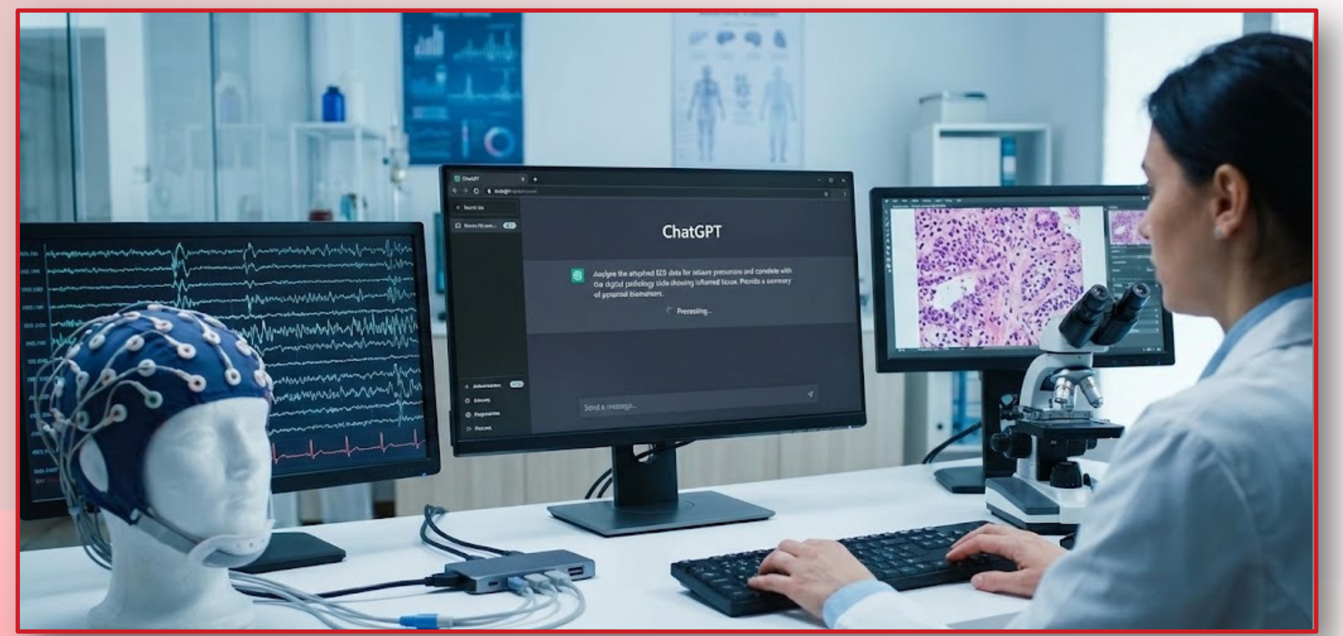
Assessing Visual Reasoning of Multimodal Language Models in Biomedical Applications



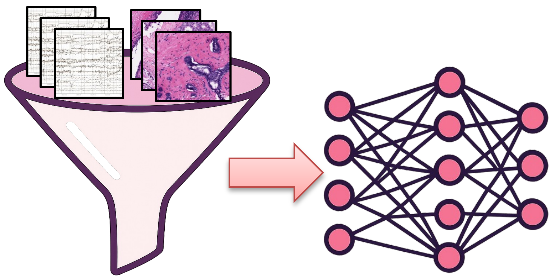
S. Purba, A. Melles, D. Hackel, I. Obeid and J. Picone

Neural Engineering Data Consortium
Temple University

NEURAL ENGINEERING
DATA CONSORTIUM

Motivation

Current Challenges: Deep Learning (CNNs, ViTs)



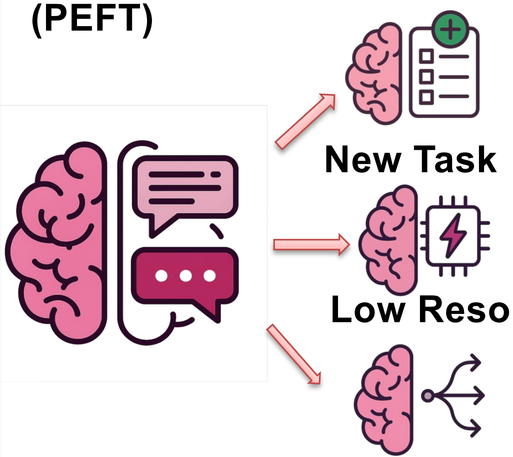
Large Labeled Datasets

Task Specific Models

- × Needs massive data
- × Lacks flexibility
- × Limited explainability

Multimodal Large Language Models (MLLMs)

In-Context Learning (ICL)
Parameter Efficient Fine-tuning (PEFT)



New Task

Low Resource

Generate Reasoning

- × No retraining is needed

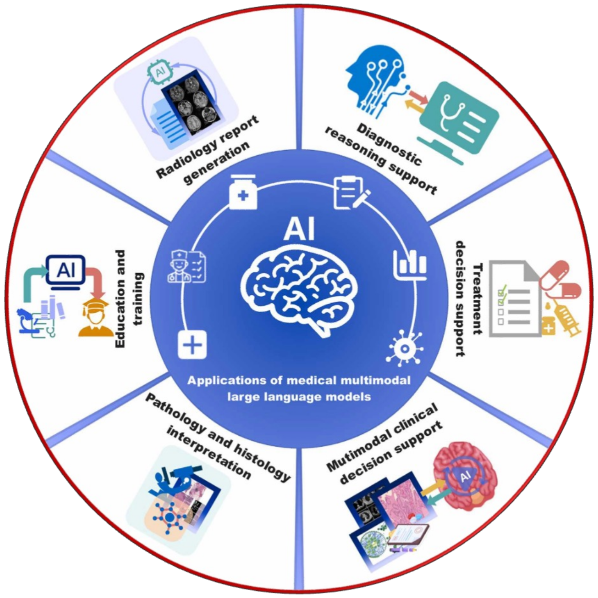
Benchmarking Need

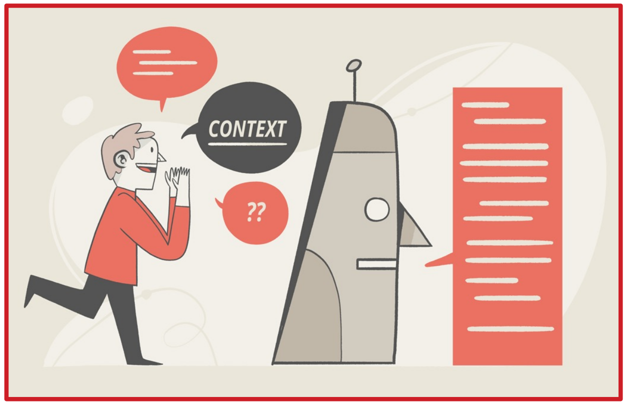
- Systematic benchmarking of MLLMs vs. domain-specific models
- Statistically Rigorous Comparisons (Z-test)

Impacts: MLLMs show promise as baselines; PEFT significantly enhances performance & clinical utility.

In-Context Learning (ICL)

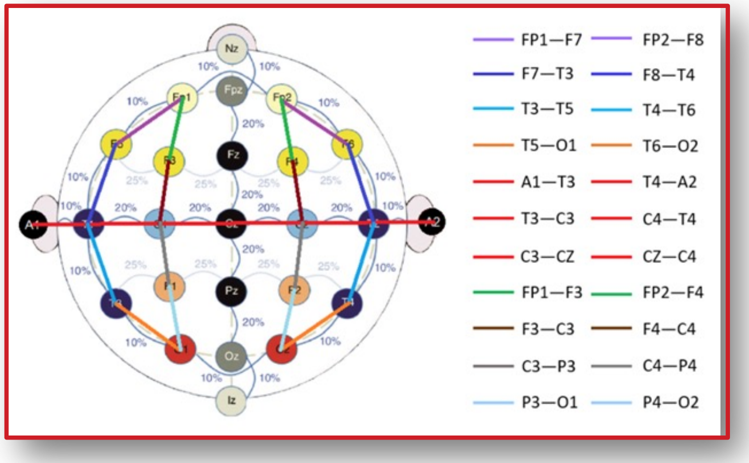
- Models perform new tasks by interpreting instructions and examples in the prompt, without updating model weights.
- For vision, MLLMs can:
  - Accept both images + text as input
  - Produce class labels and natural language reasoning as output.
- As a result, organizations can deploy powerful models without the cost and time of domain-specific fine-tuning.


Radiology report generation
Diagnostic reasoning support
Treatment decision support
Mutimodal clinical decision support
Pathology and histology interpretation
Education and training
AI
Applications of medical multimodal large language models



- In biomedical imaging, this allows:
  - Zero-shot classification with only guidelines and a few examples
  - Rapid adaptation to new image types and label sets.
- Prior work shows GPT-4V can match or surpass specialized networks on some histopathology tasks and radiology question answering under carefully designed prompts.
- This makes large models more clinically useful, especially in settings with limited annotations or rapidly changing imaging protocols.

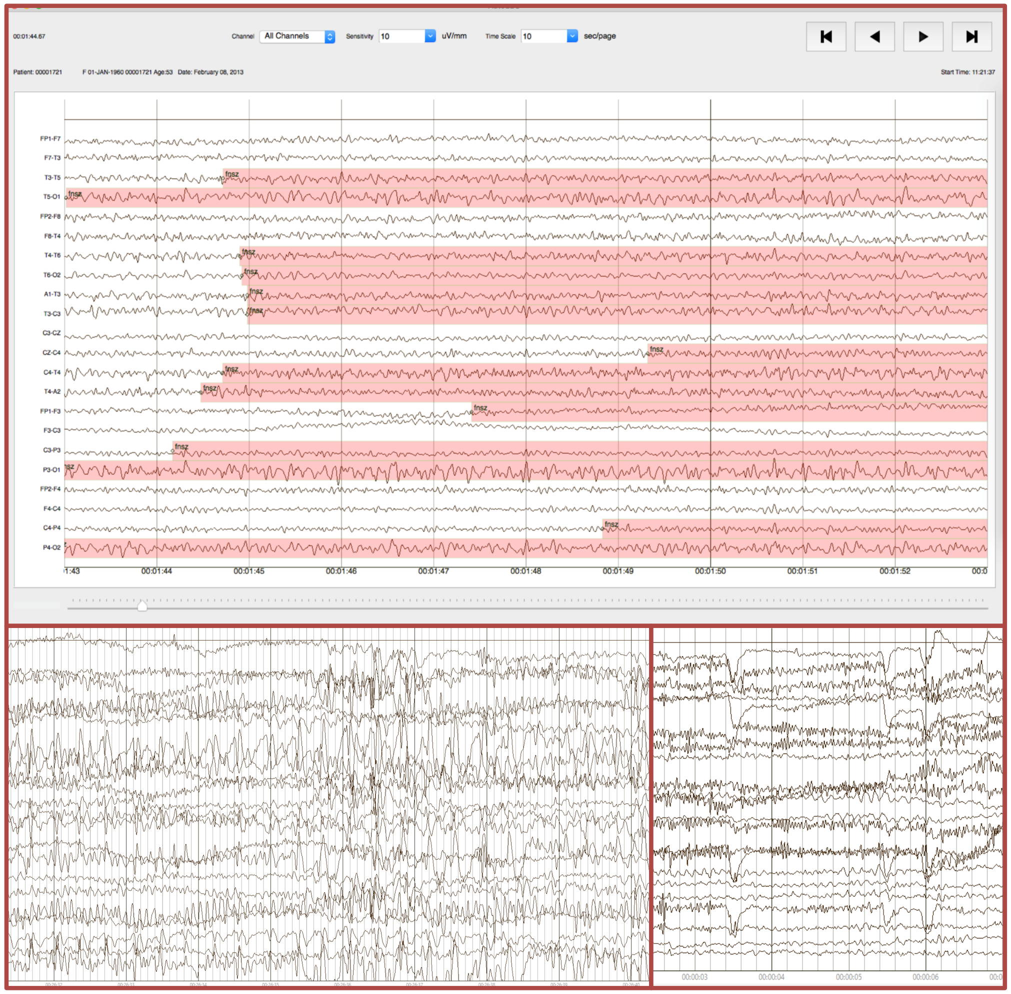
The Natus Continuous EEG Corpus



- **Data source: Natus Ambulatory EEG Corpus (NAEG v1.0.0).**
- **104 EEG images sampled to maximize patient diversity and seizure/background variability.**
- **Annotation follows TUH EEG Seizure Corpus guidelines:**
  - ❑ **25 event types including multiple seizure subtypes and non-epileptic artifacts.**
  - ❑ **Events annotated with start/stop times and channel localization.**
- **Key seizure criteria:**
  - ❑ **Morphology (e.g., spike-and-slow-wave)**
  - ❑ **Rhythmicity and synchrony**
  - ❑ **Evolution over time**
  - ❑ **Duration (≥ 3 s for absence; ≥ 10 s for other seizures).**
- **Decade of experience in EEG annotation and more than 11K subscribers of the TUH EEG dataset.**
- **For this study, non-seizure events are collapsed into background (bckg); seizure classes include generalized seizure (gnsz), focal non-specific seizure (fnsz), absence seizure (absz).**

The EEG Annotation Process

- **EEG signals are annotated with our open-sourced NEDC EEG annotation tool (Python-based, integrated with cohort retrieval).**
- **We take screenshots of the EEG signals and use these image frames in our experiments.**
- **Do not use raw time-series EEG because:**
  - ❑ **To test the visual reasoning capability of MLLMs using EEG images**
  - ❑ **Mirroring how neurologists visually inspect EEGs to decide whether a seizure is present or not.**
- **Before querying the MLLMs, we carefully designed the prompts so that no ground-truth label names or hints were exposed, ensuring no label leakage into the model inputs.**



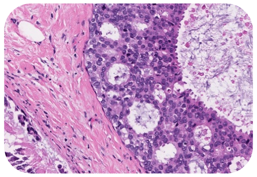
00:01:44.67

Channel: All Channels Sensitivity: 10 uV/mm Time Scale: 10 sec/page

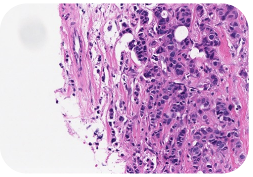
Patient: 00001721 F 01-JAN-1960 00001721 Age:53 Date: February 08, 2013 Start Time: 11:21:37

The Digital Pathology Annotation Process

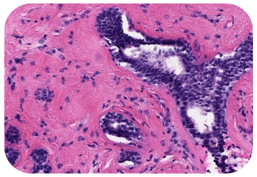
- **Data source:** TUH Digital Pathology (TUDP) Corpus of breast tissue slides.
- **Label taxonomy:**
  - ❑ Non-cancerous: normal (norm), background (bckg), null, artifacts (artf)
  - ❑ Cancerous: ductal carcinoma in situ (dcis), invasive ductal carcinoma (indc)
  - ❑ Neoplastic-associated/benign: non-neoplastic (nneo), inflammation (infl), suspicious (susp)
- **Imaging characteristics:**
  - ❑ Image patches extracted from whole slides
  - ❑ Window size: 1024x1024, frame size: 512x512
  - ❑ Curated to capture diverse architectures and clear label-defining features.
- **Study subset:** 10–12 images per label, total 101 images across 9 classes.



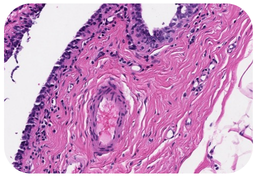
dcis



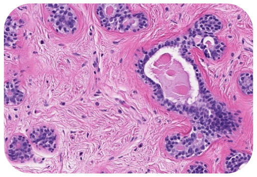
indc



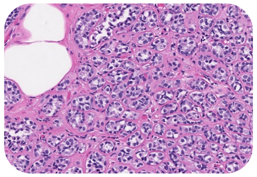
nneo



infl

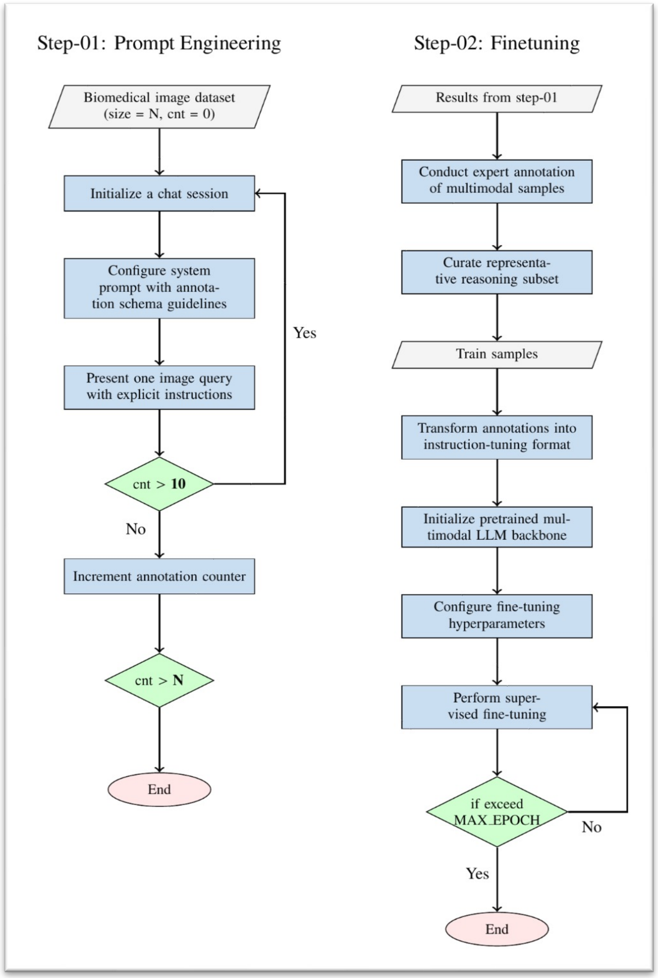


norm



susp

Experimental Design



Step-01: Prompt Engineering

Biomedical image dataset (size = N, cnt = 0)

Initialize a chat session

Configure system prompt with annotation schema guidelines

Present one image query with explicit instructions

cnt > **10**

Yes

No

Increment annotation counter

cnt > **N**

End

Step-02: Finetuning

Results from step-01

Conduct expert annotation of multimodal samples

Curate representative reasoning subset

Train samples

Transform annotations into instruction-tuning format

Initialize pretrained multimodal LLM backbone

Configure fine-tuning hyperparameters

Perform supervised fine-tuning

if exceed MAX_EPOCH

No

Yes

End

- **Two-stage workflow:**
  - ❑ **Prompt engineering + zero-shot evaluation with ChatGPT o3-mini-high.**
  - ❑ **Parameter-efficient fine-tuning (PEFT) of Qwen2-VL using expert-validated reasoning.**
- **Stage 1 (prompt engineering):**
  - ❑ **Load EEG/DPATH images and guidelines.**
  - ❑ **Initialize chat with system instructions.**
  - ❑ **Present one image per query with the structured JSON prompt.**
  - ❑ **Collect model label + reasoning; have experts review and curate correct reasoning samples.**
- **Stage 2 (fine-tuning):**
  - ❑ **Convert curated samples into instruction-tuning format (system / user / assistant messages).**
  - ❑ **Initialize Qwen2-VL-7B Instruct backbone.**
  - ❑ **Apply LoRA-based PEFT and train on expert-validated reasoning examples.**
  - ❑ **Evaluate fine-tuned Qwen and compare to ChatGPT zero-shot, ResNet, and ViT baselines.**

Case Study: EEG I (Tasks & Quantitative Results)

- **Goal:** Assess visual reasoning of ChatGPT o3-mini-high model for EEG images
- Three experiments using images from NAEG:

Exp. 1 – Single-frame, 4-way classification

- Classes: gnsz, fnsz, absz, bckg
- 104 images (26 per class)
- ChatGPT o3-mini-high accuracy: 25%.

Exp. 2 – Single-frame, binary classification

- Labels collapsed to “seizure” vs “no seizure”
- Same 104 images
- Accuracy improves to 49%.

Exp. 3 – Temporal-context, 4-way classification

- Input: 3 consecutive frames (pre-ictal, ictal, post-ictal) per case
- 100 cases (75 seizure, 25 background; 300 images)
- Accuracy: 35%.

- Temporal context helps but does not solve main challenges; absence seizures remain especially difficult.

Case Study: EEG II (Qualitative Reasoning & Failure Modes)

✓ STRENGTHS IN REASONING

- **Rich temporal descriptions:** Onset, buildup, widespread, abrupt termination.
- **Clinically relevant concepts:** Spike-and-wave morphology, synchronization, evolution.
- **Clinically Interpretable:** Reasoning often valid even if final label is wrong.

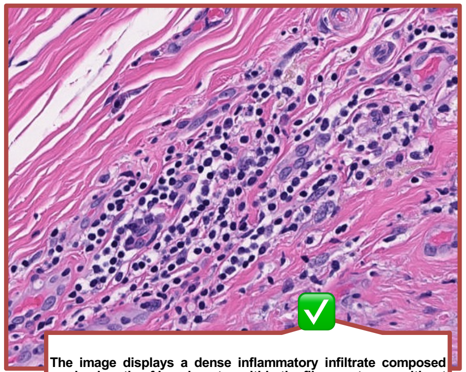
X KEY FAILURE MODES

- **Absence seizures (absz):** Misclassified as GNSZ or artifacts; underestimates 3 Hz spike-slow & abruptness.
- **Background (bckg) with Artifacts:** Eye blinks misclassified as Focal Seizures (fnsz).
- **Rigid Expectations:** Difficulty with low-amplitude rhythmic patterns & lack of robustness.

ANNOTATOR PERSPECTIVE: Partial understanding of EEG features, lacks robustness & temporal nuance. Reasoning is often clinically interpretable even when the final label is wrong.

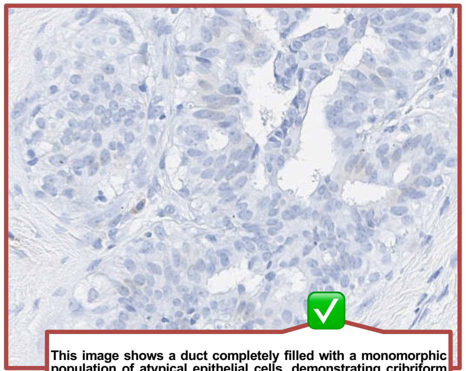
Case Study: DPATH I (Tasks & Quantitative Results)

- **Goal:** Assess visual reasoning of ChatGPT o3-mini-high model for DPATH images
- **DPATH evaluation setup:**
  - ❑ 101 high-resolution pathology patches.
  - ❑ For each image, model outputs:
    - Cancer vs non-cancer decision
    - One of nine labels
    - Pathology-informed reasoning.
- **ChatGPT o3-mini-high performance:**
  - ❑ Cancer vs non-cancer accuracy: 54%.
  - ❑ Nine-way label accuracy: 48%.
  - ❑ In ~40% of cases, reasoning is judged histologically accurate and well-aligned with the image.
- Even with moderate accuracy, the model demonstrates **clinically meaningful reasoning** ability—suggesting strong potential for decision support systems in pathology.



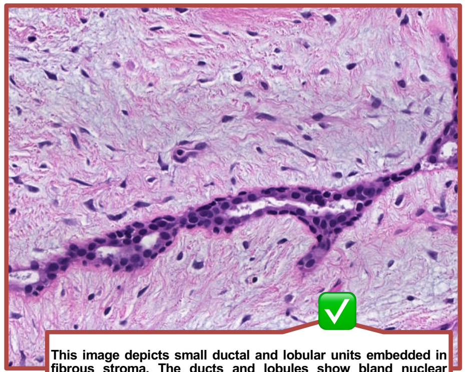
✓

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.



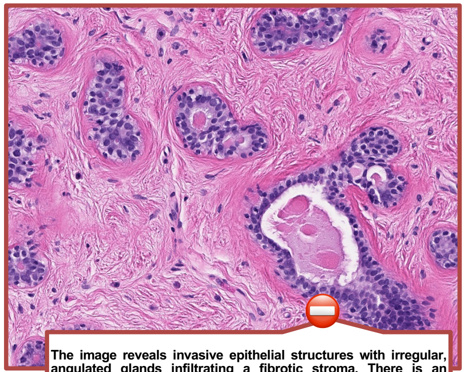
✓

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).



✓

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.



⛔

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.

Case Study: DPATH II (Reasoning Quality and Error Patterns)

✓ STRENGTHS IN REASONING

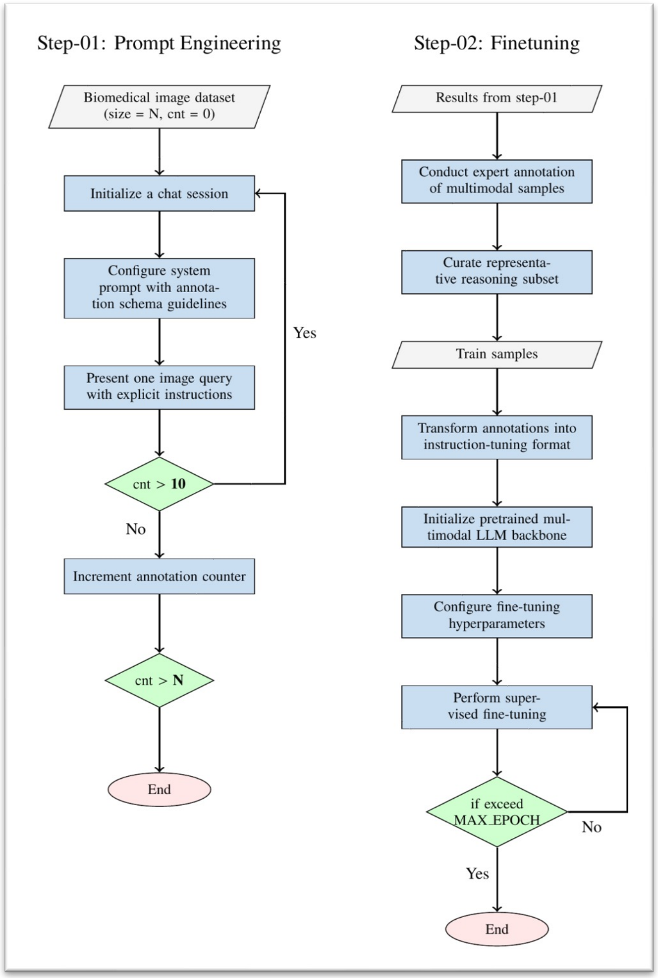
- **Correctly distinguishes DCIS vs INDC:** Recognizes preservation vs. loss of basement membrane, confinement vs. stromal invasion.
- **Normal tissue recognitions:** Notes dual cell layers, bland nuclei, preserved architecture, and unremarkable stroma.
- **Strong differential reasoning:** Correctly rules out malignancy by noting the absence of key cancer features.

X KEY FAILURE MODES

- **Misses secondary context** (e.g,. background artifacts)
- **Benign structures misclassified as Malignant** (overcalling ‘cancer’/’suspicious’)
- **Mislabeling** due to stain differences or subtle morphology changes.
- **Reasoning** describes DCIS/INDC patterns even for normal/benign images.

Annotator Perspective: Model may succeed on one instance and fall on a similar one, indicating limited robustness across subtle variations.

Fine-Tuning Approach (PEFT with Qwen2-VL)



Step-01: Prompt Engineering

Biomedical image dataset (size = N, cnt = 0)

Initialize a chat session

Configure system prompt with annotation schema guidelines

Present one image query with explicit instructions

cnt > **10**

Yes

No

Increment annotation counter

cnt > **N**

End

Step-02: Finetuning

Results from step-01

Conduct expert annotation of multimodal samples

Curate representative reasoning subset

Train samples

Transform annotations into instruction-tuning format

Initialize pretrained multimodal LLM backbone

Configure fine-tuning hyperparameters

Perform supervised fine-tuning

if exceed MAX_EPOCH

No

Yes

End

- **Goal: leverage expert-validated reasoning to improve an open-source MLLM. Used Qwen2-VL-7B Instruct as the base model.**
- **Training data:**
  - ❑ **EEG: 25 correct reasoning texts (bckg: 13, seiz: 12 after label collapsing).**
  - ❑ **DPATH: 48 correct reasoning texts (label-imbalanced: indc, bckg, norm, infl, dcis, nneo).**
- **Instruction-tuning format:**
  - ❑ **System: describes role (e.g., “medical professional specializing in cancer detection”).**
  - ❑ **User: includes the image and structured query prompt with required JSON output.**
  - ❑ **Assistant: provides ground-truth label and expert-validated reasoning.**
- **PEFT – Low rank adaptation (LoRA)**
  - ❑ **Instead of updating all weights, LoRA adds small low-rank adapter matrices to selected layers and only trains those.**
  - ❑ **~2.5M trainable params out of ~8.3B total (~0.03% updated).**
  - ❑ **Preserves the general knowledge of the base model while specializing it for EEG and DPATH reasoning.**

Results I (Analysis – EEG)

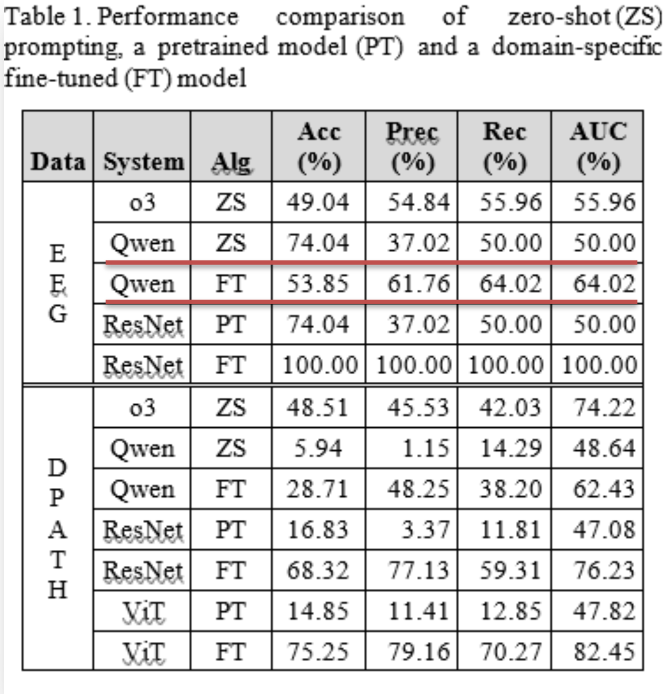


Table 1. Performance comparison of zero-shot (ZS) prompting, a pretrained model (PT) and a domain-specific fine-tuned (FT) model

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

- **Models compared on EEG (binary seizure vs background):**
  - ❑ **ChatGPT o3-mini-high (ZS), Qwen ZS, Qwen FT (LoRA), ResNet PT (ImageNet) and ResNet FT (EEG domain-specific)**
- **Key quantitative findings (EEG):**
  - ❑ **Qwen ZS:**
    - ➢ **Accuracy ≈ 74% (predicts most samples as “seiz”).**
    - ➢ **High accuracy but low precision (~37%), reflecting strong class bias.**
  - ❑ **Qwen FT:**
    - ➢ **Accuracy ≈ 54%, but precision improves to ~62% and recall ~64%.**
    - ➢ **Indicates a better calibrated classifier than Qwen ZS.**
- **Statistical analysis:**
  - ❑ **Pairwise Z-tests show significant improvements of Qwen FT over its zero-shot and some baseline counterparts at 99% confidence.**

Results II (Analysis – DPATH)

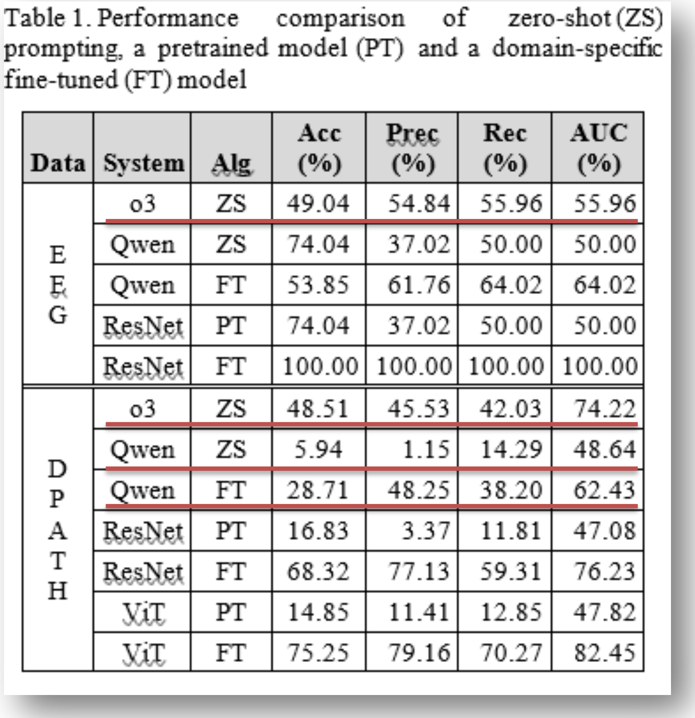


Table 1. Performance comparison of zero-shot (ZS) prompting, a pretrained model (PT) and a domain-specific fine-tuned (FT) model

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

- **Models and settings (DPATH, 6-way classification):**
  - ❑ **Multimodal: o3-mini-high (ZS), Qwen ZS, Qwen FT (LoRA)**
  - ❑ **Vision baselines: ResNet PT/FT, ViT PT/FT**
- **Zero-shot vs fine-tuned MLLMs (Qwen):**
  - ❑ **Qwen ZS: very low accuracy (5.94%) and AUC (48.64%), indicating difficulty on pathology images without adaptation.**
  - ❑ **Qwen FT (LoRA): accuracy improves to 28.71% and AUC to 62.43%, a large gain from fine-tuning on just 48 expert-validated reasoning samples.**
- **o3-mini-high ZS vs pretrained CNN/ViT baselines:**
  - ❑ **o3 ZS: accuracy 48.51%, AUC 74.22%.**
  - ❑ **Pretrained ResNet PT: AUC 47.08%; ViT PT: AUC 47.82%—both substantially below o3 ZS.**
  - ❑ **Pairwise Z-tests confirm o3's AUC advantage over these pretrained baselines is statistically significant at 99% confidence, making o3 ZS a strong reference baseline for future DPATH studies.**
- **For both domains, MLLMs are better baseline models than deep learning-based models.**

Summary

- Introduced a two-step framework:
  - ❑ Zero-shot prompt engineering with ChatGPT o3-mini-high.
  - ❑ PEFT fine-tuning of Qwen2-VL using expert-curated reasoning.
- Developed a structured prompting strategy that uses annotation guidelines to provide rich in-context information to MLLMs.
- Benchmarked MLLMs against ResNet and ViT on two biomedical image tasks (EEG & DPATH).
- Performed expert review of model reasoning to evaluate clinical interpretability.
- Applied pairwise Z-tests to quantify the statistical significance of performance differences between models.
- Findings:
  - ❑ MLLMs can deliver moderate zero-shot performance with clinically meaningful reasoning.
  - ❑ PEFT can yield statistically significant improvements.
  - ❑ Domain-specific supervised models still lead in accuracy, but MLLMs offer flexibility and explainability.

Future Directions

- **Scale up datasets for both EEG and DPATH to move beyond proof-of-concept:**
  - ❑ **Larger, more diverse images**
  - ❑ **More balanced label distributions.**
- **Implement k-fold cross-validation and more robust evaluation protocols.**
- **Develop a systematic prompt-engineering framework to reduce human bias and improve reproducibility.**
- **Explore architectural enhancements and training methods enabling better:**
  - ❑ **Temporal reasoning for EEG**
  - ❑ **Handling of complex, multi-structure pathology images.**
- **Integrate model-generated reasoning directly into clinical workflows as structured, verifiable explanations.**
- **Investigate hybrid systems combining domain-specific vision backbones with MLLM-style reasoning for best-of-both-worlds performance.**

Acknowledgements

- 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.
- The title slide’s image was generated by the Gemini Nano Banana tool using the following prompt: ‘Generate an image showing a physician who is an expert in EEG signal interpretation and digital pathology, working with ChatGPT on her computer.’

References

1. 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*.

2. 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*.

3. 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*.

4. 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*.

5. 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*.

6. I. Obeid and J. Picone, “Machine Learning Approaches to Automatic Interpretation of EEGs,” in Signal Processing and Machine Learning for Biomedical Big Data, 1st ed., E. Sejdik and T. Falk, Eds., Boca Raton, Florida, USA: Taylor & Francis Group, 2018, p. 30. doi: *10.1201/9781351061223*.

7. N. Capp, C. Campbell, T. Elseify, I. Obeid, and J. Picone, “Optimizing EEG Visualization Through Remote Data Retrieval,” 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, 2018, pp. 1–2. doi: *10.1109/SPMB.2018.8615613*.

8. N. Shawki et al., “The Temple University Digital Pathology Corpus,” in Signal Processing in Medicine and Biology: Emerging Trends in Research and Applications, 1st ed., I. Obeid, I. Selesnick, and J. Picone, Eds., New York City, New York, USA: Springer, 2020, pp. 67–104. doi: *10.1007/978-3-030-36844-9*.