

Assessing Visual Reasoning of Multimodal Language Models in Biomedical Applications

S. Purba, A. Melles, D. Hackel, I. Obeid and J. Picone

Neural Engineering Data Consortium, Temple University, Philadelphia, Pennsylvania, USA
{sadia.afirin.purba, anne-mai.melles, dmitry.hackel, iobeid, picone}@temple.edu

Abstract—Recent advances in multimodal large language models (MLLMs) have opened new possibilities for biomedical image interpretation without task-specific training. This study explores the zero-shot visual reasoning capabilities of a leading MLLM, the ChatGPT vision model, for two challenging biomedical image classification tasks: electroencephalogram (EEG) signal interpretation and digital pathology (DPATH) image diagnosis. In this work, datasets of single-frame and three-frame EEG images and breast cancer pathology patches were used to benchmark performance. We show that while zero-shot MLLMs lag specialized models in accuracy, ChatGPT’s vision model delivers moderate performance and meaningful explanations compared to popular supervised computer vision models (ViT, ResNet). We also apply parameter-efficient fine-tuning (PEFT) to an open-source MLLM (the Qwen model) to improve accuracy across both domains. We find that off-the-shelf ChatGPT (o3-mini-high) can serve as a strong baseline model for biomedical tasks, highlighting the potential for model adaptation through lightweight supervised fine-tuning. The integration of AI-generated reasoning can enhance explainability and decision-making in clinical contexts.

Keywords—multimodal large language models, generative artificial intelligence, EEG, digital pathology

1. INTRODUCTION

The interpretation of biomedical images, including EEG multichannel signals and high resolution digital histopathological images, remains a challenging task that is central to clinical practice. This complex task requires domain expertise, large amounts of high-quality annotated data, and robust modeling approaches. Modern systems using deep learning architectures, such as convolutional neural networks (CNNs) [1][2] and vision transformers (ViT) [3] have improved classification accuracy in domains such as cancer detection and seizure classification [4][5]. However, these approaches demand vast amounts of labeled training data and domain-specific fine-tuning, limiting their scalability and flexibility in low-resource settings. They also lack an ability to accurately segment data and localize findings, which prevents widespread use in clinical settings.

In contrast, MLLMs that integrate vision and language inputs promise zero-shot generalization: the ability to perform tasks outside their training domain, also known as in-context learning, when provided only with prompt-based guidance. These models can classify images using concise textual instructions or expert-provided examples,

without requiring additional task-specific training. In biomedical applications, in-context learning enables these models to generalize to unseen image types and diagnostic categories simply by presenting a few annotated examples or a detailed guideline along with some example images. In contrast, conventional supervised learning requires labeled datasets and explicit model fine-tuning for each new classification task. This paradigm shift offers a powerful new mechanism for knowledge transfer and supports flexible deployment of AI-driven decision support in clinical applications.

Recent literature provides empirical support for the zero-shot potential of MLLMs for biomedical image classification tasks. The systematic evaluation of GPT-4V demonstrated that in-context learning is sufficient to match or even outperform specialized neural networks on histopathology classification tasks involving colorectal and breast tumor images [6]. Zhu et al. [7] showed diagnostic accuracy of 77.01% for USMLE-style questions using GPT-4V and achieved an AUC of 76.80% in detecting abnormality in chest radiography with zero-shot prompting. Researchers also confirmed that vision-language approaches can handle classification, segmentation, report generation, and visual question answering (QA) with zero-shot performance across different domains such as histopathology and radiology with minimal domain-specific supervision [8].

Existing literature indicates that MLLMs, even without training on medical images, can achieve reasonable zero-shot accuracy and produce interpretable, clinically relevant explanations. However, performance depends strongly on prompt design, context, and few-shot examples. Although zero-shot accuracy is generally lower than that of domain-specific models, this approach greatly reduces the need for large, labeled datasets and costly fine-tuning.

In contrast to earlier work focused mainly on qualitative or accuracy-based comparisons, our study introduces statistical rigor to the MLLM evaluation process. We apply pairwise Z-tests to measure the significance of performance differences between models, showing that improvements from parameter-efficient fine-tuning (PEFT) are statistically significant at a 99% confidence interval. This approach provides a reliable and reproducible framework for evaluating MLLMs in biomedical imaging.

2. DATASET DEVELOPMENT

In this study, we evaluate the zero-shot visual reasoning capability of a ChatGPT vision model (o3-mini-high) [9] on two demanding biomedical classification tasks: seizure detection in EEG signals [10] and breast histopathology [5]. To begin, we briefly describe each corpus to underscore the importance of accurate segmentation in these tasks.

2.1. THE NATUS AMBULATORY CORPUS

A total of 104 EEG recordings were randomly selected from the Natus Ambulatory EEG Corpus (NAEG: v1.0.0) [12] maximizing the diversity of patients. This sampling method ensures that the dataset captures inter-patient variability in seizure and background activity. NAEG was used because these are continuous recordings that have been carefully annotated for seizure events.

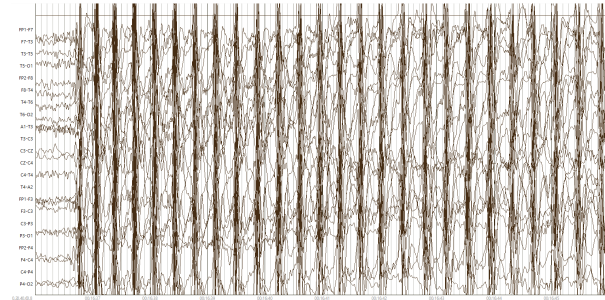
For this study, the EEG annotation process follows the guidelines developed by Ochal et al. [11] for the TUH EEG Seizure Corpus. EEG signals are annotated using a set of over 25 event types, including seizure subtypes (e.g., focal non-specific, tonic-clonic) and non-epileptic artifacts (e.g., muscle activity, eye blinks). Events are annotated with precise start/stop times and localized to specific channels.

Key criteria for identifying seizures include morphology (e.g., spike-and-slow-wave), rhythmicity, synchrony across channels of interest, evolution over time, and duration (> 3 secs for absence seizures or > 10 sec for other types of seizures). In this study, non-seizure events (e.g., artifacts, background rhythms) are grouped under a unified class named background (bckg).

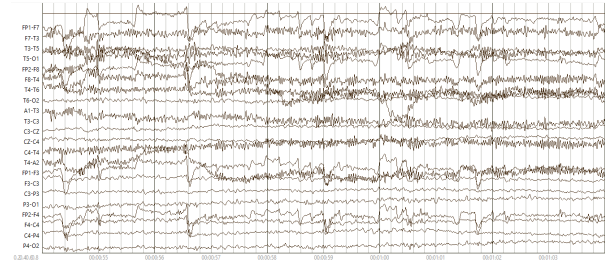
Figure 1(a) shows a typical example of an absence seizure. This is one of three seizure types we asked ChatGPT to identify. Absence seizures are characterized by an abrupt onset and a 3 Hz spike-and-slow-wave morphology. A typical example of “bckg” is also shown in Figure 1(b). This class includes any activity that is not ictal. In this example, we observe eye movements along with muscle artifacts.

2.2. THE TUH DIGITAL PATHOLOGY CORPUS

The TUH Digital Pathology (TUDP) Corpus [13] was developed to support machine learning in breast cancer pathology by providing a standardized, well-informed annotation of digital pathology slides. Annotation is divided into three main categories: non-cancerous, carcinogenic and benign structures. Non-cancerous labels include norm (normal ducts/lobules with intact architecture), bckg (background tissue such as stroma and adipose tissue), null (indistinguishable tissue), and artf (artifact like pen marks or dust). Cancerous labels include dcis (ductal carcinoma in situ characterized by high nuclear density and preserved ductal boundaries)



(a) a typical absence seizure (absz)



(b) a typical example of background (bckg)

Figure 1. Typical examples of annotated events in TUSZ

and indc (invasive ductal carcinoma, showing cellular growth into surrounding tissue, unconfined, and disorganized). Neoplastic-associated or benign labels encompass nneo (non-neoplastic changes such as fibrosis, hyperplasia, and intraductal papilloma), infl (inflammatory response to abnormalities), and susp (suspicious regions with atypical features). Images were created by an annotation team of trained undergraduate annotators, with oversight from pathologists. The staining methods include hematoxylin and eosin (H&E) and immunohistochemistry (IHC), which help differentiate tissue architecture.

The annotation process is described in detail in [5][14]. JPEG image patches that capture the annotated regions were used for model development. We selected 10-12 representative JPEG images per annotation label, resulting in a total of 101 images. This dataset was carefully curated to ensure diversity of histological stain, variation in tissue architecture, and clear depiction of the defining features of each label. Examples of three types of labels are shown in Figure 2.

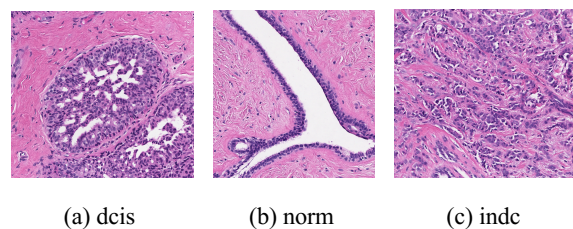


Figure 2. Typical examples of annotated events in TUDP

3. EXPERIMENTAL DESIGN

We designed our experimentation in two stages: (1) prompt engineering, and (2) parameter-efficient fine-tuning of an open-source MLLM. The end-to-end workflow is shown in Figure 3. First, we loaded biomedical image datasets and initialize a chat session with the annotation guidelines. Next, we presented one image query at a time with explicit instructions and incremented the annotation counter until we reach the target sample size. We then had experts review the responses from the chat sessions, curate a representative reasoning subset as a training set for fine-tuning and convert it into instruction tuning format. Finally, we initialized a pretrained multimodal LLM backbone, configure fine-tuning hyperparameters, and performed supervised fine-tuning across multiple epochs until the model completes the specified number of iterations.

For prompt engineering, we uploaded the annotation guideline, followed by a general system message. Next, we uploaded a query image with a specific prompt to guide ChatGPT's output. Our query prompt was:

Query Prompt: ' As an output, give me a dictionary as follows:

{*"label"*: label, *"reasoning"*: "Correct and accurate medical reasoning to classify the image, think of yourself as a cancer physician/EEG technologist, and give reasoning."}. Do not give any extra output. I repeat, do not give extra output. This is an important task for me. You will be penalized if you give the wrong label."

Empirical evidence [15] exists that suggests including specific phrases in the query prompt improves ChatGPT's response. This motivated us to add a warning: "You will be penalized if you give an incorrect label" at the end of the query prompt.

To evaluate the impact of fine-tuning, we conducted both quantitative and qualitative analysis against prompt-based zero-shot (ZS) responses as well as traditional deep learning models.

3.1. CASE STUDY: EEG

To evaluate ChatGPT o3neteu-mini-high's visual reasoning capabilities in EEG classification, we designed three experiments using annotated images from NAEg. Each experiment involved prompting ChatGPT with EEG screenshots and asking it to perform classification.

EXP. 1: SINGLE-FRAME, FOUR-WAY CLASSIFICATION

In this experiment, we classified single-frame EEG images into one of four classes: (1) generalized seizure (gnsz), (2) focal non-specific seizure (fnsz), (3) absence seizure (absz), or (4) background (bckg). We utilized 104 randomly sampled images (26 per class) and obtained 25% accuracy.

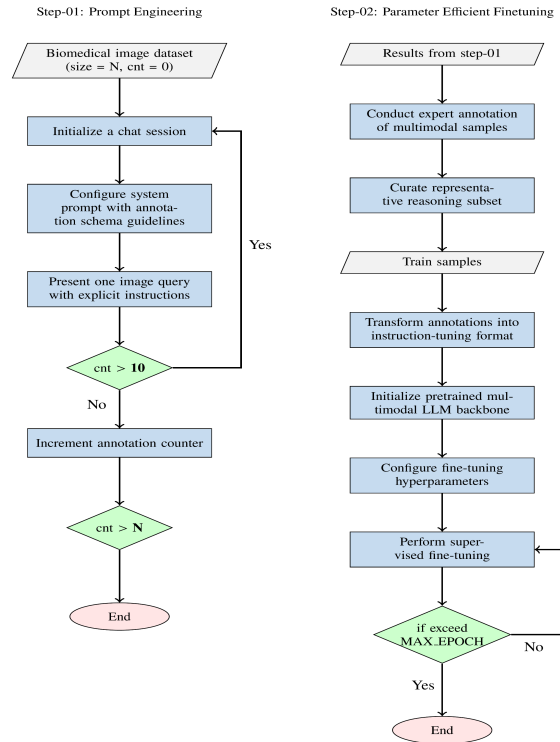


Figure 3. Workflow for prompt engineering and parameter-efficient fine-tuning in biomedical image reasoning

ChatGPT struggled to classify absz correctly, often mislabeling an event as musc (muscle artifact) or gped (generalized periodic epileptiform discharges), due to its expectation of seizure evolution. However, absz typically starts and stops abruptly without evolution. Gnsz were frequently misclassified as artf (artifact) or musc (muscle artifact), especially when the image came from the mid-ictal phase, where artifact content is high. Fnsz were frequently missed due to the difficulty detecting low-amplitude rhythmic activity, leading to misclassifications such as slow (slowing) or artf (artifact). Bckg were sometimes misinterpreted as fnsz, particularly when eye movements or blinks were present.

We hypothesize that the lack of temporal context in single frames limited ChatGPT's ability to recognize seizure evolution (pre- and postictal phases)

EXP. 2: SINGLE-FRAME, BINARY CLASSIFICATION

To analyze how the o3-mini-high model performs when we have binary cases, we use the same 104 images to classify either "seizure" or "no seizure" and obtained 49% accuracy. Performance improved compared to the first experiment due to the reduced classification complexity. Removing seizure subtype differentiation reduces confusion and allows accurate identification of clear ictal vs. non-ictal signals. Simplifying the label space helped ChatGPT focus on broad seizure indicators

like rhythmicity and morphology, improving performance.

EXP. 3: TEMPORAL-CONTEXT, FOUR-WAY CLASSIFICATION

To identify how zero-shot prompting is affected by context, we provided ChatGPT with three sequential EEG screenshots (preictal, ictal, and postictal) and asked for one of four labels: gnsz, fnsz, absz and bckg. For conducting this experiment, we used 300 images across 100 cases (75 seizure, 25 background) and obtained an accuracy of 35%.

Accuracy improved over the first experiment, confirming the benefit of temporal context. Gnsz were better recognized due to visible evolution. However, absence seizures remained a challenge. Despite an increase in context, abrupt onsets were frequently missed, and absz were confused with gnsz. Context helped in some fnsz cases where the buildup of focal rhythmic activity could be seen over time.

Temporal input improved classification, but challenges remained, particularly in capturing absz morphology and onset cues. This suggests that future prompts might benefit from emphasizing absz characteristics explicitly or adjusting visualization (e.g., channel sensitivity).

ChatGPT achieved a high degree of accurate reasoning when classifying gnsz as follows:

"The EEG demonstrates a clear evolution of rhythmic epileptiform activity across nearly all channels, beginning around 00:02:00 with low-amplitude rhythmic theta activity that builds in amplitude and frequency into high-amplitude, rhythmic discharges by 00:02:03 – 00:02:05. This widespread symmetrical involvement without localization suggests a generalized onset. The morphology — progressing from lower-voltage rhythmic waves to high-voltage rhythmic spike-and-wave bursts and then abruptly terminating — fits the pattern of a generalized seizure. The absence of a tight 3 Hz spike-and-wave pattern excludes typical absence, and there is no focal evolution to indicate a partial seizure. Therefore, the appropriate classification is a generalized non-specific seizure (GNSZ)." (Experiment 1)

ChatGPT's reasoning was inadequate when classifying absz as gnsz:

"The high-amplitude spike-and-wave bursts emerge nearly simultaneously across bilateral frontal, temporal, and posterior channels with clear synchrony and rhythmicity. This pattern is characteristic of a generalized epileptic seizure rather than a focal discharge or artifact, as it does not remain confined to a single region, nor does it display the slower, asymmetric morphology of eye-movement or muscle-artifact." (Experiment 1)

Although all absence seizures are generalized, as they involve synchronous activity across all channels, ChatGPT should have been precise in this case. Absence

seizures are specifically characterized by 3 Hz spike-and-slow-wave discharges and an abrupt onset. Therefore, this seizure should not be classified merely as a generalized seizure, but more specifically as an absence seizure as stated in the guidelines.

ChatGPT was nonviable when classifying bckg as fnsz and gave the following reasoning:

"These three snapshots show rhythmic spike-and-slow-wave complexes that arise focally — most prominently in the left frontotemporal leads (Fp1–F7, F7–T3) — and then evolve in frequency and morphology over more than 10 seconds before spreading. The clear focal onset with subsequent evolution without immediate bilateral synchronous involvement is characteristic of a focal non-specific seizure (fnsz)." (Experiment 3)

The activity that ChatGPT classified as rhythmic spike-and-slow-wave complexes in the frontal channels is, in fact, simply eye blinks – normal background activity for the patient. While eye blinks can appear rhythmic, they lack evolution and typically do not exceed 2.5 Hz, which is necessary for seizure classification. These distinctions are clearly outlined in the guidelines, indicating that ChatGPT misinterpreted the features in this case.

ANNOTATOR ANALYSIS

From an annotator's perspective, ChatGPT's reasoning consistently demonstrated partial understanding of EEG patterns – highlighting concepts like spike-wave morphology, synchrony, and regional distribution. However, its limitations include: (1) rigid assumptions about seizure evolution (especially problematic for absz), (2) difficulty distinguishing low-frequency artifacts from ictal activity, and (3) lack of adaptability to typical EEG conventions such as sensitivity adjustments.

Despite these shortcomings, the model's structured reasoning outputs were often clinically interpretable and occasionally insightful.

3.2. CASE STUDY: DPATH

To assess the performance of the multimodal language model within the DPATH framework, 101 high-resolution images and standardized annotation guidelines were used. The model is prompted to generate three outputs: (1) whether the image contains cancer or not, (2) the appropriate diagnostic label from the predefined classes, and (3) a concise, pathology-informed reasoning for its decision. No additional output and utilization of outside sources were permitted.

After review of 101 unknown images, o3-mini-high returned accurate cancer vs non-cancerous prediction 54% of the time and achieved correct classification of the nine diagnostic labels in 48% of the cases. In 40% of the reasoning outputs, evaluation shows that o3-mini-high is successful at providing detailed, histologically accurate

rationales that directly correspond to the structures visible in the image. For example, in one correct prediction of both label and cancerous vs non-cancerous inflammation, the model accurately stated:

“The image displays a dense inflammatory infiltrate composed predominantly of lymphocytes within the fibrous stroma, without any clear evidence of ductal structures or atypical epithelial proliferation. There is no architectural distortion, mitotic activity, or cytologic atypia suggestive of malignancy. This pattern is consistent with a benign inflammatory process, not cancer.”

The response not only identified the dominant cellular population and its context within stroma but also appropriately ruled out malignant characteristics through negative reasoning, demonstrating an understanding of the histopathological features associated with abnormal tissue architecture and lymphocyte presence.

In some cases, the o3-mini-high model correctly distinguished between cancerous cells confined within a duct (dcis), compared to cancer invasion of surrounding tissue (indc). This distinction represents an important diagnostic feature of breast cancer, and it was unexpected that an MLLM could learn to recognize it without explicit task-specific training. In one reasoning, the model correctly identified the differentiation between the borders of the basement membrane, surrounding tissue and cancerous cells, stating:

“This image shows a duct completely filled with a monomorphic population of atypical epithelial cells, demonstrating cribriform and solid architecture with clear preservation of the basement membrane — classic features of DCIS. The nuclei are enlarged, irregular, and hyperchromatic, consistent with neoplastic changes. There is no evidence of stromal invasion, so this is non-invasive but still classified as cancer (in situ).”

The model’s ability to generate decisions and demonstrate accurate identification of subtle histomorphological features is notable. The ability to differentiate the ductal unit from the surrounding stroma, as well as distinct cellular boundaries and tissue organization, reflects a level of understanding that closely aligns with annotators’ perspective. When describing a normal ductal formation, the model properly identifies three separate aspects of normal ductal formations to contextualize its decision:

“This image depicts small ductal and lobular units embedded in fibrous stroma. The ducts and lobules show bland nuclear morphology with clear luminal spaces, no architectural distortion, and preservation of the dual cell layers (epithelial and myoepithelial). The stromal background is unremarkable. There are no cytologic atypia or features of neoplasia. This is consistent with normal breast tissue histology.”

However, the model’s output was not uniformly reliable. Several responses included an improper description of

tissue architecture. A common error pattern involved the misclassification of benign, non-neoplastic structures as malignant cells. Identifying any abnormality as either suspicious (susp label or labeling it as cancer). In these incorrect outputs, o3-mini-high’s reasoning accurately described the characteristics associated with its predicted label but failed to reflect the actual features present in the image. When doing the analysis of a single normal duct, ChatGPT o3 incorrectly predicted dcis and properly described ductal carcinoma in situ, but the models’ reasoning did not speak to the image. One reasoning showing this is:

“The image reveals invasive epithelial structures with irregular, angulated glands infiltrating a fibrotic stroma. There is an absence of a continuous myoepithelial layer, nuclear atypia, and loss of normal ductal architecture, consistent with invasive ductal carcinoma (indc). These features indicate malignant cells breaching the basement membrane and invading surrounding tissue.”

In this instance, though the model correctly identified cellular atypia, it frequently failed to distinguish between benign and malignant processes. ChatGPT o3’s predictions on the DPATH dataset generally fell into five categories: (1) correct identification of the primary tissue structure; (2) correct recognition of secondary features such as background or artifact, while failing to classify the main diagnostic structure; (3) misclassification of non-neoplastic structure as malignant; (4) incorrect labeling of an image with different or darker immunohistochemical stains as cancerous; and (5) complete mischaracterization of the content and label.

It is important to note that all the images only contained a single morphological structure generally surrounded by background stromal or adipose tissue. These errors may prevent it from generalizing to complex contexts, and its error rate is expected to increase if asked to analyze larger images or whole slide images with a greater frequency of morphological features.

ANNOTATOR ANALYSIS

From an annotator’s perspective, instances in which ChatGPT correctly identifies pathological images are impressive, particularly due to the depth and accuracy of its reasoning. However, despite evaluating images of the same label and structurally similar features, the model frequently succeeds on one instance while failing on another. This suggests that the ChatGPT interpretative framework lacks generalizability across subtle changes in tissue architecture, falling into one of the error categories above. The reasoning behind the errors made does not align with the structural element in the image. Therefore, its ability to characterize pathological images appears to vary substantially across cases. The inconsistencies show limited model interpretability, emphasizing alignment between human expertise and computational reasoning.

4. FINE-TUNING USING EXPERT REASONING

We use the expert-validated subset from the previous step as a parameter-efficient fine-tuning training dataset for the Qwen2-VL [16] model. Following best practices for instruction-tuning large multimodal models [17], each training example was structured using a conversational style: a system message provided the model with the instruction of the classification task and valid label set, a user message supplied the tissue image and a query prompt, and an assistant message containing the correct reasoning response as validated by the domain experts and true class label. For both the EEG and DPATH domains, we have followed the same data formatting and training technique. One example of the preprocessed training sample from DPATH dataset is given below:

System Prompt: "You are a medical professional specialize in detecting cancer from pathology images. Based on the provided pathology images and query, you detect if the image contains label:{CLS_NAMES}. Also you provide a factual reasoning about your decision."

Query Prompt: "Given the tissue image, classify it into one of the following categories: {CLS_NAMES}. Provide a valid JSON dictionary as output with two keys: 'label': the correct classification label for the digital pathology tissue image. 'reasoning': detailed clinical reasoning supporting the classification, written as a pathologist would explain, referencing histologic patterns, spatial distribution of cells and structures, cellular morphology, and other relevant tissue features. Output Format Example: {'label': 'dcis', 'reasoning': 'The image shows enlarged, atypical nuclei and increased cell density within ductal structures, without stromal invasion—findings consistent with ductal carcinoma in situ.'}"

Assistant Response: {'label': 'right class', 'reasoning': 'valid reasoning text'}"

We train the base model using a technique called Low-Rank Adaptation (LoRA) [18], which is an effective parameter-efficient fine-tuning technique for large language and vision models. Instead of updating all the model's parameters, LoRA only adds and updates small, trainable adapters inside specific parts of the network. For this experiment, we used Qwen2-VL-7B Instruct [16] as our base model, loaded it in an efficient 4-bit quantized format to optimize memory usage, and applied LoRA adapters to its cross-attention layers. The configuration for LoRA was set to a rank of 8, alpha of 16, and dropout of 0.05, balancing training efficiency and capacity. After applying LoRA, 2.5M parameters were set as trainable out of the total approximately 8.3B parameters, meaning that only about 0.03% of the model's parameters were updated during fine-tuning.

We trained the model for 20 epochs, using a batch size of one and gradient accumulation steps of two. The learning rate was set at $2e-04$, a value chosen for stable training in low-data regions. We also enable gradient checkpointing

and mixed-precision training (using bfloat16 precision), both of which helped keep memory usage low and speed up training. All the training and evaluation were done on a single NVIDIA GPU where the process took 12 GB of memory, along with about 4 GB of system RAM. Each epoch took around 2,212 seconds to complete. During response generation, we used a max token size of 1,024, and each response took around 48 seconds.

In Table 1, we summarize several key insights about the comparative strengths of MLLMs (o3-mini-high and Qwen) and traditional computer vision models (ResNet, ViT) on complex biomedical classification tasks. For EEG, we use two classes by collapsing "gnsz", "absz" and "fnsz" into "seiz", alongside "bckg". From the prompt engineering (Step-1), we obtained expert validated 25 correct reasoning texts generated by o3-mini-high model. In these training samples, the label distribution is: bckg (13) and seiz (12).

To evaluate the robustness of the observed performance differences, we present pairwise Z-tests across all model combinations for both datasets in Figure 4. The darker cells represent higher p-values while "n.s." denotes non-significant results. For EEG, the fine-tuned Qwen model significantly outperformed both its zero-shot variant ($p = 0.0024$) and the pretrained ResNet baseline ($p = 0.0024$). For DPATH, most inter-model comparisons were significant at a 99% confidence interval, demonstrating clear separability among multimodal and domain-specific models.

For EEG, the accuracy of the zero-shot (ZS) variant of Qwen, which was 74% (77/104), matches the pretrained version of ResNet. We used a ResNet model pretrained on ImageNet. However, fine-tuning performed worse

Table 1. Performance comparison of zero-shot (ZS) prompting, a pretrained model (PT) and a domain-specific fine-tuned (FT) model

Data	System	Alg	Acc (%)	Prec (%)	Rec (%)	AUC (%)
EEG	o3	ZS	49.04	54.84	55.96	55.96
	Qwen	ZS	74.04	37.02	50.00	50.00
	Qwen	FT	53.85	61.76	64.02	64.02
	ResNet	PT	74.04	37.02	50.00	50.00
	ResNet	FT	100.00	100.00	100.00	100.00
DPATH	o3	ZS	48.51	45.53	42.03	74.22
	Qwen	ZS	5.94	1.15	14.29	48.64
	Qwen	FT	28.71	48.25	38.20	62.43
	ResNet	PT	16.83	3.37	11.81	47.08
	ResNet	FT	68.32	77.13	59.31	76.23
	ViT	PT	14.85	11.41	12.85	47.82
	ViT	FT	75.25	79.16	70.27	82.45

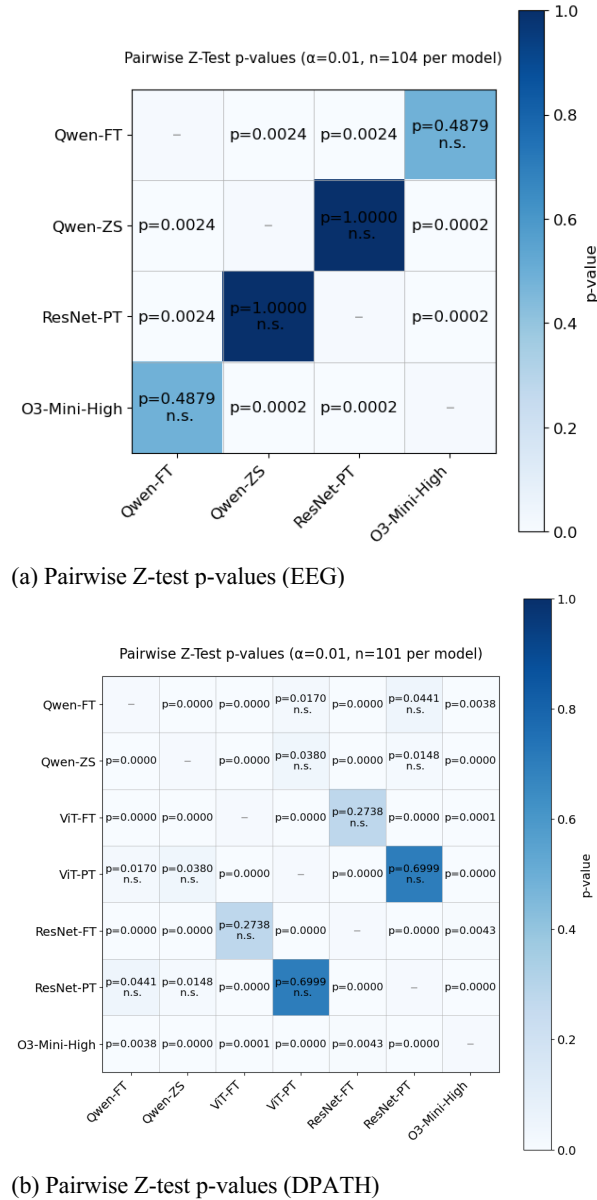


Figure 4. Statistical significance tests

than the ZS version for the Qwen model because the training samples were limited. During evaluation, we used 104 samples, with the majority belonging to “seiz” (77 samples). The Qwen ZS version predicted all evaluation samples as “seiz,” leading to high accuracy. In contrast, the fine-tuned version correctly predicted 33 samples as “seiz” and 23 samples as “bckg”, with the remaining samples misclassified. The precision score for the fine-tuned and zero-shot versions was 61% and 37%, respectively, which suggests that the fine-tuned Qwen model is a better classifier than the ZS version.

For DPATH, we observe a similar pattern, with accuracy improving after fine-tuning (from 6% ZS to 29%). From

Step-1, we collected 48 correct reasoning texts, with the majority belonging to “indc” (14), followed by “bckg” (10), “norm” (8), “infl” (7), “dcis” (6), and “nneo” (3). Both the ZS and FT versions show a bias toward “dcis” as it was the most accurate class prediction during evaluation, while “nneo” was the most misclassified class, which is expected due to the limited number of training samples. The AUC score of the Qwen fine-tuned version (62%) is also comparable to that of the ResNet-FT model (76%), suggesting that PEFT can achieve statistically significant discriminative capability even with limited training data.

Another interesting observation for DPATH dataset is that the zero-shot (ZS) performance of o3-mini-high and Qwen is comparable to, and in some cases exceeds, that of the pretrained ViT and ResNet models. Similar to the MLLM models, ViT and ResNet are trained on natural images rather than biomedical images. The o3-mini-high ZS model achieved an AUC of 74%, which is substantially higher than the pretrained ResNet (47%) and ViT (48%) models. Pairwise Z-test analysis also confirms that this improvement is statistically significant at the 99% confidence interval, indicating the robustness of the observed difference. Consequently, the o3-mini-high ZS model should be considered as a baseline for future biomedical image classification studies, besides conventional architectures such as ResNet and ViT.

In both datasets, the domain-specific deep learning models perform best, as they were trained with a large number of EEG and DPATH images from classification tasks. We developed these domain-specific models upon using our open-source EEG [19] and digital pathology toolboxes [13].

5. SUMMARY

In this paper, we have:

- introduced a two-step approach combining prompt engineering and parameter-efficient fine-tuning to assess and enhance the reasoning abilities of MLLMs;
- developed a structured ZS prompting framework that uses annotation guidelines to provide ChatGPT with additional context;
- benchmarked MLLMs against domain-specific supervised vision models (ResNet and ViT);
- conducted expert reviews of the models’ reasoning outputs to evaluate interpretability and provide actionable clinical insights;
- performed pairwise Z-test analyses to assess statistical significance between models, confirming that improvements from PEFT are significant at the 99% confidence level.

Our study demonstrates that multimodal large language models (MLLMs) such as ChatGPT-o3 and Qwen can perform ZS biomedical image classification and generate clinically relevant reasoning, even without domain-

specific training. Our findings show that, while MLLMs currently lag supervised models in raw accuracy, they can provide statistically significant classification results along with insightful and human-interpretable explanations. In particular, the ChatGPT model excelled in digital pathology tasks, showing clear differentiation between cancerous and non-cancerous features in breast tissue images.

However, the model demonstrated reduced effectiveness in interpreting EEG signals, likely due to the need for temporal and contextual reasoning. The reason behind this difference is that DPATH images offer visually rich spatially organized patterns that align well with the capabilities of MLLMs, while EEG interpretation requires nuanced context and an understanding of underlying signal dynamics, which current visual language models do not fully capture. Experimenting further with MLLMs for EEG analysis remains valuable, as advances in model architecture and prompt design could improve their ability to reason over signal data.

Our findings validate that the improvements achieved through parameter-efficient fine-tuning are statistically meaningful within the limits of small-sample evaluation. However, given the dataset sizes (EEG = 104, DPATH = 101), we explicitly frame these results as a proof-of-concept rather than conclusive evidence. We propose that future work should integrate model-generated reasoning as a structured, explainable output, enabling experts to validate and fine-tune model decisions. In addition, we plan to expand the dataset scale and diversity, implement k -fold cross-validation, and develop a systematic prompt-engineering framework to reduce human bias and enhance reproducibility.

ACKNOWLEDGMENTS

This material is based on work supported by several organizations over the years including the National Science Foundation (grants nos. 2211841 and 1726188 and 1925494), the Temple University Catalytic Collaborative Funding Initiative and most recently by the Pennsylvania Breast Cancer Coalition Breast and Cervical Cancer Research Initiative. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations.

REFERENCES

- [1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, Nevada, USA, 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.
- [2] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in *Proceedings of the International Conference on Machine Learning (ICML)*, K. Chaudhuri and R. Salakhutdinov, Eds., in *Proceedings of Machine Learning Research*, vol. 97. Long Beach, California, USA: PMLR, May 2019, pp. 6105–6114. url: <http://proceedings.mlr.press/v97/tan19a.html>.
- [3] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, Vienna, Austria: OpenReview.net, 2021, pp. 1–21. url: <https://iclr.cc/virtual/2021/oral/3458>.
- [4] V. Khalkhali, N. Shawki, V. Shah, M. Golmohammadi, I. Obeid, and J. Picone, "Low Latency Real-Time Seizure Detection Using Transfer Deep Learning," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, I. Obeid, I. Selesnick, and J. Picone, Eds., Philadelphia, Pennsylvania, USA: IEEE, 2021, pp. 1–7. doi: 10.1109/SPMB52430.2021.9672285.
- [5] D. Hackel et al., "Enabling Microsegmentation: Digital Pathology Corpora for Advanced Model Development," in *Signal Processing in Medicine and Biology: Applications of Artificial Intelligence in Medicine and Biology*, vol. 1, New York City, New York, USA: Springer, 2026, p. 50. [Online]. Available: https://isip.piconepress.com/publications/book_sections/2026/springer/dpath/ (in publication).
- [6] D. Ferber et al., "In-context learning enables multimodal large language models to classify cancer pathology images," *Nature Communications*, vol. 15, no. 1, p. 10104, Nov. 2024, doi: 10.1038/s41467-024-51465-9.
- [7] L. Zhu et al., "Step into the era of large multimodal models: a pilot study on ChatGPT-4V(ision)'s ability to interpret radiological images," *Int J Surg*, vol. 110, no. 7, pp. 4096–4102, Mar. 2024, doi: 10.1097/JS9.0000000000001359.
- [8] R. Al Saad et al., "Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook," *J Med Internet Res*, vol. 26, p. e59505, Sep. 2024, doi: 10.2196/59505.
- [9] "OpenAI o3-mini." Accessed: Aug. 01, 2025. [Online]. Available: <https://openai.com/index/openai-o3-mini/>.
- [10] J. Picone, "The Temple University Hospital EEG Corpus," in *International Congress of Clinical Neurophysiology, Geneva, Switzerland: International Federation of Clinical Neurophysiology*, 2022, p. 1. url: www.isip.piconepress.com/publications/presentations_misc/2022/iccn/tuh_eeg/.
- [11] D. Ochal, S. Rahman, S. Ferrell, T. Elseify, I. Obeid, and J. Picone, "The Temple University Hospital EEG Corpus: Annotation Guidelines," Temple University, Philadelphia, Pennsylvania, USA, 2020. url: www.isip.piconepress.com/publications/reports/2020/tuh_eeg/annotations.
- [12] A.-M. Melles et al., "Annotation of Ambulatory EEGs," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, Philadelphia, Pennsylvania, USA: IEEE, Dec. 2024, pp. 1–4. doi: 10.1109/SPMB62441.2024.10842264.
- [13] M. Bagritsevich, J. Picone, and I. Obeid, "The TUH Digital Pathology Corpus." [Online]. Available: www.isip.piconepress.com/projects/nedc/html/tuh_dpath/.
- [14] S. S. Shalamzari et al., "Big Data Resources for Digital Pathology," in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium*, Philadelphia, Pennsylvania, USA: IEEE, 2023, pp. 1–19. doi: 10.1109/SPMB59478.2023.10372721.
- [15] C. Li et al., "Large Language Models Understand and Can be Enhanced by Emotional Stimuli," Nov. 12, 2023, arXiv: arXiv:2307.11760. doi: 10.48550/arXiv.2307.11760.
- [16] A. Yang et al., "Qwen2 Technical Report," Sep. 10, 2024, arXiv: arXiv:2407.10671. doi: 10.48550/arXiv.2407.10671.
- [17] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," in *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine,

- Eds., Curran Associates, Inc., 2023, pp. 34892–34916. doi: [10.48550/arXiv.2304.08485](https://doi.org/10.48550/arXiv.2304.08485).
- [18] E. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” *International Conference on Learning Representations*, 2022. doi: [10.48550/arXiv.2106.09685](https://doi.org/10.48550/arXiv.2106.09685).
- [19] S. Rahman, M. Miranda, I. Obeid, and J. Picone, “Software and Data Resources to Advance Machine Learning Research in Electroencephalography,” in *Proceedings of the IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, I. Obeid, I. Selesnick, and J. Picone, Eds., Philadelphia, Pennsylvania, USA: IEEE, 2019, pp. 1–4. doi: [10.1109/SPMB47826.2019.9037851](https://doi.org/10.1109/SPMB47826.2019.9037851).