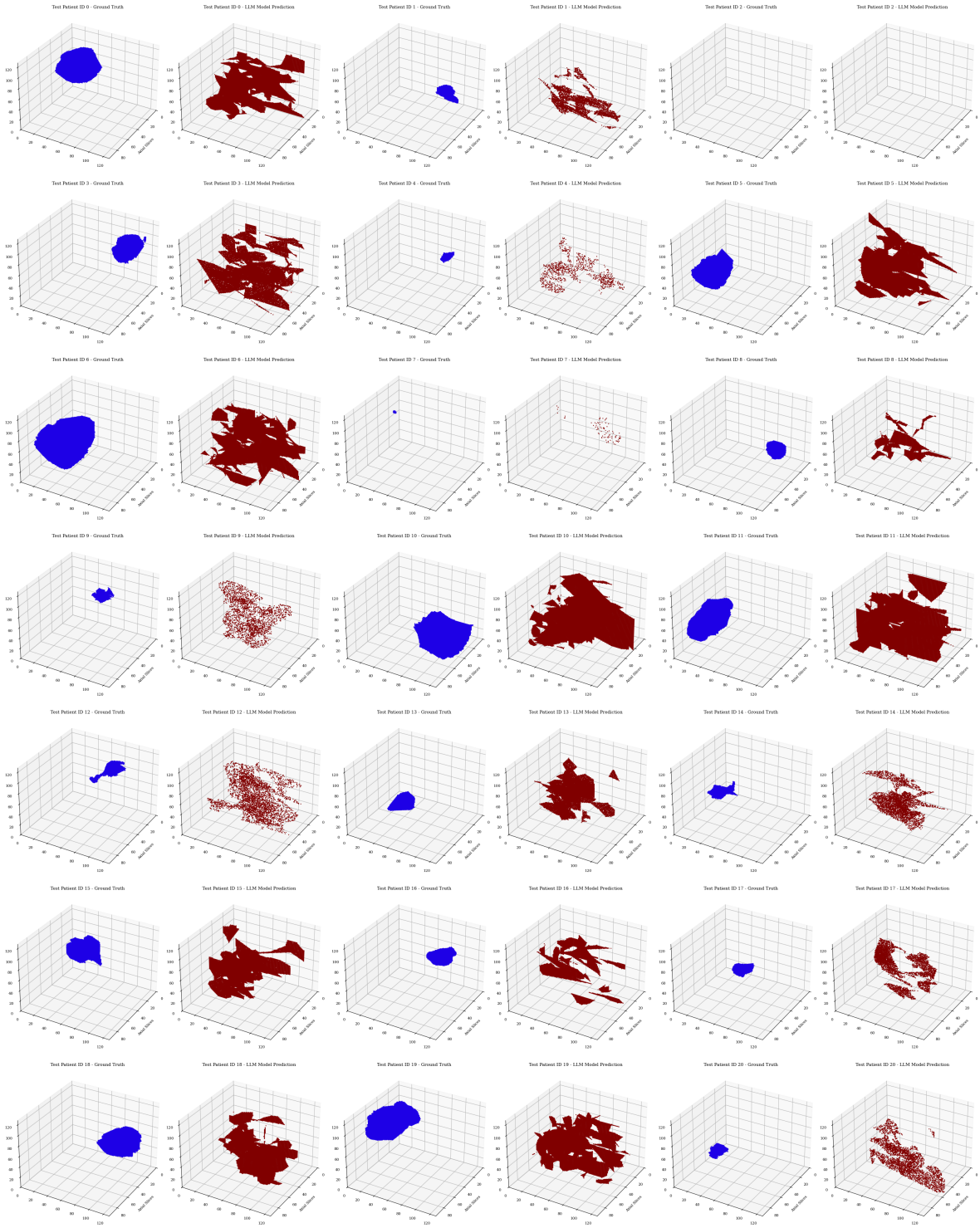


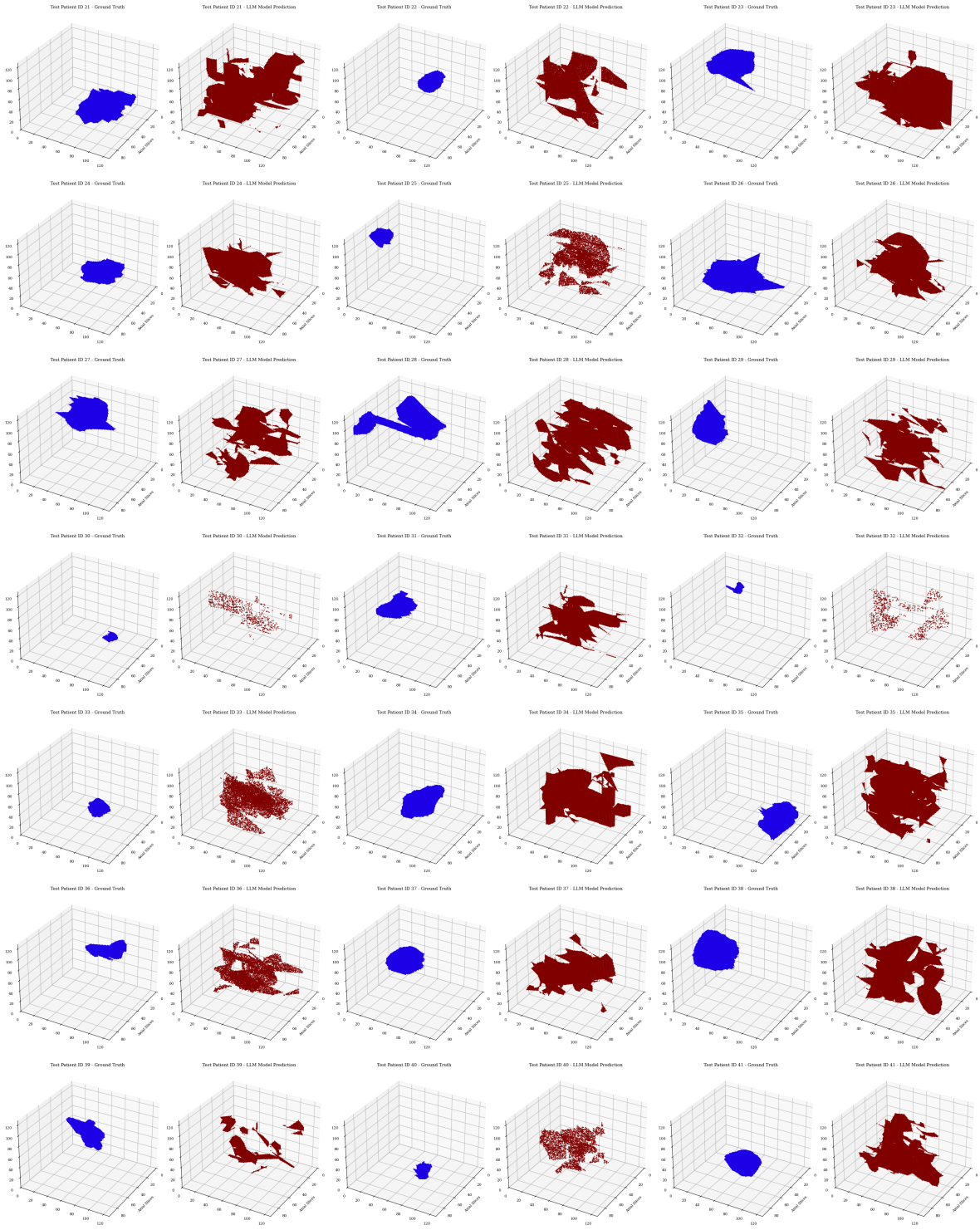

Test Patient ID 21 - Ground Truth
Test Patient ID 21 - LLM Model Prediction
Test Patient ID 22 - Ground Truth
Test Patient ID 22 - LLM Model Prediction
Test Patient ID 23 - Ground Truth
Test Patient ID 23 - LLM Model Prediction
Test Patient ID 24 - Ground Truth
Test Patient ID 24 - LLM Model Prediction
Test Patient ID 25 - Ground Truth
Test Patient ID 25 - LLM Model Prediction
Test Patient ID 26 - Ground Truth
Test Patient ID 26 - LLM Model Prediction
Test Patient ID 27 - Ground Truth
Test Patient ID 27 - LLM Model Prediction
Test Patient ID 28 - Ground Truth
Test Patient ID 28 - LLM Model Prediction
Test Patient ID 29 - Ground Truth
Test Patient ID 29 - LLM Model Prediction
Test Patient ID 30 - Ground Truth
Test Patient ID 30 - LLM Model Prediction
Test Patient ID 31 - Ground Truth
Test Patient ID 31 - LLM Model Prediction
Test Patient ID 32 - Ground Truth
Test Patient ID 32 - LLM Model Prediction
Test Patient ID 33 - Ground Truth
Test Patient ID 33 - LLM Model Prediction
Test Patient ID 34 - Ground Truth
Test Patient ID 34 - LLM Model Prediction
Test Patient ID 35 - Ground Truth
Test Patient ID 35 - LLM Model Prediction
Test Patient ID 36 - Ground Truth
Test Patient ID 36 - LLM Model Prediction
Test Patient ID 37 - Ground Truth
Test Patient ID 37 - LLM Model Prediction
Test Patient ID 38 - Ground Truth
Test Patient ID 38 - LLM Model Prediction
Test Patient ID 39 - Ground Truth
Test Patient ID 39 - LLM Model Prediction
Test Patient ID 40 - Ground Truth
Test Patient ID 40 - LLM Model Prediction
Test Patient ID 41 - Ground Truth
Test Patient ID 41 - LLM Model Prediction

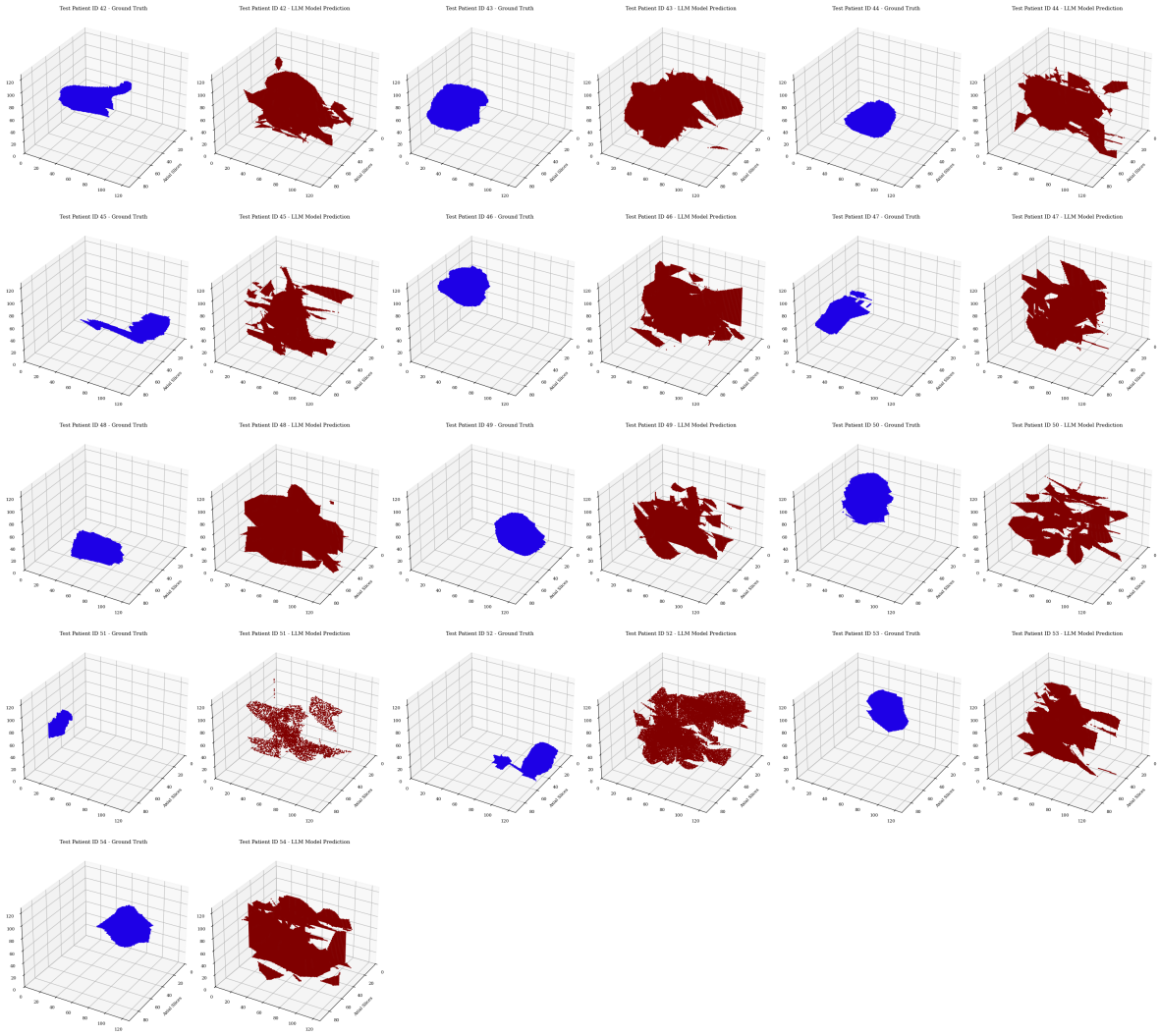

Test Patient ID 42 - Ground Truth
Test Patient ID 42 - LLM Model Prediction
Test Patient ID 43 - Ground Truth
Test Patient ID 43 - LLM Model Prediction
Test Patient ID 44 - Ground Truth
Test Patient ID 44 - LLM Model Prediction
Test Patient ID 45 - Ground Truth
Test Patient ID 45 - LLM Model Prediction
Test Patient ID 46 - Ground Truth
Test Patient ID 46 - LLM Model Prediction
Test Patient ID 47 - Ground Truth
Test Patient ID 47 - LLM Model Prediction
Test Patient ID 48 - Ground Truth
Test Patient ID 48 - LLM Model Prediction
Test Patient ID 49 - Ground Truth
Test Patient ID 49 - LLM Model Prediction
Test Patient ID 50 - Ground Truth
Test Patient ID 50 - LLM Model Prediction
Test Patient ID 51 - Ground Truth
Test Patient ID 51 - LLM Model Prediction
Test Patient ID 52 - Ground Truth
Test Patient ID 52 - LLM Model Prediction
Test Patient ID 53 - Ground Truth
Test Patient ID 53 - LLM Model Prediction
Test Patient ID 54 - Ground Truth
Test Patient ID 54 - LLM Model Prediction

I Fine-Tuned LLM Polygon Segmentation

This section presents the bounding polygon segmentation visualizations generated by the fine-tuned LLM model for all 55 test patients.




Test Patient ID 21 - Ground Truth
Test Patient ID 21 - LLM Model Prediction
Test Patient ID 22 - Ground Truth
Test Patient ID 22 - LLM Model Prediction
Test Patient ID 23 - Ground Truth
Test Patient ID 23 - LLM Model Prediction
Test Patient ID 24 - Ground Truth
Test Patient ID 24 - LLM Model Prediction
Test Patient ID 25 - Ground Truth
Test Patient ID 25 - LLM Model Prediction
Test Patient ID 26 - Ground Truth
Test Patient ID 26 - LLM Model Prediction
Test Patient ID 27 - Ground Truth
Test Patient ID 27 - LLM Model Prediction
Test Patient ID 28 - Ground Truth
Test Patient ID 28 - LLM Model Prediction
Test Patient ID 29 - Ground Truth
Test Patient ID 29 - LLM Model Prediction
Test Patient ID 30 - Ground Truth
Test Patient ID 30 - LLM Model Prediction
Test Patient ID 31 - Ground Truth
Test Patient ID 31 - LLM Model Prediction
Test Patient ID 32 - Ground Truth
Test Patient ID 32 - LLM Model Prediction
Test Patient ID 33 - Ground Truth
Test Patient ID 33 - LLM Model Prediction
Test Patient ID 34 - Ground Truth
Test Patient ID 34 - LLM Model Prediction
Test Patient ID 35 - Ground Truth
Test Patient ID 35 - LLM Model Prediction
Test Patient ID 36 - Ground Truth
Test Patient ID 36 - LLM Model Prediction
Test Patient ID 37 - Ground Truth
Test Patient ID 37 - LLM Model Prediction
Test Patient ID 38 - Ground Truth
Test Patient ID 38 - LLM Model Prediction
Test Patient ID 39 - Ground Truth
Test Patient ID 39 - LLM Model Prediction
Test Patient ID 40 - Ground Truth
Test Patient ID 40 - LLM Model Prediction
Test Patient ID 41 - Ground Truth
Test Patient ID 41 - LLM Model Prediction


Test Patient ID 42 - Ground Truth
Test Patient ID 42 - LLM Model Prediction
Test Patient ID 43 - Ground Truth
Test Patient ID 43 - LLM Model Prediction
Test Patient ID 44 - Ground Truth
Test Patient ID 44 - LLM Model Prediction
Test Patient ID 45 - Ground Truth
Test Patient ID 45 - LLM Model Prediction
Test Patient ID 46 - Ground Truth
Test Patient ID 46 - LLM Model Prediction
Test Patient ID 47 - Ground Truth
Test Patient ID 47 - LLM Model Prediction
Test Patient ID 48 - Ground Truth
Test Patient ID 48 - LLM Model Prediction
Test Patient ID 49 - Ground Truth
Test Patient ID 49 - LLM Model Prediction
Test Patient ID 50 - Ground Truth
Test Patient ID 50 - LLM Model Prediction
Test Patient ID 51 - Ground Truth
Test Patient ID 51 - LLM Model Prediction
Test Patient ID 52 - Ground Truth
Test Patient ID 52 - LLM Model Prediction
Test Patient ID 53 - Ground Truth
Test Patient ID 53 - LLM Model Prediction
Test Patient ID 54 - Ground Truth
Test Patient ID 54 - LLM Model Prediction