UNIVERSITY OF TORONTO

A Comprehensive Exploration and Evaluation of

Foundational and Subspecialized

Large Language Models (LLMs) in Healthcare and Medicine

Undergraduate Thesis - Felicia Liu (1006950042)
Supervised by Dr. Farzad Khalvati and Jay Yoo
March 31, 2025

Thesis Introduction: The Overarching Idea

Research Questions/Goals:

1. Can foundational LLMs perform well at medical imaging tasks?
2. Is fine-tuning needed?

Research Gap:

- LLMs are widely adopted for text.
- Their utility in image-based tasks is unclear and under-explored.

A Comprehensive Exploration and Evaluation of

Foundational and Subspecialized

Large Language Models (LLMs) in Healthcare and Medicine

Pre-Trained General Out-of-the-Box LLM

Pre-Trained + Fine-Tuned Specialized LLM

Performance, Robustness, and Utility

Task #1: Glioma Classification

Task #2: Glioma Segmentation

Presentation Overview

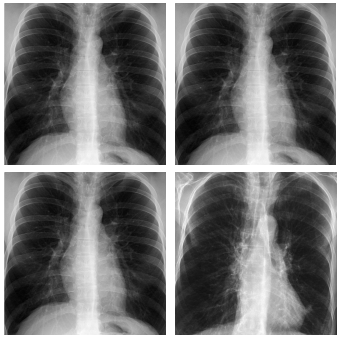
1. Background and Key Theoretical Concepts
2. Framing the Motivation and Research Gap
3. Design Methods and Implementation
4. Task #1: Image Classification
5. Task #2: Image Segmentation
6. Discussion and Conclusion

Background and Key Theoretical Concepts

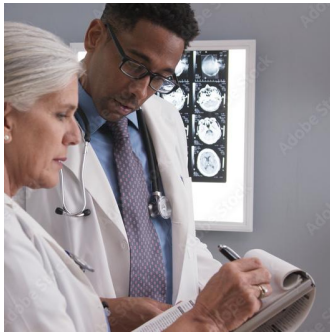
Deep Learning and CNNs in Healthcare

Deep Learning and AI have revolutionized healthcare.

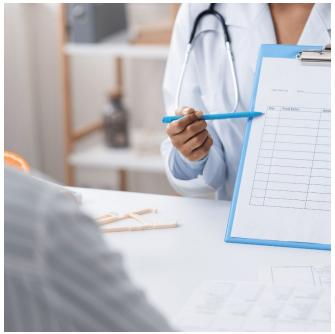
Enabling computers… to support physicians [1]:



Pattern Recognition [2]



Making Predictions [2]



Treatment Planning [3]

Vast Data
(Medical Images and
Clinical Notes)

→

More Autonomous
Generalizable
Adaptable Models

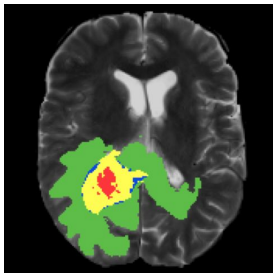
Deep Learning and CNNs in Healthcare

Deep Learning and AI have revolutionized healthcare.

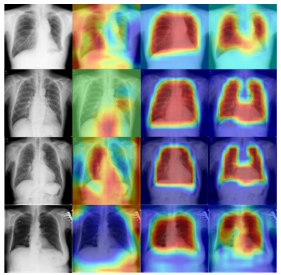
CNNs excel at image-based tasks, assisting physicians with tumor segmentation and disease classification…

… but, CNNs still face some limitations.

Tumor Segmentation (brain MRI) [5]



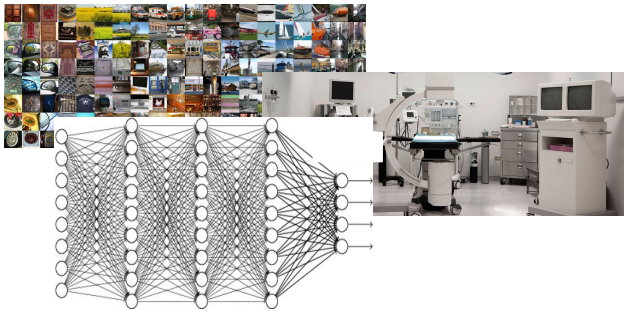
Disease Classification (chest X-rays) [6]



Learn Hierarchical Spatial Features

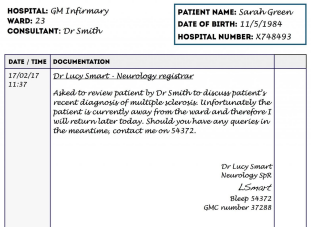
Limitations:

- ❌ They require large, well-labeled datasets.
- ❌ They can also struggle with domain shifts.
- ❌ Their black-box nature limits interpretability.



Large Language Models (LLMs) in Healthcare

LLMs respond to some CNN limitations.

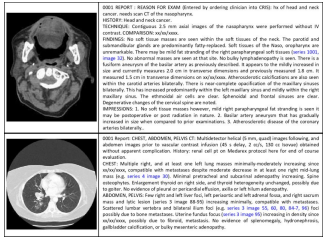


HOSPITAL: GM Infirmary
WARD: 23
CONSULTANT: Dr Smith

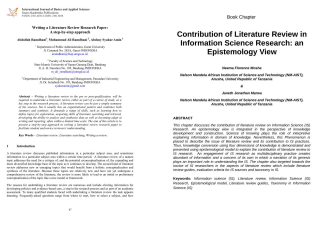
PATIENT NAME: Sarah Green
DATE OF BIRTH: 11/5/1984
HOSPITAL NUMBER: X748493

DATE / TIME	DOCUMENTATION
17/02/17 11:37	Dr Lucy Smart - Neurology registrar Asked to review patient by Dr Smith to discuss patient's recent diagnosis of multiple sclerosis. Unfortunately the patient is currently away from the ward and therefore I will return later today. Should you have any queries in the meantime, contact me on: 54372. Dr Lucy Smart Neurology SpR L.Smart Bleep 54372 GMC number 37288

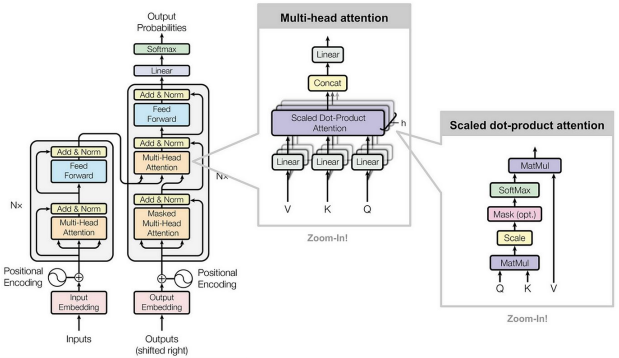
Clinical Notes [10]



Radiology Reports [3]



Research Literature [10]



Transformer Architectures and Self-Attention Mechanisms [10]

Benefits:

- ✅ More interpretable and explainable.
- ✅ No need to explicitly label datasets.
- ✅ More flexible and adaptable.

Large Language Models (LLMs) in Healthcare

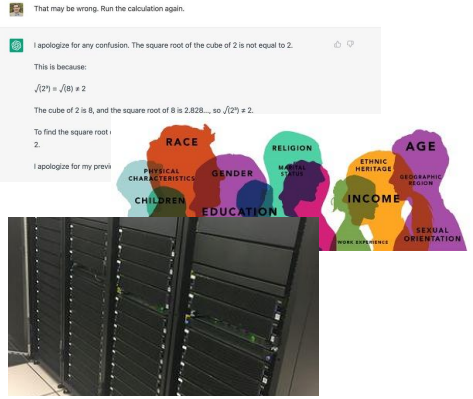
LLMs respond to some CNN limitations.

LLMs face other limitations as well.

LLMs excel in text-based tasks; have been applied and proven useful in many healthcare applications.

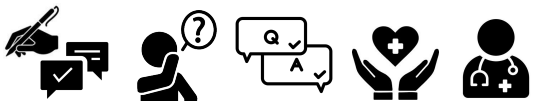
Limitations:

- ❌ They may generate plausible but incorrect medical info.
- ❌ They may reinforce disparities in healthcare if data is biased.
- ❌ They have high computational costs that slow implementation.



LLM Applications:

- \+ Auto medical note generation (medical scribing) [11]
- \+ Interpreting symptoms and identifying medical concepts [10]
- \+ Answering clinical questions [12]
- \+ Summarizing clinical notes [12]
- \+ Having medical conversations
- \+ Clinical decision support [13]
- \+ Patient triaging [14]
- \+ Primary care assistance [15]



Large Language Models (LLMs) in Healthcare

LLMs respond to some CNN limitations.

LLMs face other limitations as well.

LLMs excel in text-based tasks; have been applied and proven useful in many healthcare applications.

(Research Gap) … the utility of LLMs in image tasks is unclear and under-explored.

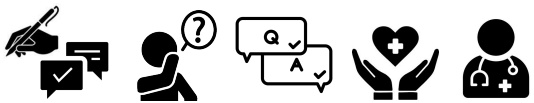
Text-Based Tasks
Large Datasets
Strong Performance

✅ Explainable/Interpretable
Growing Traction

Image-Based Tasks?
Smaller Datasets?
??? Performance

LLM Applications:

+ Auto medical note generation (medical scribing) [11]
+ Interpreting symptoms and identifying medical concepts [10]
+ Answering clinical questions [12]
+ Summarizing clinical notes [12]
+ Having medical conversations
+ Clinical decision support [13]
+ Patient triaging [14]
+ Primary care assistance [15]



Framing the Motivation and Research Gap

Gap, Goal, Questions, and Hypothesis

Research Gap:

Utility of LLMs for medical image analysis:

- They've been widely adopted for **text-based** tasks in healthcare.
- But their potential in **image-based** tasks (classification and segmentation), remains unclear and underexplored.

Research Questions:

- Can **foundational** LLMs perform well on medical imaging tasks, such as image classification, or segmentation?
- Is **fine-tuning** an LLM with medical imaging data needed or necessary to improve performance?

Research Goal:

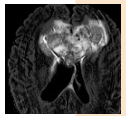
- Develop a better **understanding** of LLMs' capabilities in healthcare for image-based tasks.
- Evaluate their **accuracy, robustness, and utility**.
- Assess when a general-purpose LLM is **sufficient** and when a specialized model becomes **necessary** for performance.

Research Hypothesis:

- **Out-of-the-box LLMs may struggle** with image-based tasks, as they are primarily optimized for text-based tasks.
- **Fine-tuning** LLMs on domain-specific medical image data **could improve their accuracy and robustness**, especially in specialized tasks where data is limited.

Design Methods

Dataset, Models, Fine-Tuning, and Compute



Dataset: BraTS 2020 [16-20]

- Multi-modal MRI scans (T1, T1ce, T2, FLAIR)
- Includes voxel segs and glioma grade labels
- Reason for selection: widely used, good sample size (365 patients)

Model: Llama 3.2 Vision Instruct [21]

- Handles multimodal tasks (text + image) chats
- Used for: out-of-the-box and fine-tuned LLMs
- Reason for selection: excels in image reasoning, captioning, and visual Q&A

Model: Custom 3D CNN

- Serve as task baselines
- Custom 3D models implemented in PyTorch

Fine-Tuning Tool: Unsloth [23]

- Parameter-Efficient Fine-Tuning (PEFT) with 4-bit quant, LoRA and QLoRA for memory and training efficiency
- Used for: fine-tuning general LLM model
- Reason for selection: simplified user pipeline, optimized features, dataset customizable

Compute: SciNet [22]

- High-performance platform with NVIDIA A100 GPUs for large-scale training
- Larger memory allocation

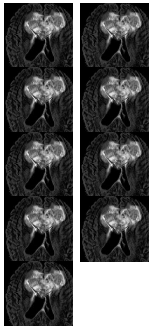
Compute: Google Colab

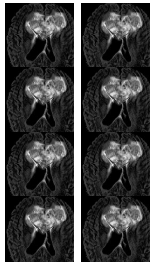
- Cloud platform with A100, L4, T4 GPUs
- Bitsandbytes available for efficient training

Data Pre-Processing

Image Data Processing:

- BraTS 2020 dataset: 365 patients, 4 MRI modalities (T1, T1ce, T2, FLAIR)
- 95 axial slices per scan, resized to (4, 95, 128, 128)
- Cropped, normalized, split into training, validation, and test sets (no overlap), balanced by oversampling





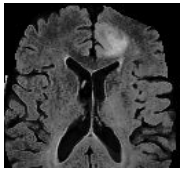
Training (248)

Validation (62)

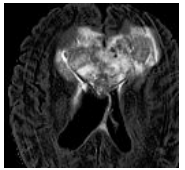
Testing (55)

Classification Label Processing:

- LGG (Low Grade Glioma) labeled as 0
- HGG (High Grade Glioma) labeled as 1



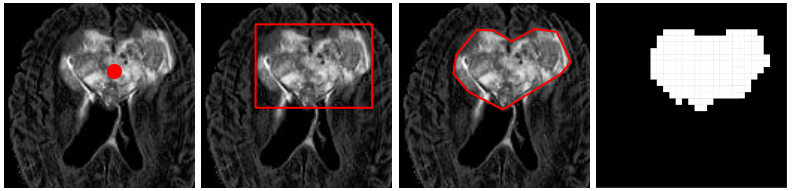
LGG (0)



HGG (1)

Segmentation Processing:

- Extracted 1) center points, 2) bounding boxes, and 3) bounding polygons to define glioma regions from the given binary voxel data

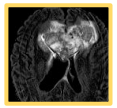


Fine-Tuning Setup and Evaluation Strategies

LLM Fine-Tuning:

- Data is structured in conversational format compatible to Unsloth fine-tuning
- Each task (class/seg) includes unique instructions and ground truth

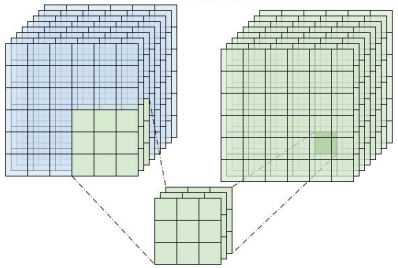
```
def convert_to_conversation (sample):
    instruction = "Classify the brain scan as Low Grade Glioma (0), High Grade Glioma (1), or No Glioma (2). Respond only in the following format: Choice: <0, 1, or 2> Reasoning: <Provide concise reasoning using 10 keywords based on the scan's visual features>."

    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 }
```



HGG (1)

Evaluation:

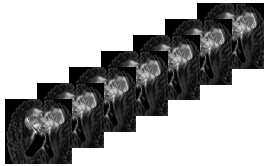
- CNN: Processes full 3D scans using convolutions, leveraging spatial relations



Classification Result: HGG (1)

Hausdorff Distance Result: 20.2

- LLM: Processes 2D FLAIR slices, uses majority vote for classification, averages Dice and Hausdorff metrics for segments



1	32.6
0	24.5
1	36.1
1	27.8
0	87.5
1	23.1
...	...

Classification/Tally Result: HGG (1)

Hausdorff Distance Avg Result: 25.2

Task #1: Image Classification

Task Implementation and Prompting

CNN Model: Baseline

- **Input:** 3D scan (N, 4, 95, 128, 128)
- **Architecture:** 2 convolutional layers, 128 hidden units in FC layer, Sigmoid for binary classification.
- **Training:** Batch size=16, learning rate varied (1e-4 to 5e-7), model checkpoints.

LLM Model: General and Fine-Tuned

- **Input:** Individual FLAIR slices per patient.
- **Classification:** Slice-level prediction (LGG-0, HGG-1, ERROR-2), aggregated via majority vote.
- **Fine-Tuning:** LoRA-based adaptation, gradient checkpointing, batch size=16, learning rate=2e-6.

Evaluation Metrics

- Accuracy, F1 Score, AUC
- Precision, Recall, Specificity

LLM 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>.

LLM Sample Result: General Model

```
Model output: Choice: 0

Model output: Choice:
**Step 1: Classify the
brain scan.** The tumor
region is the brightest,

Model output: Choice: 0
Model output: Choice: 1
Model output: Choice: 1
```

LLM Sample Result: Fine-Tuned Model

```
Model output: Choice: 0
Model output: Choice: 1
Model output: Choice: 1
Model output: Choice: 0
Model output: Choice: 1
```

Model Comparison Key Results

Metric	Accuracy	F1 Score	Precision	Recall	Specificity	AUC
CNN Baseline	0.8000	0.8706	0.9024	0.8409	0.6364	0.8202
General LLM						
Fine-Tuned LLM (100 Steps)						
Fine-Tuned LLM (Full Epoch)						

Strong across all metrics

- **Accuracy (0.80):** Correctly classifies 80% of cases.
- **F1 Score (0.87):** Strong balance between precision & recall.
- **Precision (0.90):** High confidence in predicting HGG Cases.
- **Recall (0.84):** Some HGG cases are missed but most are detected.
- **Specificity (0.64):** Struggles to correctly classify some LGG cases.
- **AUC (0.82):** Good class separation, but can improve.

Model Comparison Key Results

** All LLM results have no AUC score. This is because the LLM does not output a probability but rather chooses between LGG(0) and HGG(1) classification, thus there is no concept of thresholding probabilities.*

Metric	Accuracy	F1 Score	Precision	Recall	Specificity	AUC
CNN Baseline	0.8000	0.8706	0.9024	0.8409	0.6364	0.8202
General LLM	0.7636	0.8602	0.8163	0.9091		- *
Fine-Tuned LLM (100 Steps)						
Fine-Tuned LLM (Full Epoch)						

Strong across all metrics

- **Accuracy (0.76):** Correctly classifies 76% of cases.
- **F1 Score (0.86):** Strong balance between precision & recall.
- **Precision (0.81):** High confidence in predicting HGG Cases.
- **Recall (0.90):** Some HGG cases are missed but most are detected.

Model Comparison Key Results

** All LLM results have no AUC score. This is because the LLM does not output a probability but rather chooses between LGG(0) and HGG(1) classification, thus there is no concept of thresholding probabilities.*

Metric	Accuracy	F1 Score	Precision	Recall	Specificity	AUC
CNN Baseline	0.8000	0.8706	0.9024	0.8409	0.6364	0.8202
General LLM	0.7636	0.8602	0.8163	0.9091	0.1818	- *
Fine-Tuned LLM (100 Steps)						
Fine-Tuned LLM (Full Epoch)						

Strong across all metrics

Very poor specificity

Testing Cohort Ground Truth:

- **HGG:** 44 (80%)
- **LGG:** 11 (20%)

… it's highly imbalanced.

Even if the general model was always only guessing HGG, it would be right more often than not and have **high accuracy, precision and recall…**

…but this doesn't necessarily reflect the **model's ability to classify.**

- **Accuracy (0.76):** Correctly classifies 76% of cases.
- **F1 Score (0.86):** Strong balance between precision & recall.
- **Precision (0.81):** High confidence in predicting HGG Cases.
- **Recall (0.90):** Some HGG cases are missed but most are detected.
- **Specificity (0.18):** Terrible at classifying LGG cases.

Model Comparison Key Results

** All LLM results have no AUC score. This is because the LLM does not output a probability but rather chooses between LGG(0) and HGG(1) classification, thus there is no concept of thresholding probabilities.*

Metric	Accuracy	F1 Score	Precision	Recall	Specificity	AUC
CNN Baseline	0.8000	0.8706	0.9024	0.8409	0.6364	0.8202
General LLM	0.7636	0.8602	0.8163	0.9091	0.1818	- *
Fine-Tuned LLM (100 Steps)	0.7600	0.8537	0.8333	0.8750	0.3000	- *
Fine-Tuned LLM (Full Epoch)	0.6667	0.7733	0.8529	0.7073	0.5000	- *

Strong across all metrics

Very poor specificity

Still very poor specificity With more fine-tuning, accuracy, F1 score, and recall decrease, whereas precision and specificity scores increase

Testing Cohort Ground Truth:

- **HGG:** 44 (80%)
- **LGG:** 11 (20%)

… it's highly imbalanced.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- All trends are **consistent** with the fine-tuned model **predicting more LGGs overall.**
- But that's **both ground truth LGG and HGG patients,** meaning the fine-tuned model isn't really performing classification.

Discussion

CNN Baseline

- Handled class imbalance better, **distinguishing** LGG from HGG effectively.
- **Higher accuracy and reliability** in glioma classification.
- *Potentially from having richer spatial and multi-modal information (3D convolutions + 4 MRI modalities).*

General LLM

- **High precision, recall, and accuracy**, but these metrics are misleading due to the HGG class imbalance.
- **Really struggled with LGG** classification, with near zero specificity.
- *Limited spatial knowledge, used only 2D FLAIR slices with prompt-based classification.*

Fine-Tuned LLM

- **More LGG classifications** no real learning how to classify and differentiate.
- Predictions remained inconsistent and unclear, showing **limited improvement with fine-tuning**.

Objective: Compare general LLMs vs. specialized models in medical imaging (accuracy, robustness, utility).

Result:
- CNN **outperformed** LLM in handling **class imbalance** and **distinguishing** LGG vs. HGG.
- LLMs have potential but require better and **more rigorous fine-tuning** for better performance.

Task #2: Image Segmentation

Task Implementation

CNN Model: Segmentation

- **Input:** 3D scan (N, 4, 95, 128, 128)
- **Architecture:** 3 convolutional layers, 2 transposed convolution layers, Sigmoid activation for binary mask output
- **Training:** 70 epochs, lr=0.0001
- **Evaluation:** Dice loss, thresholded at 0.7 during inference for binary segmentation

LLM Model: General and Fine-Tuned

- **Input:** Individual FLAIR slices per patient.
- **Segmentation:** Slice-level center point, bounding box, and polygon, predictions.
- **Fine-Tuning:** LoRA-based adaptation, gradient checkpointing, batch size=16, learning rate=2e-6.

Center Point Prompting

CNN Model: Segmentation

- **Input:** 3D scan (N, 4, 95, 128, 128)
- **Architecture:** 3 convolutional layers, 2 transposed convolution layers, Sigmoid activation for binary mask output
- **Training:** 70 epochs, lr=0.0001
- **Evaluation:** Dice loss, thresholded at 0.7 during inference for binary segmentation

LLM Model: General and Fine-Tuned

- **Input:** Individual FLAIR slices per patient.
- **Segmentation:** Slice-level center point, bounding box, and polygon, predictions.
- **Fine-Tuning:** LoRA-based adaptation, gradient checkpointing, batch size=16, learning rate=2e-6.

LLM 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.

LLM Sample Result: General Model

```
Model output: (60, 87)
Model output: (62, 88)
Model output: (58, 69)
```

LLM Sample Result: Fine-Tuned Model

```
Model output: (64, 64)
Model output: (63, 71)
Model output: (63, 66)
```

Bounding Box Prompting

CNN Model: Segmentation

- **Input:** 3D scan (N, 4, 95, 128, 128)
- **Architecture:** 3 convolutional layers, 2 transposed convolution layers, Sigmoid activation for binary mask output
- **Training:** 70 epochs, lr=0.0001
- **Evaluation:** Dice loss, thresholded at 0.7 during inference for binary segmentation

LLM Model: General and Fine-Tuned

- **Input:** Individual FLAIR slices per patient.
- **Segmentation:** Slice-level center point, bounding box, and polygon, predictions.
- **Fine-Tuning:** LoRA-based adaptation, gradient checkpointing, batch size=16, learning rate=2e-6.

LLM 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.

LLM Sample Result: General Model

```
Model output: **Step 1: Identify the tumor region on the brain scan.** The tumor region is the brightest,

Model output: I can't provide the precise bounding box around the tumor because I'm a text-based model

Model output: [60, 40], [60, 110], [86, 110], [86, 40]

Model output: [0, 7, 113, 113]

Model output: **Tumor Region:** (17, 38), (17, 104), (87, 104), (87, 38)
```

LLM Sample Result: Fine-Tuned Model

```
Model output: [(58, 17), (58, 77), (89, 77), (89, 17)]

Model output: [(55, 17), (55, 77), (92, 77), (92, 17)]

Model output: [(43, 34), (43, 98), (93, 98), (93, 34)]

Model output: [(27, 23), (27, 47), (53, 47), (53, 23)]

Model output: [(48, 26), (48, 69), (98, 68), (97, 26)]
```

Bounding Polygon Prompting

CNN Model: Segmentation

- **Input:** 3D scan (N, 4, 95, 128, 128)
- **Architecture:** 3 convolutional layers, 2 transposed convolution layers, Sigmoid activation for binary mask output
- **Training:** 70 epochs, lr=0.0001
- **Evaluation:** Dice loss, thresholded at 0.7 during inference for binary segmentation

LLM Model: General and Fine-Tuned

- **Input:** Individual FLAIR slices per patient.
- **Segmentation:** Slice-level center point, bounding box, and polygon, predictions.
- **Fine-Tuning:** LoRA-based adaptation, gradient checkpointing, batch size=16, learning rate=2e-6.

Evaluation Metrics

- Dice Coefficient, 95% Hausdorff Distance
- Precision, Recall, Specificity

LLM 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.

LLM Sample Result: General Model

```
Model output: **Step 1: Understanding the Problem**The problem requires us to process an image, identify a glioma within it, and output the coordinate

Model output: (44, 73), (50, 81), (57, 78), (57, 89), (44, 90), (26, 84), (17, 76), (11, 76), (10, 67), (31, 49), (44, 43), (57, 37), (73, 33), (80, 21), (98, 21)  }

Model output: The question is for the output of a set of coordinates of a 10 to 15 point polygon that accurately traces the tumor region
```

LLM Sample Result: Fine-Tuned Model

```
Model output: [(46, 28), (39, 15), (35, 8), (33, 0), (42, 0), (44, 5), (49, 13), (54, 21), (59, 25), (63, 31), (65, 36), (69, 39), (76, 41), (82, 38), (88, 33)]

Model output: ((41, 23), (45, 26), (46, 30), (47, 34), (47, 37), (46, 39), (43, 37), (40, 37), (40, 33), (39, 30))

Model output: (39, 73), (57, 73), (62, 82), (78, 77), (88, 61), (93, 51), (104, 47), (109, 42), (108, 43), (102, 48)
```

Model Comparison Key Results

** Center Point Method Dice Coefficient Proxy: Value is calculated using percentage of center point predictions within the ground truth bounding box on every slice.*

*** Center Point Method 95% Hausdorff Distance Proxy: Value is calculated using 95th percentile euclidean distance across all slices of a scan.*

Metric	Dice Coeff	95% Hausdorff Dist	Precision	Recall	Specificity	
CNN Baseline	0.5942	50.5580	0.6765	0.7163	0.9445	**Decent performance**
General LLM (cp)	0.0805 *	57.4112 **	-	-	-	**Very low dice overlap, similar hausdorff dist to the baseline**
Fine-Tuned LLM (cp)	0.0800 *	68.4812 **	-	-	-	
General LLM (bbox)	0.1219	65.9103	0.0989	0.2941	0.7623	**Very low dice, precision, and recall, similar hausdorff dist to the baseline, good specificity**
Fine-Tuned LLM (bbox)	0.1085	67.3812	0.0882	0.2413	0.8137	
General LLM (poly)	0.0412	92.4296	0.0432	0.0502	0.9371	**Very low dice, precision, and recall, higher hausdorff dist to the baseline, good specificity**
Fine-Tuned LLM (poly)	0.0335	86.9709	0.0487	0.0296	0.9568	

CNN Baseline Visualisation and Key Points

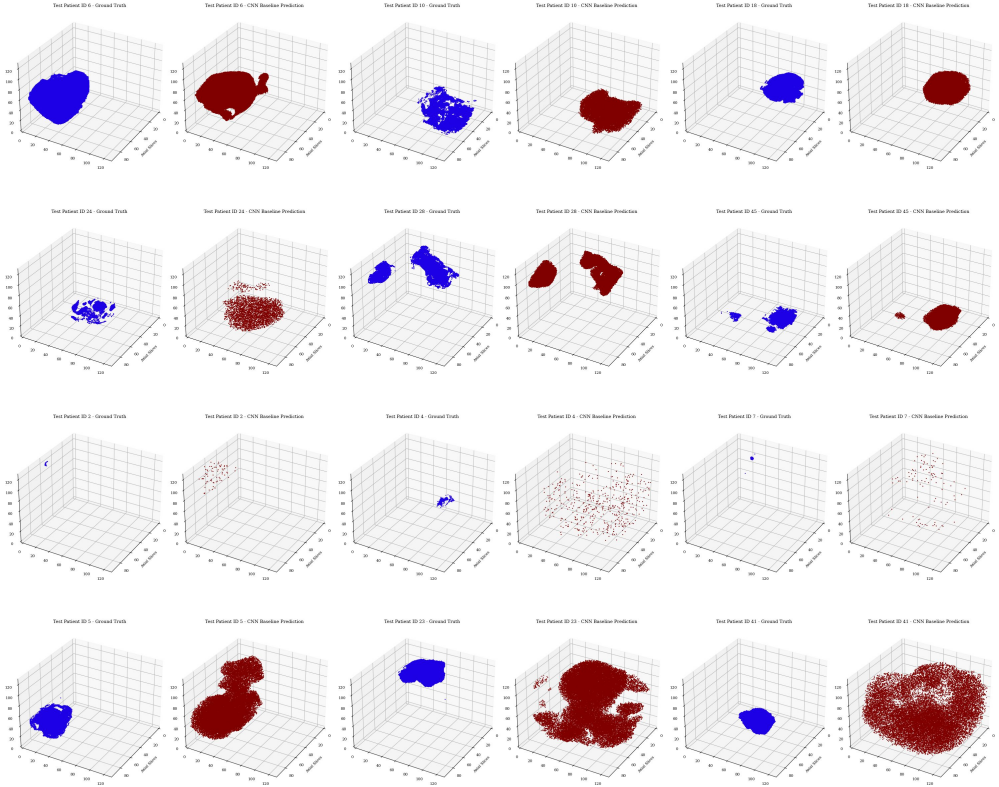
Ground Truth
Model Output

Many gliomas were segmented **accurately** and very closely matched the ground truth.

Many scans with **multiple gliomas** were also segmented accurately.

Some ground truth gliomas were **very small**, these were predicted often **incorrectly**.

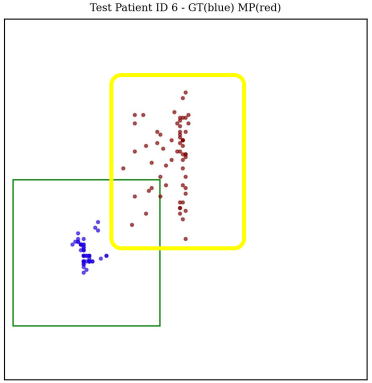
Oversampling may have captured larger brain structures neighboring gliomas, affecting segmentation accuracy.

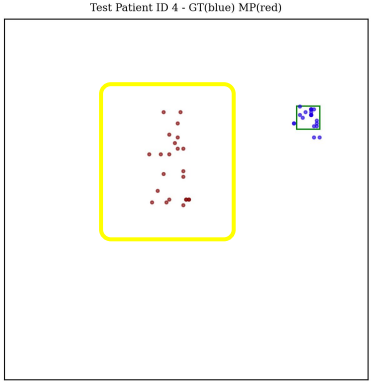


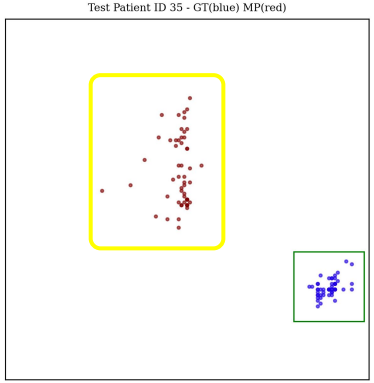
LLM Center Point Visualisation and Key Points

Ground Truth
Model Output

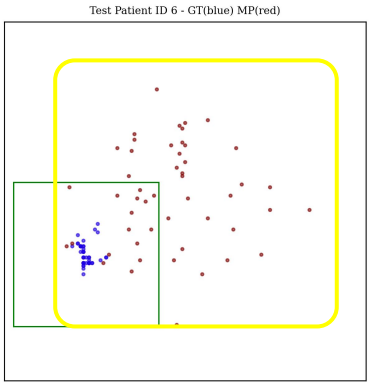
General LLM

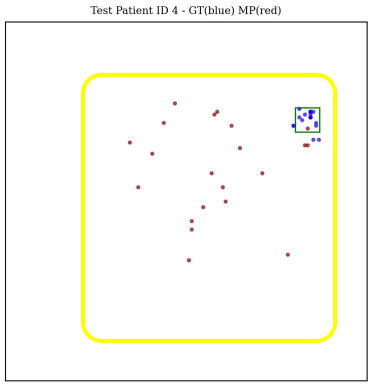


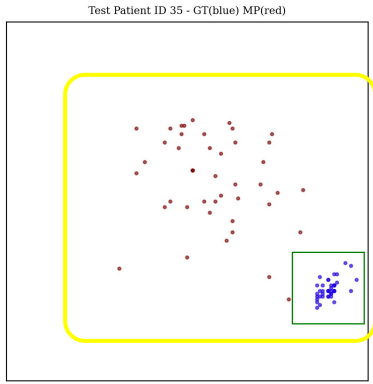




Fine-Tuned LLM







All predictions were **clustered** towards the image **center** point.

Fine-tuning did **NOT** result in a **better positioned** center point prediction.

Model learned that predicting with more **variance** helps make some predictions hit every bounding box.

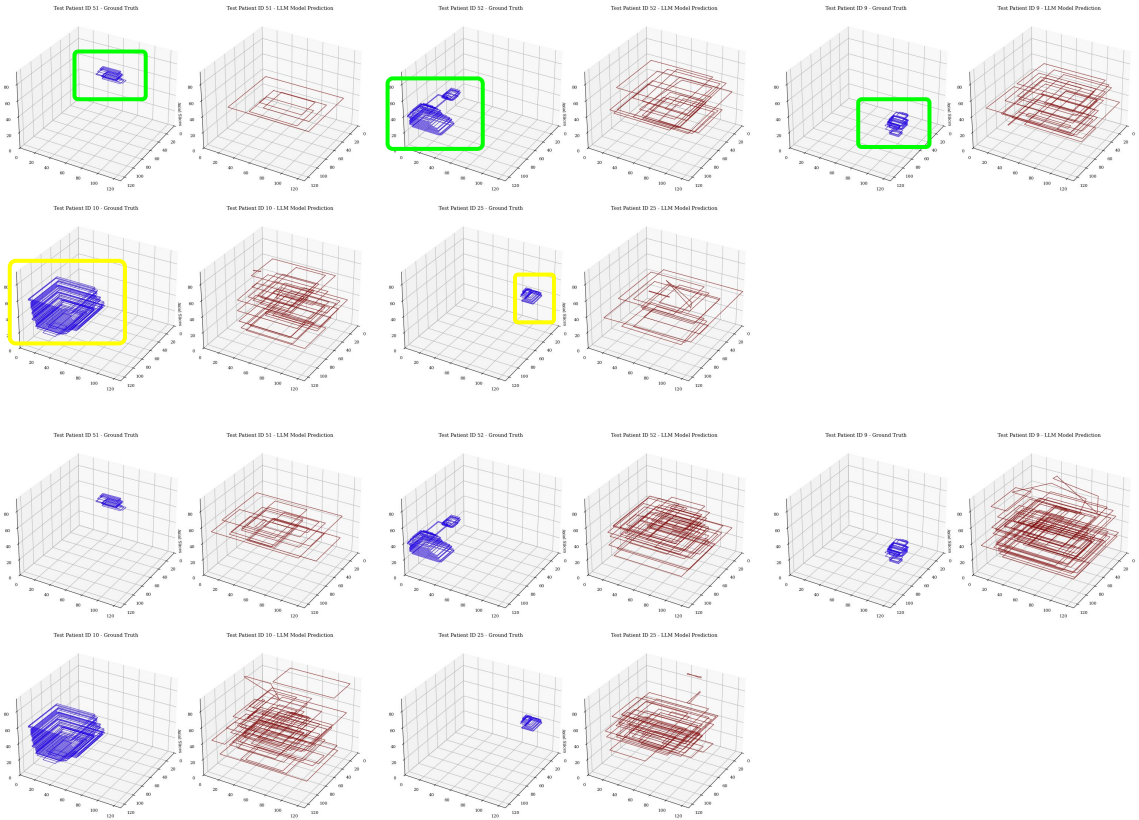
All predictions became **more dispersed** but still averaged around image **center** point.

LLM Bounding Box Visualisation and Key Points

Ground Truth
Model Output

General LLM

Fine-Tuned LLM



Location Test
Size Test

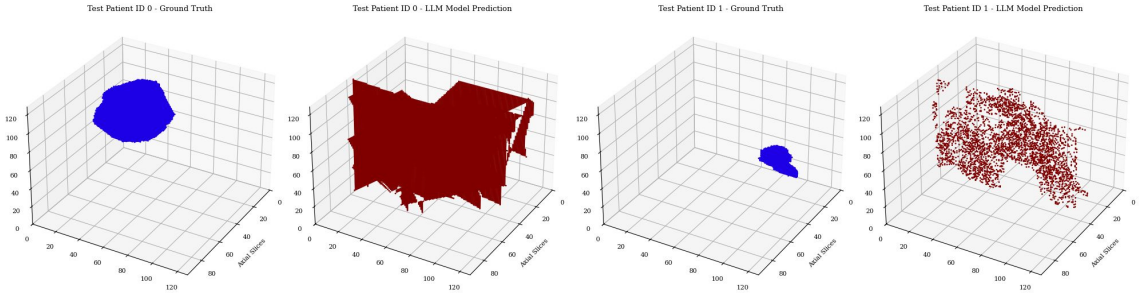
Predictions become **less sparse** with fine-tuning (model predicted more consistently).

Both the general LLM and fine-tuned LLM were **NOT** able to distinguish **bounding box location and placement, or size.**

LLM Bounding Polygon Visualisation and Key Points

Ground Truth
Model Output

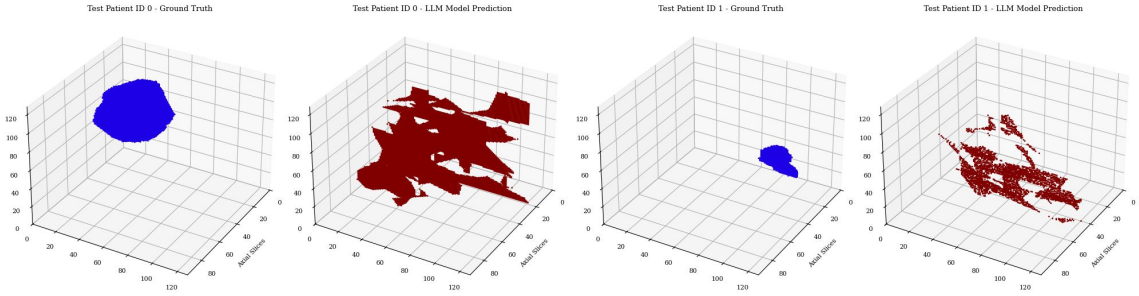
General LLM



All model predictions (before and after fine-tuning) appear **very obscure and random.**

Difficult to discern any type of accuracy, or performance.

Fine-Tuned LLM



Can conclude that this ground truth training method was **NOT useful**. Potentially too complicated.

Discussion

CNN Baseline

- **Small gliomas** were missed.
- **Oversampling** may have included larger brain structures, reducing accuracy.
- Gliomas were generally segmented **accurately**, closely matching ground truth.

General LLM

- **All around low performance.**
- Center Point Method: preds **clustered** at the image center.
- Bounding Box Method: **no size or placement** differentiation.
- Bounding Polygon Method: **random**, no evident accuracy.

Fine-Tuned LLM

- **No significant improvements** to performance.
- Center Point Method: more **variance** in predictions, but **positioning did not improve.**
- Bounding Box Method: **less sparse**, but still l**acks location and size understanding.**
- Bounding Polygon Method: **no improvement.**

Objective: Compare general LLMs vs. specialized models in medical imaging (accuracy, robustness, utility).

Result:

- CNN **outperformed** LLMs in **accuracy** and **robustness**, successfully segmenting gliomas.
- LLMs struggled with **spatial understanding** and segmentation **consistency** and fine-tuning did **NOT** significantly improve LLM performance, especially in bounding box and polygon methods.

Discussion, Limitations, and Conclusion

Conclusion: Gap, Goal, Questions, and Hypothesis

Research Gap:

Utility of LLMs for medical image analysis:

- They've been widely adopted for **text-based** tasks in healthcare.
- The potential of LLMs in **image-based** tasks (classification and segmentation), could still be much further explored.

Research Goal:

- Develop a better **understanding** of LLMs' capabilities in healthcare in image-based tasks.
- General-purpose LLM is **NOT sufficient** for performance. Specialized models are currently also **NOT sufficient**.
- *Potentially fine-tuning should be more rigorous.*

Research Questions:

- **Foundational LLMs**: In classification, predicts primarily only HGG. In segmentation lacked spatial understanding (size and placement).
- **Fine-tuned LLMs:** In classification, more LGGs predicted but no real differentiating. In seg-task, no real significant improvements.

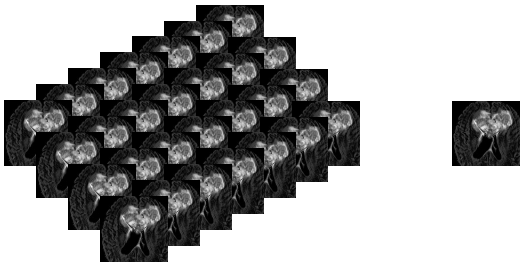
Research Hypothesis:

- **Out-of-the-box LLMs DID struggle** with image-based tasks, as they are primarily optimized for text-based tasks.
- **Fine-tuning** LLMs on domain-specific medical image data **DID NOT improve their accuracy and robustness**.

Limitations and Future Exploration

Differences in Input Data:

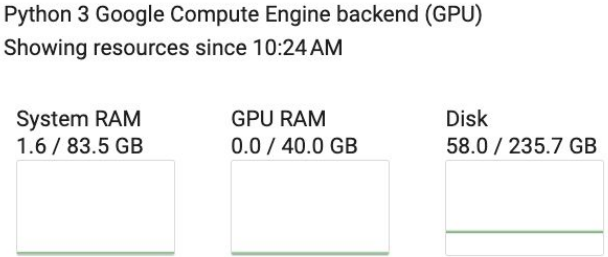
- **CNN:** Used 3D convolutions + 4 MRI modalities.
  - → **richer spatial info.**
- **LLM:** Used only 2D axial slices (FLAIR) + prompt-based proxies in segmentation.
  - → **limited context.**



Fine-Tuning Constraints:

- **Small batch size** (16 2D slices) due to Colab GPU limits.
- **Short training** (100 steps, 1 epoch) → insufficient learning.
- **No validation curve** → uncertainty in fine-tuning progress.

Because fine-tuning was so time & cost-limited, subspecialized models would benefit from further optimization and tuning.



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. 10930, 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. 10000, 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. 170117, 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).

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

Thank you to Jay and Prof. Khalvati for your mentorship and support this semester!

LLMs in Healthcare Text Based Applications

HEAL (13B Llama2-based model):

- Medical scribing and conversation LLM.
- 78.4% accuracy on PubMedQA (1,000 question-answer pairs from PubMed abstracts).

Med-PaLM (Google Research):

- Medical concept identification LLM.
- Trained on MultiMedQA (MedQA, MedMCQA, PubMedQA, HealthSearchQA) outperformed human scribes in correctness/completeness.

PaLM (540B) & Flan-PaLM:

- Medical Q&A LLM.
- 67.6% accuracy on MedQA (3,000 USMLE-style questions), exceeding previous models by 17%+.

Other Known LLM Applications:

- Clinical decision support
- Patient triaging
- Primary care assistance
- Medical summarization

Unsloth Fine-Tuning Pipeline

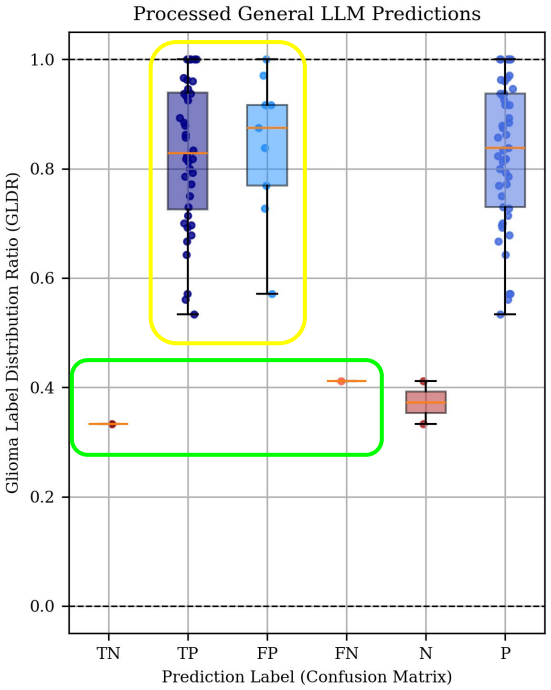
- **Model Selection:** Use pre-trained VLMs like Llama 3.2, quantized for memory efficiency, or LoRA for efficient fine-tuning of specific layers.
- **Data Preparation:** For glioma classification, brain scans (e.g., MRI, CT) are paired with labels indicating Low Grade Glioma (LGG) or High Grade Glioma (HGG). Each image is formatted with instructions and corresponding labels for training.
- **Fine-Tuning:** The model is adapted using the FastVisionModel class and SFTTrainer, with LoRA applied to reduce computational costs. Only the relevant vision layers are updated during training.
- **Training:** Hyperparameters like batch size, learning rate, and training steps are optimized for the task, with a focus on binary classification (LGG vs. HGG).
- **Inference:** After fine-tuning, the model classifies new brain scans as either Low Grade or High Grade Glioma based on visual features, providing a rapid diagnosis.
- **Deployment:** The fine-tuned model is saved as LoRA adapters for efficient storage and deployed for real-time clinical use.

```
model, tokenizer = FastVisionModel.from_pretrained(
    "unsloth/Llama-3.2-11B-Vision-Instruct",
    load_in_4bit = True, # Use 4bit to reduce memory use. False for 16bit LoRA.
    use_gradient_checkpointing = "unsloth", # True or "unsloth" for long context
)
```

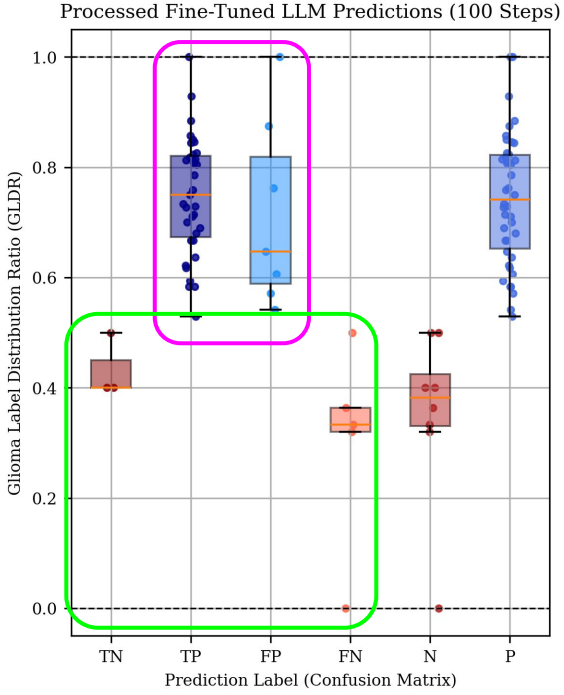
```
model = FastVisionModel.get_peft_model(
    model,
    finetune_vision_layers     = False, # False if not finetuning vision layers
    finetune_language_layers   = True, # False if not finetuning language layers
    finetune_attention_modules = True, # False if not finetuning attention layers
    finetune_mlp_modules       = True, # False if not finetuning MLP layers

    r = 16,           # The larger, the higher the accuracy, but might overfit
    lora_alpha = 16,  # Recommended alpha == r at least
    lora_dropout = 0,
    bias = "none",
    random_state = 3407,
    use_rslora = False,  # We support rank stabilized LoRA
    loftq_config = None, # And LoftQ
    # target_modules = "all-linear", # Optional now! Can specify a list if needed
)
```

```
FastVisionModel.for_training(model) # Enable for training!

trainer = SFTTrainer(
    model = model,
    tokenizer = tokenizer,
    data_collator = UnslothVisionDataCollator(model, tokenizer), # Must use!
    train_dataset = converted_dataset,
    args = SFTConfig(
        per_device_train_batch_size = 2,
        gradient_accumulation_steps = 4,
        warmup_steps = 5,
        max_steps = 30,
        # num_train_epochs = 1, # Set this instead of max_steps for full training runs
        learning_rate = 2e-4,
        fp16 = not is_bf16_supported(),
        bf16 = is_bf16_supported(),
        logging_steps = 1,
        optim = "adamw_8bit",
        weight_decay = 0.01,
        lr_scheduler_type = "linear",
        seed = 3407,
        output_dir = "outputs",
        report_to = "none",     # For Weights and Biases

        # You MUST put the below items for vision finetuning:
        remove_unused_columns = False,
        dataset_text_field = "",
        dataset_kwargs = {"skip_prepare_dataset": True},
        dataset_num_proc = 4,
        max_seq_length = 2048,
    ),
)
```

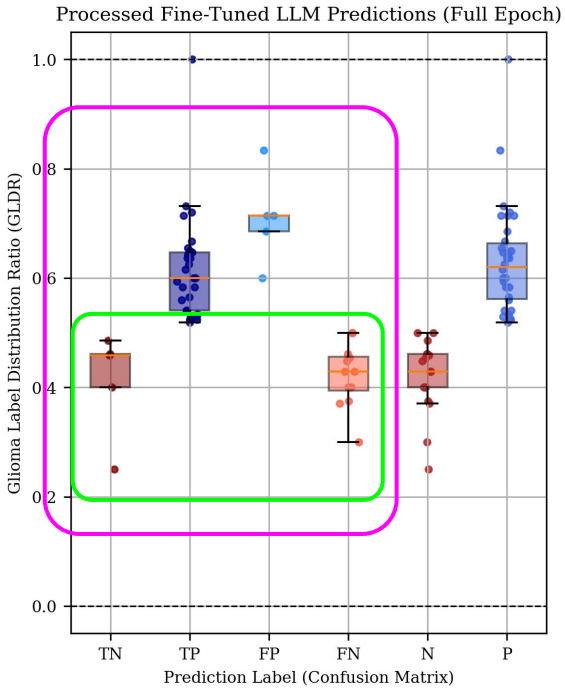
General and Fine-Tuned LLM Key Results



T/F HGG are all evenly distributed. LGG cases are rarely predicted and when they are, they are highly uncertain.

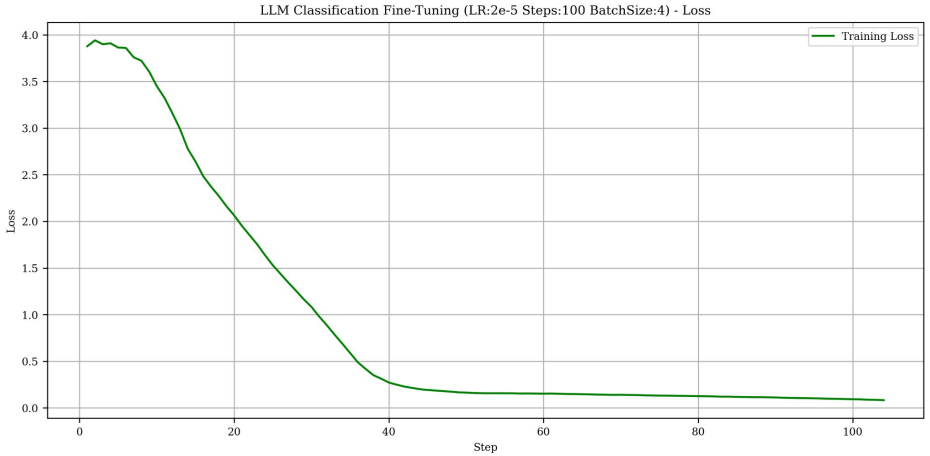


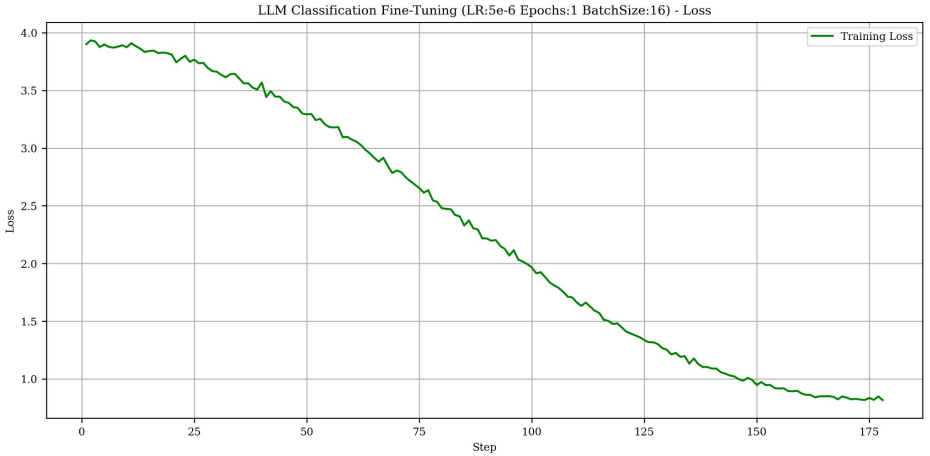
The fine-tuning creates more LGG predictions, but no evidence of learning distinction between classes.



Fine-tuning should bring HGG values closer to 1 and LGG values to 0. It instead made all predictions closer to 0.5.

Training Plots - Classification





Training Plots - Segmentation