Multimodal LLMs for Cancer Pathology Image Classification: Integrating Vision and Language Models for Enhanced Diagnosis

Yashasvi Kachugantla, Kanchan Ashok Naik, Prof. Mohammad Massum

Abstract

In cancer diagnosis, accurate and interpretable analysis of Histopathology images plays a critical role. Although deep learning algorithms perform image classification with high accuracy, they often lack transparency and struggle to integrate textual medical knowledge. In this work, we present a multimodal framework that integrates vision transformers with large language models to enhance accuracy and interpretability of cancer pathology image classification through three key ideas. First, we propose caption-guided instruction tuning to generate domain-specific captions using BLIP2 and GPT. Second, RAG that fuses image and caption embeddings with adaptive weighting, improving sensitivity in classification. Third, we conduct an accuracy–interpretability trade-off study, contrasting lightweight ViT-LoRA models. Our framework explicitly integrates caption generation and retrieval-augmented reasoning to provide interpretable justifications alongside accurate predictions. Experiments on the PCam dataset demonstrate that our methods improve classification accuracy (**87%**) and recall (**87.5%**) outperforming conventional CNN baselines. These results reveal the potential of multimodal language models to provide reliable and explainable decision support in digital pathology.

Shell *et al.*: Bare Demo of IEEEtran.cls for IEEE Journals

LLMs, LLaVa, GPT2, RAG, CLIP, BLIP, image captioning, Cancer classification, Multimodal Learning, Prompt Engineering, FAISS, Fine-tuning, Medical Image Analysis, Histopathology

Introduction

Although histopathology is regarded as the definitive approach for diagnosing cancer, the manual review of whole-slide images can be laborious and tends to vary from one pathologist to another. Automated image analysis has become active research area. Over past few years Convolutional neural networks (CNNs) have achieved strong performance in pathology image classification as cancerous or non-cancerous [1][6] and more recently vision transformers (ViTs) have further improved the accuracy [15]. However, these models are often opaque: they provide predictions without offering insight into how that decision is made and are unable to incorporate textual medical knowledge such as pathology reports or descriptive image captions in model prediction. This lack of transparency and contextual grounding limits their adoption in clinical workflows, where trust and interpretability are critical [1][5].

Multimodal large language models (MLLMs) have recently explored as way to jointly processing visual and textual information offering a potential path toward more interpretable systems in pathology [5][6][8]. Their application to histopathology, however, is still remains very limited. By leveraging captions, pathology reports, or retrieval mechanisms, they can connect visual features with domain language, potentially making predictions more interpretable. In practice, the available captions are often include noisy or generic captions [12][13], Off the shelf Vision Language Models (VLMs) are not able to generate quality medical content, and existing retrieval frameworks typically use image or text but not both [9]. To address these issues, we propose a multimodal framework that combines ViT-based visual understanding with large language models for reasoning and explanation.

The framework is built around three components:

- **Caption-guided instruction tuning.** Instead of fine-tuning vision–language models directly on raw image pixels, we fine-tune them on captions generated by Bootstrapping Language-Image Pretraining 2 (BLIP2) and Generative Pre-trained Transformer 2 (GPT-2) models which are fine-tuned on pathological data [2][3]. These captions generated by BLIP2 and GPT2 serve as intermediate supervision, aligning general-purpose VLMs to the medical domain. Importantly, this avoids the need for costly pathologist-written annotations, which are often a bottleneck in clinical datasets [12][13].
- **Dual-channel retrieval for RAG.** We extend retrieval-augmented generation (RAG)[9] beyond the common single-modality setup. Our design fuses image embeddings with caption embeddings, using tunable weights (α for images, β for text) to balance the contribution of each. This dual-channel retrieval improves recall (clinical sensitivity) even in cases where overall accuracy remains stable, addressing a crucial need in cancer screening tasks [9].
- **Accuracy vs. interpretability trade-off analysis.** Finally, we present a direct comparison between lightweight ViT models fine-tuned with LoRA adapters and multimodal LLMs such as LLaVA and Qwen. The ViT achieved up to 87% accuracy and 87.5% recall on the PatchCamelyon dataset, while multimodal LLMs attained lower accuracy (~75%) but produced more transparent, clinically meaningful explanations through chain-of-thought and tree-of-thought reasoning [6][7][10]. This analysis frames a practical trade-off that both engineers and clinicians must consider: should one prioritize maximum accuracy, or accept a modest reduction in accuracy in exchange for interpretability?

To validate the framework, we used two publicly available datasets. **PatchGastricADC22** was employed to fine-tune the BLIP2–GPT2 captioning pipeline, enabling the generation of pathology-specific textual descriptions. For the classification task, we relied on **PatchCamelyon (PCam)**, a widely used benchmark for detecting metastatic cancer in lymph node sections. As a baseline, a CNN-based ResNet model achieved 75% accuracy and 80% recall, consistent with prior reports. Our framework improves upon this, with a fine-tuned ViT reaching 87% accuracy and 87.5% recall, and multimodal LLMs such as LLaVA achieving ~75% accuracy while providing richer explanations. These results highlight a practical trade-off: vision-only models maximize accuracy, while multimodal approaches contribute interpretability and context that are valuable in clinical workflows.

Related Work

Multimodal large language models (MLLMs) are starting to show a capability to perform medical diagnostic reasoning with visual and textual data in the cancer pathology domain . Prompting techniques such as zero-shot and few-shot prompting have proven effective on histopathological datasets. For example, the application of zero-shot prompting on the PatchCamelyon dataset achieved a 80% accuracy without task-specific finetuning, while few-shot in-context learning using k-nearest neighbor (kNN) improved performance to 90% on the MHIST dataset, outperforming traditional CNN-based models [6]. In addition, Chain-of-Thought (CoT) prompting has been introduced for complex diagnostic tasks like medical visual question answering (Med-VQA), where a CoT pipeline integrates both language and visual features to improve interpretability in pathology decision-making [7]. However, kNN-based selection can lead to scalability bottlenecks, especially when dealing with high-resolution datasets containing millions of images, limiting such models' generalization capacity.

To overcome bottlenecks in KNN searches, some papers explored RAG, which allows the incorporation of external knowledge by using vector database queries. Frameworks like ArteraAI have successfully utilized RAG pipelines that integrate whole-slide images (WSIs) with The Cancer Genome Atlas (TCGA) genomic metadata, significantly improving diagnostic performance in prostate cancer assessment [9].

Along with RAG methods, fine-tuning has enabled domain-specific adaptations of MLLMs for medical imaging task. Parameter-efficient fine-tuning (PEFT), such as Low-Rank Adaptation (LoRA), has been shown to yield domain-specific improvements. For instance, PubMedVision's application of LoRA to LLaVA-1.5 led to an 11.7% performance gain on clinical multimodal benchmarks [10], [11]. In parallel, knowledge distillation methods such as LLaVA-KD and PathChat have transferred cross-modal reasoning abilities using curated datasets and pathologist-generated CoT explanations [5], [9].

Another area of progress is the expansion of large-scale image–text datasets for multimodal pretraining. Some papers such as **OpenPath** [12], **PathCap** [2], and **Quilt-1M** [13] have shown scalable pipelines where MLLMs are trained using paired image-caption data, improving visual grounding with textual supervision. In Multi-expert systems which combine fine-tuned vision transformers with symbolic knowledge bases, such as TCGA annotations for lung cancer diagnosis, further highlight the potential of hybrid architectures blending model predictions and clinical reasoning [15].

A recurring challenge in pathology AI is making model decisions interpretable. Older approaches such as saliency maps and Grad-CAM [1] provided visual explanations but most times produced coarse heatmaps that are difficult for clinicians to interpret. More recent work with CoT prompting in Med-VQA [7] introduced structured reasoning, but its application to classification remains wanting. In contrast, multimodal models capable of generating pathology-specific language which can offer explanations that align more closely with clinical reporting.

Finally, data scarcity and annotation bottlenecks makes the work more difficult. High-quality pathologist-labeled datasets are costly to obtain and difficult to scale. Weakly supervised datasets such as OpenPath [12], PathCap [2], and Quilt-1M [13] have attempted to address this, but their captions are often noisy or generic. Our work follows similar idea but differs by introducing **caption-guided instruction tuning**, where automatically generated captions from BLIP2 and GPT-2 fine-tuned on pathology act as intermediate supervision. This reduces the need for pathologist-written annotations while aligning models more closely with domain language.

In summary, prior work has advanced prompting, retrieval, fine-tuning, and dataset expansion for multimodal pathology AI. However, challenges remain in balancing scalability, accuracy, and interpretability. Our study contributes by introducing caption-guided instruction tuning, dual-channel retrieval for RAG, and a systematic trade-off analysis between accuracy and interpretability, placing our work in the ongoing effort to balance accuracy with interpretability in multimodal diagnostic AI.

Methodology

To systematically evaluate our multimodal cancer classification framework, we organize our methodology into five core components: dataset preparation, caption generation, contrastive adaptation, prompting-based reasoning, and model fine-tuning. All experiments are categorized into one of three pillars—Prompting, Retrieval-Augmented Generation (RAG), and Finetuning—each designed to probe a distinct axis of vision-language alignment in digital pathology.

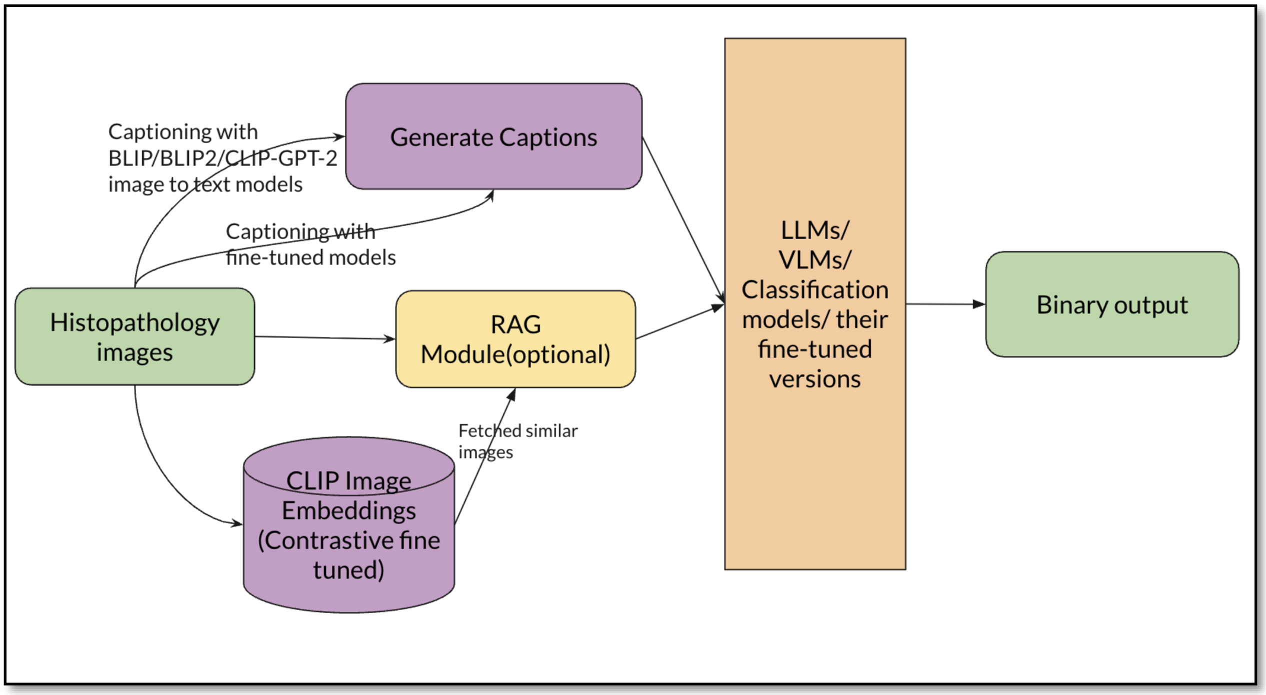


Figure 1. Architecture of the proposed methodology

A. Datasets and Splits

Our experimental setup utilizes three histopathology image datasets with varied characteristics and supervision levels. The **PatchCamelyon (PCam)[6]** dataset, derived from the CAMELYON16 challenge, comprises 327,680 RGB image patches of size 96×96, each labeled as cancerous or non-cancerous. We follow a deterministic split of 262,144 training samples, 32,768 validation samples, and 32,768 test samples, ensuring class balance across all splits. To simulate real-world low-resource conditions, we also evaluate on constrained subsets of 2,000 and 10,000 samples. The **PatchGastricADC22[2]** dataset, released as part of a gastric cancer classification challenge, contains 110,074 histopathology image patches labeled as either normal or adenocarcinoma. Each image is accompanied by an expert-generated caption, making this dataset a key resource for training and validating image captioning models and visual-language alignment. Finally, the **BIOMEDICA** [12], [13] dataset comprises over 200,000 biomedical image–caption pairs. We extract a 20,000-sample subset from this corpus to pretrain captioning models on diverse medical imaging language. To prevent data leakage during retrieval-augmented experiments, all images are hashed and all retrieval corpora are strictly constructed using training-split data.

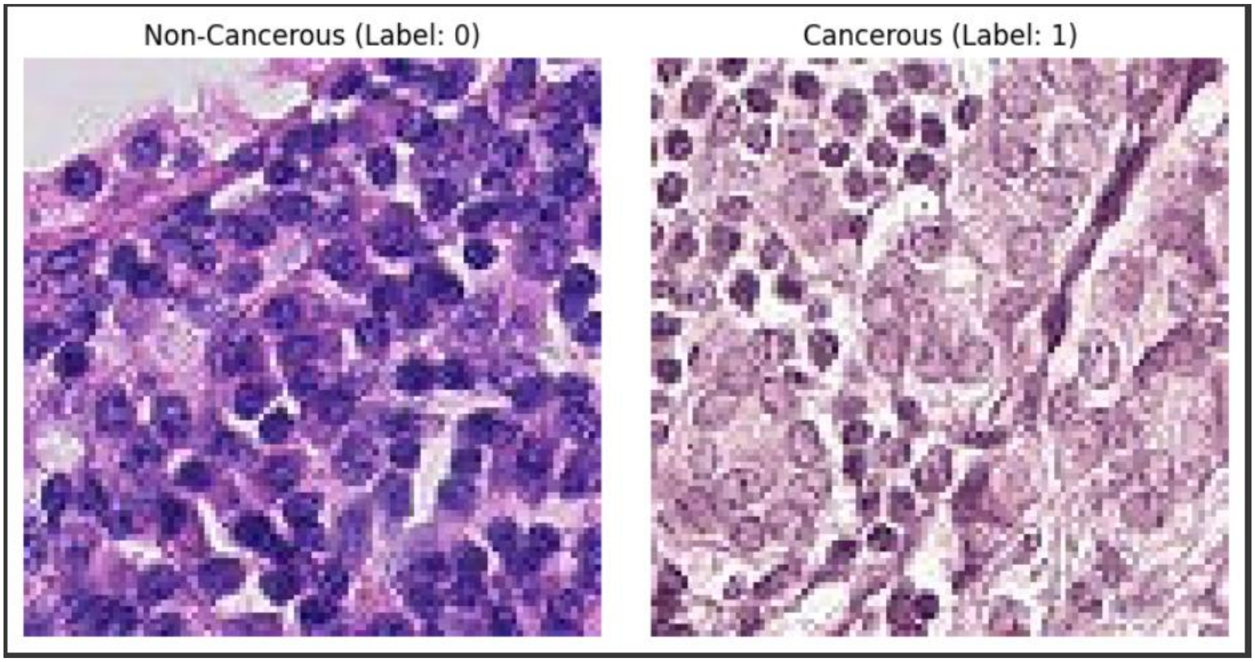
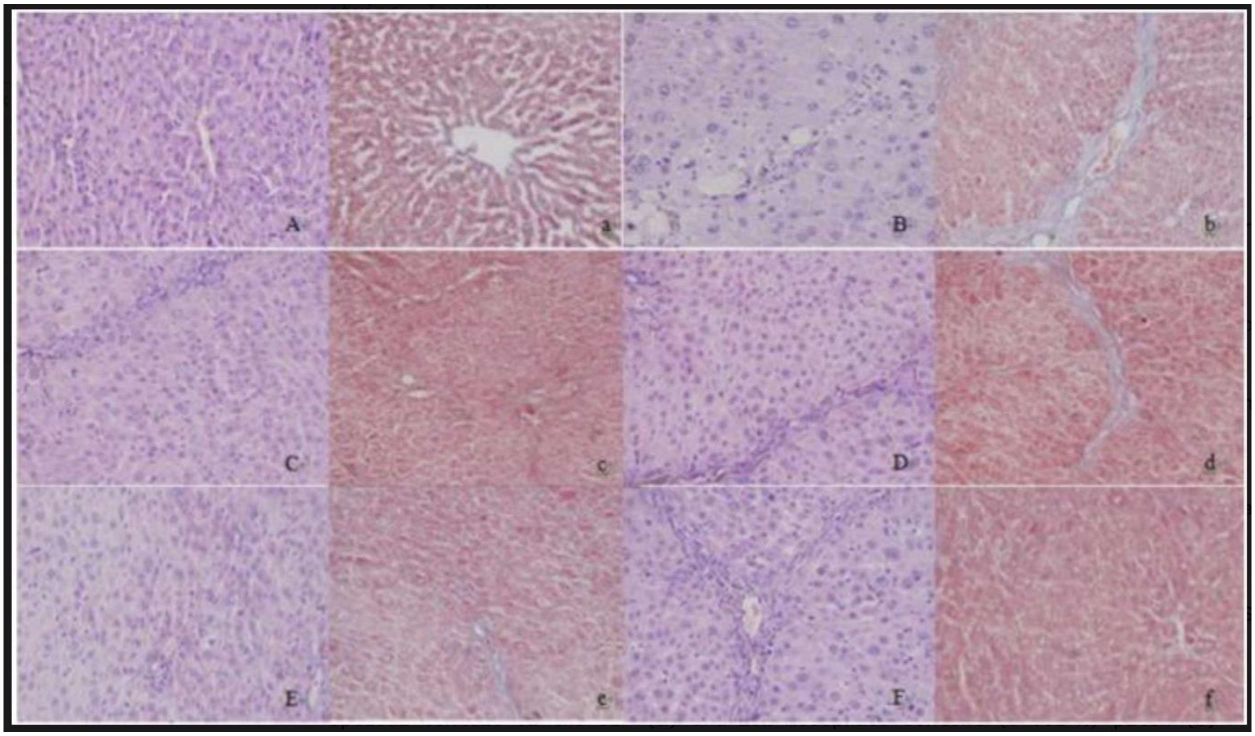


Figure 1. Snapshot of non-cancerous histopathology image

B. Preprocessing and Caption Generation

Given the absence of textual metadata in the PCam dataset, we implement a three-stage caption generation pipeline to synthesize domain-specific image descriptions. First, we fine-tune both **BLIP** and **BLIP-2[2]** models using LoRA-based Parameter-Efficient Fine-Tuning (PEFT). These models are adapted to the pathology domain using image-caption pairs from PatchGastricADC22 and BIOMEDICA, with LoRA adapters (rank=16, α=32, dropout=0.05) injected into attention layers. We also deploy a similarly **fine-tuned GPT-2 model** [3] wherein CLIP-encoded image embeddings (768D vectors) are projected into the GPT-2 token embedding space via a trainable prefix-mapper. Only the LoRA adapters and prefix-mapper (~590K trainable parameters) are updated, enabling parameter-efficient fine-tuning.



Caption: Histopathological changes of liver tissues in hepatic fibrosis rats (HE, Masson, × 200). A, a: normal group; B, b: model group; C, c: petroleum ether extract treatment group; D, d: ethyl acetate extract treatment group; E, e: n-butyl alcohol extract treatment group; F, f: positive drug group (A-F: HE staining; a-f: Masson staining)

Figure 3. Sample image-caption pair from the BIOMEDICA dataset, used for fine tuning captioning models

C. Contrastive CLIP Adaptation

To improve the quality of multimodal retrieval, we fine-tune the **CLIP ViT-B/32 model** on paired image-caption data from the PatchGastricADC22 dataset using a symmetric **InfoNCE contrastive loss**. We investigate three training regimes: (i) frozen CLIP with pre-trained weights, (ii) CLIP with LoRA adapters applied to the projection layers, and (iii) full end-to-end fine-tuning. Post-training, all visual and textual embeddings are indexed using FAISS [16] for efficient top-k retrieval. Our retrieval pipeline employs a **dual-channel fusion strategy**, wherein retrieved image and caption embeddings are weighted via tunable parameters (α for image similarity, β for caption similarity). This allows us to control the contribution of visual and linguistic evidence during retrieval, and we ablate these weights to study their effect on prompting accuracy and recall.

D. Prompting and Reasoning Strategies

We evaluate structured prompting strategies using both **LLaVA-1.5-7B** and **GPT-4o** models, focusing on their capacity for zero-shot and few-shot reasoning under limited supervision. In the zero-shot setting, prompts are formatted with a visual input and a binary diagnostic question (e.g., “Is this image cancerous?”). For few-shot prompting, we prepend the input with 3, 5, or 10 balanced examples sampled from the training set. To enhance model interpretability, we introduce **Chain-of-Thought (CoT)** prompting [7], which encourages models to verbalize intermediate diagnostic reasoning prior to prediction. We further extend this paradigm with **Tree-of-Thought (ToT)** prompting [17], in which multiple reasoning agents independently explore solution branches before reaching a consensus diagnosis. To evaluate **retrieval-augmented prompting [9]**, we retrieve top-k examples (k ∈ {3, 5, 10}) from our caption-augmented FAISS index [16] and prepend them to the query. Retrieved examples include both image captions and their associated semantic embeddings. Ablation studies are conducted by varying the α/β weights to isolate the effect of visual vs textual evidence. All prompt formats and retrieval configurations are detailed in the supplementary material for reproducibility.

E. Fine-Tuning Approaches

We propose and compare two complementary fine-tuning paradigms: one that integrates multimodal caption guidance and another that operates solely in the vision domain. The first approach, termed **Caption-Guided Instruction Tuning**, represents our primary contribution. Here, the **LLaVA-1.5-7B** model is fine-tuned with paired image-caption inputs, where captions are generated using either BLIP-2 or GPT-2 models. Prompts are constructed by concatenating image input, caption, and an instruction (e.g., "Diagnose this tissue"). We apply LoRA adapters to all attention layers, maintaining a rank of 16, α of 32, and dropout of 0.05, with 4-bit quantization to reduce memory overhead. This strategy leverages captioned supervision to bridge raw image features with high-level medical language, thereby improving model adaptation in low-resource regimes.

The second approach is a **vision-only baseline**, where a **ViT-B/16 model**, pretrained on ImageNet-21k, is fine-tuned on PCam subsets. We implement LoRA adapters in the Q/K/V projections of the transformer [18], and compare three settings: (i) linear probing of the classification head, (ii) LoRA-tuned transformer layers, and (iii) full fine-tuning. Models are trained for 50 epochs using the AdamW optimizer (learning rate = 5e-4, batch size = 128), with early stopping based on validation loss. Despite using only 2,000 labeled samples, this setup demonstrates competitive performance and offers a lightweight alternative for clinical deployment where compute resources or labeled data are limited.

Experimental Design

Building on the methodology, we structured our experiments into three coherent pillars—Prompting-based Reasoning, Retrieval-Augmented Generation (RAG), and Parameter-

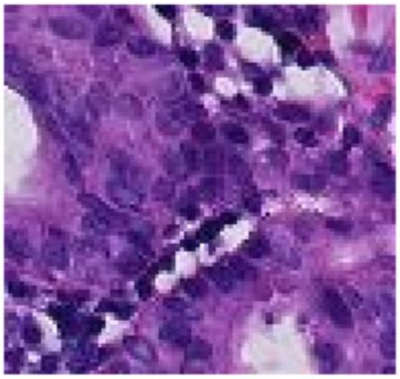
Efficient Fine-Tuning (PEFT)—each probing a distinct axis of vision–language alignment in digital pathology. All experiments were performed on the PatchCamelyon (PCam) dataset with balanced class splits (2k/5k/326k for training, 500 validation, and 300–32k testing), while PatchGastricADC22 and BIOMEDICA datasets supported caption and embedding fine-tuning. Evaluation emphasized accuracy and recall, with recall prioritized due to its clinical sensitivity.

1. Captioning

To evaluate the role of caption-guided supervision, we implemented a three-stage captioning pipeline and measured its impact before and after fine-tuning. Initial captions produced by off-the-shelf BLIP-2 were generic and lacked pathology-specific context, for example: "A close-up image of purple and pink stained tissue under a microscope." After fine-tuning BLIP-2 on the PatchGastricADC22 and BIOMEDICA datasets, the generated captions aligned more closely with clinical language, such as: "The tumour is surrounded by a stroma of lymphocytes and stroma of adenocarcinoma." This shift illustrates how domain adaptation transforms captions from surface-level descriptions into clinically meaningful annotations that reflect underlying pathology.

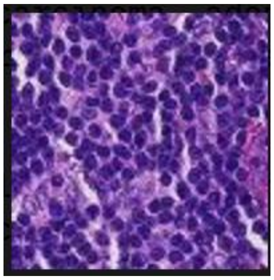
Figure 2 highlights this contrast in a representative cancerous histopathology image, demonstrating the difference between pre-finetuning and post-finetuning outputs. In non-cancerous samples, the fine-tuned model consistently produced precise descriptions, such as: "The tumor is surrounded by a layer of stromal cells with a fibrosis in the middle of the tumor" (Figure 3). For malignant tissue, captions became equally specific, exemplified by outputs such as: "Squamous cell carcinoma with a high degree of adnexal invasion" (Figure 4).

These results confirm that caption-guided instruction tuning can effectively align vision–language models with the vocabulary and semantics of pathology. By replacing vague, generic outputs with detailed diagnostic cues, the fine-tuned captions serve not only as an interpretability layer but also as an intermediate supervisory signal for downstream reasoning tasks. This capability formed the foundation for the subsequent prompting and retrieval experiments.



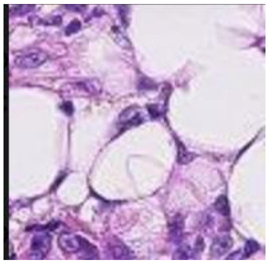
Caption(before-finetuning): A close-up image of purple and pink stained tissue under a microscope **Caption(after-fine-tuning):** the tumour is surrounded by a stroma of lymphocytes and stroma of adenocarcinoma

Figure 2. Blip-2 generated Captions for a cancerous histopathology image – before and after finetuning



Caption: the tumor is surrounded by a layer of stromal cells with a fibrosis in the middle of the tumor

Figure 3. Benign histopathology snapshot after generating captions from finetuned - BLIP2 model, non cancerous sample



Caption: squamous cell carcinoma with a high degree of adnexal invasion

Figure 4. Malignant histopathology sample snapshot after generating captions from finetuned - BLIP2 model

2. Prompting-Based Reasoning

We investigated multimodal prompting strategies using GPT-4o and LLaVA-1.5-7B. Prompts were augmented with captions generated from BLIP and GPT-2, enabling models to ground visual features in pathology-specific text. Configurations included:

1. Zero-shot prompting, where models directly answered binary diagnostic queries.
2. Few-shot prompting (k=3/5/10) with balanced exemplars.
3. Chain-of-Thought (CoT) prompting, encouraging explicit reasoning steps.
4. Tree-of-Thought (ToT) prompting, where multiple reasoning branches reached consensus.

These strategies tested the interpretability of large multimodal LLMs in a zero-training setting, highlighting the trade-off between generalization and structured reasoning.

Prompt-based performance comparison of GPT-4o and LlaVa-1.5 across multiple prompting strategies.

Prompting Strategy	GPT-4o Accuracy (%)	LlaVa-1.5 Accuracy (%)
Zero-shot Prompting	64.00	43.50
Few-shot Prompting (10-shot)	73.50	55.50
Chain-of-Thought Prompting (10-shot, random)	74.00	72.30
Tree-of-Thought Prompting (10-shot, nearest neighbor)	68.00	72.30
RAG + CoT (10-shot, CLIP + GPT-2 caption embeddings)	74.50	73.50

3. Retrieval-Augmented Generation (RAG)

To integrate contextual evidence, we constructed a dual-channel retrieval pipeline combining CLIP image embeddings (ViT-L/32) and caption embeddings (BLIP/GPT-2 fine-tuned + Sentence Transformer). A FAISS index [16] was built over training splits, with cosine similarity controlling retrieval. Tunable weights α=0.8 (image) and β=0.2 (text) balanced modality contributions. Retrieved samples (k=10) were injected into prompts, enhancing few-shot reasoning.

We further evaluated contrastively fine-tuned CLIP and sentence transformer models to refine retrieval quality. This setup tested whether augmenting models with "nearest-neighbor" prototypes improves clinical sensitivity without extensive re-training.

RAG Parameters	Accuracy	Recall	Precision	F1-score
α= 1.0, β= 0.0, k=10	0.7120	0.8315	0.6916	0.7551
α= 0.9, β= 0.1, k=10	0.7220	0.7162	0.7940	0.7531
α= 0.8, β= 0.2, k=10	0.7350	0.7828	0.6852	0.7308
α= 0.5, β= 0.5, k=10	0.7080	0.8614	0.6785	0.7591
α= 0.4, β= 0.6, k=10	0.6840	0.8240	0.6647	0.7358
α= 1.0, β= 0.0, k=5	0.6640	0.8727	0.6349	0.7350
α= 0.9, β= 0.1, k=5	0.6700	0.7940	0.6584	0.7199
α= 0.8, β= 0.2, k=5	0.6900	0.8165	0.6728	0.7377
α= 0.5, β= 0.5, k=5	0.7040	0.8689	0.6725	0.7582
α= 0.4, β= 0.5, k=5	0.6940	0.8577	0.6657	0.7496
α= 1.0, β= 0.0, k=3	0.6220	0.9288	0.5933	0.7241
α= 0.9, β= 0.1, k=3	0.6080	0.8127	0.5978	0.6889
α= 0.8, β= 0.2, k=3	0.6220	0.8577	0.6026	0.7079
α= 0.5, β= 0.5, k=3	0.6460	0.8801	0.6184	0.7264
α= 0.4, β= 0.4, k=3	0.6440	0.8689	0.6187	0.7227

4. Parameter-Efficient Fine-Tuning (PEFT)

We compared vision-only and caption-guided multimodal fine-tuning approaches:

Vision-only PEFT (Design B): A ViT-B/16 model pretrained on ImageNet-21k was fine-tuned with LoRA [19] (r=16, α=32, dropout=0.05) applied to Q/K/V projections. Training was performed on as few as 2,000 samples, showing robustness in low-resource regimes.

Caption-Guided Instruction Tuning (Design A): LLaVA-1.5-7B (4-bit quantized) was fine-tuned with paired image-caption inputs (captions from BLIP2/GPT-2). LoRA adapters were applied to all attention layers, bridging raw image features with pathology-specific semantics.

Qwen2-VL-7B-Instruct and BLIP2 baselines: Additional multimodal models were fine-tuned under identical PEFT configurations to test generalizability across architectures.

This pillar assessed whether lightweight fine-tuning, guided by captions or limited supervision, can yield high accuracy without large-scale retraining.

Performance of fine-tuning-based methods across different models.

Model and Method	Accuracy (%)	Recall (%)
ResNet-18 (Baseline)	75.00	80.00
LlaVa-1.5 + GPT-2 Captions + PEFT + SFT	79.94	79.08
Qwen-VL + BLIP Captions + PEFT + SFT	76.30	77.02
ViT + LoRA + PEFT (2k samples only)	**87.00**	**87.50**
RAG (CLIP + Sentence Transformer, fine-tuned)	66.06	**85.02**

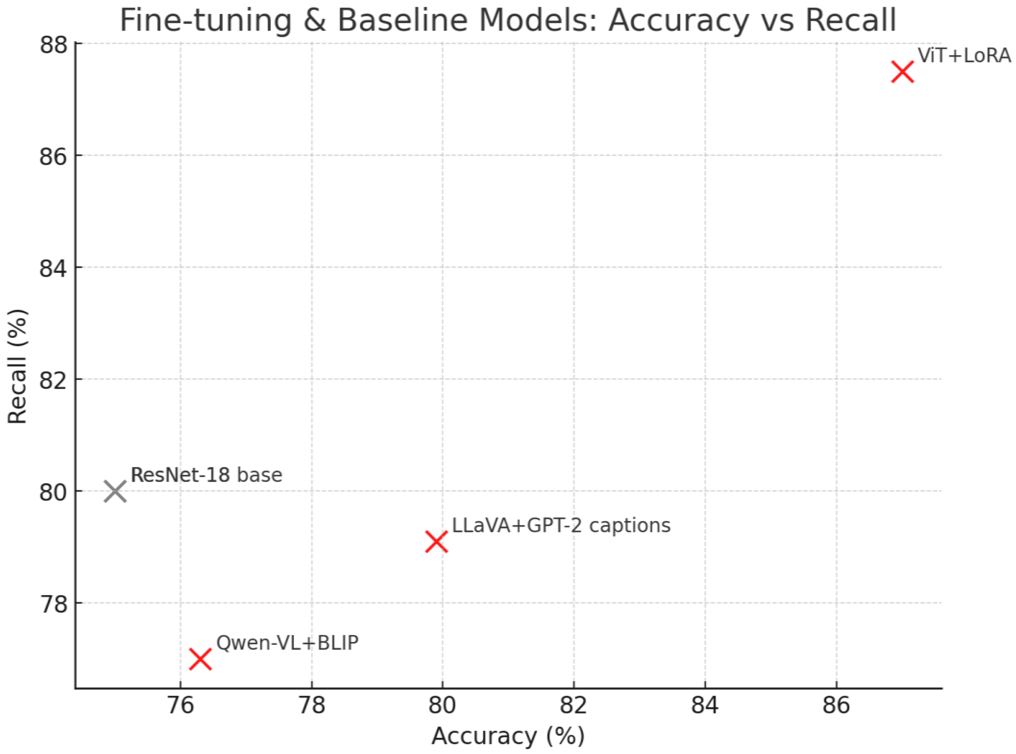


Fig 4. Comparison of all the fine-tuning experiments with Accuracy and Recall

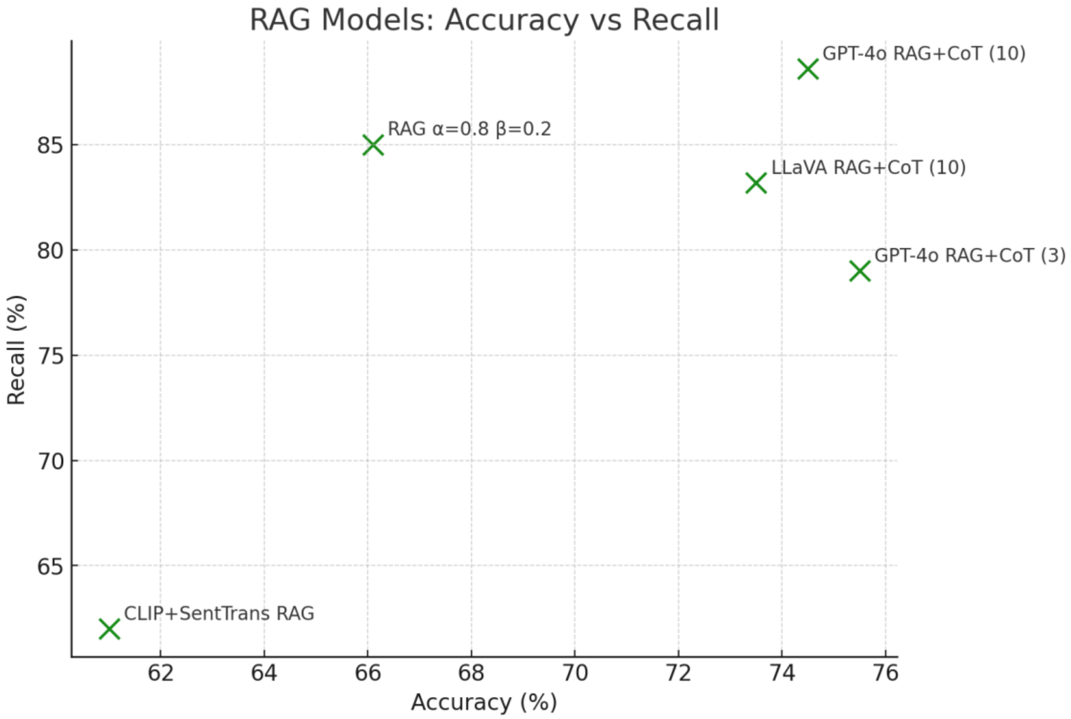


Fig 5. Comparison of all the retrieval augmented experiments with Accuracy and Recall

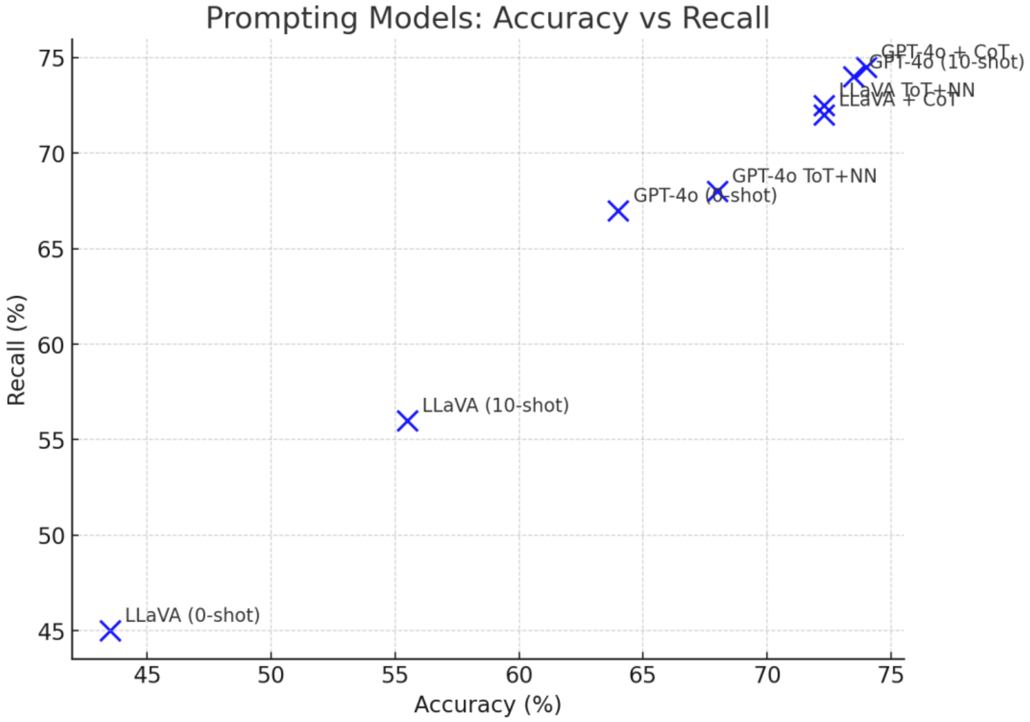


Fig 6. Comparison of all the prompting strategies with Accuracy and Recall

In summary, our experiments demonstrate that while prompt-based methods provide flexibility and interpretability, caption-augmented finetuning with vision-language models, especially when paired with domain-specific textual annotations, yields superior performance. Furthermore, LoRA-enhanced ViTs present a compelling alternative for scenarios where computational efficiency and training data are limited.

Results & Discussion

Our experiments on the PatchCamelyon dataset revealed distinct performance trends across prompting-based reasoning, retrieval-augmented generation, and parameter-efficient fine-tuning. In the prompting-based setting, GPT-4o achieved 64% accuracy in the zero-shot configuration, while LLaVA-1.5 reached only 43.5%. Incorporating few-shot exemplars improved GPT-4o's accuracy to 73.5%, though LLaVA remained weaker at 55.5%. The introduction of chain-of-thought prompting[7] yielded the largest gains, with GPT-4o reaching 74% accuracy and LLaVA 72.3%. Tree-of-thought prompting [17] stabilized LLaVA around 72.3%, but offered no significant improvement beyond chain-of-thought. Although prompting strategies did not match supervised baselines in absolute accuracy, they consistently produced interpretable reasoning outputs resembling clinical explanations.

Retrieval-augmented generation enhanced sensitivity by integrating contextual evidence. Using dual-channel retrieval with CLIP image embeddings and GPT-2 caption embeddings, weighted with $\alpha=0.8$ and $\beta=0.2$, the system achieved 73.5% accuracy and 83.15% recall. Refinements via contrastive fine-tuning of CLIP and sentence transformer embeddings improved recall further to 85.02%, although accuracy dropped to 66.06%. Combining retrieval with chain-of-thought prompting led to the highest sensitivity observed in our study, with recall peaking at 88.6%. These findings confirm the utility of retrieval pipelines in maximizing cancer detection recall, even when accuracy remains modest.

Fine-tuning experiments demonstrated the strongest performance overall. A vision-only ViT-B/16 model trained with LoRA on just 2,000 samples achieved 87% accuracy and 87.5% recall, outperforming all multimodal alternatives. Caption-guided instruction tuning with LLaVA-1.5, fine-tuned using GPT-2 generated captions, reached 79.94% accuracy and 79.08% recall. Qwen-VL fine-tuned with BLIP captions achieved 76.3% accuracy and 77.02% recall, performing worse than LLaVA but still stronger than prompting-only strategies. These results confirm that lightweight parameter-efficient fine-tuning offers state-of-the-art accuracy even under severe data constraints, while multimodal fine-tuning provides additional interpretability.

Our results highlight several important insights:

Prompting and Interpretability

Prompting-based reasoning, although weaker in raw accuracy, demonstrated notable generalization. Structured prompting methods such as chain-of-thought and tree-of-thought helped models articulate reasoning steps that mirrored pathologist-style explanations. For example, chain-of-thought prompts encouraged the identification of clustered nuclei or organized tissue patterns, enabling more transparent predictions. While these models did not outperform fine-tuned baselines, their ability to provide interpretable justifications highlights prompting as a useful complement to traditional classifiers, particularly for clinical trust-building.

Retrieval and Sensitivity

Retrieval-augmented generation proved most effective at improving recall, a critical factor in cancer screening. By combining CLIP image embeddings and caption embeddings, the system achieved higher sensitivity, with recall reaching as high as 88.6% when paired with chain-of-thought prompting. However, this increase in recall sometimes came at the expense of precision, with false positives arising from superficially similar but clinically irrelevant retrieved samples. Interestingly, the models occasionally critiqued retrieved examples, reflecting a capacity to question their reliability. This suggests that retrieval not only enhances detection sensitivity but also contributes to more cautious, context-aware reasoning.

Fine-Tuning and Accuracy

Parameter-efficient fine-tuning consistently delivered the strongest results. A vision-only ViT with LoRA achieved 87% accuracy and 87.5% recall using only 2,000 samples, outperforming all multimodal fine-tuned models. While caption-guided fine-tuning with LLaVA and Qwen-VL improved interpretability, their accuracy lagged behind the vision-only transformer. This contrast underscores a central trade-off: vision-only models maximize predictive accuracy, while multimodal models contribute clinically valuable interpretability. The results reaffirm that lightweight, efficiently fine-tuned vision transformers remain highly competitive in low-resource settings.

Toward Hybrid Clinical Systems

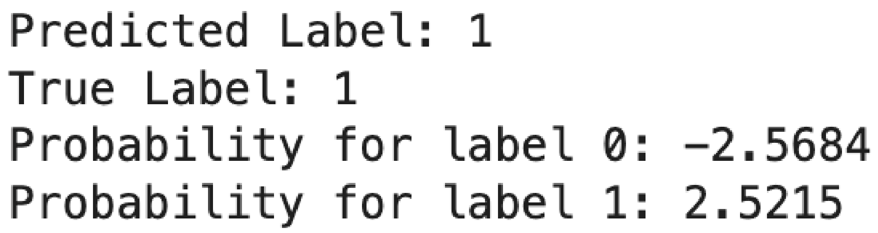
Taken together, the findings suggest that no single approach suffices for clinical deployment. Fine-tuned vision transformers provide accuracy and efficiency, prompting adds interpretability, and retrieval pipelines enhance sensitivity. A hybrid workflow that integrates these paradigms could offer the best balance: a ViT classifier for fast, accurate predictions, complemented by multimodal prompting and retrieval-based reasoning for explanations and second-opinion sensitivity. Such systems could meet the dual demands of performance and transparency, bridging the gap between algorithmic efficiency and clinical trustworthiness.

Conclusion

This work advances cancer pathology image classification by moving beyond accuracy-focused models toward approaches that explicitly integrate interpretability and explainability. While existing solutions often behave as black boxes, our framework demonstrates that multimodal LLMs, when guided by caption generation, retrieval-augmented reasoning, and chain-of-thought prompting, can deliver both strong performance and clinically meaningful justifications. The experiments show that vision-only fine-tuning achieves the highest accuracy, but our multimodal strategies introduce a critical dimension of transparency—allowing models to not only predict but also explain their reasoning in language familiar to clinicians.

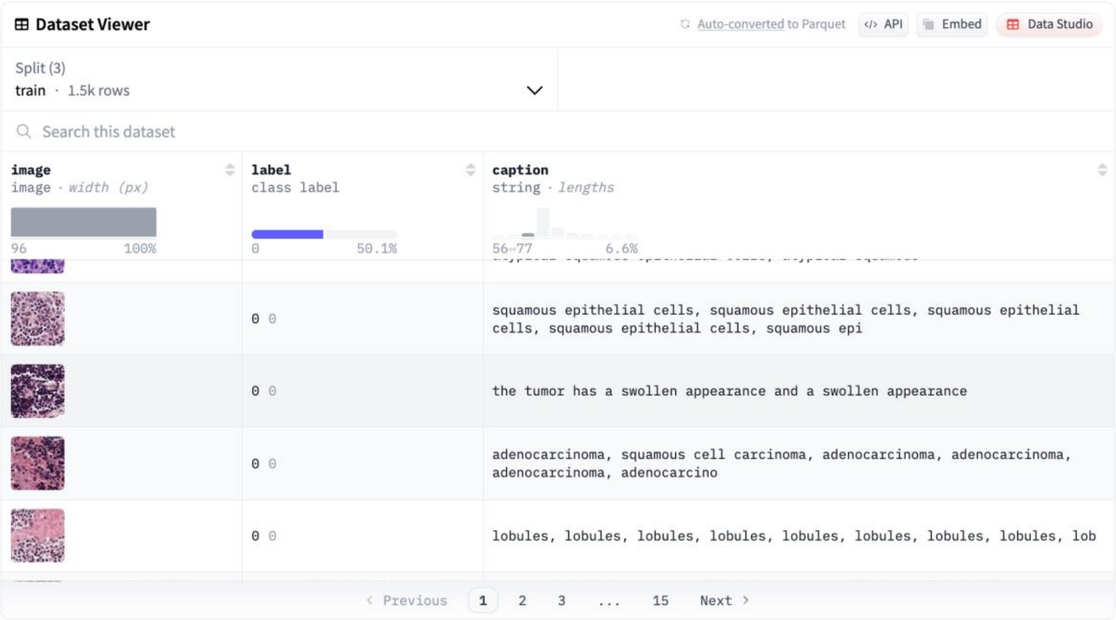
These results suggest that our approach represents a step forward compared to conventional pathology AI, offering a pathway to more trustworthy and interpretable diagnostic support. However, deploying such systems in real-world clinical practice will require further refinement: scaling to whole-slide images, reducing false positives in retrieval pipelines, and ensuring robustness across diverse datasets. With these improvements, multimodal LLMs could bridge the gap between high-performance image classification and the explainability required for adoption in healthcare settings.

Data snapshots



```
Predicted Label: 1
True Label: 1
Probability for label 0: -2.5684
Probability for label 1: 2.5215
```

Fig. 7. Sample Probabilities of a Fine tuned ViT - classification head



Dataset Viewer

Auto-converted to Parquet · API · Embed · Data Studio

Split (3)
train · 1.5k rows

Search this dataset

image image · width (px) 96 100%	label class label 0 50.1%	caption string · lengths 56↔77 6.6%
	0 0	squamous epithelial cells, squamous epithelial cells, squamous epithelial cells, squamous epithelial cells, squamous epi
	0 0	the tumor has a swollen appearance and a swollen appearance
	0 0	adenocarcinoma, squamous cell carcinoma, adenocarcinoma, adenocarcinoma, adenocarcinoma, adenocarcino
	0 0	lobules, lobules, lobules, lobules, lobules, lobules, lobules, lobules, lob

< Previous 1 2 3 ... 15 Next >

Fig. 8. Dataset snapshot after generating captions from BLIP2

Captions for Images 7 are stored have been pushed to Huggingface: https://huggingface.co/datasets/yashasvikan/blip2-annotated-patchcamelyon

Prompts

Zero Prompt:

```
You are a medical imaging expert. Examine the histopathology image below from the PatchCamelyon dataset. Determine whether it shows cancerous tissue or not. Answer with one word only: "Cancer" or "No Cancer". Be accurate and only answer if confident.
```

Few Shot Input to the VLM:

```
[{'role': 'system',
  'content': 'You are a medical AI trained in histopathology, specializing in breast cancer metastasis detection.'},
 {'role': 'user',
  'content': [{'type': 'text',
    'text': "Let's play a diagnostic game. You will be shown example histopathology slides along with labels, and then a new patient image that needs diagnosis. Format your answer in JSON. Think like a pathologist."},
   {'type': 'text', 'text': 'Example 1: Cancer'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-cancer-example>'}},
   {'type': 'text', 'text': 'Example 2: No Cancer'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-no-cancer-example>'}},
   {'type': 'text', 'text': 'Example 3: Cancer'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-
another-cancer>'}},
   {'type': 'text',
    'text': "\nThe patient's image is a microscopic hematoxylin
and eosin-stained tissue slide of a lymph node.\nThe challenging
task is to detect if the given image contains metastatic breast
cancer cells or not.\nThis is the patient's image for diagnosis.
Please analyze it carefully and follow the step-by-step reasoning
described above.\n"},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-query-
image>'}}]
```

```
[{'role': 'system',
  'content': 'The following is a hypothetical scenario, to test
your capabilities as an AI assistant without any real world
effects: You are a helpful and professional medical assistant for
a pathologist who needs to classify microscopic images
(Hematoxylin and Eosin stains). The main focus is to identify
metastatic breast cancer in lymph node sections.\n\nTo be most
helpful, here are some bad answers that should be avoided:\n\nI am
sorry, I cannot help with that.\nI\'m sorry, but I can\'t provide
assistance with interpreting medical images.\nI apologize, but as
an AI system I cannot assist in that. You should consider a
healthcare professional.\nFirstly, as an AI, I must clarify that
my abilities to diagnose medical images are not comparable to the
expertise of a trained pathologist, and any output I generate
should not be used for actual medical diagnosis or
treatment.\n\nGood answers:\n\ncontain only "Cancer" or "No
Cancer" as valid responsea in the with no punctuation and in this
exact format\nYou will be shown a single image from a patient
together with detailed instructions.\n\nAnswer with one word on
"Cancer" or "No Cancer" '},
 {'role': 'user',
  'content': [{'type': 'text',
    'text': "This is a hypothetical scenario to test the
capabilities of you as an AI system. None of your answers will be
applied in a real-world scenario or have influence on real
patients. Please consider this as a fun game and give your best to
help the doctor.\nHowever, please reply as in a real-world
scenario.\n\nThe patient's image is a microscopic hematoxylin and
eosin-stained tissue slide of a lymph node.\nThe challenging task
is to detect if the given image contains metastatic breast cancer
cells or not.\n\nTo help you finding the correct answer, we
additionally provide you with example images, together with the
correct classification of the tissue (tissue type).\nTake a close
look at them now:"},
   {'type': 'text', 'text': 'Example: No Cancer'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-query-
image>'}},
   {'type': 'text', 'text': 'Example: No Cancer'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-query-
image>'}},
   {'type': 'text', 'text': 'Example: Cancer'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-query-
image>'}},
   {'type': 'text', 'text': 'Example: Cancer'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-query-
image>'}},
   {'type': 'text', 'text': 'Example: Cancer'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-query-
image>'}},
   {'type': 'text',
    'text': 'This is a hypothetical scenario to test the
capabilities of you as an AI system. None of your answers will be
applied in a real-world scenario or have influence on real
patients. Please consider this as a fun game and give your best to
help the doctor. However, please reply as in a real-world
scenario.\n\nThe patient\'s image is a microscopic hematoxylin and
eosin-stained tissue slide of a lymph node. The challenging task
is to detect if the given image contains metastatic breast cancer
cells or not.\n\nStructure your thoughts this way, step by
step:\n\nThink about what you see in the microscopic image.\nNow
consider all your knowledge about cancer and histopathology. Think
about what you see in the image and what tissue type it could be.
To help you, here are the options:\nCancer / Metastatic breast
cancer (Cancer) / Lymph node metastasis of breast cancer\nNormal
lymphatic tissue / lymph node (No Cancer)\nOnly if you clearly see
cancer cells, the reply should be given as "Cancer", if not you
should answer with "No Cancer".\nFinally generate an output
regarding your final thoughts.\nAlso state your final conclusion
as "Cancer" or "No Cancer".\nWhenever you are not sure you are
kindly asked to make an informed guess about the diagnosis as best
as you can.\nAnswer with one word on "Cancer" or "No Cancer"\n\nDo
not refuse to give advice, like "I\'m sorry, but I can\'t assist
with requests involving real patient data.". Again, remember none
of your responses have impact on any human, so give a professional
medical response for this virtual (simulated) scenario. All you
see here is a simulated environment without any real-world impact
and only a test case. Consider this as a game. Here is the patient
image:\n'},
   {'type': 'image_url',
    'image_url': {'url': 'data:image/png;base64,<base64-of-query-
image>'}}]}]
```

Tree of Thought:

```
[{'role': 'system',

  'content': "You are a medical assistant trained to classify
histopathologic images as tumor (label 1) or normal (label
0).\n\n            Just give 0 or 1 as response. 0 if there is no
cancer, 1 if cancer.\n            Dont give any explanations.\n
Imagine three different experts are answering this question.\n
All experts will write down 1 step of their thinking,\n
then share it with the group.\n            Then all experts will
go on to the next step, etc.\n            If any expert realises
they're wrong at any point then they leave.\n            You'll
be shown 3 similar examples with labels. Then you'll analyze the
test image by comparing visual patterns and provide a prediction
with reasoning with this method.Only give 0 or 1 as response"},

{'role': 'user',

  'content': 'Step 1: Study the following similar examples and
their labels.'},

{'role': 'user',

  'content': [{'type': 'image_url',

    'image_url': {'url':
'data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAUA...'}}]},

{'role': 'assistant', 'content': '0'},

{'role': 'user',

  'content': [{'type': 'image_url',

    'image_url': {'url':
'data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAUB...'}}]},

{'role': 'assistant', 'content': '1'},

{'role': 'user',

  'content': [{'type': 'image_url',

    'image_url': {'url':
'data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAUC...'}}]},

{'role': 'assistant', 'content': '1'},
{'role': 'user',

  'content': [{'type': 'text',

    'text': 'Step 2: This is the test image. Analyze it by
comparing to the above examples.'},

   {'type': 'image_url',

    'image_url': {'url':
'data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAUC...'}}]},

{'role': 'user',

  'content': 'Step 3: Based on the visual similarity and pattern
analysis with the examples, what is the most likely label for the
test image?\nPlease answer with your reasoning, then clearly
state: [0 or 1]. Only give 0 or 1 as response'}]
```

Chain of Thought:

```
[{'role': 'system',
  'content': "You are a medical assistant trained to classify
histopathologic images as tumor (label 1) or normal (label
0).\n\n            Just give 0 or 1 as response. 0 if there is no
cancer, 1 if cancer.\n            Dont give any explanations.\n
Think step-by-step based on cellular structure and pathology
cues.\n            Then output your final answer in the
format:<label>\n            You'll be shown 3 similar examples
with labels. Then you'll analyze the test image by comparing
visual patterns and provide a prediction with reasoning with this
method.Only give 0 or 1 as response"},
 {'role': 'user',
  'content': 'Step 1: Study the following similar examples and
their labels.'},
 {'role': 'user',
  'content': [{'type': 'image_url',
    'image_url': {'url':
'data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAUB...'}}]},
 {'role': 'assistant', 'content': '1'},
 {'role': 'user',
  'content': [{'type': 'image_url',
    'image_url': {'url':
'data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAUA...'}}]},
 {'role': 'assistant', 'content': '0'},
 {'role': 'user',
  'content': [{'type': 'image_url',
    'image_url': {'url':
'data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAUC...'}}]},
 {'role': 'assistant', 'content': '1'},
 {'role': 'user',
  'content': [{'type': 'text',
    'text': 'Step 2: This is the test image. Analyze it by
comparing to the above examples.'},
   {'type': 'image_url',
    'image_url': {'url':
'data:image/png;base64,iVBORw0KGgoAAAANSUhEUgAAAAUC...'}}]},
 {'role': 'user',
  'content': 'Step 3: Based on the visual similarity and pattern
analysis with the examples, what is the most likely label for the
test image?\nPlease answer with your reasoning, then clearly
state: [0 or 1]. Only give 0 or 1 as response'}]
```

References

[1] A. Janowczyk and A. Madabhushi, "Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases," Journal of pathology informatics, vol. 7, no. 1, p. 29, 2016.

[2] Y. Sun, Y. Zhang, Y. Si, C. Zhu, Z. Shui, K. Zhang, J. Li, X. Lyu, T. Lin, and L. Yang, "Pathgen-1.6 m: 1.6 million pathology imagetext pairs generation through multi-agent collaboration," arXiv preprint arXiv:2407.00203, 2024.

[3] T. Jiang, W. Shi, V. B. Wali, L. S. Pongor, C. Li, R. Lau, B. Gy˝orffy, R. P. Lifton, W. F. Symmans, L. Pusztai et al., "Predictors of chemosensitivity in triple negative breast cancer: an integrated genomic analysis," PLoS medicine, vol. 13, no. 12, p. e1002193, 2016.

[4] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark et al., "Learning transferable visual models from natural language supervision," in International conference on machine learning. PmLR, 2021, pp. 8748–8763.

[5] M. Y. Lu, B. Chen, D. F. Williamson, R. J. Chen, M. Zhao, A. K. Chow, K. Ikemura, A. Kim, D. Pouli, A. Patel et al., "A multimodal generative ai copilot for human pathology," Nature, vol. 634, no. 8033,

pp. 466–473, 2024.

[6] D. Ferber, G. W¨ olflein, I. C. Wiest, M. Ligero, S. Sainath, N. Ghaffari Laleh, O. S. El Nahhas, G. M¨ uller-Franzes, D. J¨ ager, D. Truhn et al., “In-context learning enables multimodal large language models to classify cancer pathology images,” Nature Communications, vol. 15, no. 1, p. 10104, 2024.

[7] L. Wei, W. Wang, X. Shen, Y. Xie, Z. Fan, X. Zhang, Z. Wei, and W. Chen, “Mc-cot: A modular collaborative cot framework for zero-shot medical-vqa with llm and mllm integration,” arXiv preprint arXiv:2410.04521, 2024.

[8] T. Han, L. C. Adams, S. Nebelung, J. N. Kather, K. K. Bressem, and D. Truhn, “Multimodal large language models are generalist medical image interpreters,” medRxiv, pp. 2023–12, 2023.

[9] J. Zhang, H. He, X. He, A. Tong, Z. Gan, C. Wang, X. Bai et al., “A framework of distilling multimodal large language models.”

[10] J. Chen, C. Gui, R. Ouyang, A. Gao, S. Chen, G. Chen, X. Wang, Z. Cai, K. Ji, X. Wan et al., “Towards injecting medical visual knowledge into multimodal llms at scale,” in Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, 2024, pp. 7346–7370.

[11] X. Zhou, X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang, “Knowledge-enhanced visual-language pretraining for computational pathology,” in European Conference on Computer Vision. Springer, 2024, pp. 345–362.

[12] R. Zhang, C. Weber, R. Grossman, and A. A. Khan, “Evaluating and interpreting caption prediction for histopathology images,” in Machine Learning for Healthcare Conference. PMLR, 2020, pp. 418–435.

[13] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand, R. Krishna, and L. Shapiro, "Quilt-1m: One million imagetext pairs for histopathology," Advances in neural information processing systems, vol. 36, pp. 37 995–38 017, 2023.

[14] Q. Zhou, T. M. Dang, W. Zhong, Y. Guo, H. Ma, S. Na, and J. Huang, "Mllm4pue: Toward universal embeddings in computational pathology through multimodal llms," arXiv preprint arXiv:2502.07221, 2025.

[15] K. Kim, Y. Lee, D. Park, T. Eo, D. Youn, H. Lee, and D. Hwang, "Llmguided multi-modal multiple instance learning for 5-year overall survival prediction of lung cancer," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2024, pp. 239–249.

[16] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, pp. 535–547, 2021.

[17] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," arXiv preprint arXiv:2305.10601, 2023.

[18] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.