Multimodal LLMs Outperform Pathology Foundation Models in Cross-Domain Histological Similarity

Yishu Zhang **Yun Li** **Daiwei Zhang***
University of North Carolina at Chapel Hill
yzhang4@unc.edu, yun_li@med.unc.edu, daiwei@unc.edu

Abstract

State-of-the-art pathology foundation models, trained on millions of histology tiles, can fail to preserve tissue similarity when comparisons cross slide or institution boundaries. We show that general-purpose multimodal LLMs, without being trained as pathology foundation models, match or outperform these specialized models in cross-domain histological similarity judgments. Using a relative similarity framework that we release as the MOSAIC (Model Similarity Assessment across Institutions and Cohorts) benchmark, we evaluate 20 models across 6 datasets and find that pathology encoders often rank same-institution, different-disease tiles as more similar than same-disease, different-institution tiles, a clinically dangerous failure mode invisible to standard within-domain evaluations. LLMs appear less susceptible to this failure, likely because they perform semantic visual comparison of morphology and tissue architecture rather than relying on shortcut features tied to acquisition context. Scaling training data does not resolve the problem for pathology encoders, implicating the learning objective rather than data coverage. Our results expose a fundamental robustness gap in current pathology foundation models and establish multimodal LLMs as a viable alternative for cross-institutional retrieval, dataset harmonization, and multi-site quality control. The MOSAIC dataset is available at github.com/davidzhanglab/mosaic.

1 Introduction

Large-scale self-supervised pretraining has transformed computational pathology through visual foundation models [9, 52, 15, 49, 37]. Trained on millions of histology image tiles, these models extract general-purpose representations that support a broad spectrum of downstream tasks, including tissue classification, cancer subtyping, and biomarker inference [14, 5]. For these representations to support safety-critical applications such as cross-institutional image retrieval, few-shot adaptation, and multi-site clinical trials, the learned embedding space must cluster images by disease-relevant morphology rather than by acquisition context.

In this work, we expose a fundamental failure mode of current pathology foundation models: their embedding spaces can remain strongly entangled with technical variation, causing them to rank same-institution, different-disease tiles as more similar than same-disease, different-institution tiles. This failure is invisible to standard within-domain evaluations (linear probes, k-NN) yet directly undermines the retrieval and transfer tasks that motivate foundation model development. Concretely, if a pathologist queries a multi-site case database, a model with this failure mode will return color-matched tissue from the same scanner rather than morphologically equivalent disease from another laboratory, with direct consequences for diagnostic support and clinical decision-making.

Recent work has confirmed that pathology foundation model embeddings encode site-level technical signatures more strongly than biological class [28, 10, 29], and that standard stain normalization does not fully remove these shortcuts [25, 24, 7]. However, prior evaluations rely on embedding geometry

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

or downstream classifiers, making them inapplicable to generative models. Standard task-specific probes (linear classifiers, k-NN) can further obscure the problem by achieving high accuracy through cohort-specific correlations [11, 33].

To expose and quantify this failure mode, we introduce a relative similarity framework, inspired by representational similarity analysis [30] and human similarity judgments [42, 36], that tests whether a model ranks a biologically similar tile from a different source (slide or institution) above a biologically dissimilar tile that shares the same acquisition context. This formulation can be applied uniformly to both embedding-based models and generative LLMs. Using this framework, we demonstrate that multimodal LLMs, without being trained as pathology foundation models, match or outperform specialized pathology encoders in cross-domain settings, serving as an existence proof that the failure is specific to current training paradigms rather than inherent to the task.

Our contributions are: (1) We expose a clinically dangerous robustness gap in state-of-the-art pathology foundation models that is invisible to standard evaluations. (2) We introduce a relative similarity evaluation framework that directly quantifies cross-domain generalization. (3) We demonstrate that multimodal LLMs serve as an existence proof that robust cross-domain similarity is achievable without domain-specific training. (4) We release MOSAIC, a benchmark suite including comparison triplets, evaluation code, and baseline outputs, spanning 6 histopathology cohorts for evaluating cross-domain generalization in computational pathology.

2 Related Work

Robustness Evaluation for Pathology Models. Kömen et al. [28] demonstrated that pathology foundation model embeddings retain hospital-specific signatures that persist across principal components even after stain normalization. De Jong et al. [10] formalized this with the Robustness Index, which quantifies whether embedding neighborhoods are dominated more by biological labels or medical center labels, finding that nearly all evaluated models encode site identity more strongly than tissue type. PathoROB [29] scaled this analysis to 20 models across 34 centers, establishing systematic benchmarking of embedding robustness. Additional work has examined scanner-specific [47] and staining-specific [7] domain shifts. Plismbench [16] and SCORPION [6] benchmark scanner and staining robustness, which models and training objectives such as H0-mini [16], GenBio-PathFM [26], and ScanGen [6] aim to improve. These benchmarks measure invariance on registered same-tissue pairs, whereas MOSAIC tests discrimination across different slides and classes, a harder and clinically relevant task. All of these approaches operate on embedding geometry or downstream classifiers, requiring access to continuous vector representations. Our relative similarity framework complements these methods by reformulating robustness evaluation as a comparative judgment task, enabling direct comparison between embedding-based models and generative multimodal LLMs on the same task.

Domain Shift Mitigation in Computational Pathology. Multiple strategies have been proposed to reduce the impact of technical variation on histopathology models. Stain normalization methods transform images to a canonical color space [24], while stain augmentation approaches like RandStainNA [44] train models to be invariant to stain variation by sampling diverse styles during training. Style transfer methods such as STRAP [50] use medically irrelevant artistic images as style sources to learn domain-agnostic representations. However, these methods target color and style variation, whereas the robustness studies above show that residual site information extends beyond color to deeper structural artifacts (tissue thickness, cutting angle, scanner optics) that current mitigation strategies do not fully address [28, 46]. Our work is diagnostic rather than mitigative: rather than modifying the input or training to remove domain shift, we evaluate whether models can make correct similarity judgments despite it, which in turn measures how well mitigation strategies achieve the invariance they target.

Representational Similarity and Human Judgment. Representational similarity analysis (RSA) [30] provides a framework for comparing model representations by correlating dissimilarity matrices, independent of task-specific performance. Human similarity judgments offer a principled ground truth for evaluating representational geometry: Hebart et al. [23] collected 1.5 million odd-one-out triplet judgments to reveal the dimensional structure of object representations, and Roads and Love [42] enriched ImageNet with human similarity data. Muttenthaler et al. [35] showed that aligning neural network representations with human triplet judgments improves downstream generalization, and subsequent work demonstrated that such alignment transfers across abstraction levels [36]. Our relative similarity framework adapts the triplet evaluation paradigm to histopathol-

ogy, using tissue-class structure as the ground truth rather than human perceptual judgments, and specifically tests whether similarity is preserved across domain boundaries. We operationalize this as MOSAIC, a cross-domain triplet benchmark for tissue-level biological similarity.

Multimodal LLMs as Visual Judges. The LLM-as-a-judge paradigm [51] has established that language models can serve as reliable evaluators for text generation, with pairwise comparison being more reliable than direct scoring. Chen et al. [8] extended this to the multimodal setting, showing that models like GPT-4V exhibit human-like discernment in pairwise visual comparisons. Our work applies this paradigm to a specialized scientific domain: rather than judging aesthetic quality or prompt adherence, we use multimodal LLMs to judge biological similarity between histopathology tiles, a task requiring domain-specific visual reasoning about cell morphology and tissue architecture.

Generative Multimodal LLMs in Histopathology. Recent studies have begun to evaluate whether general-purpose multimodal LLMs can perform histopathology reasoning without task-specific training. Ferber et al. [13] showed that GPT-4V with in-context examples can approach specialized models on selected histopathology classification tasks. However, large-scale pathology vision-language benchmarks report only moderate accuracy and suggest that models may exploit textual question or answer priors rather than consistently rely on visual evidence [18]. These findings suggest that general-purpose MLLMs have visual reasoning capabilities but require stronger task structure for fine-grained histomorphologic interpretation. Our work shifts the focus from direct classification to probing whether MLLMs can make relational judgments of histological similarity, a task that provides stronger visual grounding by requiring explicit comparison between images rather than generating a class label from a single image.

3 Method

3.1 Relative Similarity Analysis

We propose a relative similarity evaluation framework that tests whether vision models preserve correct tissue-type similarity structure in histopathology images (Figure 1a). Given a set of tiles $\mathcal{T} = \{t_1, \ldots, t_N\}$ with associated class labels $y_i \in \mathcal{C}$ and a model-induced similarity function $s : \mathcal{T} \times \mathcal{T} \to \mathbb{R}$, we construct comparison tuples (t_r, t_p, t_n) where the reference t_r and positive t_p share the same class ($y_r = y_p$) and the negative t_n belongs to a different class ($y_n \neq y_r$). The model is evaluated on whether it correctly ranks the positive candidate as more similar to the reference than the negative:

$$\mathbb{1}\big[s(t_r, t_p) > s(t_r, t_n)\big]. \tag{1}$$

Relative similarity accuracy, the proportion of correctly ordered comparisons, directly measures whether the model's similarity function respects the semantic class structure.

3.1.1 Scope of Comparison

We define three comparison configurations (Figure 1b) that introduce progressively stronger domain shifts, each testing a different aspect of the model's similarity structure.

Within-slide. All three tiles come from the same slide s: $t_r, t_p, t_n \in \mathcal{T}_s$. This tests whether the model can discriminate tissue classes in the absence of batch effects, providing an upper bound on class-discrimination ability.

Cross-slide. The reference t_r and negative t_n come from the same slide s, while the positive t_p comes from a different slide s' of the same class: $t_r, t_n \in \mathcal{T}_s$, $t_p \in \mathcal{T}_{s'}$, $s \neq s'$. The model must recognize that a tile from a different slide but the same tissue class is more similar to the reference than a tile from the same slide but a different class. In all cross-slide comparisons, the positive comes from a different patient than the reference. This tests whether tissue-type similarity generalizes across slide-level variation.

Cross-institution. The reference t_r and negative t_n come from the same institution I, while the positive t_p comes from a different institution I': $t_r, t_n \in \mathcal{T}_I$, $t_p \in \mathcal{T}_{I'}$, $I \neq I'$. This is the most demanding configuration, testing whether the model can see past institution-level batch effects (scanner differences, staining protocols, tissue preparation) to identify tissue-type similarity.

3.1.2 Model Evaluation

Foundation model evaluation. For each comparison, we compute Euclidean distances between the reference embedding and each candidate embedding:

$$d_k = \|\mathbf{e}_{\text{ref}} - \mathbf{e}_{c_k}\|_2, \quad k \in \{1, 2\}. \tag{2}$$

The model selects the candidate with smaller distance. This directly tests whether the embedding space assigns higher similarity to same-class tiles. We use each model's standard pooled CLS feature without normalization. Images are preprocessed with each model's default pipeline (resizing and mean-std normalization) [9, 52].

LLM evaluation. Each comparison is presented as three images (reference, candidate 1, candidate 2) along with a text prompt instructing the model to identify which candidate is more biologically similar to the reference. The prompt directs the model to focus on cell morphology (cell size, shape, nuclear features) and tissue architecture (glandular organization, stromal patterns, cell density), while ignoring non-biological differences such as staining intensity, imaging artifacts, or tissue preparation variations. Images are resized to each model's native input resolution [20, 21, 40].

Accuracy. A comparison is scored as correct if the model selects the positive candidate. A response is valid when the parser recovers exactly one of the offered candidates from the LLM's output JSON. Comparisons where the model fails to return a valid response are excluded (see Appendix I for an analysis of invalid responses). Per-class accuracy aggregates over all comparisons where the reference shares a given class, and overall accuracy averages across all comparisons. We also compute per-class-pair accuracy matrices to identify which tissue type pairs are most easily confused.

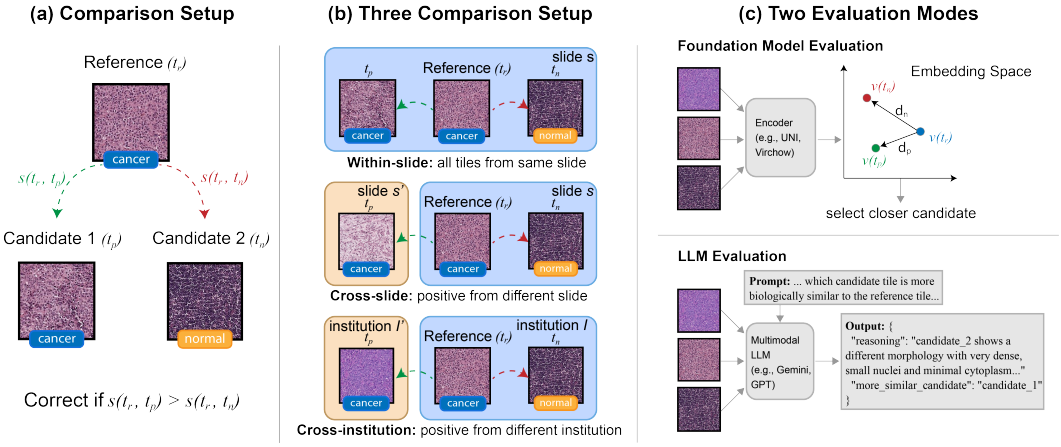


Figure 1: Schematic of the relative similarity evaluation framework.

3.2 MOSAIC Benchmark

We curate MOSAIC, a benchmark of six histopathology cohorts designed to evaluate cross-domain generalizability at two levels of domain shift. The datasets were selected to span (1) multiple organs (breast, colorectal, lung), (2) a range of tissue-class granularity (2–6 classes), (3) 9 distinct institutions across the cross-institution datasets, and (4) two distinct axes of domain shift: slide-level variation (different physical sections within a cohort) and institution-level variation (different scanners, staining protocols, and tissue preparation). HER2ST, CCTGS, and TIGER provide multi-slide cohorts for cross-slide evaluation, while CAMELYON16, TCGA, and BEETLE provide multi-institution cohorts for cross-institution evaluation. Together, these six datasets yield 13,800 comparison triplets, of which 8,400 are cross-domain (6,200 cross-slide and 2,200 cross-institution), that we release as a standardized evaluation resource alongside evaluation code and baseline model outputs.

The six component datasets are as follows. HER2ST: breast cancer dataset with 8 slides (74,275 tiles), annotated with 5 tissue classes (adipose, breast glandular, cancer, connective, immune infiltrate) [1]. CCTGS: 39 colorectal cancer slides (3,428,031 tiles) with 6 classes (adipose, lamina propria, lymphovascular invasion, muscularis propria, normal mucosa, tumor) [2]. TIGER: 40 breast cancer slides (3,063,641 tiles) from the Tumor InfiltratinG lymphocytes in breast cancER grand-challenge with 4 classes (healthy glands, necrosis, stroma, tumor) [48]. CAMELYON16: 40 breast cancer

slides (892,722 tiles) with 2 classes (tumor vs. normal) from 2 institutions (Radboud and Utrecht) [3]. TCGA: 30 lung squamous cell carcinoma slides (601,155 tiles) from The Cancer Genome Atlas [38], with annotations of invasive tumor [34], yielding 2 classes (tumor vs. normal) across 3 tissue source sites (Mayo Clinic Rochester, the International Genomics Consortium, Indivumed). BEETLE: 40 breast cancer slides (2,553,697 tiles) with 3 classes (invasive epithelium, necrosis, non-invasive epithelium) from 4 institutions (NKI, RUMC, SCH, TCGA) [32]. For each dataset, WSIs are tiled using a sliding window (512×512 pixels, stride 256) after Otsu-based tissue masking, and tiles are filtered by foreground proportion ($\geq1\%$).

For each cross-domain comparison triplet, a positive candidate shares the reference's tissue class but comes from a different slide or institution, while a negative candidate shares the reference's acquisition context but belongs to a different class. Candidates are randomly assigned to positions 1 and 2 with equal probability, so chance accuracy is 50%. Triplet sampling details, including sampling strata and counts by dataset, are in Appendix A.

3.3 Models

We evaluate three categories of models. Pathology foundation models are domain-specific vision encoders pretrained on histopathology data: UNI2H [9], H-optimus-0 [43], Midnight [27], CONCH v1.5 [12], Virchow2 [52], H-optimus-1 [4], and GenBio-PathFM [26], the latter being among the most robust pathology foundation models on existing robustness benchmarks. General foundation models include general-purpose vision encoders: DINOv2 [39], ResNet-50 [22], CLIP ViT-L/14 [41], Gemini Embedding 2 [31], and vision encoders extracted from multimodal LLM architectures (Qwen3.5-encoder [40], GLM-4.6V-encoder [19], Gemma 3-encoder [17]). LLMs are multimodal LLMs used as zero-shot judges for relative similarity: Gemini 3 Flash [20], Gemini 3.1 Flash Lite [21], GPT-5 Nano [45], Gemma 3 [17], GLM-4.6V [19], and Qwen3.5 [40]. All embedding extraction and local model inference were performed on an NVIDIA RTX 4090 GPU.

4 Results

We test whether pathology foundation models preserve biological similarity when comparisons cross domain boundaries, using multimodal LLMs as a reference point for what is achievable without domain-specific training. Figure 2 summarizes results across all six datasets, showing overall accuracy per dataset (Figure 2a) and accuracy on each tissue class within each dataset (Figure 2b). Table 1 reports overall accuracy with standard deviations for all 20 models. The pattern is consistent: pathology foundation models perform well within a domain but degrade sharply across slides and institutions, while LLMs maintain robust performance. The robustness-oriented GenBio-PathFM narrows but does not close the gap, matching the best LLM on CCTGS and CAMELYON16 but not elsewhere. Most cross-domain LLM wins are statistically significant, and insignificant wins concentrate in the within-slide setting, where the LLM margin is smallest (Appendix F). The advantage of LLMs is most pronounced in the hardest settings (cross-institution, morphologically similar classes) where foundation model shortcuts are most exposed.

4.1 Cross-Slide Relative Similarity Analysis

A robust computational pathology model should recognize the same tissue phenotype across different physical slides. However, slide-specific variation in tissue thickness, staining, and cutting artifacts can introduce technical signatures that models may exploit. We therefore evaluate cross-slide relative similarity on HER2ST, TIGER and CCTGS, testing whether models rank same-class tiles as more similar despite slide-level variation. Experiment setup details are in Appendix A.

As shown in Table 2, multimodal LLMs achieve the best overall accuracy on HER2ST and TIGER, match the robustness-oriented GenBio-PathFM on CCTGS (0.746 vs 0.752, McNemar's $p = 0.51$), and achieve the best or second-best performance for most tissue classes. The best LLM significantly outperforms every pathology foundation model on HER2ST and TIGER (McNemar's $p < 0.0001$, Appendix F). These results demonstrate the potential ability of zero-shot multimodal LLMs to look past superficial slide-specific technical artifacts and match tissues based on genuine morphological features, such as cellular organization and structural architecture.

Table 1: Overall cross-domain relative similarity accuracy (± standard deviation, bootstrapped over comparisons) on all six datasets. For each column, the best method is in **bold** and the second best is underlined. Statistical significance is reported in Table 14.

	Cross-slide			Cross-institution		
Method	HER2ST	CCTGS	TIGER	CAMELYON16	TCGA	BEETLE
Pathology foundation models						
UNI2H	0.46±0.011	0.59±0.010	0.18±0.012	0.61±0.024	0.31±0.019	0.56±0.014
H-optimus-0	0.44±0.011	0.55±0.010	0.22±0.013	0.40±0.024	0.18±0.016	0.45±0.014
H-optimus-1	0.47±0.011	0.58±0.010	0.24±0.014	0.35±0.024	0.24±0.017	0.50±0.014
Midnight	0.46±0.011	0.72±0.009	0.20±0.013	0.64±0.024	0.36±0.020	0.48±0.014
CONCH v1.5	0.45±0.011	0.56±0.010	0.21±0.013	0.56±0.025	0.36±0.020	0.47±0.014
Virchow2	0.46±0.011	0.71±0.009	0.23±0.013	0.62±0.024	0.44±0.020	0.56±0.014
GenBio-PathFM	0.46±0.011	**0.75**±0.008	0.20±0.013	0.72±0.023	0.44±0.020	0.53±0.014
General foundation models						
Gemini Emb. 2	0.48±0.011	0.67±0.009	0.36±0.015	0.61±0.024	0.50±0.020	0.56±0.014
DINOv2	0.55±0.011	0.60±0.010	0.36±0.015	0.52±0.025	0.44±0.020	0.51±0.014
ResNet-50	0.45±0.011	0.61±0.010	0.30±0.014	0.49±0.025	0.43±0.020	0.46±0.014
CLIP ViT-L/14	0.41±0.011	0.59±0.010	0.41±0.016	0.53±0.025	0.50±0.020	0.57±0.014
Qwen3.5-enc.	0.44±0.011	0.57±0.010	0.36±0.015	0.56±0.025	0.41±0.020	0.33±0.014
GLM-4.6V-enc.	0.48±0.011	0.59±0.010	0.34±0.015	0.58±0.025	0.40±0.020	0.48±0.014
Gemma 3-enc.	0.41±0.011	0.65±0.009	0.34±0.015	0.60±0.025	0.39±0.020	0.55±0.014
LLMs						
Gemini 3 Flash	0.55±0.011	**0.75**±0.009	0.35±0.015	0.70±0.023	0.55±0.020	0.54±0.014
Gemini 3.1 Flash Lite	0.53±0.011	0.72±0.009	0.37±0.015	**0.73**±0.022	0.56±0.020	**0.58**±0.014
Qwen3.5	0.55±0.011	0.74±0.009	0.40±0.016	0.68±0.025	**0.59**±0.020	0.57±0.014
GLM-4.6V	0.55±0.011	0.68±0.009	0.39±0.015	0.61±0.025	0.47±0.021	0.50±0.014
Gemma 3	0.54±0.011	0.65±0.009	0.43±0.016	0.56±0.025	0.53±0.020	0.56±0.014
GPT-5 Nano	**0.57**±0.011	0.64±0.009	**0.46**±0.016	0.59±0.025	0.52±0.020	0.57±0.014

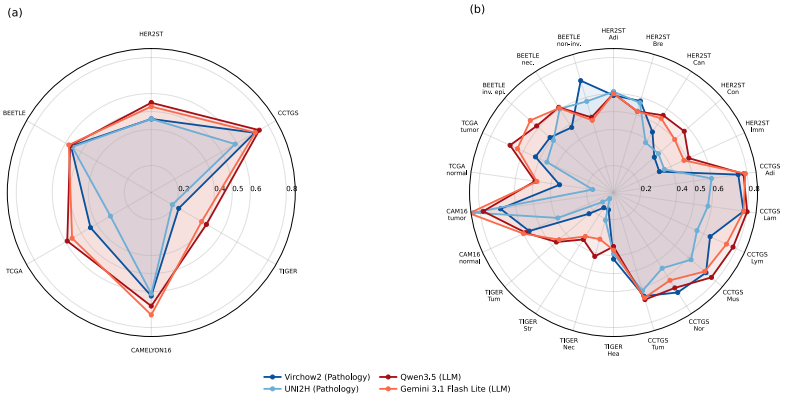


Figure 2: Accuracy for relative similarity analysis across all the experiments, for two pathology foundation models and two LLMs. (a) Overall accuracy on each dataset (cross-slide for HER2ST, CCTGS, and TIGER, cross-institution for CAMELYON16, TCGA, and BEETLE). (b) Accuracy on the sub-tasks of each dataset, i.e., each tissue class (cross-slide datasets) or each institution and class (cross-institution datasets, pooled over institutions).

4.1.1 Class-Level Confusion

To better characterize the failure modes behind this performance gap, we analyze which tissue classes are most frequently conflated in cross-slide comparisons. Per-class-pair confusion heatmaps (Appendix Figures 8, 10, and 11) show that foundation models most easily confuse morphologically similar classes. On HER2ST, for example, UNI2H achieves only 25% accuracy on the connective-vs-immune-infiltrate pair and 32% on the cancer-vs-immune-infiltrate pair, both well below chance. MLLMs improve most on these difficult pairs, where foundation models fall below chance, rather than uniformly improving all class pairs. Across all datasets, LLMs also show more uniform per-class accuracy: UNI2H accuracy on TIGER ranges from 0.09 to 0.39 across classes (spread of 0.30), while GPT-5 Nano ranges from 0.39 to 0.56 (spread of 0.17).

Table 2: Cross-slide tile relative similarity accuracy on HER2ST, CCTGS, and TIGER, broken down by positive class c_+. For each column, the best method is in **bold** and the second best is underlined. "All" refers to the overall accuracy across all classes. Class abbreviations are listed in Table 20.

	HER2ST						CCTGS							TIGER				
Method	Adi	Bre	Can	Con	Imm	All	Adi	Lam	Lym	Mus	Nor	Tum	All	Hea	Nec	Str	Tum	All
Pathology foundation models																		
UNI2H	0.61	0.57	0.38	0.38	0.36	0.46	0.60	0.58	0.56	0.62	0.55	0.62	0.59	0.39	0.21	0.09	0.13	0.19
H-optimus-0	0.59	0.53	0.36	0.40	0.32	0.44	0.59	0.56	0.57	0.59	0.51	0.51	0.55	0.40	0.20	0.17	0.17	0.22
H-optimus-1	0.52	0.53	0.49	0.39	0.40	0.47	0.68	0.65	0.60	0.52	0.46	0.59	0.58	0.41	0.20	0.19	0.21	0.24
Midnight	0.60	0.56	0.41	0.37	0.40	0.46	0.76	0.70	0.69	0.76	0.75	**0.68**	0.72	0.50	0.12	0.11	0.16	0.20
CONCH v1.5	0.43	0.52	0.51	0.41	0.38	0.45	0.47	0.59	0.58	0.66	0.53	0.55	0.56	0.32	0.19	0.18	0.20	0.21
Virchow2	0.59	**0.58**	0.45	0.35	0.33	0.46	0.75	0.78	0.64	0.73	0.71	0.65	0.71	0.42	0.15	0.15	0.23	0.23
GenBio-PathFM	0.58	0.55	0.42	0.37	0.40	0.46	0.75	**0.82**	0.76	**0.78**	**0.78**	0.64	**0.75**	0.34	0.15	0.17	0.18	0.20
General foundation models																		
Gemini Emb. 2	0.54	0.50	0.51	0.45	0.43	0.48	0.67	0.77	0.69	0.69	0.63	0.61	0.67	0.43	0.23	0.35	0.40	0.36
DINOv2	0.61	0.51	0.57	0.49	0.56	0.55	0.50	0.63	0.66	0.65	0.60	0.61	0.60	0.36	0.17	**0.45**	0.39	0.36
ResNet-50	0.55	0.54	0.45	0.33	0.41	0.45	0.49	0.70	0.66	0.65	0.62	0.60	0.61	0.28	0.16	0.36	0.34	0.30
CLIP ViT-L/14	0.47	0.50	0.37	0.37	0.36	0.41	0.48	0.65	0.66	0.69	0.55	0.60	0.59	0.42	0.43	0.35	0.43	0.41
Qwen3.5-enc.	0.63	0.50	0.38	0.32	0.36	0.44	0.59	0.61	0.56	0.56	0.55	0.54	0.57	0.41	0.52	0.29	0.30	0.36
GLM-4.6V-enc.	0.57	0.49	0.44	0.41	0.46	0.48	0.53	0.61	0.60	0.63	0.62	0.57	0.59	0.31	0.33	0.36	0.34	0.34
Gemma 3-enc.	0.54	0.49	0.40	0.31	0.34	0.41	0.63	0.75	0.72	0.66	0.60	0.60	0.65	0.40	0.19	0.29	0.45	0.34
LLMs																		
Gemini 3 Flash	0.61	0.53	0.55	0.54	0.53	0.55	**0.79**	0.79	**0.80**	**0.78**	0.68	0.67	**0.75**	0.35	0.28	0.31	0.44	0.35
Gemini 3.1 Flash Lite	0.60	0.52	0.54	0.50	0.48	0.53	**0.79**	0.78	0.74	0.72	0.63	0.66	0.72	0.37	0.32	0.34	0.45	0.37
Qwen3.5	0.60	0.52	0.56	**0.57**	0.51	0.55	0.78	0.80	0.78	0.77	0.68	0.67	0.74	0.35	0.42	0.36	0.47	0.40
GLM-4.6V	0.59	0.50	0.57	0.51	**0.58**	0.55	0.64	0.73	0.73	0.76	0.64	0.64	0.68	0.48	0.31	0.38	0.40	0.39
Gemma 3	0.55	0.54	**0.58**	0.50	0.52	0.54	0.65	0.67	0.71	0.71	0.62	0.56	0.65	0.41	0.36	0.41	**0.50**	0.43
GPT-5 Nano	**0.64**	**0.58**	0.56	0.50	0.56	**0.57**	0.60	0.68	0.66	0.73	0.62	0.60	0.64	**0.56**	**0.56**	0.39	0.41	**0.46**

4.1.2 Comparison with Within-Slide Performance

To disentangle the effect of domain shift from inherent class-discrimination limitations, we introduce a within-slide baseline on HER2ST, where the reference, positive, and negative tiles all come from the same slide ($t_r, t_p, t_n \in \mathcal{T}_s$). We then compare this baseline with cross-slide accuracy, where the positive tile comes from a different slide ($t_p \in \mathcal{T}_{s'}$, $s' \neq s$; Figure 3). In the within-slide setting, pathology foundation models dominate: Virchow2 leads at 73.5%, followed by UNI2H (72.8%). When comparisons cross slide boundaries, all models drop in accuracy, but pathology foundation models suffer the largest decline while LLMs maintain relatively higher cross-slide accuracy. This suggests that foundation models capture slide-specific features well but struggle to generalize across slides, whereas LLMs trade some within-slide precision for better cross-slide robustness. Within-slide accuracy breakdown is in Appendix B.1 (Table 5); scatter plot comparison is in Figure 7.

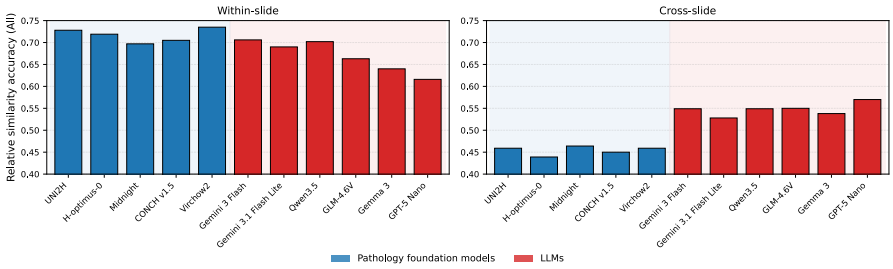


Figure 3: Within-slide (left) and cross-slide (right) relative similarity accuracy on HER2ST. Pathology foundation models lead within-slide accuracy, while LLMs achieve competitive or higher cross-slide accuracy.

4.2 Cross-Institution Relative Similarity Analysis

The cross-slide analysis shows that slide-level variation already degrades foundation model performance. Cross-institution comparisons introduce a stronger domain shift, where the positive t_p comes from a different institution I' with different scanners, staining protocols, and tissue preparation, while the reference t_r and negative t_n come from the same institution I. We evaluate cross-institutional relative similarity on CAMELYON16, TCGA, and BEETLE. Experiment setup details are in Appendix A.

Table 1 (right) reports overall accuracy on the three cross-institution datasets. LLMs generally outperform foundation models, with the advantage most pronounced on TCGA (Qwen3.5: 58.9% vs best pathology foundation model Virchow2: 44.0% and best general foundation model Gemini Emb. 2: 50.3%, McNemar's $p < 0.0001$, Appendix F). On CAMELYON16, the best LLM (Gemini 3.1

Flash Lite: 73.0%) significantly outperforms all foundation models except GenBio-PathFM (71.8%, $p = 0.67$). On BEETLE, the best LLM (57.9%) leads, although its margin over the best pathology and general foundation models is not significant. Pathology foundation models show high variance across institutions and perform worst on TCGA (H-optimus-0: 17.7%), indicating that institution-level batch effects dominate their similarity structure. When the reference and positive candidate are normal tiles, most foundation models fall below chance across datasets, while LLMs more frequently maintain above-chance accuracy. A qualitative examination of these failures makes the mechanics of this shortcut learning visually apparent (Figure 4). Foundation model embeddings often favor color-matched negative tiles from the same institution, whereas MLLMs more often recover the biological match despite staining differences. Per-institution breakdowns are in Appendix C.1 (Tables 6, 7, and 8).

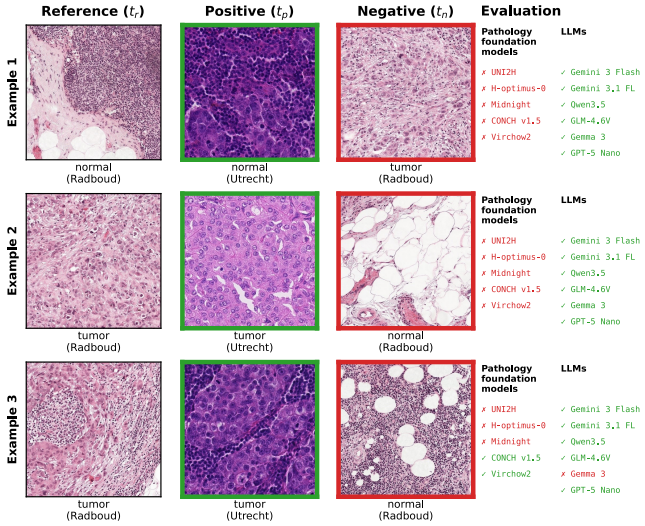


Figure 4: Example cross-institution comparisons on CAMELYON16. Each row shows one comparison: reference t_r (left), positive t_p (center, green border, same class from a different institution), negative t_n (right, red border, different class from the same institution). Correctness of each model's choice is shown on the right (✓ correct, × incorrect).

4.2.1 Comparison with Within-Institution Performance

As with the cross-slide analysis, we compare within-domain and cross-domain accuracy to quantify the effect of institution-level domain shift. Figure 5 compares within-institution accuracy (where $t_r, t_p, t_n \in \mathcal{T}_I$) with cross-institution accuracy (where $t_p \in \mathcal{T}_{I'}$, $I' \neq I$). On both CAMELYON16 and TCGA, all models fall below the diagonal, confirming that cross-institution discrimination is harder. The gap is larger on TCGA, where pathology foundation models cluster in the bottom-right with high within-institution accuracy (63–69%) but low cross-institution accuracy (18–44%). LLMs maintain higher cross-institution accuracy relative to their within-institution performance, mirroring the within-slide vs cross-slide pattern observed in Figure 3. This consistent pattern across both slide-level and institution-level domain shifts suggests that foundation models encode domain-specific features well but struggle to generalize, whereas LLMs are more robust to batch effects. Per-institution within-institution breakdowns are in Appendix C.2.

4.3 Scaling Training Data Does Not Resolve the Robustness Gap

A natural hypothesis is that the cross-domain failures of pathology foundation models reflect insufficient training data, and that models trained on more tiles would learn to separate biology from batch effects. We test this directly. Figure 6 shows relative similarity accuracy across all six datasets as a function of training data size (for pathology models) and input-token price as a proxy for deployment cost (for LLMs). Within the evaluated pathology encoders, larger reported training sets do not translate into better cross-domain similarity accuracy: models trained on over 100 million tiles perform comparably to those trained on far fewer. This suggests that training-set scale alone is insufficient and points to the learning objective as a likely driver. For LLMs, even the cheapest models consistently outperform pathology foundation models, indicating that the advantage is not driven by model scale

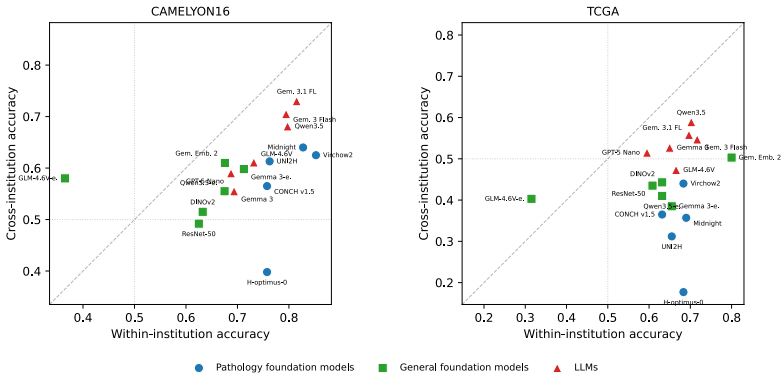


Figure 5: Within-institution vs cross-institution relative similarity accuracy on CAMELYON16 (left) and TCGA (right). Each point is a model.

or cost. General foundation models and per-dataset LLM cost-accuracy panels are in Appendix D.1. Training data sizes are listed in Appendix D.2 (Table 12).

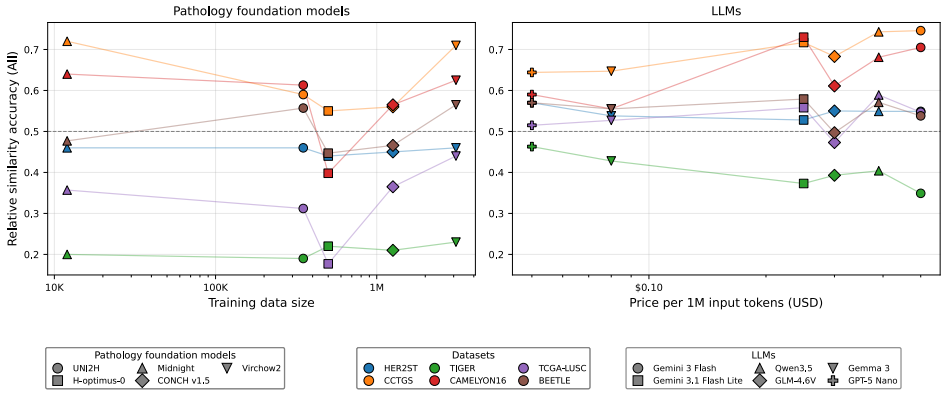


Figure 6: Relative similarity accuracy vs model cost/size across all six datasets.

4.4 Relative Similarity Predicts a Confounded Downstream Classification Task

To test whether the ranking advantage carries over to a downstream clinical task, we evaluate confounded tissue classification on CAMELYON16, TCGA, and BEETLE, following the confounded design of Kömen et al. [28]. In real cohorts a diagnosis often correlates with the collecting site, and models that exploit site-specific shortcuts can look accurate while failing to generalize [25, 10, 11]. Each query tile is matched against $N = 4$ labeled candidates. The single candidate of the query's class comes from a different institution, and the three distractors come from the query's institution, so a model that relies on the institution shortcut is misled (chance $= 0.25$). LLMs do not output feature vectors, but classification can be solved directly from their similarity judgments: embedding models select the nearest candidate (cosine similarity), and LLMs are asked which candidate shows the same tissue type. We evaluate 100 queries per dataset.

Across the 14 embedding models, relative similarity accuracy on the cross-institution triplets predicts confounded classification accuracy (Pearson $r = 0.88$, $p < 10^{-12}$, $n = 42$ model–dataset pairs). Most pathology foundation models fall near or below chance on TCGA and BEETLE, while the best LLM matches or exceeds both the best pathology and the best general foundation model on every dataset (Table 3, per-model results in Appendix G).

4.5 Ablation Studies

To assess the contribution of the biology-focused prompt, we compare it against a minimal prompt that simply asks which candidate is more similar to the reference without any domain-specific guidance (full prompt text in Appendix E.1). The biology-focused prompt instructs the LLM to focus on cell morphology and tissue architecture while ignoring non-biological differences such as staining intensity and imaging artifacts. Table 4 reports accuracy for Gemini 3 Flash and GPT-5 Nano on

Table 3: Confounded nearest-neighbor classification accuracy ($\pm$ standard deviation, 100 queries per dataset, $N = 4$ candidates, chance $= 0.25$). For foundation models, the best model per column is shown with its name. Per-model results are in Table 18.

Model	CAMELYON16	TCGA	BEETLE	Avg
Best pathology FM	0.610$\pm$0.049 (GenBio-PathFM)	0.290$\pm$0.045 (GenBio-PathFM)	0.310$\pm$0.046 (Virchow2)	0.373$\pm$0.026 (GenBio-PathFM)
Best general FM	0.400$\pm$0.049 (Gemini Emb. 2)	0.250$\pm$0.043 (Gemini Emb. 2)	0.430$\pm$0.050 (CLIP ViT-L/14)	0.343$\pm$0.027 (Gemini Emb. 2)
Gemini 3 Flash	0.588$\pm$0.050	0.354$\pm$0.048	0.410$\pm$0.049	0.450$\pm$0.028
Gemini 3.1 Flash Lite	**0.646**$\pm$0.048	0.330$\pm$0.047	**0.480**$\pm$0.050	**0.485**$\pm$0.028
Qwen3.5	0.549$\pm$0.052	**0.394**$\pm$0.049	0.430$\pm$0.050	0.458$\pm$0.029
GPT-5 Nano	0.216$\pm$0.044	0.247$\pm$0.044	0.261$\pm$0.047	0.242$\pm$0.026

HER2ST (cross-slide) and CAMELYON16 (cross-institution). For both models and both datasets, the biology-focused prompt yields higher accuracy than the minimal prompt. Even the minimal prompt outperforms all pathology foundation models except GenBio-PathFM on CAMELYON16 (0.668 vs at most 0.640). We also asked the LLM to identify the odd one out instead of the more similar candidate. On CAMELYON16 cross-institution, Gemini 3 Flash scores 0.725 with this odd-one-out prompt, matching the default prompt (0.705) and above every pathology foundation model (Appendix E.1). This highlights a practical advantage of LLMs: similarity criteria can be steered at inference time, whereas foundation model embeddings are fixed after pretraining.

Table 4: Effect of LLM prompt on relative similarity accuracy.

	HER2ST (cross-slide)		CAMELYON16 (cross-inst.)	
Prompt	Gemini 3 Flash	GPT-5 Nano	Gemini 3 Flash	GPT-5 Nano
Biology-focused (default)	0.549	0.570	0.705	0.590
Minimal	0.510	0.560	0.668	0.555

5 Conclusion

Pathology foundation models can appear robust under standard within-domain evaluations while failing to preserve biological similarity across institutions, a clinically consequential gap that scaling training data alone does not resolve. Multimodal LLMs, without domain-specific training, match or outperform these specialized models in cross-domain settings, demonstrating that robust histological similarity is achievable today and highlighting the potential of LLMs as a tool for mitigating batch effects in computational pathology. We release MOSAIC, a cross-domain benchmark spanning 6 histopathology cohorts, to support future development of models that prioritize biological signal over acquisition-context shortcuts. Our evaluation is limited to H&E-stained images. Future work could extend this framework to other pathology imaging modalities such as immunohistochemistry and multiplexed imaging, as well as spatial biology data where cross-platform batch effects pose similar challenges.

Acknowledgments and Disclosure of Funding

D.Z. is supported by an NVIDIA Academic Grant. The authors declare no competing interests.

References

[1] Alma Andersson, Ludvig Larsson, Linnea Stenbeck, Fredrik Salmén, Anna Ehinger, Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Roden, Alex Swarbrick, Åke Borg, Jonas Frisén, Camilla Engblom, and Joakim Lundeberg. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nature Communications*, 12(1):6012, 2021. doi: 10.1038/s41467-021-26271-2.

[2] Duygu Arslan, Sina Sehlaver, Erce Guder, Mehmet Arda Temena, Alper Bahcekapili, Umut Ozdemir, Duriye Ozer Turkay, Gunes Guner, Servet Guresci, Cenk Sokmensuer, et al. Colorectal cancer tumor grade segmentation: A new dataset and baseline results. *Heliyon*, 11(4), 2025. doi: 10.1016/j.heliyon.2025.e42467.

[3] Babak Ehteshami Bejnordi, Mitko Veta, Paul J van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM van der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection

of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585.

[4] Bioptimus. H-optimus-1. Hugging Face model card, 2025.

[5] Gabriele Campanella, Shengjia Chen, Manbir Singh, Ruchika Verma, Silke Muehlstedt, Jennifer Zeng, Aryeh Stock, Matt Croken, Brandon Veremis, Abdulkadir Elmas, Ivan Shujski, Noora Neittaanmäki, Kuan-lin Huang, Ricky Kwan, Jane Houldsworth, Adam J. Schoenfeld, and Chad Vanderbilt. A clinical benchmark of public self-supervised pathology foundation models. *Nature Communications*, 16:3640, 2025. doi: 10.1038/s41467-025-58796-1.

[6] Gianluca Carloni, Biagio Brattoli, Seongho Keum, Jongchan Park, Taebum Lee, Chang Ho Ahn, and Sergio Pereira. Pathology foundation models are scanner sensitive: Benchmark and mitigation with contrastive scangen loss. In *MedAGI 2025: Third International Workshop on Foundation Models for General Medical AI (MICCAI 2025)*, pages 44–53. Springer, 2025. doi: 10.1007/978-3-032-07845-2_5.

[7] Binghao Chai, Jianan Chen, Paul Cool, Fatine Oumlil, Anna Tollitt, David Steiner, Tapabrata Chakraborti, and Adrienne M Flanagan. Impact of tissue staining and scanner variation on the performance of pathology foundation models: a study of sarcomas and their mimics. *The Journal of Pathology: Clinical Research*, 12, 2026. doi: 10.1002/2056-4538.70080.

[8] Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. MLLM-as-a-Judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *International Conference on Machine Learning*, pages 6562–6595, 2024.

[9] Richard J Chen, Tong Ding, Ming Y Lu, Drew F K Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30:850–862, 2024. doi: 10.1038/s41591-024-02857-3.

[10] Edwin D. de Jong, Eric Marcus, and Jonas Teuwen. Current pathology foundation models are unrobust to medical center differences. *arXiv preprint arXiv:2501.18055*, 2025.

[11] Taher Dehkharghanian, Azam Asilian Bidgoli, Abtin Riasatian, Pooria Mazaheri, Clinton J. V. Campbell, Liron Pantanowitz, H. R. Tizhoosh, and Shahryar Rahnamayan. Biased data, biased AI: Deep networks predict the acquisition site of TCGA images. *Diagnostic Pathology*, 18(1): 67, 2023. doi: 10.1186/s13000-023-01355-3.

[12] Tong Ding, Sophia J Wagner, Andrew H Song, Richard J Chen, Ming Y Lu, Andrew Zhang, Anurag Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31:3749–3761, 2025. doi: 10.1038/s41591-025-03982-3.

[13] Dyke Ferber, Georg Wölflein, Isabella C Wiest, Marta Ligero, Srividhya Sainath, Narmin Ghaffari Laleh, Omar SM El Nahhas, Gustav Müller-Franzes, Dirk Jäger, Daniel Truhn, and Jakob Nikolas Kather. In-context learning enables multimodal large language models to classify cancer pathology images. *Nature Communications*, 15(1):10104, 2024. doi: 10.1038/s41467-024-51465-9.

[14] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Axel Camara, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, 2023. doi: 10.1101/2023.07.21.23292757.

[15] Alexandre Filiot, Paul Jacob, Alice Mac Kain, and Charlie Saillard. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173*, 2024. doi: 10.48550/arXiv.2409.09173.

[16] Alexandre Filiot, Nicolas Dop, Oussama Tchita, Auriane Riou, Rémy Dubois, Thomas Peeters, Daria Valter, Marin Scalbert, Charlie Saillard, Geneviève Robin, and Antoine Olivier. Distilling foundation models for robust and efficient models in digital pathology. In *Medical Image*

Computing and Computer Assisted Intervention – MICCAI 2025, volume 15966 of *Lecture Notes in Computer Science*, pages 162–172. Springer, 2025. doi: 10.1007/978-3-032-04981-0_16.

[17] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-Bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gael Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Róbert Istvan Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, et al. Gemma 3 technical report. *ArXiv*, abs/2503.19786, 2025. doi: 10.48550/arXiv.2503.19786.

[18] Nauman Ullah Gilal, Rachida Zegour, Khaled Al-Thelaya, Erdener Özer, Marco Agus, Jens Schneider, and Sabri Boughorbel. Pathvlm-eval: Evaluation of open vision language models in histopathology. *Journal of Pathology Informatics*, 18:100455, 2025. doi: 10.1016/j.jpi.2025.100455.

[19] GLM-V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Haochen Li, Jiale Zhu, Jiali Chen, Jiaxing Xu, Jiazheng Xu, Jing Chen, Jinghao Lin, Jinhao Chen, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Ruiliang Lyu, Shangqin Tu, Sheng Yang, Shengbiao Meng, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wei Jia, Wenkai Li, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyu Zhang, Xinyue Fan, Xuancheng Huang, Yadong Xue, Yanfeng Wang, Yanling Wang, Yanzi Wang, Yifan An, Yifan Du, Yiheng Huang, Yilin Niu, Yiming Shi, Yu Wang, Yuan Wang, Yuanchang Yue, Yuchen Li, Yusen Liu, Yutao Zhang, Yuting Wang, Yuxuan Zhang, Zhao Xue, Zhengxiao Du, Zhenyu Hou, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. GLM-4.5V and GLM-4.1V-Thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025.

[20] Google DeepMind. Gemini 3 Flash model card, 2025. URL `https://deepmind.google/models/model-cards/gemini-3-flash/`.

[21] Google DeepMind. Gemini 3.1 Flash-Lite model card, 2026. URL `https://deepmind.google/models/model-cards/gemini-3-1-flash-lite/`.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. doi: 10.1109/cvpr.2016.90.

[23] Martin N Hebart, Charles Y Zheng, Francisco Pereira, and Chris I Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgments. *Nature Human Behaviour*, 4:1173–1185, 2020. doi: 10.1038/s41562-020-00951-3.

[24] Md. Ziaul Hoque, Anja Keskinarkaus, Pia Nyberg, and Tapio Seppänen. Stain normalization methods for histopathology image analysis: A comprehensive review and experimental comparison. *Information Fusion*, 102:101997, 2024. doi: 10.1016/j.inffus.2023.101997.

[25] Frederick M. Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I. Olopade, Jakob N. Kather, Nicole Cipriani, Robert L. Grossman, and Alexander T. Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12(1):4423, 2021. doi: 10.1038/s41467-021-24698-1.

[26] Saarthak Kapse, Mehmet Aygün, Elijah Cole, Emma Lundberg, Le Song, and Eric P. Xing. GenBio-PathFM: A state-of-the-art foundation model for histopathology. *bioRxiv*, 2026. doi: 10.64898/2026.03.17.712534. Preprint.

[27] Mikhail Karasikov, Joost van Doorn, Nicolas Känzig, Melis Erdal Cesur, Hugo Mark Horlings, Robert Berke, Fei Tang, and Sebastian Otálora. Training state-of-the-art pathology foundation models with orders of magnitude less data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume LNCS 15967, pages 573–583. Springer Nature Switzerland, October 2025. doi: 10.1007/978-3-032-04984-1_55.

[28] Jonah Kömen, Hannah Marienwald, Jonas Dippel, and Julius Hense. Do histopathological foundation models eliminate batch effects? a comparative study. *arXiv preprint arXiv:2411.05489*, 2024. doi: 10.48550/arXiv.2411.05489.

[29] Jonah Kömen, Edwin D. de Jong, Julius Hense, Hannah Marienwald, Jonas Dippel, Philip Naumann, Eric Marcus, Lukas Ruff, Maximilian Alber, Jonas Teuwen, Frederick Klauschen, and Klaus-Robert Müller. Towards Robust Foundation Models for Digital Pathology, 2025.

[30] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

[31] Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, et al. Gemini embedding: Generalizable embeddings from gemini. *arXiv preprint arXiv:2503.07891*, 2025. doi: 10.48550/arXiv.2503.07891.

[32] Carlijn Lems, Leslie Tessier, John-Melle Bokhorst, Mart van Rijthoven, Witali Aswolinskiy, Matteo Pozzi, Natalie Klubickova, Suzanne Dintzis, Michela Campora, Maschenka Balkenhol, et al. A multicentric dataset for training and benchmarking breast cancer segmentation in h&e slides. *arXiv preprint arXiv:2510.02037*, 2025. doi: 10.48550/arXiv.2510.02037.

[33] Siyu Lin, Haowen Zhou, Mark Watson, Ramaswamy Govindan, Richard J Cote, and Changhuei Yang. Impact of stain variation and color normalization for prognostic predictions in pathology. *Scientific Reports*, 15(1):2369, 2025. doi: 10.1038/s41598-024-83267-w.

[34] Chiara Loeffler and Jakob Nikolas Kather. Manual tumor annotations in tcga, August 2021. URL `https://doi.org/10.5281/zenodo.5320076`.

[35] Lukas Muttenthaler, Lorenz Linhardt, Jonas Dippel, Robert A Vandermeulen, Katherine L Hermann, Andrew Kyle Lampinen, and Simon Kornblith. Improving neural network representations using human similarity judgments. In *Neural Information Processing Systems*, 2023. doi: 10.48550/arXiv.2306.04507.

[36] Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew Kyle Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647:349–355, 2025. doi: 10.1038/s41586-025-09631-6.

[37] Dmitry Nechaev, Alexey Pchelnikov, and Ekaterina Ivanova. Hibou: A family of foundational vision transformers for pathology. *arXiv preprint arXiv:2406.05074*, 2024. doi: 10.48550/arXiv.2406.05074.

[38] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489:519–525, 2012. doi: 10.1038/nature11404.

[39] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. doi: 10.48550/arXiv.2304.07193.

[40] Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL `https://qwen.ai/blog?id=qwen3.5`.

[41] Alec Radford et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[42] Brett D Roads and Bradley C Love. Enriching ImageNet with human similarity judgments and psychological embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3546–3556, 2021. doi: 10.1109/CVPR46437.2021.00355.

[43] Charlie Saillard, Rodolphe Jenatton, Felipe Llinares-López, Zelda Mariet, David Cahané, Eric Durand, and Jean-Philippe Vert. H-optimus-0, 2024. URL `https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0`.

[44] Yiqing Shen, Yulin Luo, Dinggang Shen, and Jing Ke. RandStainNA: Learning stain-agnostic features from histology slides by bridging stain augmentation and normalization. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 212–221, 2022. doi: 10.1007/978-3-031-16434-7_21.

[45] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, et al. OpenAI GPT-5 System Card. *arXiv preprint arXiv:2601.03267*, 2026. doi: 10.48550/arXiv.2601.03267.

[46] David Tellez, Geert Litjens, Péter Bándi, Wouter Bulten, John-Melle Bokhorst, Francesco Ciompi, and Jeroen van der Laak. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. *Medical Image Analysis*, 58:101544, 2019. doi: 10.1016/j.media.2019.101544.

[47] Erik Thiringer, Fredrik Gustafsson, Kajsa Ledesma Eriksson, and Mattias Rantalainen. Scanner-induced domain shifts undermine the robustness of pathology foundation models. *arXiv preprint arXiv:2601.04163*, 2026.

[48] Mart van Rijthoven, Witali Aswolinskiy, Leslie Tessier, Maschenka Balkenhol, Joep M A Bogaerts, Damien Drubay, Laura Comerma Blesa, Dieter Peeters, Elisabeth Specht Stovgaard, Anne-Vibeke Lænkholm, et al. Tumor-infiltrating lymphocytes in breast cancer through artificial intelligence: biomarker analysis from the results of the TIGER challenge. *medRxiv preprint medRxiv:2025.02.28.25323078*, 2025. doi: 10.1101/2025.02.28.25323078.

[49] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi Weerasinghe, Bill J. Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181–188, 2024. doi: 10.1038/s41586-024-07441-w.

[50] Rikiya Yamashita, Jin Long, Snikitha Banda, Jeanne Shen, and Daniel Rubin. Learning domain-agnostic visual representation for computational pathology using medically-irrelevant style transfer augmentation. *IEEE Transactions on Medical Imaging*, 40(12):3945–3954, 2021. doi: 10.1109/TMI.2021.3101985.

[51] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023.

[52] Eric Zimmermann, Eugene Vorontsov, Julian Viret, Adam Casson, Michal Zelechowski, George Shaikovski, Neil Tenenholtz, James Hall, David Klimstra, Razik Yousfi, Thomas Fuchs, Nicolo Fusi, Siqi Liu, and Kristen Severson. Virchow2: Scaling self-supervised mixed magnification models in pathology. *arXiv preprint arXiv:2408.00738*, 2024.

A Experiment Setup

A.1 Within-Slide Comparisons

All three tiles come from the same slide s: $t_r, t_p, t_n \in \mathcal{T}_s$, with t_r and t_p sharing class c_+ and t_n from class c_-. For HER2ST, 20 comparisons are sampled per (slide, c_+, c_-) combination. The positive and negative are randomly assigned to positions 1 and 2 with equal probability, so chance accuracy is 50%.

A.2 Cross-Slide Comparisons

For each ordered class pair (c_+, c_-) with $c_+ \neq c_-$, we sample 100 comparisons where the reference t_r has class c_+, the positive $t_p \in \mathcal{T}_{s'}$ has class c_+ from a different slide ($s' \neq s$), and the negative $t_n \in \mathcal{T}_s$ has class c_- from the same slide as the reference. HER2ST (5 classes) yields 20 ordered pairs for 2000 total comparisons. CCTGS (6 classes, 39 slides) yields 30 pairs for 3000 total, and TIGER (4 classes, 40 slides) yields 12 pairs for 1200 total. The positive and negative are randomly assigned to positions 1 and 2 with equal probability, so chance accuracy is 50%.

A.3 Within-Institution Comparisons

All three tiles come from the same institution I: $t_r, t_p, t_n \in \mathcal{T}_I$, with t_r and t_p sharing class c_+ and t_n from class c_-. Tiles may come from different slides within the institution. For each (institution, c_+, c_-) combination, 100 comparisons are sampled.

A.4 Cross-Institution Comparisons

For each (institution I, c_+) combination, 100 comparisons are sampled where the reference $t_r \in \mathcal{T}_I$ has class c_+, the positive $t_p \in \mathcal{T}_{I'}$ has class c_+ from a different institution ($I' \neq I$), and the negative $t_n \in \mathcal{T}_I$ has class $c_- \neq c_+$ from the same institution (with c_- drawn uniformly from available classes). CAMELYON16 has 2 institutions (Radboud with 21 slides, Utrecht with 19 slides) and 2 classes (tumor vs normal, assigned by computing tumor fraction from pixel-level masks with threshold $>$50%), yielding 4 combinations and 400 total comparisons. TCGA has 3 tissue source sites (TSS-22, TSS-56, TSS-66, each with 10 slides) and 2 classes, yielding 6 combinations and 600 total. BEETLE has 4 institutions (NKI, RUMC, SCH, TCGA, each with 10 slides) and 3 classes (invasive epithelium, necrosis, non-invasive epithelium derived from pixel-level segmentation masks), yielding 12 combinations and 1200 total.

B Cross-Slide Supplementary Results

B.1 Within-Slide Accuracy

Table 5: Within-slide tile relative similarity accuracy on HER2ST, broken down by tissue class.

Method	Adipose	Breast gl.	Cancer	Connective	Immune inf.	All
Pathology foundation models						
UNI2H	**0.775**	**0.759**	0.664	**0.757**	0.712	0.728
H-optimus-0	0.709	0.691	0.682	0.748	**0.756**	0.719
Midnight	0.706	0.732	0.670	0.695	0.700	0.697
CONCH v1.5	0.706	0.736	0.691	0.689	0.724	0.705
Virchow2	0.741	0.718	**0.755**	0.739	0.709	**0.735**
General foundation models						
Gemini Emb. 2	0.616	0.591	0.720	0.691	0.641	0.662
DINOv2	0.681	0.523	0.689	0.566	0.674	0.633
ResNet-50	0.675	0.691	0.689	0.700	0.676	0.687
Qwen3.5-enc.	0.734	0.605	0.650	0.591	0.647	0.644
GLM-4.6V-enc.	0.662	0.582	0.670	0.623	0.671	0.646
Gemma 3-enc.	0.641	0.614	0.720	0.632	0.688	0.664
LLMs						
Gemini 3 Flash	0.728	0.659	0.716	0.718	0.685	0.706
Gemini 3.1 Flash Lite	0.744	0.609	0.707	0.727	0.621	0.690
Qwen3.5	0.713	0.623	0.705	0.745	0.682	0.702
GLM-4.6V	0.684	0.568	0.675	0.671	0.676	0.663
Gemma 3	0.697	0.600	0.691	0.609	0.588	0.640
GPT-5 Nano	0.669	0.573	0.634	0.577	0.624	0.616

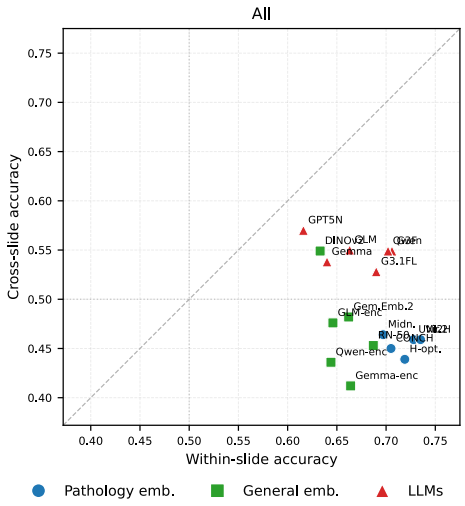


Figure 7: Within-slide vs cross-slide relative similarity accuracy on HER2ST (overall accuracy). Each point is a method, colored/marked by category (pathology foundation models = blue circles, general foundation models = green squares, LLMs = red triangles) and labeled with an abbreviated name. Dashed diagonal: $y = x$; dotted gray lines: chance (0.5). All methods fall below the diagonal, confirming that cross-slide discrimination is harder. Pathology foundation models cluster in the lower-right (high within-slide, low cross-slide), while LLMs maintain higher cross-slide accuracy relative to their within-slide performance.

B.2 Per-Class Pair Confusion Heatmaps

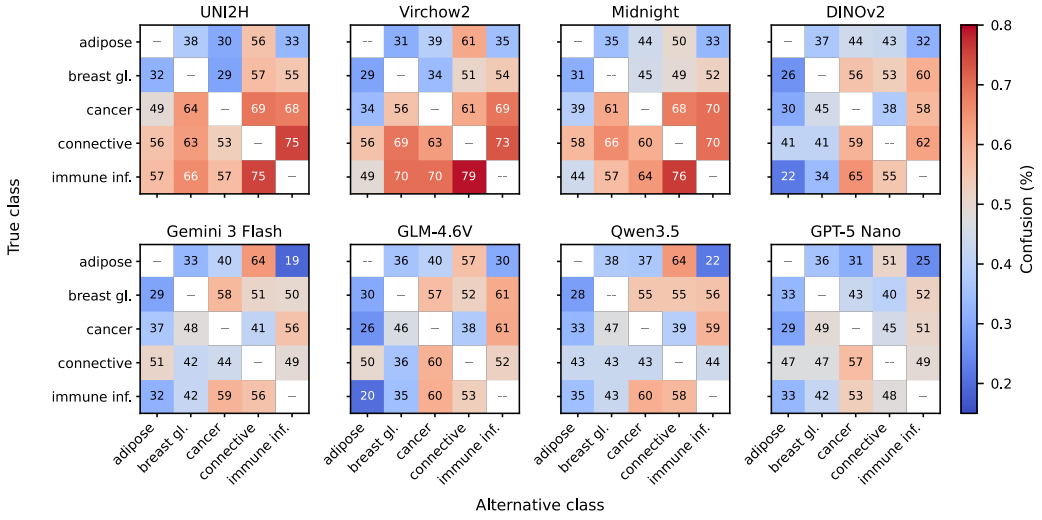


Figure 8: Per-class pair confusion for cross-slide comparisons on HER2ST. Rows indicate the positive class c_+ (shared by reference t_r and positive t_p) and columns the negative class c_-. Higher values (warmer colors) indicate greater confusion.

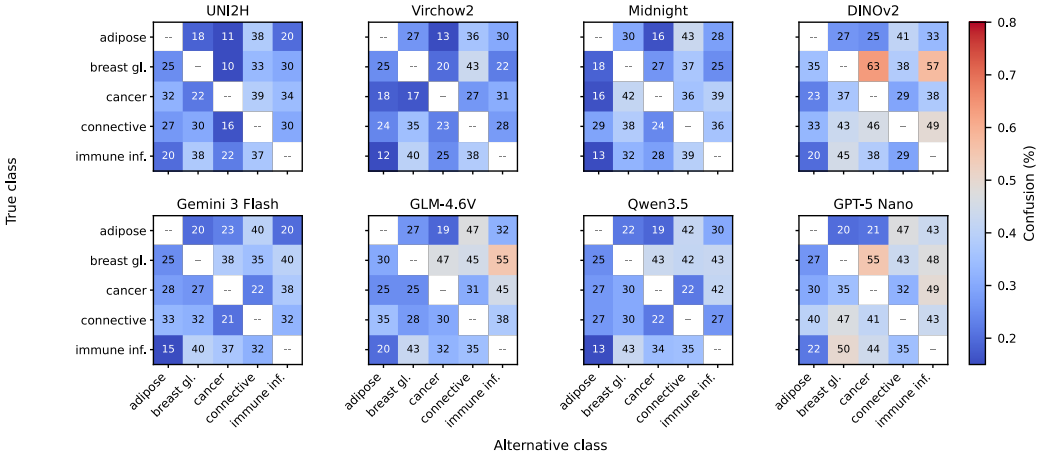


Figure 9: Per-class pair confusion for within-slide comparisons on HER2ST. Rows indicate the positive class c_+ and columns the negative class c_-. Higher values (warmer colors) indicate greater confusion.

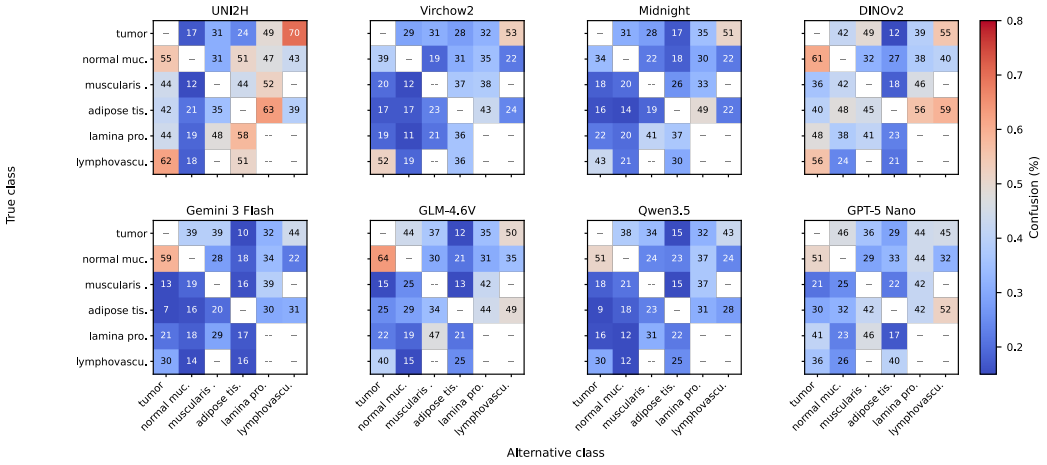


Figure 10: Per-class pair confusion for cross-slide comparisons on CCTGS. Rows indicate the positive class c_+ and columns the negative class c_-. Higher values (warmer colors) indicate greater confusion.

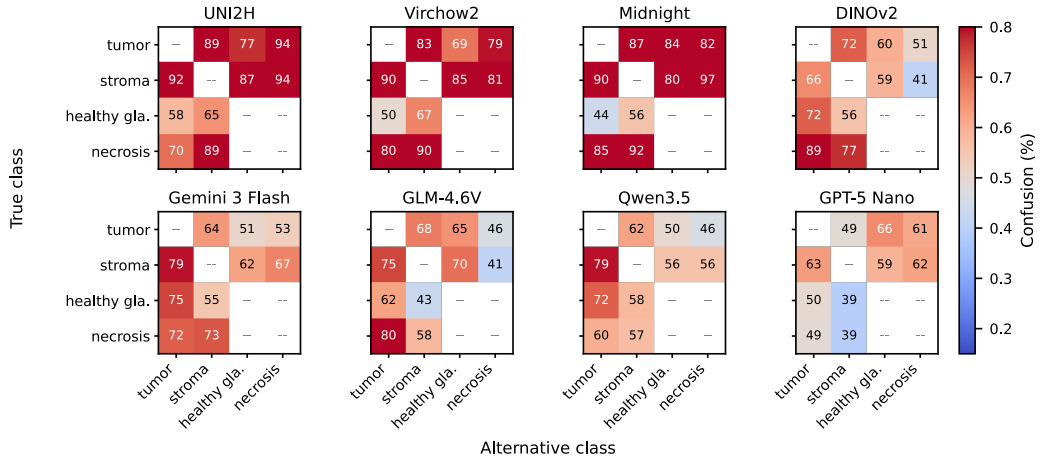


Figure 11: Per-class pair confusion for cross-slide comparisons on TIGER. Rows indicate the positive class c_+ and columns the negative class c_-. Higher values (warmer colors) indicate greater confusion.

C Cross-Institution Supplementary Results

C.1 Per-Institution Breakdown

Table 6: Cross-institution tile relative similarity accuracy on CAMELYON16.

Method	Radboud normal	Radboud tumor	Utrecht normal	Utrecht tumor	All
Pathology foundation models					
UNI2H	0.340	0.770	0.440	**0.900**	0.613
H-optimus-0	0.390	0.310	0.490	0.400	0.398
H-optimus-1	0.420	0.210	0.500	0.280	0.352
Midnight	**0.610**	0.670	0.550	0.730	0.640
CONCH v1.5	0.460	0.630	0.410	0.760	0.565
Virchow2	0.500	0.660	0.630	0.710	0.625
GenBio-PathFM	0.530	0.780	**0.670**	0.890	0.718
General foundation models					
Gemini Emb. 2	0.510	0.800	0.400	0.730	0.610
DINOv2	0.370	0.750	0.250	0.690	0.515
ResNet-50	0.380	0.700	0.330	0.560	0.492
CLIP ViT-L/14	0.370	0.680	0.390	0.690	0.532
Qwen3.5-enc.	0.340	0.630	0.480	0.770	0.555
GLM-4.6V-enc.	0.430	0.750	0.410	0.730	0.580
Gemma 3-enc.	0.510	0.730	0.410	0.740	0.598
LLMs					
Gemini 3 Flash	0.590	0.790	0.550	0.890	0.705
Gemini 3.1 Flash Lite	0.600	**0.820**	0.600	**0.900**	**0.730**
Qwen3.5	0.593	0.700	0.581	0.840	0.681
GLM-4.6V	0.479	**0.820**	0.414	0.720	0.611
Gemma 3	0.480	0.600	0.420	0.720	0.555
GPT-5 Nano	0.550	0.680	0.410	0.720	0.590

Table 7: Cross-institution tile relative similarity accuracy on TCGA.

Method	TSS-22 normal	TSS-22 tumor	TSS-56 normal	TSS-56 tumor	TSS-66 normal	TSS-66 tumor	All
Pathology foundation models							
UNI2H	0.190	0.540	0.220	0.450	0.090	0.380	0.312
H-optimus-0	0.020	0.040	0.200	0.300	0.200	0.300	0.177
H-optimus-1	0.030	0.030	0.260	0.470	0.180	0.450	0.237
Midnight	0.110	0.080	**0.470**	0.590	0.300	0.590	0.357
CONCH v1.5	0.300	0.410	0.300	0.480	0.300	0.400	0.365
Virchow2	0.330	0.470	0.300	0.530	0.430	0.580	0.440
GenBio-PathFM	0.440	0.670	0.290	0.550	0.290	0.390	0.438
General foundation models							
Gemini Emb. 2	0.490	0.630	0.360	0.580	0.390	0.570	0.503
DINOv2	0.350	0.490	0.300	0.590	0.410	0.520	0.443
ResNet-50	0.450	0.510	0.250	0.390	0.370	0.640	0.435
CLIP ViT-L/14	0.420	0.530	0.410	0.550	0.430	0.640	0.497
Qwen3.5-enc.	0.330	0.410	0.350	0.490	0.330	0.550	0.410
GLM-4.6V-enc.	0.350	0.340	0.460	0.500	0.330	0.440	0.403
Gemma 3-enc.	0.140	0.390	0.380	0.600	0.260	0.540	0.385
LLMs							
Gemini 3 Flash	0.490	0.750	0.360	0.610	0.390	**0.680**	0.547
Gemini 3.1 Flash Lite	**0.680**	0.790	0.400	0.560	0.360	0.560	0.558
Qwen3.5	0.609	**0.794**	0.455	**0.630**	0.424	0.626	**0.589**
GLM-4.6V	0.413	0.544	0.299	0.540	0.378	0.660	0.473
Gemma 3	0.490	0.640	0.430	0.580	**0.500**	0.520	0.527
GPT-5 Nano	0.580	0.540	0.360	0.590	0.370	0.650	0.515

Table 8: Cross-institution tile relative similarity accuracy on BEETLE.

Method	NKI inv.	NKI nec.	NKI non-inv.	RUMC inv.	RUMC nec.	RUMC non-inv.	SCH inv.	SCH nec.	SCH non-inv.	TCGA inv.	TCGA nec.	TCGA non-inv.	All
Pathology foundation models													
UNI2H	0.650	0.750	0.800	0.620	0.670	0.670	0.300	0.310	0.320	0.400	**0.680**	0.520	0.557
H-optimus-0	0.600	0.570	0.430	**0.800**	0.410	0.490	0.410	0.310	0.390	0.500	0.170	0.280	0.447
H-optimus-1	0.490	0.510	0.620	0.740	0.590	0.740	0.340	0.180	0.400	0.530	0.370	0.530	0.503
Midnight	0.350	0.740	0.860	0.120	0.630	0.700	0.140	0.420	0.590	0.190	0.290	0.700	0.477
CONCH v1.5	0.450	0.470	0.450	0.700	0.710	0.590	0.390	0.370	0.360	0.360	0.310	0.430	0.466
Virchow2	0.390	0.810	0.740	0.660	0.520	**0.850**	0.600	0.170	0.480	0.420	0.420	**0.720**	0.565
GenBio-PathFM	0.570	0.610	**0.890**	0.600	0.590	0.670	0.380	0.220	0.530	0.400	0.300	0.550	0.526
General foundation models													
Gemini Emb. 2	0.640	0.670	0.570	0.690	0.710	0.520	0.460	0.630	0.400	0.690	0.360	0.420	0.563
DINOv2	0.710	0.500	0.600	0.510	0.190	0.350	**0.620**	0.480	0.550	**0.800**	0.370	0.440	0.510
ResNet-50	0.480	0.360	0.360	0.370	0.450	0.470	0.490	0.470	0.470	0.650	0.490	0.490	0.463
CLIP ViT-L/14	0.620	0.430	0.440	0.720	**0.770**	0.450	0.380	0.560	0.540	0.760	0.630	0.510	0.568
Qwen3.5-enc.	0.590	0.230	0.340	0.460	0.160	0.260	0.290	0.430	0.460	0.400	0.170	0.160	0.329
GLM-4.6V-enc.	0.550	0.400	0.670	0.600	0.380	0.510	0.460	0.350	0.410	0.660	0.350	0.430	0.481
Gemma 3-enc.	0.650	0.530	0.470	0.630	0.750	0.520	0.540	0.480	0.410	0.730	0.320	0.530	0.547
LLMs													
Gemini 3 Flash	**0.760**	0.810	0.540	0.790	0.660	0.520	0.430	0.460	0.360	0.680	0.160	0.290	0.538
Gemini 3.1 Flash Lite	0.650	0.810	0.610	0.760	0.660	0.450	0.510	0.650	0.450	0.730	0.310	0.360	**0.579**
Qwen3.5	0.620	**0.820**	0.590	0.620	**0.770**	0.590	0.460	**0.660**	0.360	0.760	0.200	0.400	0.571
GLM-4.6V	0.707	0.500	0.600	0.660	0.354	0.379	0.550	0.530	0.560	0.720	0.020	0.378	0.497
Gemma 3	0.600	0.670	0.510	0.640	0.650	0.490	0.610	0.630	0.600	0.660	0.200	0.400	0.555
GPT-5 Nano	0.710	0.510	0.720	0.550	0.590	0.540	**0.620**	0.560	**0.700**	0.590	0.300	0.450	0.570

C.2 Within-Institution Accuracy

C.2.1 CAMELYON16

Table 9: Within-institution tile relative similarity accuracy on CAMELYON16.

Method	Radboud normal	Radboud tumor	Utrecht normal	Utrecht tumor	All
Pathology foundation models					
UNI2H	0.610	**0.970**	0.530	0.940	0.763
H-optimus-0	0.570	0.890	0.680	0.890	0.758
Midnight	**0.730**	**0.970**	0.680	0.930	0.828
CONCH v1.5	0.610	0.950	0.560	0.910	0.758
Virchow2	0.710	0.960	**0.790**	**0.950**	**0.853**
General foundation models					
DINOv2	0.450	0.840	0.370	0.870	0.633
ResNet-50	0.460	0.820	0.460	0.760	0.625
Qwen3.5-enc.	0.390	0.910	0.610	0.790	0.675
GLM-4.6V-enc.	0.240	0.420	0.330	0.470	0.365
Gemma 3-enc.	0.600	0.860	0.510	0.880	0.713
LLMs					
Gemini 3 Flash	0.690	0.920	0.650	0.919	0.794
Gemini 3.1 Flash Lite	**0.730**	0.930	0.690	0.910	0.815
Qwen3.5	0.705	0.908	0.674	0.902	0.797
GLM-4.6V	0.625	0.907	0.505	0.870	0.732
Gemma 3	0.596	0.727	0.590	0.860	0.693
GPT-5 Nano	0.620	0.780	0.580	0.770	0.688

C.2.2 TCGA

Table 10: Within-institution tile relative similarity accuracy on TCGA.

Method	TSS-22 normal	TSS-22 tumor	TSS-56 normal	TSS-56 tumor	TSS-66 normal	TSS-66 tumor	All
Pathology foundation models							
UNI2H	0.730	0.760	0.430	0.690	0.490	0.830	0.655
H-optimus-0	0.720	0.820	0.450	0.730	0.460	**0.920**	0.683
Midnight	0.760	0.750	0.520	0.650	**0.570**	0.890	0.690
CONCH v1.5	0.710	0.600	0.450	0.690	0.470	0.870	0.632
Virchow2	0.780	0.720	0.500	0.720	0.540	0.840	0.683
General foundation models							
DINOv2	0.610	0.790	0.430	0.710	0.470	0.780	0.632
ResNet-50	0.660	0.700	0.410	0.650	0.440	0.790	0.608
Qwen3.5-enc.	0.660	0.810	0.500	0.640	0.500	0.680	0.632
GLM-4.6V-enc.	0.200	0.410	0.320	0.460	0.160	0.340	0.315
Gemma 3-enc.	0.760	0.800	0.470	0.660	0.560	0.680	0.655
LLMs							
Gemini 3 Flash	0.810	0.810	0.460	**0.840**	0.510	0.870	**0.717**
Gemini 3.1 Flash Lite	**0.830**	0.760	0.490	0.800	0.550	0.750	0.697
Qwen3.5	0.773	0.800	**0.535**	0.723	**0.570**	0.816	0.702
GLM-4.6V	0.747	**0.828**	0.429	0.710	0.458	0.810	0.666
Gemma 3	0.700	0.790	0.510	0.650	0.540	0.710	0.650
GPT-5 Nano	0.700	0.670	0.460	0.550	0.520	0.670	0.595

C.2.3 BEETLE

Table 11: Within-institution tile relative similarity accuracy on BEETLE.

Method	NKI inv.	NKI nec.	NKI non-inv.	RUMC inv.	RUMC nec.	RUMC non-inv.	SCH inv.	SCH nec.	SCH non-inv.	TCGA inv.	TCGA nec.	TCGA non-inv.	All
Pathology foundation models													
UNI2H	0.930	0.940	**0.980**	0.770	**1.000**	0.945	0.945	0.905	0.740	0.785	**1.000**	0.780	0.893
H-optimus-0	0.885	0.950	0.965	0.845	1.000	0.910	**0.985**	0.795	0.715	0.815	1.000	**0.795**	0.888
Midnight	0.940	**0.965**	0.950	0.870	1.000	0.945	0.930	**0.905**	0.770	0.905	0.995	0.755	**0.910**
CONCH v1.5	0.840	0.960	0.805	0.870	0.995	0.905	0.970	0.825	0.640	0.765	0.975	0.755	0.859
Virchow2	**0.945**	0.955	0.970	0.835	1.000	**0.975**	0.980	0.870	**0.800**	0.790	0.975	0.765	0.905
General foundation models													
Gemini Emb. 2	0.835	0.965	0.765	0.815	0.980	0.800	0.890	0.830	0.595	0.910	1.000	0.705	0.841
DINOv2	0.835	0.930	0.780	0.675	0.955	0.760	0.900	0.625	0.575	0.915	0.800	0.690	0.787
ResNet-50	0.755	0.950	0.735	0.720	0.990	0.725	0.835	0.705	0.570	0.805	0.975	0.670	0.786
Qwen3.5-enc.	0.805	0.950	0.785	0.720	0.885	0.715	0.815	0.745	0.555	0.870	0.985	0.710	0.795
GLM-4.6V-enc.	0.675	0.885	0.820	0.630	0.945	0.760	0.840	0.650	0.575	0.745	0.825	0.710	0.755
Gemma 3-enc.	0.775	0.955	0.795	0.895	0.975	0.800	0.850	0.695	0.600	0.930	0.935	0.755	0.830
LLMs													
Gemini 3 Flash	0.900	0.955	0.855	0.899	0.985	0.780	0.870	0.730	0.600	0.895	0.990	0.775	0.853
Gemini 3.1 Flash Lite	0.900	0.965	0.810	0.835	1.000	0.745	0.895	0.765	0.610	0.820	0.970	0.745	0.838
Qwen3.5	0.838	0.964	0.820	**0.919**	1.000	0.781	0.899	0.893	0.648	0.848	0.985	0.758	0.863
GLM-4.6V	0.798	0.943	0.868	0.850	1.000	0.795	0.895	0.719	0.582	**0.934**	0.980	0.749	0.843
Gemma 3	0.716	0.934	0.807	0.778	0.969	0.769	0.850	0.735	0.565	0.873	0.880	0.665	0.794
GPT-5 Nano	0.780	0.850	0.845	0.810	0.925	0.775	0.740	0.545	0.565	0.725	0.630	0.780	0.748

C.3 Per-Class Pair Confusion

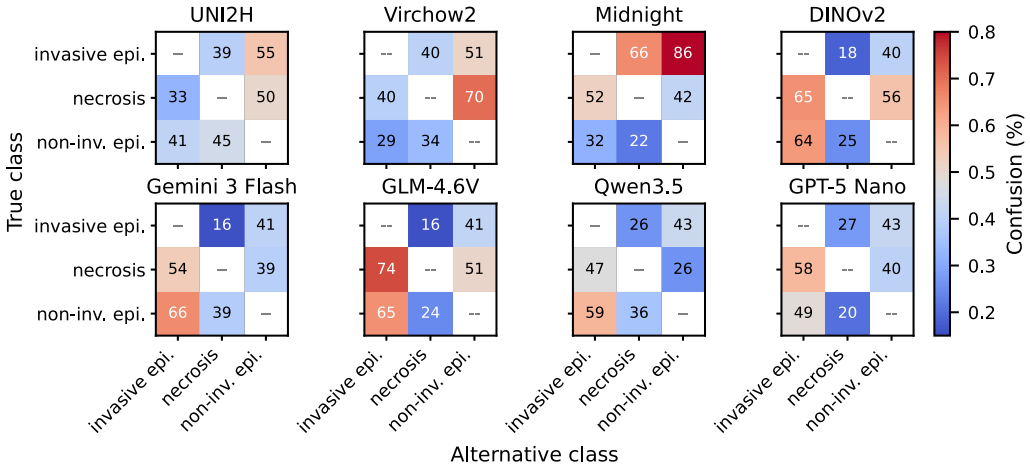


Figure 12: Per-class pair confusion for cross-institution comparisons on BEETLE. Rows indicate the positive class c_+ and columns the negative class c_-. Higher values (warmer colors) indicate greater confusion.

D Model Cost and Training Data

D.1 Accuracy vs Model Cost

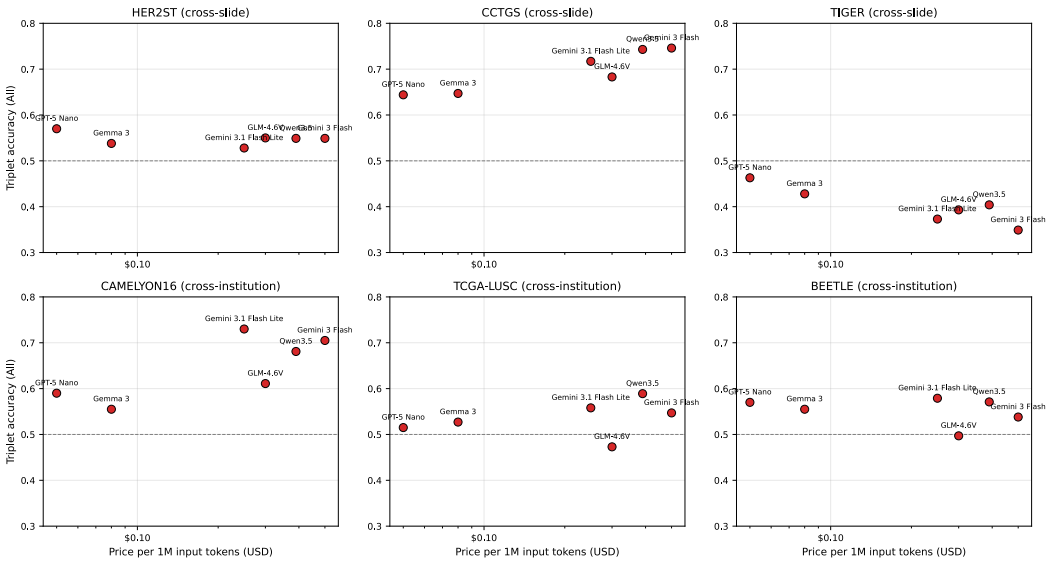


Figure 13: LLM relative similarity accuracy vs input-token price. Top row: cross-slide accuracy on HER2ST, CCTGS, and TIGER (Table 2). Bottom row: cross-institution accuracy on CAMELYON16, TCGA, and BEETLE (Tables 6, 7, and 8). Each point is one LLM. Dashed line indicates chance (0.5).

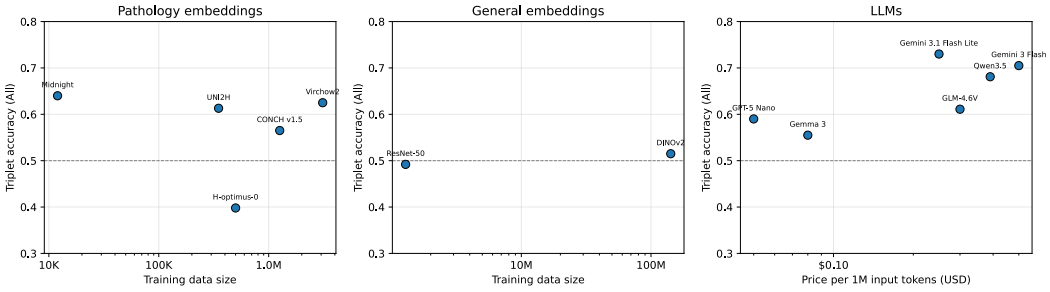


Figure 14: CAMELYON16 cross-institution relative similarity accuracy vs model cost. Left: pathology foundation model accuracy vs training data size. Center: general foundation model accuracy vs training data size (Gemini Emb. 2 training size estimated at $\sim$1B). Right: LLM accuracy vs price per million input tokens. Dashed line indicates chance (50%).

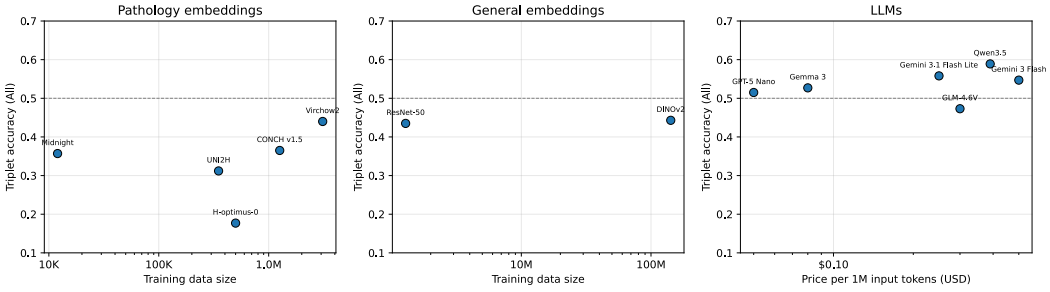


Figure 15: TCGA cross-institution relative similarity accuracy vs model cost. Left: pathology foundation model accuracy vs training data size. Center: general foundation model accuracy vs training data size. Right: LLM accuracy vs price per million input tokens. Dashed line indicates chance (50%).

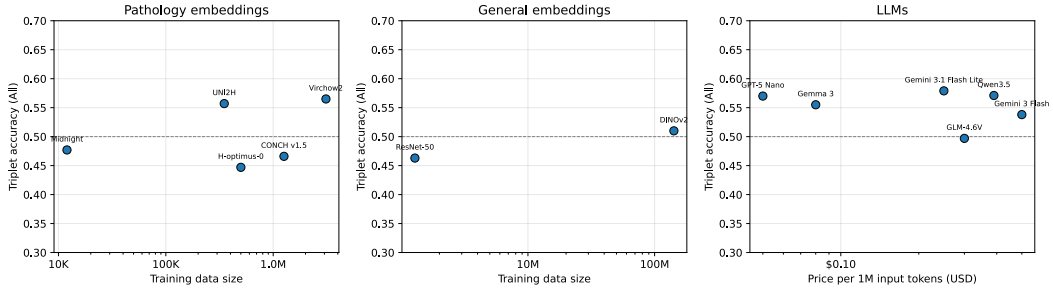


Figure 16: BEETLE cross-institution relative similarity accuracy vs model cost. Left: pathology foundation model accuracy vs training data size. Center: general foundation model accuracy vs training data size. Right: LLM accuracy vs price per million input tokens. Dashed line indicates chance (50%).

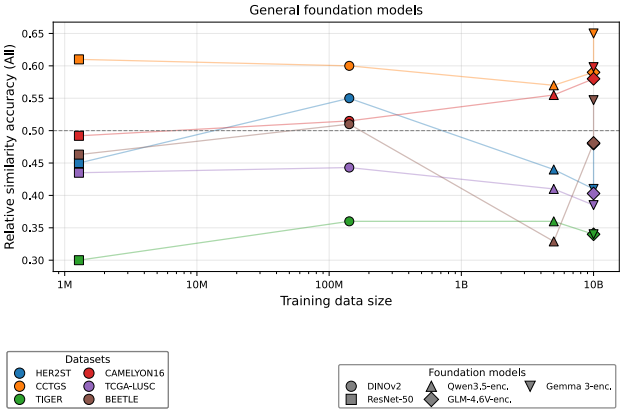


Figure 17: General foundation model relative similarity accuracy vs training data size across all six datasets. Color indicates dataset; marker shape indicates method. Lines connect methods within the same dataset. Dashed line indicates chance (0.5). Training data sizes for LLM-derived encoders are estimates based on the upstream vision encoder's pretraining data (Table 12).

D.2 Training Data Sources

Table 12: Training data sizes for foundation models used in cost/size vs accuracy analysis (Figure 6). Sizes marked with † are estimates based on the upstream vision encoder's pretraining data, as the multimodal training data size is not separately disclosed.

Model	Training size	Source	URL	Quote
Pathology foundation models				
UNI2H	>350K WSIs	HuggingFace model card	https://huggingface.co/MahmoodLab/UNI2-h	"Over 200 million image tiles sampled from over 350k diverse H&E and IHC slides sourced from Mass General Brigham."
H-optimus-0	>500K WSIs	HuggingFace model card	https://huggingface.co/bioptimus/H-optimus-0	"The model is a 1.1B parameter vision transformer trained on a proprietary collection of more than 500,000 H&E stained whole slide histology images."
Midnight	12K WSIs	Tolkach et al. (2025)	https://arxiv.org/abs/2504.05186	"We trained our first FM on the 12k TCGA WSIs alone."
CONCH v1.5	1.26M pairs	Ding et al. (2024)	https://arxiv.org/abs/2411.19666	"CONCHv1.5, an extended version of CONCH, which was trained with 1.26 million image-caption pairs using the CoCa training objective."
Virchow2	3.1M WSIs	Zimmermann et al. (2024)	https://arxiv.org/abs/2408.00738	"each trained with 3.1 million histopathology whole slide images"
General foundation models				
Gemini Emb. 2†	∼1B images	Lee et al. (2025)	https://arxiv.org/abs/2503.07891	Training data size not disclosed; estimated at ∼1B based on the Gemini model family.
DINOv2	142M images	Oquab et al. (2024)	https://arxiv.org/abs/2304.07193	"We gathered a small but diverse corpus of 142M images to validate our approach."
ResNet-50	1.28M images	ILSVRC 2012	https://www.image-net.org/challenges/LSVRC/2012/	"The training data, the subset of ImageNet containing the 1000 categories and 1.2 million images"
Qwen3.5-enc.†	∼5B images	DFN-5B model card	https://huggingface.co/apple/DFN5B-CLIP-ViT-H-14-378	"trained on a 5B subset of a pool of 43B uncurated image-text pairs"
GLM-4.6V-enc.†	∼10B pairs	Yang et al. (2025)	https://arxiv.org/abs/2507.01006	"an initial pool of over 10 billion image-text pairs from diverse sources"
Gemma 3-enc.†	∼10B pairs	SigLIP model card; Gemma 3 report	https://arxiv.org/abs/2503.19786	"a 400M variant of the SigLIP encoder"; SigLIP pre-trained on WebLI (∼10B image-text pairs).

D.3 LLM Input-Token Prices

Table 13: LLM input-token prices used in cost vs accuracy analysis (Figure 6).

Model	Price per 1M input tokens (USD)
Gemini 3 Flash	$0.50
Gemini 3.1 Flash Lite	$0.25
Qwen3.5	$0.39
GLM-4.6V	$0.30
Gemma 3	$0.08
GPT-5 Nano	$0.05

E Ablation Studies

E.1 Prompt Ablation

We compare two prompts for LLM-based relative similarity evaluation. In both cases, the LLM receives three images (one reference and two candidates) and must select which candidate is more similar to the reference.

Biology-focused prompt (default).

> Attached are three H&E histology image tiles: one reference tile (reference_tile) and two candidate tiles (candidate_tile_1, candidate_tile_2). Your task is to determine which candidate tile is more biologically similar to the reference tile. When assessing similarity, focus on biological differences:
>
> - Cell morphology (e.g., cell size, shape, nuclear features)
> - Tissue architecture (e.g., glandular organization, stromal patterns, cell density)
>
> Ignore non-biological differences such as imaging artifacts, stain intensity, resolution, or tissue preparation variations.

Minimal prompt (ablation).

> Attached are three images: one reference image (reference_tile) and two candidate images (candidate_tile_1, candidate_tile_2). Which candidate is more similar to the reference?

Both prompts require the LLM to output a JSON response with a `more_similar_candidate` field. The biology-focused prompt provides domain-specific guidance on which features to attend to and which to ignore, while the minimal prompt contains no such guidance.

Odd-one-out prompt. We also revised the biology-focused prompt to ask which candidate is the odd one out (i.e., less similar to the reference) instead of which candidate is more similar. On CAMELYON16 cross-institution, Gemini 3 Flash scores 0.725 with the odd-one-out prompt, matching the default prompt (0.705) and above the best pathology foundation model (GenBio-PathFM, 0.718).

F Statistical Significance

We computed standard deviations of accuracy by bootstrapping over comparisons (Table 1) and assessed each pairwise model difference with a two-sided paired McNemar's test at $p < 0.05$. All models within a dataset evaluate the same comparison tuples, so outcomes are paired. For LLMs that fail to return a valid response on some comparisons, McNemar's test uses only the subset where both models returned valid responses. Bootstrap confidence intervals are computed with 10,000 resamples over the sampled comparisons. For McNemar's test, we use the exact binomial test when the number of discordant pairs is below 25, and the chi-squared test with continuity correction otherwise. The pairwise comparisons are exploratory (no multiple-comparison correction), so p-values should be interpreted accordingly.

Table 14 reports the paired McNemar's test of the best LLM on each dataset against every foundation model, and Table 15 summarizes the comparison with the best pathology and general foundation models. Given the large number of models, sub-tasks, and pairwise comparisons, Table 16 summarizes the remaining tests: for each LLM, it reports the percentage of its wins that are statistically significant, where a win is a sub-task (an overall accuracy or a column of the per-class and per-institution tables) on which the LLM is more accurate than a foundation model. A high proportion of LLM wins in cross-domain tasks are statistically significant (e.g., 74 to 75% for Gemini 3 Flash). Insignificant LLM wins concentrate in the within-slide (in-domain) setting, where pathology foundation models stay competitive and the LLM margin is smallest. Table 17 additionally reports bootstrap 95% confidence intervals on HER2ST and CAMELYON16.

Table 14: Statistical significance of the overall accuracies in Table 1. Each cell is the two-sided paired McNemar's test p-value of the best LLM on that dataset (second row) against the foundation model in that row. p-values ≥ 0.05 (difference not significant) are in **bold**.

	Cross-slide			Cross-institution		
Method	HER2ST	CCTGS	TIGER	CAMELYON16	TCGA	BEETLE
Best LLM	GPT-5 Nano	Gemini 3 Flash	GPT-5 Nano	Gemini 3.1 Flash Lite	Qwen3.5	Gemini 3.1 Flash Lite
Pathology foundation models						
UNI2H	<0.001	<0.001	<0.001	<0.001	<0.001	**0.239**
H-optimus-0	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
H-optimus-1	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Midnight	<0.001	0.021	<0.001	<0.001	<0.001	<0.001
CONCH v1.5	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Virchow2	<0.001	<0.001	<0.001	<0.001	<0.001	**0.463**
GenBio-PathFM	<0.001	**0.511**	<0.001	**0.672**	<0.001	0.002
General foundation models						
Gemini Emb. 2	<0.001	<0.001	<0.001	<0.001	<0.001	**0.324**
DINOv2	**0.153**	<0.001	<0.001	<0.001	<0.001	<0.001
ResNet-50	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
CLIP ViT-L/14	<0.001	<0.001	0.005	<0.001	<0.001	**0.531**
Qwen3.5-enc.	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
GLM-4.6V-enc.	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Gemma 3-enc.	<0.001	<0.001	<0.001	<0.001	<0.001	**0.051**

Table 15: Statistical significance across all six datasets. For each dataset we report the best-accuracy model per category (overall accuracy) and paired McNemar's p-values for the best LLM vs. the best pathology foundation model (pFM) and vs. the best general-purpose encoder (Gen). "cs" = cross-slide, "ci" = cross-institution.

Dataset	Best LLM	Best pFM	LLM vs pFM p	Best Gen. encoder	LLM vs Gen p
HER2ST (cs)	GPT-5 Nano (0.570)	GenBio-PathFM (0.465)	<0.0001	DINOv2 (0.549)	0.153
CCTGS (cs)	Gemini 3 Flash (0.746)	GenBio-PathFM (0.752)	0.511	Gemini Emb. 2 (0.671)	<0.0001
TIGER (cs)	GPT-5 Nano (0.463)	Virchow2 (0.226)	<0.0001	Qwen3.5-enc. (0.360)	<0.0001
CAMELYON16 (ci)	Gemini 3.1 F. Lite (0.730)	GenBio-PathFM (0.718)	0.672	Gemini Emb. 2 (0.610)	<0.0001
TCGA (ci)	Qwen3.5 (0.589)	Virchow2 (0.440)	<0.0001	Gemini Emb. 2 (0.503)	<0.0001
BEETLE (ci)	Gemini 3.1 F. Lite (0.579)	Virchow2 (0.565)	0.463	Gemini Emb. 2 (0.563)	0.324

Table 16: Percentage of LLM wins over pathology foundation models (pFM) and general foundation models (Gen) that are statistically significant (paired McNemar's test, $p < 0.05$). Within-domain: within-slide comparisons on HER2ST. Cross-domain: cross-slide and cross-institution comparisons on all six datasets.

	Within-domain		Cross-domain	
LLM	vs pFM	vs Gen	vs pFM	vs Gen
Gemini 3 Flash	9	52	74	75
Gemini 3.1 Flash Lite	0	62	76	66
Qwen3.5	9	45	78	73
Gemma 3	0	7	76	59
GLM-4.6V	0	20	74	55

Table 17: Bootstrap 95% confidence intervals for overall accuracy on HER2ST (cross-slide, $n = 2000$) and CAMELYON16 (cross-institution, $n = 400$). Models are grouped by category.

	HER2ST (cross-slide)			CAMELYON16 (cross-inst.)		
Method	Acc.	95% CI	n	Acc.	95% CI	n
Pathology foundation models						
UNI2H	0.459	[0.437, 0.480]	2000	0.613	[0.565, 0.660]	400
H-optimus-0	0.439	[0.416, 0.461]	2000	0.398	[0.350, 0.445]	400
Midnight	0.464	[0.442, 0.485]	2000	0.640	[0.593, 0.685]	400
CONCH v1.5	0.450	[0.428, 0.472]	2000	0.565	[0.517, 0.613]	400
Virchow2	0.459	[0.436, 0.480]	2000	0.625	[0.578, 0.672]	400
GenBio-PathFM	0.465	[0.443, 0.486]	2000	0.718	[0.672, 0.762]	400
H-optimus-1	0.466	[0.444, 0.488]	2000	0.352	[0.305, 0.400]	400
General foundation models						
Gemini Emb. 2	0.482	[0.461, 0.504]	2000	0.610	[0.562, 0.657]	400
DINOv2	0.549	[0.528, 0.571]	2000	0.515	[0.465, 0.562]	400
ResNet-50	0.453	[0.431, 0.474]	2000	0.492	[0.443, 0.540]	400
Qwen3.5-enc.	0.436	[0.414, 0.458]	2000	0.555	[0.505, 0.605]	400
GLM-4.6V-enc.	0.476	[0.454, 0.497]	2000	0.580	[0.530, 0.630]	400
Gemma 3-enc.	0.412	[0.391, 0.434]	2000	0.598	[0.550, 0.645]	400
LLMs						
Gemini 3 Flash	0.549	[0.527, 0.572]	2000	0.705	[0.660, 0.748]	400
Gemini 3.1 Flash Lite	0.528	[0.506, 0.550]	2000	0.730	[0.685, 0.775]	400
Qwen3.5	0.549	[0.527, 0.571]	2000	0.681	[0.633, 0.729]	354
GLM-4.6V	0.550	[0.529, 0.572]	1995	0.611	[0.559, 0.659]	370
Gemma 3	0.538	[0.516, 0.559]	2000	0.555	[0.507, 0.603]	400
GPT-5 Nano	0.570	[0.547, 0.591]	2000	0.590	[0.540, 0.637]	400

G Confounded Downstream Classification

Table 18 reports the confounded nearest-neighbor classification accuracy of every model (Section 4.4).

Table 18: Confounded nearest-neighbor classification accuracy (± standard deviation, 100 queries per dataset, $N = 4$ candidates, chance $= 0.25$). The best model per column is in **bold**.

Model	CAMELYON16	TCGA	BEETLE	Avg
Pathology foundation models				
UNI2H	0.590±0.049	0.100±0.030	0.220±0.041	0.303±0.024
H-optimus-0	0.190±0.039	0.070±0.026	0.250±0.043	0.170±0.021
Midnight	0.480±0.050	0.160±0.037	0.260±0.044	0.300±0.025
CONCH v1.5	0.390±0.049	0.170±0.038	0.230±0.042	0.263±0.025
Virchow2	0.460±0.050	0.170±0.038	0.310±0.046	0.313±0.026
GenBio-PathFM	0.610±0.049	0.290±0.045	0.220±0.041	0.373±0.026
H-optimus-1	0.180±0.038	0.050±0.022	0.270±0.044	0.167±0.021
General foundation models				
Gemini Emb. 2	0.400±0.049	0.250±0.043	0.380±0.049	0.343±0.027
DINOv2	0.340±0.047	0.240±0.043	0.340±0.047	0.307±0.026
ResNet-50	0.270±0.044	0.200±0.040	0.240±0.043	0.237±0.024
Qwen3.5-enc.	0.260±0.044	0.160±0.037	0.250±0.043	0.223±0.024
GLM-4.6V-enc.	0.350±0.048	0.170±0.038	0.290±0.045	0.270±0.025
Gemma 3-enc.	0.390±0.049	0.130±0.034	0.310±0.046	0.277±0.025
CLIP ViT-L/14	0.350±0.048	0.230±0.042	0.430±0.050	0.337±0.027
LLMs				
Gemini 3 Flash	0.588±0.050	0.354±0.048	0.410±0.049	0.450±0.028
Gemini 3.1 Flash Lite	**0.646**±0.048	0.330±0.047	**0.480**±0.050	**0.485**±0.028
Qwen3.5	0.549±0.052	**0.394**±0.049	0.430±0.050	0.458±0.029
GPT-5 Nano	0.216±0.044	0.247±0.044	0.261±0.047	0.242±0.026

H Robustness to Distance Metric

Foundation model comparisons use the raw (unnormalized) pooled CLS feature of each model with Euclidean distance. Table 19 re-scores every embedding model with cosine similarity, which is equivalent to Euclidean distance on L2-normalized embeddings. The mean absolute change in overall accuracy is at most 0.015 on every dataset. On every dataset, the best pathology foundation model under cosine similarity stays below the best LLM, so no conclusion changes.

Table 19: Overall relative similarity accuracy with Euclidean distance on raw embeddings (left) vs cosine similarity (right). Mean $|\Delta|$: mean absolute difference across models.

Method	HER2ST	CCTGS	TIGER	CAMELYON16	TCGA	BEETLE
Pathology foundation models						
UNI2H	0.459 / 0.468	0.589 / 0.558	0.185 / 0.198	0.613 / 0.682	0.312 / 0.223	0.557 / 0.488
H-optimus-0	0.439 / 0.443	0.552 / 0.556	0.222 / 0.222	0.398 / 0.403	0.177 / 0.155	0.447 / 0.453
Midnight	0.464 / 0.460	0.724 / 0.720	0.203 / 0.210	0.640 / 0.647	0.357 / 0.362	0.477 / 0.477
CONCH v1.5	0.450 / 0.449	0.560 / 0.564	0.214 / 0.213	0.565 / 0.550	0.365 / 0.363	0.466 / 0.470
Virchow2	0.459 / 0.454	0.714 / 0.704	0.226 / 0.222	0.625 / 0.620	0.440 / 0.427	0.565 / 0.570
General foundation models						
Gemini Emb. 2	0.482 / 0.482	0.671 / 0.671	0.357 / 0.357	0.610 / 0.610	0.503 / 0.503	0.563 / 0.563
DINOv2	0.549 / 0.547	0.602 / 0.598	0.357 / 0.358	0.515 / 0.510	0.443 / 0.445	0.510 / 0.507
ResNet-50	0.453 / 0.457	0.611 / 0.631	0.299 / 0.275	0.492 / 0.495	0.435 / 0.435	0.463 / 0.450
Qwen3.5-enc.	0.436 / 0.415	0.569 / 0.570	0.360 / 0.322	0.555 / 0.550	0.410 / 0.403	0.329 / 0.388
GLM-4.6V-enc.	0.476 / 0.489	0.590 / 0.587	0.338 / 0.341	0.580 / 0.590	0.403 / 0.400	0.481 / 0.478
Gemma 3-enc.	0.412 / 0.410	0.649 / 0.649	0.339 / 0.342	0.598 / 0.605	0.385 / 0.380	0.547 / 0.543
Mean $\|\Delta\|$	0.006	0.007	0.009	0.012	0.013	0.015

I Invalid LLM Responses

A response is valid when the parser recovers exactly one of the offered candidates from the LLM's output JSON. The valid-response rate is at least 99% for the top LLMs on every dataset. Open-weight models hosted by third-party providers have lower valid-response rates (usually around 90%). Most invalid responses come from reasoning models that were not allocated a sufficient reasoning

budget, which caused the output to be truncated before the answer was produced. Allocating a larger reasoning budget fixes this issue. Scoring invalid responses as incorrect rather than excluding them never changes the best LLM on any dataset, because the top models are nearly 100% valid, and only reorders the open-weight LLMs. On the comparisons where an LLM returned an invalid response, the best pathology foundation model scores only 0.47 to 0.67, close to its overall accuracy on the same datasets, so these comparisons are not systematically easy for foundation models.

J Class Abbreviations

Table 20: Tissue class abbreviations used in Table 2.

Dataset	Abbreviation	Full name
HER2ST	Adi	Adipose
	Bre	Breast glandular
	Can	Cancer
	Con	Connective
	Imm	Immune infiltrate
CCTGS	Adi	Adipose
	Lam	Lamina propria
	Lym	Lymphovascular invasion
	Mus	Muscularis propria
	Nor	Normal mucosa
	Tum	Tumor
TIGER	Hea	Healthy glands
	Nec	Necrosis
	Str	Stroma
	Tum	Tumor

K Model Licenses

Table 21: Licenses for models used in this paper.

Model	License
Pathology foundation models	
UNI2H	CC-BY-NC-ND 4.0
H-optimus-0	Apache 2.0
Midnight	MIT
CONCH v1.5	CC-BY-NC-ND 4.0
Virchow2	CC-BY-NC-ND 4.0
H-optimus-1	CC-BY-NC-ND 4.0
GenBio-PathFM	GenBio AI Community License
General foundation models	
DINOv2	Apache 2.0
ResNet-50	BSD 3-Clause
CLIP ViT-L/14	MIT
Gemini Embedding 2	Google API Terms of Service
Qwen3.5-encoder	Apache 2.0
GLM-4.6V-encoder	MIT
Gemma 3-encoder	Gemma Terms of Use
LLMs	
Gemini 3 Flash	Google API Terms of Service
Gemini 3.1 Flash Lite	Google API Terms of Service
GPT-5 Nano	OpenAI Terms of Service
Gemma 3	Gemma Terms of Use
GLM-4.6V	MIT
Qwen3.5	Apache 2.0

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction (Section 1) clearly state four contributions: (1) exposing a robustness gap in pathology foundation models, (2) introducing a relative similarity evaluation framework, (3) demonstrating that multimodal LLMs achieve robust cross-domain similarity, and (4) releasing a benchmark suite. All claims are supported by the experimental results in Section 4.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The conclusion discusses limitations: the evaluation is restricted to H&E-stained images, and future work could extend the framework to other pathology imaging modalities (e.g., immunohistochemistry, multiplexed imaging) and spatial biology data.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results or proofs. It is an empirical study evaluating model robustness via a relative similarity framework.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the evaluation framework (Section 3), all datasets with preprocessing details (Section 3.2), all evaluated models by name and version, triplet sampling procedures (Appendix A), and the LLM prompt text (Appendix E.1). The benchmark triplets, evaluation code, and baseline outputs will be released.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The MOSAIC dataset is available at `github.com/davidzhanglab/mosaic`, as stated in the abstract. The paper describes the evaluation framework in sufficient detail for independent reimplementation.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper describes all evaluation details: dataset composition and preprocessing (Section 3.2), triplet construction and sampling procedures (Appendix A), embedding distance computation (Section 3), and LLM prompt design (Appendix E.1). No model training is performed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: Standard deviations of overall accuracy are reported for all models on all six datasets (Table 1), and paired McNemar's tests for all six datasets are reported in Appendix F. Key pairwise comparisons are tested for statistical significance and referenced in the main text.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports that all embedding extraction and local model inference were performed on an NVIDIA RTX 4090 GPU (Section 3). LLM costs are reported via input-token prices (Appendix D.3).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available datasets and pre-trained models for evaluation purposes. No new data collection involving human subjects was conducted. The work aims to improve the safety and reliability of computational pathology models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the clinical danger of relying on foundation models whose embeddings are entangled with acquisition context (Section 1), highlighting negative societal impacts of deploying non-robust models in multi-site clinical settings. Positive impacts are discussed in the conclusion, where LLMs are proposed as complementary tools for cross-institutional retrieval, dataset harmonization, and quality control.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The released assets (comparison triplets, evaluation code, baseline outputs) are evaluation tools that do not pose risks for misuse. No new models or scraped datasets are released.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and models are cited with their original publications (Sections 3.2 and 3). Model licenses are listed in Appendix K (Table 21).

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The benchmark is documented in the paper: triplet construction procedures (Sections 3 and 3.2), sampling details (Appendix A), and baseline results. Documentation and evaluation code will accompany the released benchmark.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve research with human subjects. All data used consists of publicly available histopathology image datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs are a core component of the methodology. Section 3 describes how multimodal LLMs are used as zero-shot judges for relative similarity evaluation, and Section 4 reports their performance. All LLM models are named with specific versions.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.