** Interim Report **

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

Submitted by

Felicia Liu (1006950042)

Supervised by

Dr. Farzad Khalvati

Jay Yoo

University of Toronto

Faculty of Applied Science and Engineering

January 20, 2025

Contents

List of Figures

List of Tables

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 is 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 is to compare general and subspecialized (fine-tuned) models. We will assess whether general LLMs, such

as Large Language Model Meta AI (Llama) Models, can perform healthcare tasks effectively without fine-tuning, or if specialized training on niche datasets significantly improves performance. The second objective is 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 will contribute to a more informed understanding of when general models suffice and when subspecialized models are 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

Notably, CNNs have become central to many image-based healthcare applications due to their exceptional performance in medical imaging tasks.

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.

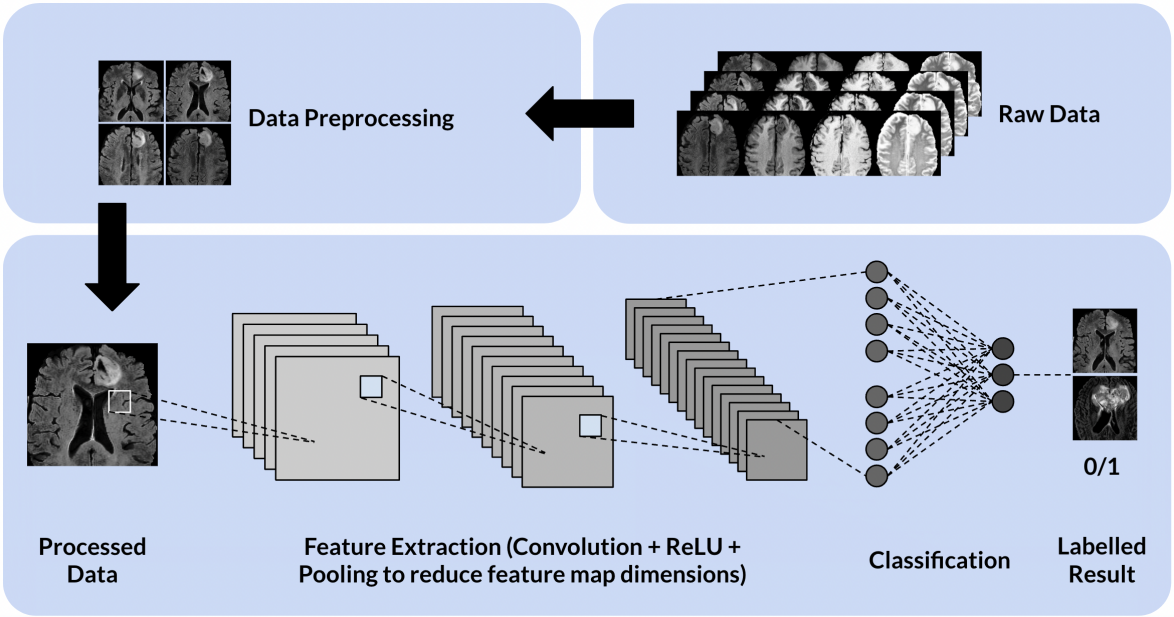


Figure 1: Diagram of a standard CNN pipeline, illustrating the flow from data collection and preprocessing, through convolutional and pooling layers for feature extraction, to a fully connected layer that generates a binary classification output.

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].

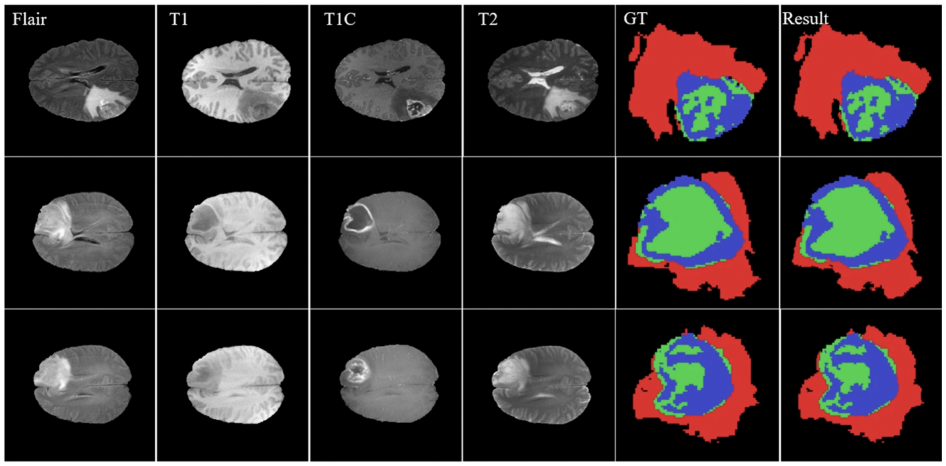


Figure 2: Example of results from brain tumor segmentation using C-CNN enhanced (blue), core (green), and edema (red) regions segmented. Source: [5]

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 and efficacy 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. Moreover, CNNs generally struggle to integrate data from different sources, such as audio or sensor data, which may also be critical for diagnosis.

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.

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 represents visually a standard transformer architecture.

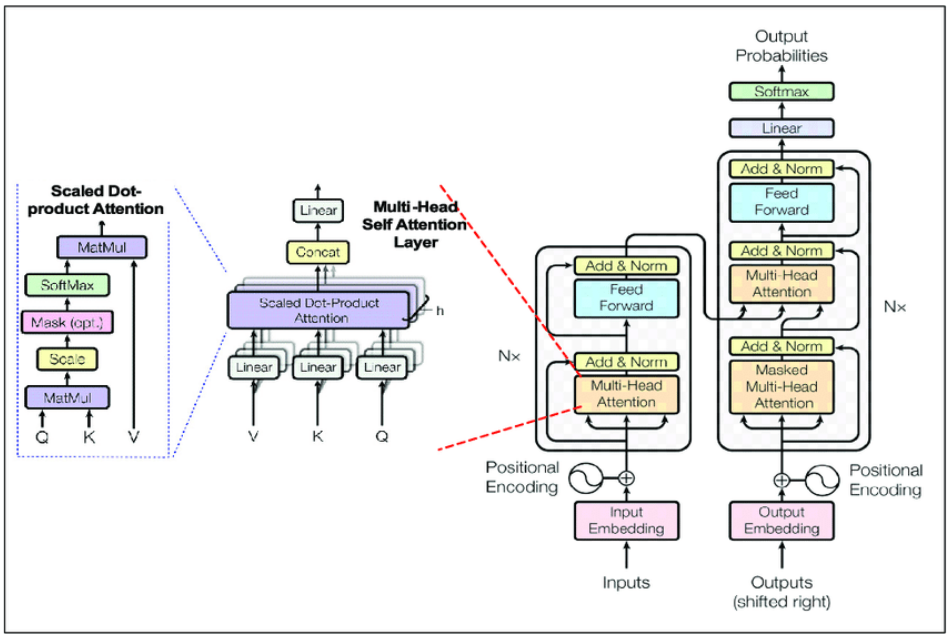


Figure 3: Diagram of a standard transformer architecture, showcasing key components such as multi-head self-attention, positional encoding, and skip connections that enable efficient processing of sequential data. Source: [10]

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 Progress to Date

3.1 Methodology

To thoroughly investigate the capabilities of medical LLMs, we aim to evaluate their performance across a variety of medical tasks. This approach allows for a deeper understanding of how factors such as the choice of general LLM, the type of dataset (particularly non-text data like medical images), and the specific medical task may influence model performance.

The first task in this study focuses on classifying brain gliomas using the Multimodal Brain Tumor Segmentation Challenge 2020 (BraTS2020) dataset. For this evaluation, we compare the performance of the Llama 3.2 Vision Instruct LLM model, both in its out-of-the-box form (called "general LLM") and its subspecialized form fine-tuned with Unsloth (called "fine-tuned LLM), against a custom-built, small-scale 3D CNN, which serves as a baseline due to this deep learning model's proven effectiveness in medical image classification, providing a comparable benchmark.

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 BraTS2020 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 [16]–[20].

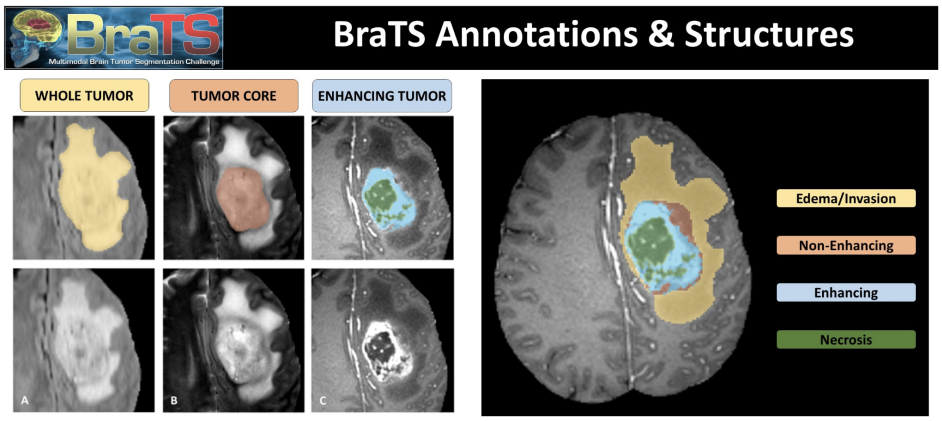


Figure 4: The BraTS dataset provides multi-modal MRI scans with expert-annotated brain tumor segmentations for the classification of gliomas. Source: [16]–[20]

3.1.2 Llama-3.2-11B-Vision-Instruct (General LLM Model)

The Llama 3.2 Vision Instruct model was selected due to its high performance in NLP and multimodal tasks, and 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. The model benefits from large-scale pretraining, allowing it to process medical images effectively. By testing its out-of-the-box performance, we can assess how well a general model handles a specialized medical task (glioma classification) and understand the adaptability of general LLMs to the medical domain [21].

3.1.3 Classification Task

The goal of this task is for the model to classify each patient as either an HGG or LGG patient using their MRI scans and corresponding ground truth labels [16]–[20]. Due to differences in input processing, the CNN and LLM approaches differed. The CNN utilized full 3D convolutions across all four imaging modalities (T1, T1ce, T2, FLAIR), capturing comprehensive spatial information. In contrast, the LLM, limited to 2D inputs, processed individual axial slices from only the FLAIR scans. Predictions for each slice (HGG or LGG) were tallied, and the patient's final classification was determined by majority vote (winner-takes-all) across all slice predictions, shown in Figure 5.

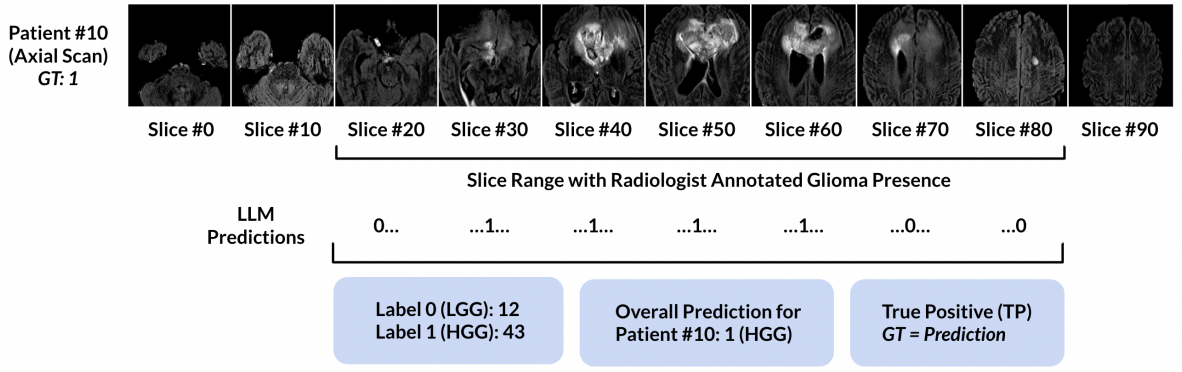


Figure 5: Diagram illustrates how predictions for each patient were made by the LLM. Only FLAIR scans were evaluated.

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 [22].

3.1.5 Unsloth Finetuning

Unsloth was chosen as a fine-tuning tool for its simplified user pipeline and optimized features that enable efficient training of LLMs. It is a lightweight framework designed to accelerate training while reducing memory usage, making it ideal for running on limited hardware, such as Google Colab GPUs. Unsloth integrates with popular packages like Hugging Face Transformers, bitsandbytes (for 8-bit and 4-bit quantization), and Parameter-Efficient Fine-Tuning (PEFT). Its ability to handle fine-tuning on user-specified specialized datasets allows for a customisable approach to adapting the general LLM to medical image classification. Additionally, its efficiency and resource-saving capabilities make it a practical choice for tasks like text classification, generation, and instruction tuning [23].

3.2 Procedure and Implementation

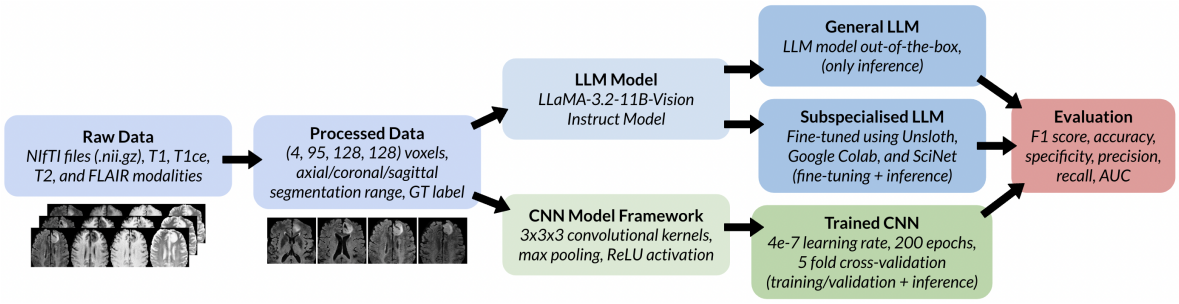


Figure 6: Full methodology pipeline from raw data to evaluation and comparison.

3.2.1 Initial Data Processing

The dataset consists of 365 patients, each with four 128x128 scans, totaling 1460 scans. The scans are processed as integer arrays (0-256 per pixel), normalized, and centered around the glioma using their radiologist annotated segmentation. For the CNN, the scans are resized to a consistent shape of (4, 95, 128, 128). The dataset is split into separate cohorts: 310 training, 62 validation (for each of 5-folds), and 55 test samples (15%). Labels were extracted and assigned as LGG=0 and HGG=1. The segmentation layout includes non-zero values for glioma voxels, with axial, coronal, and sagittal ranges extracted for each patient (same segmentation range for all four imaging modes of a patient). 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 7 and Figure 8.

Figure 7: Visualisation for the 4 modalities of scans (T1, T1ce, T2, FLAIR) with an LGG patient. Source: [16]–[20]

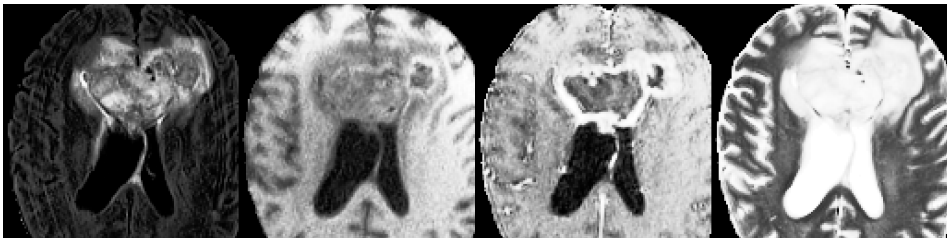


Figure 8: Visualisation for the 4 modalities of scans (T1, T1ce, T2, FLAIR) with an HGG patient. Source: [16]–[20]

3.2.2 CNN 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.

3.2.3 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. Since the LLM was limited to processing 2D axial slices, predictions for each slice were aggregated, and the final patient classification was determined using a majority vote (winner-takes-all) across all slice predictions. The general LLM model prompt is given in Figure 9 and a sample output response from the general LLM is given in Figure 10.

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 9: LLM general model prompt sample.

```
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 10: 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.

3.2.4 LLM Subspecialised Model Fine-tuning

Unsloth Vision Fine-tuning uses parameter-efficient techniques like Low-Rank Adaptation (LoRA) and Quantized Low-Rank Adaptation (QLoRA) to adapt large vision models without updating all parameters. The base model remains frozen while small trainable adapters are updated to learn task-specific features, reducing memory usage and computation. QLoRA further improves efficiency by applying 4-bit quantization to model weights, significantly lowering memory demands while maintaining performance. The bitsandbytes library is essential for enabling low-bit precision operations, such as 8-bit and 4-bit quantization, which drastically reduce memory consumption and accel-

erate computation. It provides optimized matrix multiplication and gradient updates for quantized models, making fine-tuning large models more feasible.

Fine-tuning is currently at a standstill as I'm experiencing issues building bitsandbytes on SciNet's MIST system. SciNet support reported that despite recent bug fixes, they have yet to test this library and the build issue remains unresolved. As a result, Unsloth cannot be downloaded, built, or imported correctly into my SciNet environment. In the meantime, I've shifted to using Google Colab, where Unsloth installs and runs properly. I've formatted the dataset appropriately to be able to be passed in for fine-tuning and the same training and validation cohorts as the CNN are being used to maintain consistency in model training and inference evaluation. Unfortunately, Colab's free version imposes strict limits on runtime, RAM, and GPU access, restricting training to small batch sizes (under 4) and sessions shorter than 3 hours (sometimes shorter). High memory usage often causes runtime errors, preventing me from saving small fine-tuned model weights and being able to evaluate using the full test cohort to check if the LLM is learning. Once the environment issues with SciNet or resource limitations on Colab are resolved, the fine-tuned model will be tested on the designated cohort to compare its performance with the CNN.

3.2.5 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 11, 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.

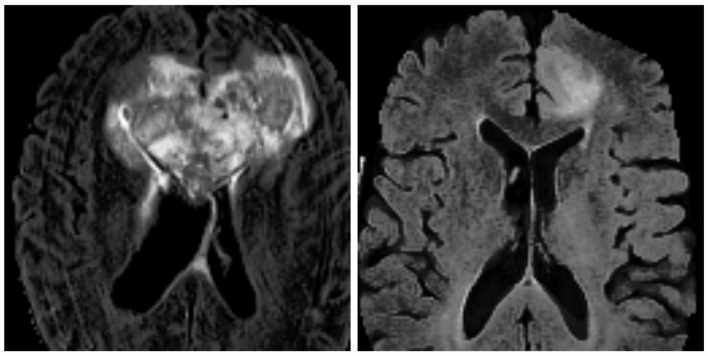


Figure 11: FLAIR scan slices of High-Grade Glioma (HGG, left) and Low-Grade Glioma (LGG, right). These images clearly illustrate the distinct features of each condition, and the LLM is expected to predict them accurately. Source: [16]–[20]

3.2.6 Chosen 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, in contrast, assesses the model's ability to correctly identify negative cases, reflecting how well it avoids false positives. The F1 score, the harmonic mean of precision and recall, provides a balanced measure, 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.

3.3 Results

3.3.1 CNN Classification Performance

The CNN training results are presented below, where various learning rates and epoch counts were tested to identify the most stable training configuration and optimal test set performance. The best-performing model, with a learning rate of 4e-7 and 200 epochs, is used as the baseline and its performance is presented below.

Table 1: Performance metrics for the CNN model (learning rate = 4e-7, epochs = 200) selected as the baseline.

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

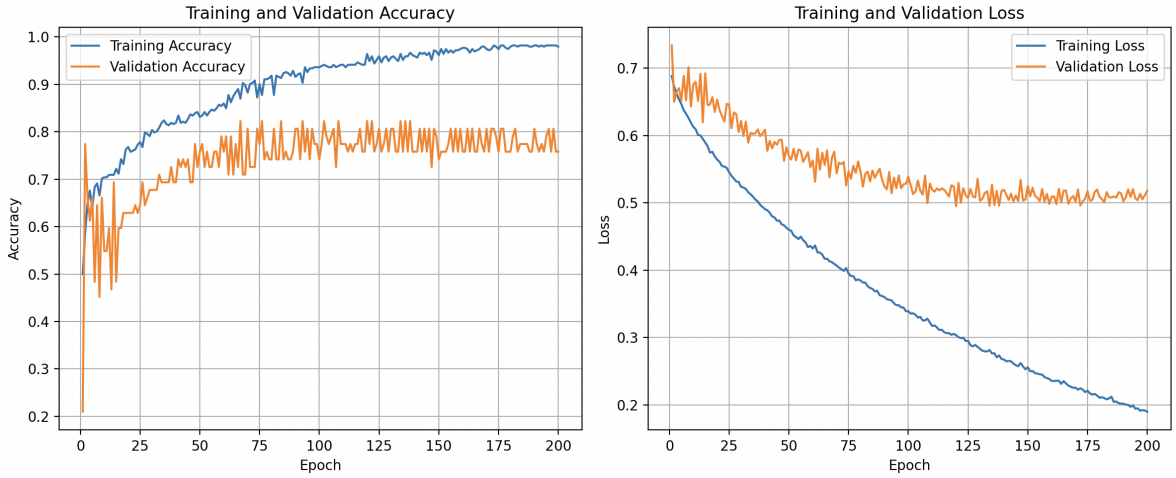


Figure 12: Training and validation accuracy (left) and loss (right) curves for the CNN model (learning rate = 4e-7, epochs = 200) showing stable training with consistent trends. The validation curve remains steady without a rise, indicating no overfitting of the model.

After 200 epochs, the model achieved impressive performance in the training set with an accuracy of 0.9677, F1 score of 0.9792, precision of 1.0000, and recall of 0.9592. 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.

Other models trained are listed below but were not selected due to unstable convergence, with validation accuracy fluctuating too widely. Despite this, they still performed well, achieving AUC scores between 69%-75% and F1 scores between 81%-88%.

Table 2: Performance metrics for the CNN model (learning rate = 1e-4, epochs = 50). selected as the baseline.

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

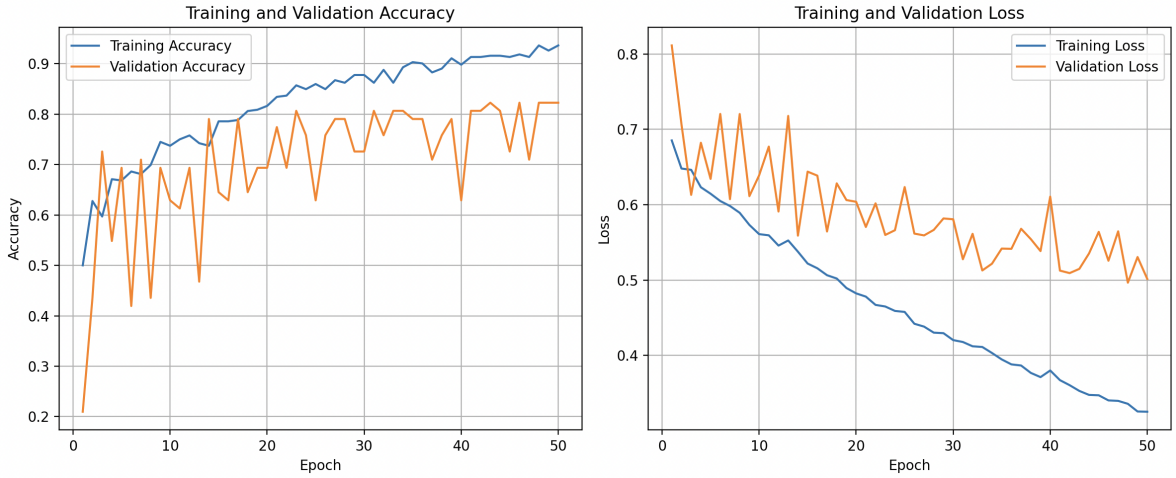


Figure 13: Training and validation accuracy (left) and loss (right) curves for the CNN model (learning rate = 1e-4, epochs = 50).

Table 3: Performance metrics for the CNN model (learning rate = 1e-6, epochs = 80) selected as the baseline.

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

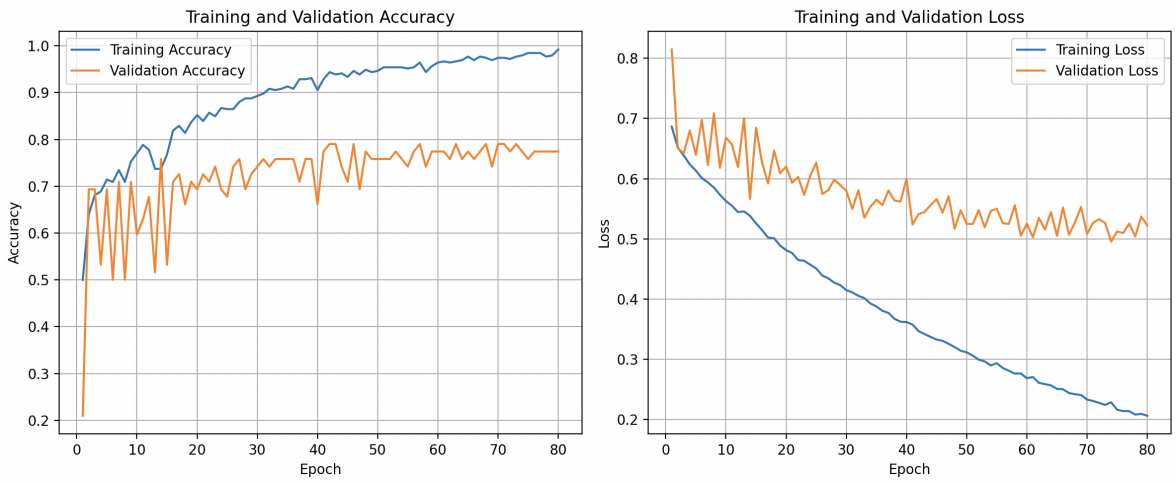


Figure 14: Training and validation accuracy (left) and loss (right) curves for the CNN model (learning rate = 1e-6, epochs = 80).

Table 4: Performance metrics for the CNN model (learning rate = 5e-7, epochs = 100) selected as the baseline.

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

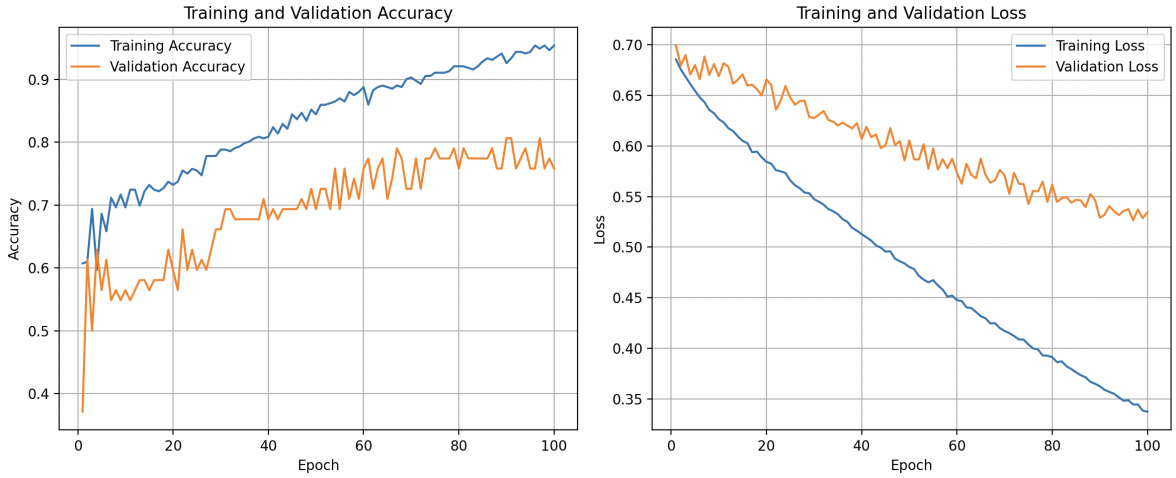


Figure 15: Training and validation accuracy (left) and loss (right) curves for the CNN model (learning rate = 5e-7, epochs = 100).

3.3.2 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 model is expected to improve consistency by adapting to task-specific features, reducing prediction variability for the same input.

3.3.3 LLM 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 the figures, the vertical axis is modelled 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 5: Classification metrics for the general LLM model's performance 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
AUC	N/A	N/A	N/A
Specificity	0.0000	0.0000	0.0909

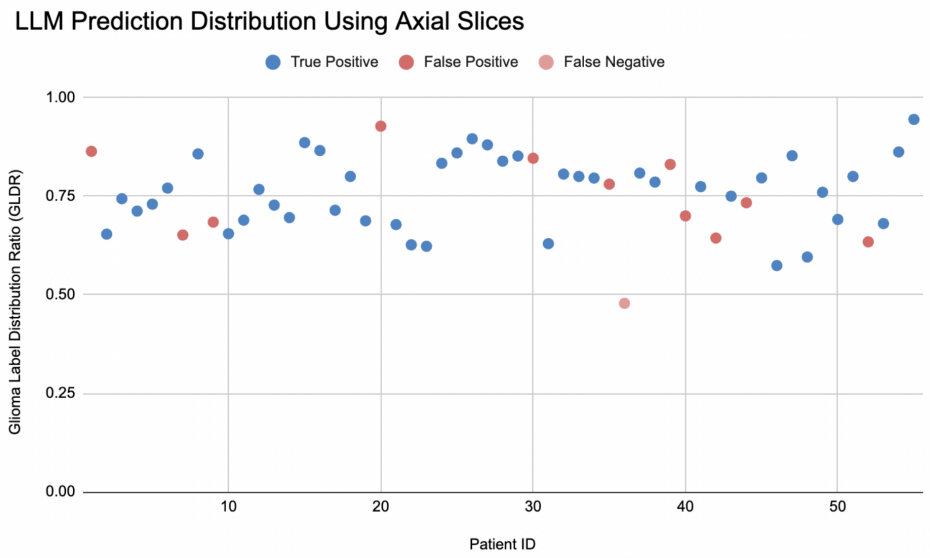


Figure 16: Visual representation of the ratio of HGG to LGG labels across each test patient's axial scan slices. One patient scan was predicted as LGG, which was a False Negative. Note how False Positives are distributed randomly over the HGG prediction range (ratios between 0.5 and 1.0) among the True Positives.

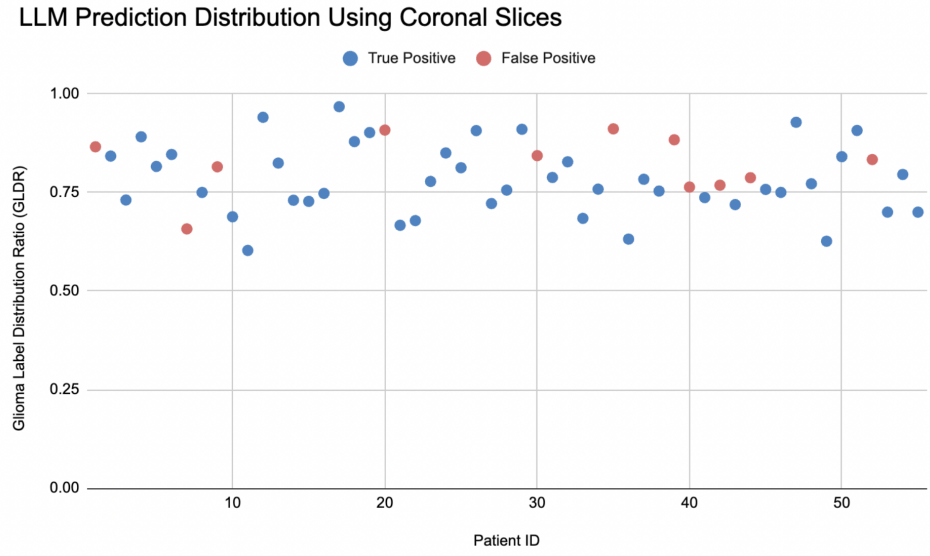


Figure 17: Visual representation of the ratio of HGG to LGG labels across each test patient's coronal scan slices. No patient scans were predicted as LGG, which is incorrect.

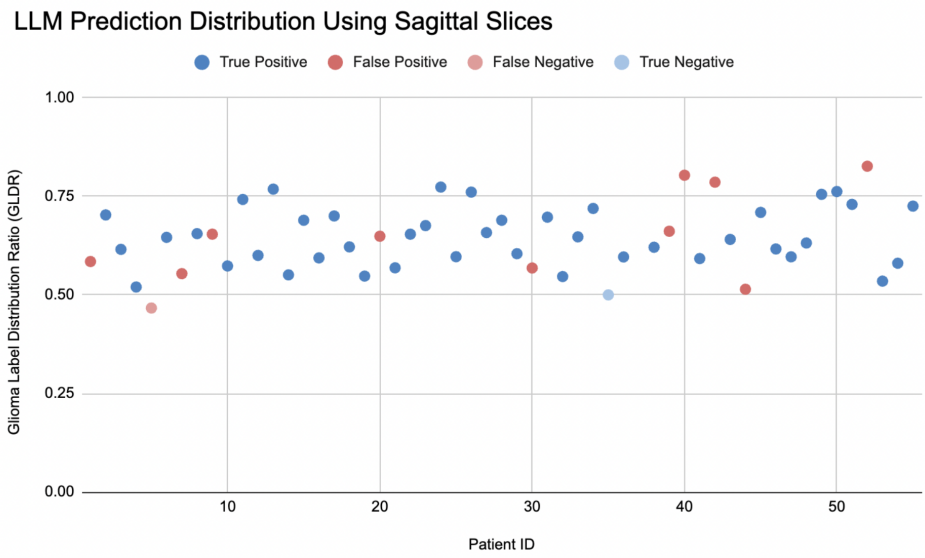


Figure 18: Visual representation of the ratio of HGG to LGG labels across each test patient's sagittal scan slices. Two patient scans were predicted as LGG, one was a False Negative, and the other a True Negative.

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.

3.4 Discussion

3.4.1 Accuracy, Robustness, and Utility Comparison

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.

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 without proper fine-tuning. The CNN's stronger performance in differentiating glioma types underscores the advantage of domain-specific models in handling class imbalance and nuanced medical data. Looking forward, fine-tuning the LLM model will hopefully address its shortcomings by enabling the model to learn glioma-specific features, thereby improving specificity and overall classification performance. This will allow for a direct comparison between the general and fine-tuned LLM, providing deeper insights into the benefits and limitations of adapting general models for specialized healthcare applications.

3.4.2 Limitations and Sources of Bias and Error

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. Addressing this inconsistency is critical in the fine-tuning process as glioma-specific balanced data is expected to improve the LLM's consistency and task-specific accuracy.

3.4.3 Conclusion

In the progress to date, the CNN has demonstrated higher accuracy, robustness, and reliability in classifying glioma types compared to the general-purpose LLM, particularly in handling class imbalances and distinguishing between LGG and HGG gliomas. While the LLM showed potential, its out-of-the-box performance was limited by low specificity and inconsistent predictions. Fine-tuning is expected to address these limitations and unlock the LLM's utility for specialized medical imaging tasks.

4 Future Work

4.1 Next Steps

4.1.1 Fine-tuning a Subspecialised LLM Model Using Unsloth

The primary next step for this exploration is to finalize the fine-tuning of the general LLM model using the Unsloth framework. Due to issues with building the bitsandbytes

library on SciNet's MIST system as outlined in the procedures, I have shifted to using Google Colab, where Unsloth runs successfully, however, runtime, storage, and GPU resource access issues are still hindering active progress. Once either the resource limitations on Colab or the environment issues with SciNet are resolved, the fine-tuned model will be tested on the designated cohort to compare its performance with the CNN. During fine-tuning, hyperparameters such as learning rate, batch size, and the number of training epochs will be adjusted to identify the optimal configuration and improve accuracy and generalization in glioma detection for the LLM. If further technical challenges arise with the Unsloth framework, alternative fine-tuning methods or domain-specific pre-trained models may be explored instead.

4.1.2 Performing More Consistency Tests

The consistency test provided valuable insight into the general-purpose LLM's prediction instability, particularly its tendency to misclassify LGG cases as HGG. While this pattern appears consistent with the overall performance results, expanding the test with more LGG and HGG samples will help confirm whether this trend holds across a broader set of inputs. The same consistency test will also eventually be conducted on the fine-tuned model to evaluate whether fine-tuning improves prediction stability.

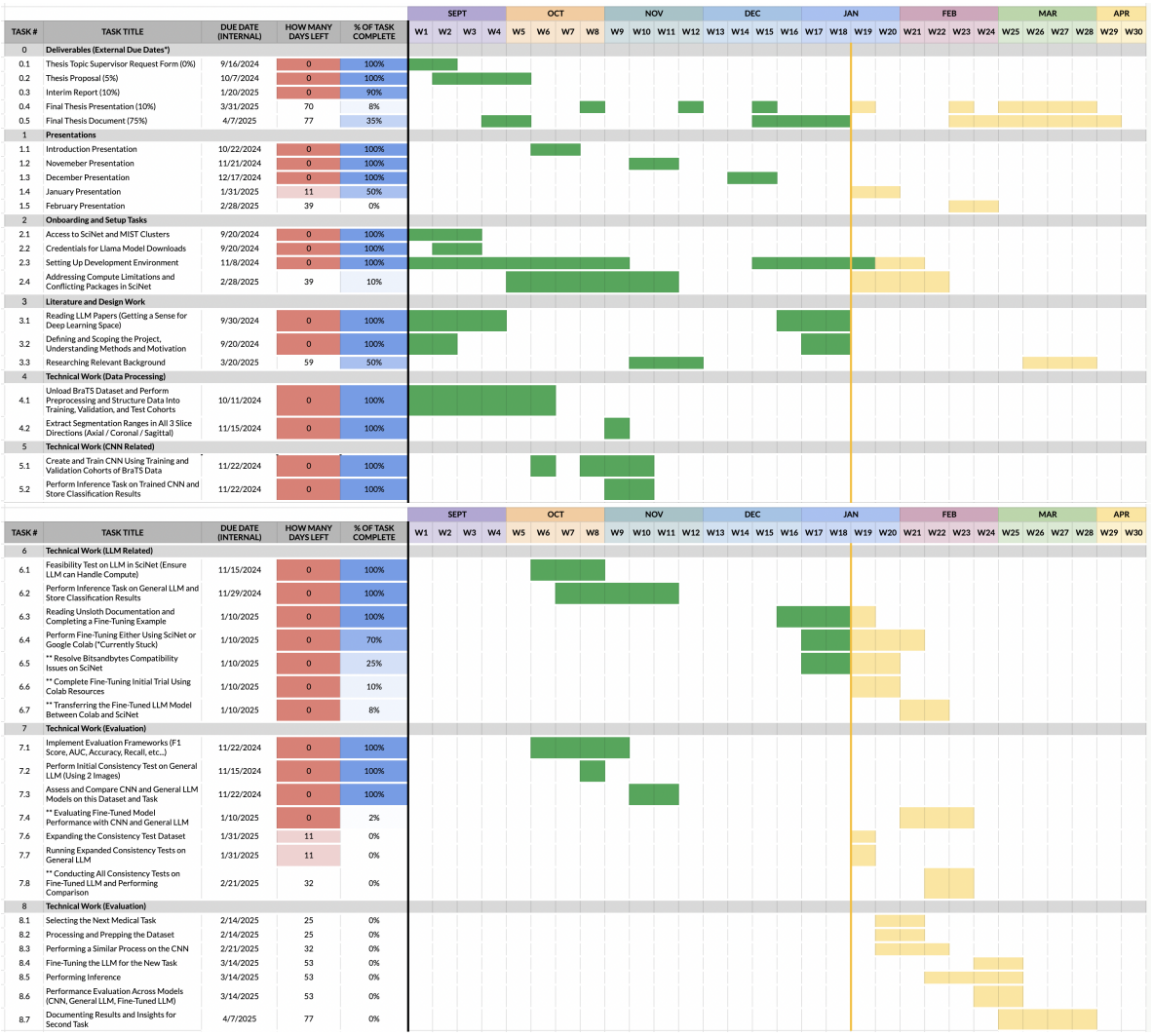
4.1.3 Selecting the Following Medical Task and Dataset

Once the comparison between the CNN, general LLM, and fine-tuned LLM for glioma classification is complete, the next step is to explore a second medical task. This could involve using the BraTS 2021 dataset for a segmentation task, for example, or switching to an entirely different dataset, such as MIMIC-CXR, the largest open-source dataset of chest X-rays paired with radiology reports [24]. 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.

4.2 Timeline

4.2.1 Gantt Chart

The Gantt Chart presented in Figure 19 summarizes my current relevant milestones and concisely presents a timeline for my future work. The yellow vertical line marks the current point in time. Completed tasks are marked as green squares and future tasks are marked as yellow squares.



TASK #	TASK TITLE	DUE DATE (INTERNAL)	HOW MANY DAYS LEFT	% OF TASK COMPLETE
0	**Deliverables (External Due Dates*)**			
0.1	Thesis Topic Supervisor Request Form (0%)	9/16/2024	0	100%
0.2	Thesis Proposal (5%)	10/7/2024	0	100%
0.3	Interim Report (10%)	1/20/2025	0	90%
0.4	Final Thesis Presentation (10%)	3/31/2025	70	8%
0.5	Final Thesis Document (75%)	4/7/2025	77	35%
1	**Presentations**			
1.1	Introduction Presentation	10/22/2024	0	100%
1.2	Novemeber Presentation	11/21/2024	0	100%
1.3	December Presentation	12/17/2024	0	100%
1.4	January Presentation	1/31/2025	11	50%
1.5	February Presentation	2/28/2025	39	0%
2	**Onboarding and Setup Tasks**			
2.1	Access to SciNet and MIST Clusters	9/20/2024	0	100%
2.2	Credentials for Llama Model Downloads	9/20/2024	0	100%
2.3	Setting Up Development Environment	11/8/2024	0	100%
2.4	Addressing Compute Limitations and Conflicting Packages in SciNet	2/28/2025	39	10%
3	**Literature and Design Work**			
3.1	Reading LLM Papers (Getting a Sense for Deep Learning Space)	9/30/2024	0	100%
3.2	Defining and Scoping the Project, Understanding Methods and Motivation	9/20/2024	0	100%
3.3	Researching Relevant Background	3/20/2025	59	50%
4	**Technical Work (Data Processing)**			
4.1	Unload BraTS Dataset and Perform Preprocessing and Structure Data Into Training, Validation, and Test Cohorts	10/11/2024	0	100%
4.2	Extract Segmentation Ranges in All 3 Slice Directions (Axial / Coronal / Sagittal)	11/15/2024	0	100%
5	**Technical Work (CNN Related)**			
5.1	Create and Train CNN Using Training and Validation Cohorts of BraTS Data	11/22/2024	0	100%
5.2	Perform Inference Task on Trained CNN and Store Classification Results	11/22/2024	0	100%
6	**Technical Work (LLM Related)**			
6.1	Feasibility Test on LLM in SciNet (Ensure LLM can Handle Compute)	11/15/2024	0	100%
6.2	Perform Inference Task on General LLM and Store Classification Results	11/29/2024	0	100%
6.3	Reading Unsloth Documentation and Completing a Fine-Tuning Example	1/10/2025	0	100%
6.4	Perform Fine-Tuning Either Using SciNet or Google Colab (*Currently Stuck)	1/10/2025	0	70%
6.5	** Resolve Bitsandbytes Compatibility Issues on SciNet	1/10/2025	0	25%
6.6	** Complete Fine-Tuning Initial Trial Using Colab Resources	1/10/2025	0	10%
6.7	** Transferring the Fine-Tuned LLM Model Between Colab and SciNet	1/10/2025	0	8%
7	**Technical Work (Evaluation)**			
7.1	Implement Evaluation Frameworks (F1 Score, AUC, Accuracy, Recall, etc...)	11/22/2024	0	100%
7.2	Perform Initial Consistency Test on General LLM (Using 2 Images)	11/15/2024	0	100%
7.3	Assess and Compare CNN and General LLM Models on this Dataset and Task	11/22/2024	0	100%
7.4	** Evaluating Fine-Tuned Model Performance with CNN and General LLM	1/10/2025	0	2%
7.6	Expanding the Consistency Test Dataset	1/31/2025	11	0%
7.7	Running Expanded Consistency Tests on General LLM	1/31/2025	11	0%
7.8	** Conducting All Consistency Tests on Fine-Tuned LLM and Performing Comparison	2/21/2025	32	0%
8	**Technical Work (Evaluation)**			
8.1	Selecting the Next Medical Task	2/14/2025	25	0%
8.2	Processing and Prepping the Dataset	2/14/2025	25	0%
8.3	Performing a Similar Process on the CNN	2/21/2025	32	0%
8.4	Fine-Tuning the LLM for the New Task	3/14/2025	53	0%
8.5	Performing Inference	3/14/2025	53	0%
8.6	Performance Evaluation Across Models (CNN, General LLM, Fine-Tuned LLM)	3/14/2025	53	0%
8.7	Documenting Results and Insights for Second Task	4/7/2025	77	0%

Figure 19: Gantt Chart shows current milestones achieved and future tasks organized by various types of technical work, presentations, deliverable writing, etc.

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] 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).

[17] 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.

[18] 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.

[19] 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).

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

[21] *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).

[22] *Mist - SciNet Users Documentation*. [Online]. Available: https://docs.scinet.utoronto.ca/index.php/Mist (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] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, *et al.*, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," en, *Scientific Data*, vol. 6, no. 1, p. 317, Dec. 2019, Publisher: Nature Publishing Group, ISSN: 2052-4463. DOI: 10.1038/s41597-019-0322-0. [Online]. Available: https://www.nature.com/articles/s41597-019-0322-0 (visited on 09/29/2024).