Supplementary Data

Table S1 | Overview of VLMs for Whole-Slide Histopathology: Dataset Tasks, and Evaluation Metrics

Model	Training Dataset	Task	Test Dataset	Metrics used
PRISM[1]	Internal Dataset	Report generation	Hold-out	-
PRISM2[2]	Internal Dataset	VQA	TCGA	AUC
HistoGPT[3]	Internal Dataset	Report generation (Dermatopathology)	Hold-out, Münster-3H, Münster-1K, Mayo Clinic, Radboud, Queensland, Linköping, TCGA, CPTAC	Jaccard Index, Cosine Similarity, Weighted F1, Precision, Recall, Accuracy Score,
Sengupta & Brown, 2023[4]	GTEx	Report generation	Hold-out	BLEU-4, METEOR, ROUGE-L
MR-ViT[5]	Internal Dataset	Report generation (Kidney and Colon)	Hold-out	BLEU-n, METEOR, ROUGE-L, F1 Score, Total accuracy
CPath-Omni[6]	PathCap, PathInstruct-200K , VALSET-TCGA, VALSET-WNS, VALSET-CHA, KIRC, CocaHis, PAIP23, BNCB, CATCH, PAIP21, MIDOG22, KICH, CAMEL, Gleason-CNN, OCELOT, TCGA, and others	Report generation and VQA	Hold-out	BLEU-n, ROUGE-L
ANTONI-α[7]	HistAI	VQA	Hold-out	Precision, Recall, F1
Kim et al., 2024[8]	K-MEDION	Report generation	Hold-out	Composite ranking score (BLEU-4, ROUGE-L, BioLLM score, Jaccard similarity)
HistGen[9]	TCGA+Others	Report generation	Hold-out	BLEU-n, METEOR, ROUGE-L
Hu et al., 2025[10]	Gastric-3300, Gastric-online, GastricADC	Report generation (gastric WSI)	Hold-out	BLEU-n, METEOR, ROUGE
PolyPath[11]	Internal Dataset	Report generation	Hold-out	ROUGE-L and METEOR
SlideChat[12]	TCGA	VQA	Hold-out, BCNB	BLEU-n, ROUGE-L,

				LLM-assigned Score
WSI-LlaVA[13]	TCGA	VQA	Hold-out, BCNB, CPTAC-NSCLC	BLEU-n, ROUGE-L, METEOR, LLM-derived Precision and Recall, Accuracy
WSI-VQA[14]	TCGA-BRCA	VQA	Hold-out	BLEU-1, BLEU-4, METEOR, ROUGE
MI-Gen[15]	TCGA-BRCA	Report generation	Hold-out	BLEU-n, METEOR, ROUGE
ALPaCA[16]	TCGA & GTEx	VQA	Hold-out	Accuracy, F1-Score, CIDEr, LLM-based judger assessment
HistoSelect[17]	TCGA	VQA	Hold-out, BCNB, CPTAC-NSCLC	Accuracy, BLEU-n, ROUGE-L, LLM-derived Precision and Recall
PathReasoner-R1[18]	TCGA	VQA	Hold-out, BCNB, CPTAC-NSCLC	BLEU-n, METEOR, ROUGE, LLM-as-judge

Table S2 | NLG scores across 3 external cohort (Aggregated Overall, CPTAC, PANDA and BRACS)

Metric	Overall		
	PRISM	**SlideFlame**	**Corrected P Value**
BLEU-2	0.0873 [0.0840-0.0906]	0.0252 [0.0243-0.0260]	< 0.001
ROUGE-L	0.2096 [0.2051-0.2138]	0.0778 [0.0761-0.0794]	< 0.001
METEOR	0.2282 [0.2220-0.2344]	0.1528 [0.1492-0.1565]	< 0.001
BERTScore	0.3390 [0.3357-0.3423]	0.2276 [0.2252-0.2298]	< 0.001

Metric	Breast (BRACS)		
	PRISM	**SlideFlame**	**Corrected P Value**
BLEU-2	0.0104 [0.0086-0.0123]	0.0689 [0.0656-0.0723]	< 0.001
ROUGE-L	0.2096 [0.2051-0.2138]	0.1820 [0.1788-0.1855]	< 0.001
METEOR	0.2282 [0.2220-0.2344]	0.2099 [0.2045-0.2157]	< 0.001
BERTScore	0.2274 [0.2241-0.2306]	0.4005 [0.3968-0.4043]	< 0.001

Metric	Prostate (PANDA)		
	PRISM	**SlideFlame**	**Corrected P Value**

BLEU-2	0.0911 [0.0808-0.1011]	0.0238 [0.0216-0.0262]	< 0.001
ROUGE-L	0.1838 [0.1697-0.1972]	0.0879 [0.0830-0.0933]	< 0.001
METEOR	0.2207 [0.2007-0.2416]	0.2166 [0.2021-0.2316]	0.074
BERTScore	0.3581 [0.3438-0.3734]	0.2748 [0.2661-0.2835]	< 0.001

Metric	**CPTAC**		
	Brain		
	PRISM	**SlideFlame**	**Corrected P Value**
BLEU-2	0.0006 [0.0002-0.0011]	0.0014 [0.0011-0.0016]	< 0.001
ROUGE-L	0.0033 [0.0011-0.0059]	0.0085 [0.0070-0.0099]	< 0.001
METEOR	0.0047 [0.0017-0.0082]	0.0182 [0.0151-0.0215]	< 0.001
BERTScore	0.2250 [0.2208-0.2288]	0.1502 [0.1487-0.1516]	< 0.001
	Lung		
BLEU-2	0.1125 [0.1077-0.1175]	0.0195 [0.0187-0.0204]	< 0.001
ROUGE-L	0.2697 [0.2638-0.2756]	0.0682 [0.0664-0.0700]	< 0.001
METEOR	0.3388 [0.3304-0.3476]	0.1538 [0.1488-0.1587]	< 0.001
BERTScore	0.3840 [0.3795-0.3884]	0.2004 [0.1982-0.2026]	< 0.001
	Pancreas		
BLEU-2	0.0577 [0.0531-0.0625]	0.0178 [0.0164-0.0192]	< 0.001
ROUGE-L	0.1619 [0.1553-0.1691]	0.0749 [0.0719-0.0780]	< 0.001
METEOR	0.1190 [0.1123-0.1259]	0.1475 [0.1380-0.1572]	< 0.001
BERTScore	0.3206 [0.3151-0.3260]	0.2091 [0.2060-0.2125]	< 0.001
	Kidney		
BLEU-2	0.1960 [0.1839-0.2084]	0.0501 [0.0473-0.0528]	< 0.001
ROUGE-L	0.3724 [0.3574-0.3871]	0.1258 [0.1210-0.1305]	< 0.001
METEOR	0.3745 [0.3542-0.3958]	0.2658 [0.2544-0.2783]	< 0.001

BERTScore	0.4354 [0.4233-0.4470]	0.2788 [0.2738-0.2836]	< 0.001
	Uterine Corpus		
BLEU-2	0.0240 [0.0222-0.0260]	0.0035 [0.0032-0.0038]	< 0.001
ROUGE-L	0.1052 [0.0987-0.112]	0.0199 [0.0183-0.0215]	< 0.001
METEOR	0.1110 [0.1032-0.1191]	0.0410 [0.0379-0.0442]	< 0.001
BERTScore	0.2621 [0.2575-0.2668]	0.1568 [0.1551-0.1586]	< 0.001

Table S3 | Normalised bNLI score across 3 external cohort (Aggregated Overall, CPTAC, PANDA and BRACS)

Model	**Overall**		
	PRISM	**SlideFlame**	**Corrected P Value**
bNLI	40.43 [37.12-43.74]	44.63 [41.55-47.71]	< 0.001

Model	**BRACS**		
	PRISM	**SlideFlame**	**Corrected P Value**
bNLI	36.37 [32.46-40.27]	27.37 [23.15-31.59]	< 0.001

Model	**PANDA**		
	PRISM	**SlideFlame**	**Corrected P Value**
bNLI	36.33 [32.91-39.75]	42.58 [38.41-46.75]	< 0.001

CPTAC			
Model	**Brain**		
	PRISM	**SlideFlame**	**Corrected P Value**
bNLI	8.91 [4.87-12.95]	46.85 [35.36-58.35]	< 0.001
	Lung		
bNLI	54.29 [50.65-57.92]	42.75 [39.93-45.58]	< 0.001
	Pancreas		
bNLI	48.44 [46.03-50.85]	47.81 [45.34-50.28]	0.1978

	Kidney		
bNLI	22.35 [17.10-27.59]	55.47 [50.56-60.38]	< 0.001
	Uterine Corpus		
bNLI	41.70 [37.22-46.18]	46.56 [42.32-50.81]	0.0017

Table S4 | Balanced accuracy (BA) and Macro-F1 across all cohorts (Aggregated Overall, CPTAC, PANDA and BRACS), derived from the LLM-based label-extraction framework for PRISM and SlideFlame.

Model	Overall		
	PRISM	**SlideFlame**	**Corrected P Value**
BA	0.2590 [0.2494-0.2687]	0.3847 [0.3580-0.4098]	< 0.001
Macro-F1	0.2827 [0.2721-0.2928]	0.3820 [0.3684-0.3948]	< 0.001

Model	PANDA (Prostate)		
	PRISM	**SlideFlame**	**Corrected P Value**
BA	0.6395 [0.5802-0.6938]	0.5413 [0.4695-0.6162]	0.1903
Macro-F1	0.5327 [0.4698-0.5898]	0.5643 [0.5023-0.6279]	0.4422

Model	BRACS (Breast)		
	PRISM	**SlideFlame**	**Corrected P Value**
BA	0.2973 [0.2627-0.3328]	0.2703 [0.2442-0.3000]	0.4411
Macro-F1	0.2673 [0.2307-0.3024]	0.2240 [0.1961-0.2534]	0.1419

CPTAC			
Brain (GBM)			
Metric	**PRISM**	**SlideFlame**	**Corrected P Value**
BA	0.0173 [0.0065-0.0304]	0.8608 [0.8282-0.8913]	< 0.001
Macro-F1	0.0341 [0.0129-0.0590]	0.9252 [0.9060-0.9425]	< 0.001
Lung			
BA	0.4206 [0.4099-0.4312]	0.4310 [0.4139-0.4527]	0.4649
Macro-F1	0.4749 [0.4655-0.4837]	0.4542 [0.4359-0.4773]	0.1419

Pancreas			
BA	0.3918 [0.3478-0.4358]	0.6741 [0.6312-0.7187]	< 0.001
Macro-F1	0.5219 [0.4741-0.5645]	0.7587 [0.7201-0.7959]	< 0.001
Kidney			
BA	0.0651 [0.0577-0.0726]	0.4181 [0.3215-0.5095]	< 0.001
Macro-F1	0.0974 [0.0879-0.1072]	0.3447 [0.3085-0.3834]	< 0.001
Uterine Corpus			
BA	0.1818 [0.1666-0.1963]	0.1815 [0.1674-0.1969]	0.9810
Macro-F1	0.2325 [0.2138-0.2497]	0.2192 [0.2014-0.2381]	0.2914

Table S5 | Hallucination rates in reports generated by PRISM and SlideFlame across all cohorts (Aggregated Overall, CPTAC, PANDA and BRACS).

Error-type	Overall		
	PRISM	**SlideFlame**	**Corrected P Value**
Context Mismatch	25.91% [24.85-27.02]	9.13% [8.40-9.88]	< 0.001
Case-level Overreach	18.08% [17.08-19.05]	4.37% [3.84-4.89]	< 0.001

Error-type	PANDA (Prostate)		
	PRISM	**SlideFlame**	**Corrected P Value**
Context Mismatch	16.97% [13.48-20.46]	8.37% [5.81-10.93]	< 0.001
Case-level Overreach	8.13% [5.58-10.69]	0.46% [0.0-1.16]	< 0.001

Error-type	BRACS (Breast)		
	PRISM	**SlideFlame**	**Corrected P Value**
Context Mismatch	14.99% [12.06-17.91]	3.29% [1.82-4.93]	< 0.001
Case-level Overreach	8.40% [6.03-10.96]	1.64% [0.73-2.74]	< 0.001

CPTAC			
Error-type	**Brain**		
	PRISM	**SlideFlame**	**Corrected P Value**
Context Mismatch	58.91% [54.55-63.69]	3.91% [2.17-5.86]	< 0.001

Case-level Overreach	28.26% [24.56-32.39]	0.0	< 0.001
Lung			
Context Mismatch	12.09% [10.72-13.50]	5.69% [4.70-6.72]	< 0.001
Case-level Overreach	16.47% [14.91-18.07]	0.42% [0.18-0.70]	< 0.001
Pancreas			
Context Mismatch	28.30% [24.44-32.35]	18.75% [15.44-22.05]	< 0.001
Case-level Overreach	9.74% [7.35-12.31]	0.36% [0.0-0.91]	< 0.001
Kidney			
Context Mismatch	52.52% [49.34-55.60]	9.78% [7.80-11.75]	< 0.001
Case-level Overreach	33.07% [30-36.04]	0.0	< 0.001
Uterine Corpus			
Context Mismatch	24.23% [21.40-27.06]	17.55% [15.06-20.15]	< 0.001
Case-level Overreach	17.21% [14.83-19.70]	26.72% [23.78-29.55]	< 0.001

Table S6 | Technical comparison highlighting parameter count, trainable parameters, GPU memory usage, and training dataset size for SlideFlame and PRISM.

Features	PRISM	SlideFlame
Total Parameters	557695233	507669017
Trainable Parameters	210931969	204306989
Idle GPU Memory	2.081 GiB	1.907 GiB
Training Dataset Size	587,196	17,336

Table S7 | Discriminative performance of NLG and bNLI based evaluation metrics for identifying LLM-assigned diagnostically correct reports

Metric	PRISM	SlideFlame
	AUC [95% CI]	
BLEU-2	0.7779 [0.7663-0.7890]	0.6148 [0.6003-0.6293]
METEOR	0.8257 [0.8149-0.8367]	0.6182 [0.6044-0.6318]
ROUGE-L	0.7896 [0.7786-0.8016]	0.5981 [0.5832-0.6134]
BERTScore	0.7317 [0.7190-0.7442]	0.5644 [0.5489-0.5800]
bNLI	0.8833 [0.8745-0.8917]	0.8168 [0.8052-0.8276]

You are a senior consultant neuropathologist auditing an AI-generated microscopic description from a SINGLE H&E whole-slide image (WSI) of brain tissue. Your task is NOT to improve or rewrite the report. Your task is to EXTRACT structured, auditable signals from the text so that model performance can be quantified.

GENERAL SCOPE RULES
Use ONLY the content of the provided report. Treat this strictly as a SINGLE H&E slide. Do NOT assume molecular testing, immunohistochemistry (IHC), radiology, or multi-block integration unless explicitly stated.

TARGET LABEL SET (GBM)
You must output ONE of the following labels:

"GBM" | "BRAIN-OTHER-TUMOUR" | "BRAIN-NON-TUMOUR" | "OTHER-TUMOUR" | "OTHER-NON-TUMOUR" | "UNCERTAIN"

LABEL ASSIGNMENT RULES (SUMMARY)
GBM:
Assign ONLY IF the report explicitly states "glioblastoma"
OR clearly describes the classic H&E morphologic triad:
1) High-grade astrocytic/glial tumour morphology (marked hypercellularity, nuclear pleomorphism, atypical mitotic figures) AND
2) Tumour necrosis (especially pseudopalisading necrosis) AND
3) Microvascular proliferation (glomeruloid vascular proliferation)
All three morphologic components must be present or explicitly stated.
Necrosis alone is insufficient. Microvascular proliferation alone is insufficient. High-grade morphology alone is insufficient.
BRAIN-OTHER-TUMOUR:
Assign IF the report describes a brain tumour that is not glioblastoma, including but not limited to metastatic carcinoma, lymphoma, meningioma, oligodendroglioma, diffuse astrocytoma, anaplastic astrocytoma, ependymoma and any other CNS tumour explicitly described but not consistent with GBM
BRAIN-NON-TUMOUR:
Assign IF the report describes non-neoplastic brain tissue or a benign / reactive / inflammatory brain process, including cortex, white matter, gliosis, infarct, hemorrhage, edema, demyelination, or other non-tumour findings in brain tissue.
OTHER-TUMOUR:
Assign IF the report primarily describes a non-brain tumour or a tumour from a non-CNS organ, or a clearly non-CNS neoplasm that is not appropriate for this brain WSI context.
OTHER-NON-TUMOUR:
Assign IF the report describes non-neoplastic tissue or a benign / inflammatory / reactive process from a non-CNS organ, including other organs or tissues that are not brain.
UNCERTAIN:
Assign IF:
- Tissue origin unclear
- Description too vague
- Conflicting statements
- Insufficient morphologic detail

HALLUCINATION FLAGS
HALLUCINATION TYPE A — CASE-LEVEL OVERREACH
Flag "case_level_overreach": true if the report includes statements not supportable from a single H&E slide, including:
- IDH mutation status (e.g., “IDH-wildtype” or “IDH-mutant”)
- 1p/19q codeletion
- MGMT methylation
- ATRX, p53, Ki-67 IHC results
- WHO integrated molecular diagnosis
- Multi-slide or gross integration statements (“overall tumour”, “other sections show…”)
HALLUCINATION TYPE B — CONTEXT MISMATCH
Flag "context_mismatch": true if the report includes:
- Non-brain primary organ context
- Grading systems from other organs (Gleason, Fuhrman, FIGO, etc.)
- Morphology clearly unrelated to CNS tissue
IMPORTANT:
- Evidence spans must quote exact fragments from the report (≤12 words).
- Do NOT infer missing features.
- If necrosis or microvascular proliferation is not explicitly described, assume it is absent.
- When in doubt, choose "UNCERTAIN" rather than over-calling GBM.

Figure S1 | Brain cohort evaluation prompt for extracting diagnostic labels from pathology reports

You are a senior consultant breast pathologist performing an audit of an AI-generated microscopic description for a SINGLE breast whole-slide image (WSI) from an H&E section. Your task is NOT to improve or rewrite the report. Your task is to EXTRACT structured, auditable signals from the text so that model performance can be quantified.

GENERAL SCOPE RULES
Use ONLY the content of the provided report. Do NOT assume or infer anything not explicitly stated. Treat this as SINGLE-SLIDE text.

TARGET LABEL SET (BRACS)
You must output ONE of the following labels:

"NORMAL" | "PATHOLOGICAL_BENIGN" | "UDH" | "FEA" | "ADH" | "DCIS" | "INVASIVE_CARCINOMA" | "OTHER_TUMOUR" | "OTHER_NON_TUMOUR" | "UNCERTAIN"

LABEL ASSIGNMENT RULES (SUMMARY)
- **NORMAL:** explicitly normal breast; no lesion described
- **PATHOLOGICAL_BENIGN:** benign breast lesion without atypia
- **UDH:** explicit usual ductal hyperplasia OR classic heterogeneous epithelial hyperplasia pattern
- **FEA**: explicit flat epithelial atypia OR columnar cell lesion with atypia in flat architecture
- **ADH**: explicit atypical ductal hyperplasia OR limited monomorphic low-grade proliferation (rigid bridges/micropapillary/cribriform) without invasion
- **DCIS**: explicit DCIS OR confident in situ malignant ductal architecture (e.g., cribriform/solid/micropapillary, comedonecrosis) confined to ducts; no invasion described
- **INVASIVE_CARCINOMA**: explicit invasion OR infiltrative malignant glands/nests in desmoplastic stroma or stromal invasion described
- **OTHER_TUMOUR**: any tumour from non-breast tissue
- **OTHER_NON_TUMOUR**: any non-neoplastic (normal/benign) non-breast tissue finding
- If both DCIS and invasion are described → INVASIVE_CARCINOMA
- **UNCERTAIN**: tissue origin unclear, insufficient specificity, conflicting language, or cannot map reliably

HALLUCINATION FLAGS
HALLUCINATION TYPE A — CASE-LEVEL OVERREACH
Flag "case_level_overreach": true IF report includes:
- margin status
- tumor size
- lymph node status
- pT/pN stage
- AJCC stage
- IHC results
- multifocality across specimen
- statements implying multi-slide or gross integration

HALLUCINATION TYPE B — CONTEXT MISMATCH
Flag "context_mismatch": true IF report includes:
- primary reference to non-breast organ
- tumor types inappropriate for breast context
- grading systems inappropriate for breast H&E (e.g., Gleason, Fuhrman, WHO CNS grade)
- morphologic descriptions clearly from another organ system

IMPORTANT:
- Evidence spans must quote exact fragments (≤12 words).
- Prefer INVASIVE_CARCINOMA over DCIS if invasion is described.
- Do NOT infer invasion unless explicitly supported.
- When uncertain, choose UNCERTAIN rather than over-calling.

Figure S2 | Breast cohort evaluation prompt for extracting diagnostic labels from pathology reports

You are a senior consultant thoracic pathologist performing an audit of an AI-generated microscopic description for a SINGLE lung whole-slide image (WSI). Your task is NOT to improve or rewrite the report. Your task is to EXTRACT structured, auditable signals from the text so that model performance can be quantified.

GENERAL SCOPE RULES
Use ONLY the content of the provided report. Treat this strictly as a SINGLE H&E slide. Do NOT assume gross findings, radiology, molecular testing, staging, or multi-block integration unless explicitly stated.

TARGET LABEL SET (CPTAC-LUNG)
You must output ONE of the following labels:

"LUNG_ADENOCARCINOMA" | "LUNG_SQUAMOUS_CELL_CARCINOMA"
"LUNG_ADENOSQUAMOUS_CARCINOMA"
"LUNG_OTHER_TUMOUR" | "LUNG_NON_TUMOUR" | "OTHER_TUMOUR" | "OTHER_NON_TUMOUR" | "UNCERTAIN"

LABEL ASSIGNMENT RULES
LUNG_ADENOCARCINOMA:
Assign if: The report explicitly states lung adenocarcinoma OR Malignant gland-forming epithelial tumour is described consistent with lung adenocarcinoma OR Architectural patterns typical of adenocarcinoma are described as papillary, acinar, lepidic, micropapillary, colloid, or mixed patterns.
NOTE: All architectural subtypes (papillary, acinar, lepidic, micropapillary, colloid, mixed) must still map to LUNG_ADENOCARCINOMA.
LUNG_SQUAMOUS_CELL_CARCINOMA:
Assign if: Squamous cell carcinoma is explicitly stated OR Malignant squamous morphology is described with keratinisation, keratin pearls, intercellular bridges, or polygonal eosinophilic cells with squamoid features.
LUNG_ADENOSQUAMOUS_CARCINOMA:
Assign ONLY if BOTH: Malignant glandular (adenocarcinoma-like) component AND Malignant squamous component are explicitly described.
LUNG_OTHER_TUMOUR:
Assign if lung tumour type is described that is not adenocarcinoma or squamous carcinoma, including Small cell carcinoma, Large cell carcinoma, Neuroendocrine carcinoma, Carcinoid tumour, Metastatic carcinoma, Lymphoma OR Mesothelioma
LUNG_NON_TUMOUR:
Assign if only benign lung parenchyma is described, Inflammatory, fibrotic, reactive, or organising lung processes only or No malignant lung epithelial proliferation is described
OTHER_TUMOUR:
Assign if the report primarily describes a non-lung tumour, including any definite tumour from a non-lung primary site.
OTHER_NON_TUMOUR:
Assign if the report primarily describes non-neoplastic tissue or benign/inflammatory/reactive finding from a non-lung primary site, including normal tissue, fibrosis, inflammation, reactive change, or other non-tumour findings outside the target lung categories.
UNCERTAIN:
Assign if:
- Organ of origin unclear
- Malignancy described without clear subtype
- Conflicting morphological cues
- Insufficient morphologic detail

HALLUCINATION FLAGS
HALLUCINATION TYPE A — CASE-LEVEL OVERREACH
Flag "case_level_overreach": true if the report contains statements not supportable from a single H&E slide, including:
- Margin status
- Tumour size
- Lymph node status
- pT/pN stage
- AJCC staging
- Statements implying review of entire resection
- Molecular results (EGFR, ALK, KRAS, BRAF, PD-L1, etc.), IHC results
- Statements implying multi-slide or gross integration
HALLUCINATION TYPE B — CONTEXT MISMATCH
Flag "context_mismatch": true if:
- Non-lung primary organ described
- Grading systems from other organs used (Gleason, Fuhrman, FIGO, etc.)
- Morphologic descriptions clearly unrelated to lung tissue
IMPORTANT:
- Evidence spans must quote exact fragments from the report (≤12 words).
- Do NOT infer invasion, staging, or molecular status.
- When uncertain, choose "UNCERTAIN" rather than over-calling carcinoma.

Figure S3 | Lung cohort evaluation prompt for extracting diagnostic labels from pathology reports

You are a senior consultant renal pathologist performing an audit of an AI-generated microscopic description for a SINGLE kidney whole-slide image (WSI). Your task is NOT to improve or rewrite the report. Your task is to EXTRACT structured, auditable signals from the text so that model performance can be quantified.

GENERAL SCOPE RULES
Use ONLY the content of the provided report. Treat this strictly as a SINGLE H&E slide. Do NOT assume gross findings, radiology, molecular testing, staging, or multi-block integration unless explicitly stated.

TARGET LABEL SET (CPTAC-KIDNEY)
You must output ONE of the following labels:

"RENAL_CLEAR_CELL_CC" | "RENAL_PAPILLARY_CC" | "RENAL_ONCOCYTIC_NEOPLASM" | "RENAL_TUBULAR_HIGH_GRADE_CC" | "RENAL_UNCLASSIFIED_HIGH_GRADE_CC" | "RENAL_BENIGN_MESENCHYMAL_TUMOUR" | "RENAL_NON_TUMOUR" | "RENAL_OTHER_TUMOUR" | "OTHER_TUMOUR" | "OTHER_NON_TUMOUR" | "UNCERTAIN"

LABEL ASSIGNMENT RULES
RENAL_CLEAR_CELL_CC: Assign if: The report explicitly states clear cell renal cell carcinoma OR Clear cytoplasm is described with nested, alveolar, or solid growth and delicate vasculature, consistent with clear cell RCC.
RENAL_PAPILLARY_CC:
Assign if: Papillary renal cell carcinoma is explicitly stated OR Papillary or tubulopapillary architecture with fibrovascular cores is described, consistent with papillary RCC.
RENAL_ONCOCYTIC_NEOPLASM:
Assign if: The report explicitly states oncocytoma, chromophobe renal neoplasm, or oncocytic renal neoplasm OR Granular eosinophilic cytoplasm with nested or solid growth is described, consistent with an oncocytic renal neoplasm.
RENAL_TUBULAR_HIGH_GRADE_CC:
Assign if: The report explicitly states tubular high-grade renal cell carcinoma OR Infiltrative tubular pattern with desmoplastic stroma and high nuclear grade is described.
RENAL_UNCLASSIFIED_HIGH_GRADE_CC:
Assign if: The report explicitly states unclassified high-grade renal cell carcinoma OR A high-grade pleomorphic renal carcinoma is described without clear lineage.
RENAL_BENIGN_MESENCHYMAL_TUMOUR:
Assign if: The report explicitly states a benign mesenchymal tumour OR Mesenchymal tumour morphology is described, such as adipose tissue mixed with vessels and smooth muscle, consistent with angiomyolipoma or similar benign mesenchymal lesion.
RENAL_NON_TUMOUR:
Assign if: Only benign renal parenchyma is described, Inflammatory, fibrotic, reactive, OR other non-neoplastic renal processes only OR No malignant renal epithelial proliferation is described
RENAL_OTHER_TUMOUR:
Assign if kidney tumour type is described that is not one of the renal tumour categories above, including:
Metastatic carcinoma OR Lymphoma OR Sarcoma OR Urothelial carcinoma involving the kidney/renal pelvis OR Any other malignant neoplasm not classifiable as a renal primary tumor
OTHER_TUMOUR:
Assign if the report primarily describes a non-kidney tumour, including any definite tumour from a non-kidney primary site.
OTHER_NON_TUMOUR:
Assign if the report primarily describes non-neoplastic tissue or benign/inflammatory/reactive finding from a non-kidney primary site, including normal tissue, fibrosis, inflammation, reactive change, or other non-tumour findings outside the target kidney categories.
UNCERTAIN:
Assign if: Organ of origin unclear OR Malignancy described without clear subtype OR Conflicting morphological cues OR Insufficient morphologic detail

HALLUCINATION FLAGS
HALLUCINATION TYPE A — CASE-LEVEL OVERREAC
Flag "case_level_overreach": true if the report contains statements not supportable from a single H&E slide, including:
- Margin status
- Tumour size
- Lymph node status
- pT/pN stage
- AJCC staging
- Statements implying review of entire resection
- Molecular results, IHC results
- Statements implying multi-slide or gross integration
HALLUCINATION TYPE B — CONTEXT MISMATCH
Flag "context_mismatch": true if:
- Non-kidney primary organ described
- Grading systems from other organs used
- Morphologic descriptions clearly unrelated to kidney tissue
IMPORTANT:
- Evidence spans must quote exact fragments from the report (≤12 words).
- Do NOT infer invasion, staging, or molecular status.
- When uncertain, choose "UNCERTAIN" rather than over-calling carcinoma.

Figure S4 | Renal cohort evaluation prompt for extracting diagnostic labels from pathology reports

You are a senior consultant pancreatic pathologist performing an audit of an AI-generated microscopic description for a SINGLE pancreas whole-slide image (WSI). Your task is NOT to improve or rewrite the report. Your task is to EXTRACT structured, auditable signals from the text so that model performance can be quantified.

GENERAL SCOPE RULES
Use ONLY the content of the provided report. Treat this strictly as a SINGLE H&E slide. Do NOT assume gross findings, radiology, molecular testing, staging, or multi-block integration unless explicitly stated.

TARGET LABEL SET (CPTAC-PDA)

You must output ONE of the following labels:

"PANCREATIC_DUCTAL_ADENOCARCINOMA" | "PANCREATIC_NON_TUMOUR" | "PANCREATIC_OTHER_TUMOUR" | "OTHER_TUMOUR" | "OTHER_NON_TUMOUR" | "UNCERTAIN"

LABEL ASSIGNMENT RULES
PANCREATIC_DUCTAL_ADENOCARCINOMA:
Assign if: The report explicitly states pancreatic ductal adenocarcinoma OR Infiltrative irregular angulated glands with desmoplastic stroma and cytologic atypia are described, consistent with invasive PDAC.
PANCREATIC_NON_TUMOUR:
Assign if: Only benign pancreatic parenchyma is described OR Inflammatory, fibrotic, reactive, or organizing pancreatic processes only OR No malignant pancreatic epithelial proliferation is described
PANCREATIC_OTHER_TUMOUR:
Assign if pancreatic tumour type is described that is not pancreatic ductal adenocarcinoma, including: Neuroendocrine tumour OR Solid pseudopapillary neoplasm OR Acinar cell carcinoma OR Cystic neoplasm OR Lymphoma OR Metastatic carcinoma
OTHER_TUMOUR:
Assign if the report primarily describes a non-pancreatic tumour, including any definite tumour from a non-pancreatic primary site.
OTHER_NON_TUMOUR:
Assign if the report primarily describes non-neoplastic tissue or benign/inflammatory/reactive finding from a non-pancreatic primary site, including normal tissue, fibrosis, inflammation, reactive change, or other non-tumour findings outside the target pancreas categories.
UNCERTAIN:
Assign if: Organ of origin unclear OR Malignancy described without clear subtype OR Conflicting morphological cues OR Insufficient morphologic detail

HALLUCINATION FLAGS
HALLUCINATION TYPE A — CASE-LEVEL OVERREACH
Flag "case_level_overreach": true if the report contains statements not supportable from a single H&E slide, including:
- Margin status
- Tumour size
- Lymph node status
- pT/pN stage
- AJCC staging
- Statements implying review of entire resection
- Molecular results, IHC results
- Statements implying multi-slide or gross integration

HALLUCINATION TYPE B — CONTEXT MISMATCH
Flag "context_mismatch": true if:
- Non-pancreatic primary organ described
- Grading systems from other organs used
- Morphologic descriptions clearly unrelated to pancreas tissue

IMPORTANT:
- Evidence spans must quote exact fragments from the report (≤12 words).
- Do NOT infer invasion, staging, or molecular status.
- When uncertain, choose "UNCERTAIN" rather than over-calling carcinoma.

Figure S5 | Pancreas cohort evaluation prompt for extracting diagnostic labels from pathology reports

You are a senior consultant gynecologic pathologist performing an audit of an AI-generated microscopic description for a SINGLE endometrial whole-slide image (WSI) from an H&E section. Your task is NOT to improve or rewrite the report. Your task is to EXTRACT structured, auditable signals from the text so that model performance can be quantified.

GENERAL SCOPE RULES
- Use ONLY the content of the provided report. Do NOT assume or infer anything not explicitly stated.
- Treat this as SINGLE-SLIDE text.
- Do NOT assume staging, FIGO grade, molecular subtype, or multi-slide integration.
- If the report lacks sufficient specificity to map confidently to one UCEC category, assign "UNCERTAIN".

TARGET LABEL SET (UCEC)
You must output ONE of the following labels:
"UTERINE_CLEAR_CELL_CARCINOMA" | "UTERINE_ENDOMETRIOID_CARCINOMA" | "UTERINE_SEROUS_CARCINOMA" | "UTERINE_MUCINOUS_CARCINOMA" | "UTERINE_MIXED_CARCINOMA" | "UTERINE_NON_TUMOUR" | "OTHER_TUMOUR" | "OTHER_NON_TUMOUR" | "UNCERTAIN"

LABEL ASSIGNMENT RULES
UTERINE_CLEAR_CELL_CARCINOMA: explicit clear cell carcinoma or clear/hobnail malignant cells with tubulocystic, papillary, or solid architecture and high-grade atypia
UTERINE_ENDOMETRIOID_CARCINOMA: explicit endometrioid carcinoma or gland-forming tumour resembling endometrial glands; may include villoglandular features
UTERINE_SEROUS_CARCINOMA: explicit serous carcinoma or high-grade papillary/micropapillary architecture with marked nuclear atypia and slit-like spaces
UTERINE_MUCINOUS_CARCINOMA: explicit mucinous carcinoma or malignant glands with abundant intracellular or extracellular mucin
UTERINE_MIXED_CARCINOMA: two distinct carcinoma components explicitly described (for example clear cell + endometrioid, serous + endometrioid); if mixed components are stated, assign UTERINE_MIXED_CARCINOMA regardless of percentages
UTERINE_NON_TUMOUR: normal endometrium, proliferative/secretory phase, hyperplasia without carcinoma, or inflammatory/reactive changes only
OTHER_TUMOUR: any tumour from a non-uterine corpus organ or site, including non-endometrial gynecologic or extra-uterine primaries
OTHER_NON_TUMOUR: any non-neoplastic tissue or benign/inflammatory finding from a non-uterine corpus organ or site
UNCERTAIN: insufficient or conflicting morphologic detail, unclear organ/site, or cannot map reliably

HALLUCINATION FLAGS
HALLUCINATION TYPE A — CASE-LEVEL OVERREACH
Flag "case_level_overreach": true if the report includes:
- margins
- depth of myometrial invasion
- lymphovascular invasion extent
- lymph node status
- FIGO stage
- molecular classification (p53, MMR, POLE)
- statements implying multi-slide integration
HALLUCINATION TYPE B — CONTEXT MISMATCH
Flag "context_mismatch": true if the report includes:
- primary reference to a non-uterine organ
- tumour types inappropriate for an endometrial context
- grading systems inappropriate for endometrial H&E
- morphologic descriptions clearly from another organ system
IMPORTANT:
- Evidence spans must quote exact fragments from the report (≤12 words).
- Prefer UNCERTAIN over over-calling carcinoma.

Figure S6 | Uterine cohort evaluation prompt for extracting diagnostic labels from pathology reports

You are a senior consultant uropathologist performing an audit of an AI-generated microscopic description for a SINGLE prostate whole-slide image (WSI). Your task is NOT to improve or rewrite the report. Your task is to EXTRACT structured, auditable signals from the text so that model performance can be quantified.

GENERAL SCOPE RULES
Use ONLY the content of the provided report. Treat this strictly as a SINGLE H&E slide. Do NOT assume gross findings, radiology, molecular testing, staging, or multi-slide integration unless explicitly stated.

TARGET LABEL SET (PANDA)
You must output ONE of the following labels:

"PROSTATE_ADENOCARCINOMA" | "PROSTATE_BENIGN" | "OTHER_TUMOUR" | "OTHER_NON_TUMOUR" | "UNCERTAIN"

LABEL ASSIGNMENT RULES
PROSTATE_ADENOCARCINOMA: Assign if: The report explicitly states prostatic adenocarcinoma OR Malignant glands arising in prostatic tissue are described OR The report explicitly describes Gleason pattern, Gleason score, or Grade Group in a way that clearly indicates prostate carcinoma
PROSTATE_BENIGN: Assign if: The report explicitly describes benign prostatic tissue OR Benign prostatic glands, stroma, or parenchyma are described without malignant features OR No malignant prostatic epithelial proliferation is described
OTHER_TUMOUR: Assign if the report primarily describes a definite tumour from a non-prostatic primary site, including: Metastatic carcinoma OR Urothelial carcinoma OR Squamous cell carcinoma OR Renal cell carcinoma OR Colorectal adenocarcinoma OR Lymphoma OR Any other definite non-prostatic tumour
OTHER_NON_TUMOUR: Assign if the report primarily describes non-neoplastic tissue or benign/inflammatory/reactive findings from a non-prostatic origin, including: Normal tissue OR Fibrosis OR Inflammation OR Reactive change OR Benign tissue from another organ
UNCERTAIN: Assign if: Organ of origin is unclear OR Malignancy is described without clear prostatic or any organ attribution OR Benign tissue is described without explicit prostatic or any organ attribution OR Conflicting morphological cues are present OR Insufficient morphologic detail exists OR The report is too nonspecific to distinguish prostatic from non-prostatic tissue

HALLUCINATION FLAGS
HALLUCINATION TYPE A — CASE-LEVEL OVERREACH
Flag "case_level_overreach": true if the report contains statements not supportable from a single H&E slide, including:
- Margin status
- Extraprostatic extension
- Seminal vesicle invasion
- Lymph node status
- pT/pN stage
- Pathologic stage
- molecular information & IHC results
- Statements implying review of the entire specimen or multiple sections
- Multi-slide integration

HALLUCINATION TYPE B — CONTEXT MISMATCH
Flag "context_mismatch": true if:
- Explicit references to other organs or tissues appear in a way inconsistent with a single-slide prostate context
- Tumour types not appropriate for prostate tissue are described
- Grading or classification systems from other organs are used
- Morphologic descriptions are clearly derived from another organ system

GLEASON EXTRACTION
- If present in the report, extract Gleason information as a separate JSON object.
- Gleason score, pattern(s), and Grade Group may be recorded when explicitly mentioned.
- Do NOT infer a Gleason score from morphology alone.
- Do NOT treat Gleason score or Grade Group mention by itself as hallucination.

IMPORTANT:
- Evidence spans must quote exact fragments from the report (≤12 words).
- Do NOT infer invasion, staging, or molecular status.
- Do NOT infer Gleason score or Grade Group unless explicitly stated.
- When uncertain, choose "UNCERTAIN" rather than over-calling carcinoma or benignity.

Figure S7 | Prostate cohort evaluation prompt for extracting diagnostic labels from pathology reports

You are a senior pathologist performing quality-control editing of OCR-derived pathology text so that it reflects ONLY what can be responsibly stated from a SINGLE histologic slide (single whole-slide image). This is a DOWNGRADING task, not an enrichment task. The input text comes from an OCR-scanned pathology report and may contain recognition errors, broken words, missing punctuation, duplicated fragments, line-wrap artifacts, header/footer noise, and malformed section structure. Your first job is to reconstruct the intended meaning conservatively from the OCR text. Then perform slide-level downgrading.

WORK IN TWO INTERNAL STEPS:
Step 1: OCR reconstruction
- Denoise the OCR text into coherent pathology prose.
- Resolve obvious OCR artifacts only when the intended wording is strongly supported by context.
- Remove page headers, footers, accession metadata, pagination artifacts, and duplicated OCR fragments when clearly non-diagnostic.
- Do NOT add facts that are not recoverable from the OCR text.
- If a phrase is too corrupted to interpret confidently, omit that phrase rather than guessing.
Step 2: Slide-level downgrading
- Edit the reconstructed text so it reflects only what can be responsibly stated from a SINGLE histologic slide.

NON-NEGOTIABLE RULES:
1) **DO NOT INVENT:**
  - Do NOT add new findings, entities, grades, stages, biomarkers, or interpretations.
  - Do NOT introduce new uncertainty wording unless needed to localize a negative finding to the examined slide or to avoid overclaiming from corrupted OCR text.
  - You may ONLY reconstruct, delete, soften, or localize statements already supported by the OCR text.
2) **REMOVE CASE-/GROSS-LEVEL CONTENT:**
  - Remove any final sign-out style impression lines.
  - Remove staging, margins, lymph node status/counts, metastasis, tumor size/measurements, specimen laterality/levels, procedure names when non-microscopic, and gross descriptors.
  - Remove multi-slide phrasing (“sections show…”, “multiple sections…”, “overall features…”, “in other sections…”) or rewrite to single-slide phrasing ONLY if it does not add information.
3) **NEGATIVE FINDINGS:**
  - If the text states an absence (e.g., “no necrosis”), keep it ONLY if it can be localized to the examined slide using wording like “not identified in this section” or “not seen in the examined slide”.
  - If it is phrased as a global/case-wide exclusion, remove it.
4) **ENTITY HANDLING:**
  - If the report names a specific tumor entity, subtype, grade, or final diagnosis that cannot be definitively supported from a single slide, replace it with a MORPHOLOGIC DESCRIPTION using ONLY features already stated or clearly recoverable from the OCR text.
  - Do NOT introduce molecular classification, immunohistochemistry, resection context, or clinical synthesis.
5) **OCR UNCERTAINTY HANDLING:**
  - When OCR corruption affects a medically meaningful phrase, prefer conservative simplification over confident reconstruction.
  - Do NOT force a specific diagnosis from partially illegible text.
  - If only part of a sentence is reliable, preserve only the reliable portion.
6) **STYLE:**
  - Single paragraph.
  - Microscopic descriptive tone.
  - No bullet points, no headings, no first-person language.

Figure S8 | TCGA report preprocessing prompt: step 1, OCR-to-standardised pathology text conversion

You are a senior consultant pathologist performing quality-control review of a microscopic description intended exclusively for single-whole-slide image (WSI) annotation. Review the drafted text and revise it to ensure strict compliance with WSI-level diagnostic pathology standards.
YOUR RESPONSIBILITIES:
- Do NOT re-edit if the draft already fully complies
- Avoid rephrasing; preserve wording unless a change is required for compliance
- Remove speculative, redundant, or interpretive language
- Eliminate any case-level, gross-level, or multi-slide references
- Downgrade overconfident anatomic, extent-based, or vessel-related claims to slide-appropriate language
- Ensure all described findings are reasonably visible on a single WSI
- Preserve diagnostic accuracy without adding new findings
STRICT PROHIBITIONS:
- Do NOT mention additional sections, separate sections, other slides, or other organs
- Do NOT include negative findings outside the field of view
- Do NOT include tumor measurements, staging, margins, metastasis, or lymph node evaluation
- Do NOT include immunohistochemistry or molecular data
- Do NOT reference the original report or missing information
- Do NOT invent findings
STYLE REQUIREMENTS:
- Direct microscopic observational tone
- Concise, factual, consultant-level language
- Single paragraph, no bullet points
- No first-person language
OUTPUT REQUIREMENTS:
- Return a JSON object with keys:
  - report: the final text (unchanged if already compliant)
  - edits: a concise list of edits with brief reasons; empty list if no changes

Figure S9 | TCGA report preprocessing prompt: step 2, exclusion of gross, multi-slide and non-microscopic content

You are a senior consultant pathologist performing a compliance audit of an AI-generated microscopic description intended ONLY for annotation of a SINGLE whole-slide image (WSI). You must follow this procedure exactly.

STEP 1 — SIGN-OFF DECISION

Decide whether you would sign the report AS WRITTEN as a slide-level microscopic description, using ONLY what can be supported by the SINGLE WSI.

Return exactly one of:

- "SIGNABLE AS IS"
- "NOT SIGNABLE"

STEP 2 — REPAIR (ONLY IF "NOT SIGNABLE")

If and only if the report is NOT SIGNABLE, produce a revised report that becomes signable by applying ONLY the minimal necessary changes.

HARD SAFETY RULES (NON-NEGOTIABLE)

1) **ZERO INVENTION:**
   - Do NOT add any new findings, new entities, new structures, new locations, new qualifiers, or new interpretations.
   - You may ONLY (a) delete text, or (b) downgrade certainty of statements already present.
2) **NO CASE-LEVEL CLAIMS:**
   - Delete any statement about margins, staging, pT/pN, metastasis, lymph nodes, specimen orientation, tumor size, extent into adjacent organs/structures (e.g., pleura, adipose, skin), or “separate/other sections/blocks.”
   - Do NOT replace these with softer wording unless the original already explicitly describes the relevant structure within the field of view.
   - If in doubt, DELETE.
3) **NO NEGATIVE FINDINGS OUTSIDE FIELD:**
   - Delete statements that imply review beyond the visible slide (e.g., “no metastasis,” “no necrosis elsewhere,” “no invasion identified” unless it is strictly confined to a described structure in view).
4) **VESSEL/PLEURA RULES (STRICT):**
   - For lymphovascular invasion: if the original states “embolus in vascular lumen,” you may ONLY downgrade to “within spaces suspicious for lymphovascular invasion.”
   - For pleura/serosa/adipose invasion or extension: DO NOT assert invasion. If mentioned, change to “in close association with [structure]” ONLY if that structure is explicitly mentioned in the draft; otherwise DELETE.

STYLE RULES

- Keep the original wording whenever possible.
- Single paragraph, factual microscopic tone.
- No bullet points, no headings, no first-person.

Figure S10 | TCGA training text preprocessing prompt: step 3, signability assessment and minimal repair of slide-level microscopic descriptions

You are an expert diagnostic pathologist generating a microscopic description strictly limited to histologic features visible within a single whole-slide image (WSI). You will receive a poorly written pathology note. Your task is to synthesise a concise, accurate microscopic description using ONLY morphologic findings that can be reasonably identified on a single WSI, without inference beyond the slide.

DIAGNOSTIC LABELS:

- Diagnostic labels (e.g.,colitis, calcific sclerosis) MAY be included
- Include a diagnostic label ONLY when it is explicitly present in the source text AND clearly supported by slide-level morphology
- Do NOT invent diagnoses
- Do NOT escalate or refine diagnoses beyond what is morphologically supported
- If no diagnostic label is clearly supported, provide a purely morphologic description without assigning a diagnosis

PRIORITISE MORPHOLOGY:

- Emphasize tissue architecture, cytology, and stromal features
- Describe what is present on the slide; do not infer clinical behavior or extent
- Prefer neutral, descriptive language when diagnostic certainty is not slide-supported

INCLUDE ONLY:

- A slide-supported diagnostic label when appropriate
- Microscopic architectural features
- Cytologic features
- Stromal features
- Definite microscopic findings (e.g., necrosis, calcification, desmoplasia) ONLY when unequivocally mentioned in the text

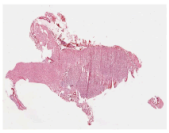
ABSOLUTE CONSTRAINTS (DO NOT VIOLATE):

- Limit content strictly to slide-level histologic observations
- Do NOT include case-level, gross-level, or multi-slide information
- Do NOT reference additional sections, separate sections, other blocks, or other organs
- Do NOT include gross description, procedural context, or clinical correlation
- Do NOT include staging, grading systems requiring case-level context, tumor size, margins, metastasis, or lymph node evaluation
- Do NOT include immunohistochemical, molecular, or biomarker data

STYLE REQUIREMENTS:

- Write in a direct microscopic observational style, as if the slide were examined firsthand
- Describe only what is present; never certify absence of features
- Use neutral, declarative diagnostic pathology language
- Do not reference the source report or missing information
- Do not use first-person language
- Output must be a single paragraph with no bullet points

Figure S11 | GTEx training text preprocessing prompt: reformatting brief pathology notes into standardised slide-level microscopic descriptions



GT: Glioblastoma

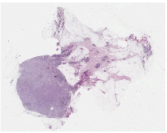
SlideFlame: *"The slide shows brain parenchyma infiltrated by a densely cellular astrocytic neoplasm with marked nuclear pleomorphism, hyperchromasia, and increased mitotic activity. Foci of microvascular proliferation and geographic necrosis are present."*

- NLG Score: BLEU-2 : 0 I BERTScore : 0.15
- bNLI Norm Score: 40.5

PRISM: *"benign fibrous tissue, no tumor in examined tissue."*

- NLG Score: BLEU-2 : 0 I BERTScore : 0.23
- bNLI Norm Score: 0.90

bNLI outcome : SlideFlame > PRISM
NLG outcome: PRISM > SlideFlame
Diagnostic outcome: SlideFlame correct



GT: Breast invasive carcinoma

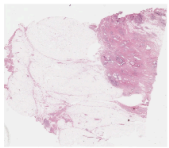
SlideFlame: *"The slide shows invasive ductal carcinoma, no special type, composed predominantly of solid nests and cords with little to no tubule formation and marked nuclear pleomorphism. An associated ductal carcinoma in situ component is present with solid architecture, high nuclear grade, and central comedo-type necrosis."*

- NLG Score: BLEU-2 : 0.05 I BERTScore : 0.36
- bNLI Norm Score: 24.9

PRISM: *"Poorly differentiated invasive ductal carcinoma with lymphovascular invasion; organ: breast."*

- NLG Score: BLEU-2 : 0.02 I BERTScore : 0.26
- bNLI Norm Score: 60.9

bNLI outcome : PRISM > SlideFlame
NLG outcome: SlideFlame > PRISM
Diagnostic outcome: Both diagnostically correct



GT: Breast atypical ductal hyperplasia

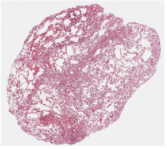
SlideFlame: *"Fibrocystic change with numerous dilated cystic spaces lined by flattened to cuboidal epithelium, some containing abundant eosinophilic granular material, and scattered calcifications including coarse brown granules and fine calcifications."*

- NLG Score: BLEU-2 : 0.02 I BERTScore : 0.33
- bNLI Norm Score: 25.8

PRISM: *"Examined tissue: Benign breast parenchyma"*

- NLG Score: BLEU-2 : 0.001 I BERTScore : 0.18
- bNLI Norm Score: 21.7

bNLI outcome : SlideFlame > PRISM
NLG outcome: SlideFlame > PRISM
Diagnostic outcome: Both diagnostically incomplete



GT: Non-tumour lung parenchyma

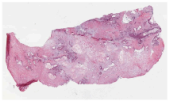
SlideFlame: *"Lung parenchyma shows markedly enlarged, irregular airspaces with attenuation of alveolar septa; residual septal walls contain dense eosinophilic collagenous fibrosis and focal scarring."*

- NLG Score: BLEU-2 : 0.02 I BERTScore : 0.20
- bNLI Norm Score: 46.7

PRISM: *"Benign lung parenchyma, no tumor present."*

- NLG Score: BLEU-2 : 0.06 I BERTScore : 0.34
- bNLI Norm Score: 97.8

bNLI outcome : PRISM > SlideFlame
NLG outcome: PRISM > SlideFlame
Diagnostic outcome: Both diagnostically correct



GT: Pancreatic ductal adenocarcinoma

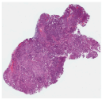
SlideFlame: *"The slide shows an infiltrative pancreatic ductal adenocarcinoma composed of irregular, angulated glands and small duct - like structures within a desmoplastic stroma, with moderate cytologic atypia and mitotic activity, and diagnostic of ductal adenocarcinoma."*

- NLG Score: BLEU-2 : 0.09 I BERTScore : 0.30
- bNLI Norm Score: 73.9

PRISM: *"Moderately differentiated metastatic adenocarcinoma from colorectal origin."*

- NLG Score: BLEU-2 : 0.05 I BERTScore : 0.38
- bNLI Norm Score: 35.9

bNLI outcome : SlideFlame > PRISM
NLG outcome: SlideFlame > PRISM
Diagnostic outcome: Both diagnostically incomplete



GT: Endometrioid adenocarcinoma

SlideFlame: *"Endometrium shows an endometrioid adenocarcinoma composed of irregular, crowded malignant glands with areas of complex architecture, lined by atypical columnar cells with nuclear enlargement and pleomorphism. Overall features are those of a moderately differentiated (grade II) endometrioid adenocarcinoma."*

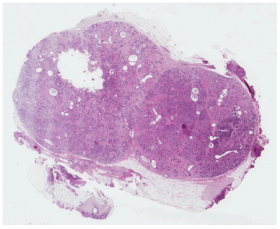
- NLG Score: BLEU-2 : 0.007 I BERTScore : 0.19
- bNLI Norm Score: 68.5

PRISM: *"Moderately differentiated adenocarcinoma of the rectum, DNA mismatch repair proteins retained."*

- NLG Score: BLEU-2 : 0 I BERTScore : 0.22
- bNLI Norm Score: 1.6

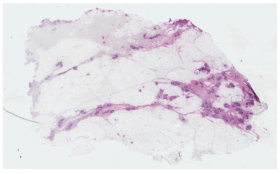
bNLI outcome : SlideFlame > PRISM
NLG outcome: PRISM > SlideFlame
Diagnostic outcome: SlideFlame correct

Figure S12 | Representative cases illustrating divergence between NLG-based and bNLI-based evaluation



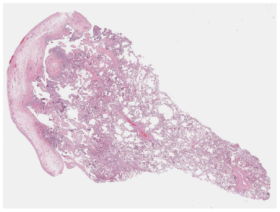
GT: Breast Pathological Benign

SlideFlame: *"Breast tissue shows invasive ductal carcinoma composed of infiltrative nests and irregular duct-forming structures with moderate nuclear pleomorphism and associated ductal carcinoma in situ exhibiting cribriform and solid patterns with intermediate-grade nuclei and focal comedo-type necrosis; findings are within spaces suspicious for lymphovascular invasion."*



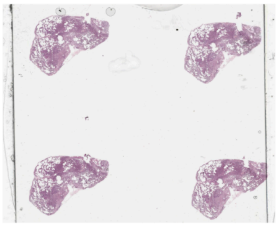
GT: Breast Pathological Benign

SlideFlame: *"The slide shows invasive lobular carcinoma composed of discohesive malignant epithelial cells infiltrating fibrous stroma in single file cords and small nests, with moderate nuclear pleomorphism and low mitotic activity. Associated ductal carcinoma in situ is present with solid and cribriform architecture, intermediate to high nuclear grade, comedo-type necrosis, and calcifications."*



GT: Non-tumour Lung Parenchyma

SlideFlame: *"Lung parenchyma shows an infiltrative malignant epithelial neoplasm composed of atypical gland-forming cells forming irregular acinar structures and solid nests with cytologic atypia."*



GT: Non-tumour Lung Parenchyma

SlideFlame: *"The slide shows a keratinizing squamous cell carcinoma composed of irregular nests and tongues of epithelium with intercellular bridges and focal keratin pearl formation, accompanied by moderate nuclear pleomorphism and increased mitotic activity."*

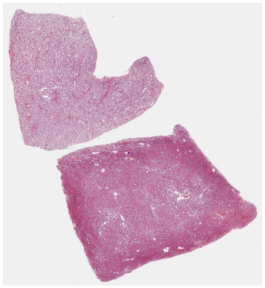
GT: Prostatic Benign Parenchyma

SlideFlame: *"The slide shows prostatic acinar adenocarcinoma composed of infiltrative, crowded small glands with areas of fused / poorly formed glandular architecture, lined by atypical epithelial cells with enlarged hyperchromatic nuclei and conspicuous nucleoli, consistent with Gleason pattern 3 with pattern 4."*

GT: Prostatic Benign Parenchyma

SlideFlame: *"The slide shows prostatic adenocarcinoma composed of infiltrative, crowded small glands with areas of fused / poorly formed glandular architecture; tumor cells have enlarged nuclei with conspicuous nucleoli and mild to moderate nuclear pleomorphism."*

Figure S13 | Examples of recurrent malignant overcalling in SlideFlame-generated reports across the breast, lung and prostate cohorts



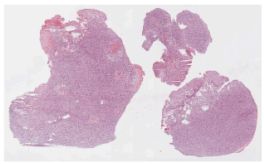
GT: *Non-tumour renal parenchyma*

SlideFlame: *"The slide shows a renal cortical neoplasm composed of nests, alveolar and trabecular aggregates of polygonal cells separated by a delicate branching capillary network, with cells displaying abundant eosinophilic cytoplasm and enlarged nuclei characteristic of these features."*

GT: *Non-tumour renal parenchyma*

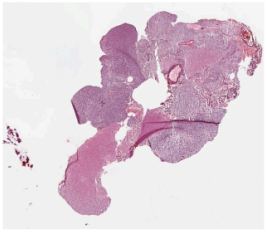
SlideFlame: *"The slide shows a renal cortical neoplasm composed of nests and alveolar / acinar groups of polygonal cells with predominantly clear cytoplasm separated by a delicate, thin walled vascular network, consistent with conventional (clear cell) renal cell carcinoma."*

Figure S14 | Examples of recurrent malignant overcalling in SlideFlame-generated reports across the renal cohorts



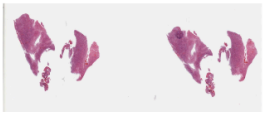
GT: *Glioblastoma*

"**PRISM:** *Diagnosed diffuse large B-cell lymphoma in gastric mucosa; CD20, BCL-2, BCL-6, MUM-1 positive.*"



GT: *Glioblastoma*

"**PRISM:** *Low-grade B-cell lymphoma with plasmacytic differentiation; organ: stomach.*"

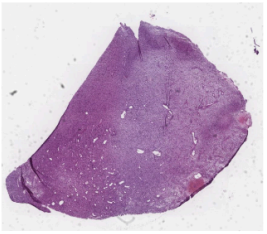


GT: *Glioblastoma*

"**PRISM:** *High-grade B-cell lymphoma with plasmacytic differentiation in left mainstem bronchus.*"

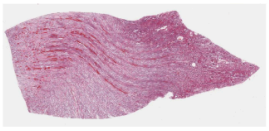
GT: *Glioblastoma*

"**PRISM:** *Examined tissue: Spindle cell sarcoma, high-grade, with extensive necrosis.*"



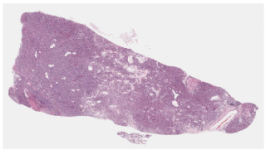
GT: *Glioblastoma*

"**PRISM:** *Spindle cell gastrointestinal stromal tumor identified in tissue sample.*"



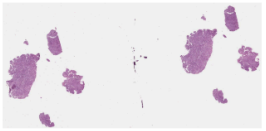
GT: *Non-tumour renal parenchyma*

"**PRISM:** *Intraductal papillary mucinous neoplasm of the pancreas.*"



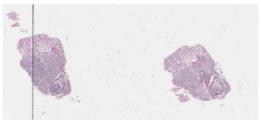
GT: *Clear cell renal cell carcinoma*

"**PRISM:** *"Spindle cell neoplasm identified as gastrointestinal stromal tumor; immunopositive for CD117, DOG1."*



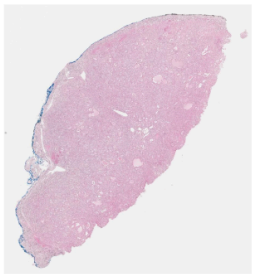
GT: *Papillary renal carcinoma*

"**PRISM:** *"Low-grade endometrioid adenocarcinoma of the uterus."*



GT: *Renal oncocytic neoplasm*

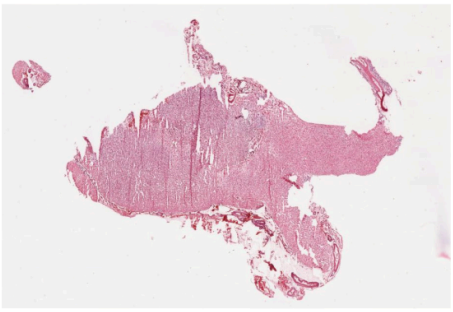
"**PRISM:** *"Examined tissue: Benign gastric mucosa."*



GT: *Renal cell tumor of Birt-Hogg-Dube syndrome*

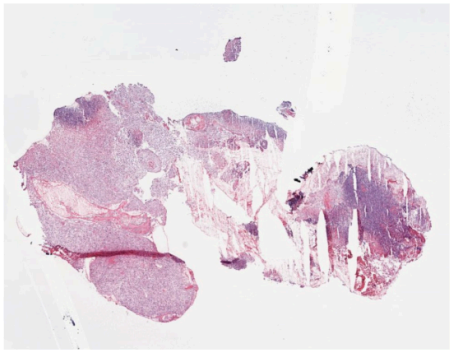
"**PRISM:** *"Metastatic clear cell renal cell carcinoma in adrenal gland."*

Figure S15 | Examples of context-mismatch cases seen in PRISM-generated reports for brain and renal cohort



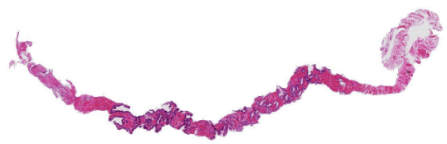
GT: *Glioblastoma*

PRISM: *"benign fibrous tissue, no tumor in examined tissue."*



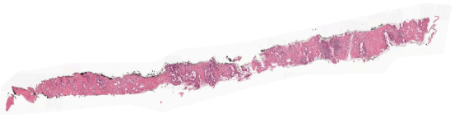
GT: *Glioblastoma*

PRISM: *"Granulation tissue with fibrinopurulent exudate; organ: unspecified."*



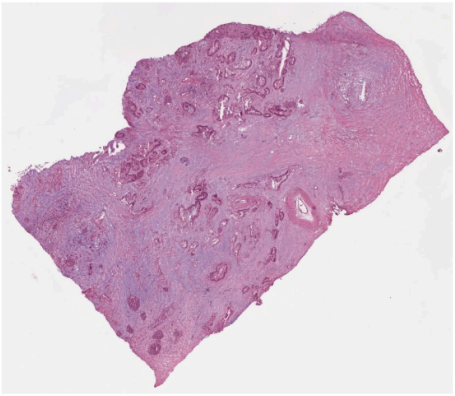
GT: *Prostatic benign tissue*

PRISM: *"Examined tissue: Benign glandular polyp fragments."*



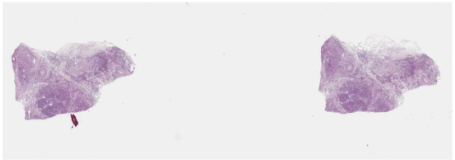
GT: *Prostatic adenocarcinoma*

PRISM: *"Moderately differentiated invasive adenocarcinoma in examined tissue."*



GT: *Pancreatic ductal adenocarcinoma*

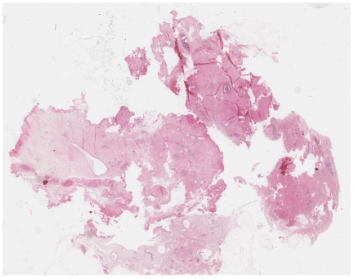
PRISM: *"Moderately differentiated adenocarcinoma in examined tissue."*



GT: *Pancreatic ductal adenocarcinoma*

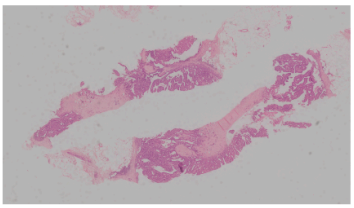
PRISM: *"Poorly differentiated carcinoma with sarcomatoid features in the examined tissue."*

Figure S16 | Examples of UNCERTAIN cases identified in PRISM-generated reports across the brain, prostate and pancreas cohorts



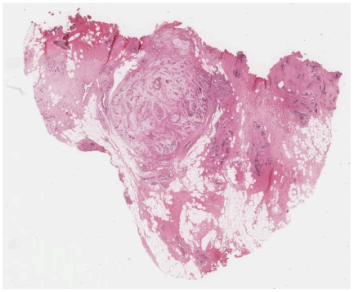
GT: *Breast pathological benign*

SlideFlame: "*The slide shows numerous cystic spaces of varying size lined by flattened to cuboidal epithelium, some containing proteinaceous debris, situated within fibrous stroma; scattered calcifications are present among the cysts and adjacent stroma.*"



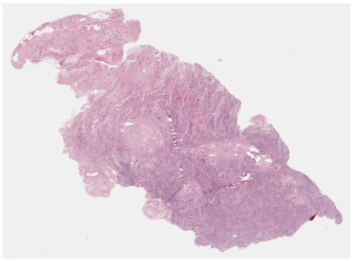
GT: *Breast invasive carcinoma*

SlideFlame: "*The slide shows complex papillary fronds with fibrovascular cores lined by atypical cuboidal to columnar epithelial cells displaying nuclear enlargement, pleomorphism, and hyperchromasia.*"



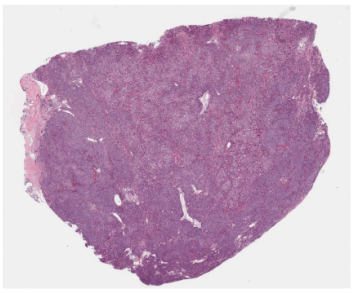
GT: *Breast pathological benign*

SlideFlame: "*The slide shows multiple cystic spaces of varying size lined by flattened to cuboidal epithelium, some appearing collapsed, situated within a dense collagenous fibrous stroma that contains scattered chronic inflammatory cells.*"



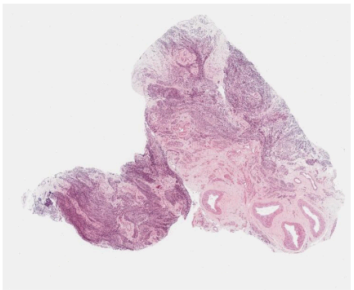
GT: *Uterine non-tumour tissue*

SlideFlame: "*The slide shows a cystic space lined by a single layer of flattened to cuboidal epithelium; the surrounding stroma contains dense collagenous fibrous tissue interspersed with bland spindle shaped fibroblasts.*"



GT: *Uterine endometrioid carcinoma*

SlideFlame: "*The slide shows a malignant epithelial neoplasm composed predominantly of solid nests and sheets with limited glandular differentiation, marked nuclear pleomorphism and hyperchromasia, brisk mitotic activity, and foci of necrosis.*"



GT: *Uterine non-tumour tissue*

SlideFlame: "*The slide demonstrates dense eosinophilic collagenous stroma forming broad, intersecting bundles that replace the native parenchyma; residual recognizable glandular and ductal structures are few and appear compressed within the fibrous background.*"

Figure S17 | Examples of UNCERTAIN cases identified in SlideFlame-generated reports across the breast and uterine cohorts.

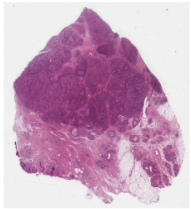
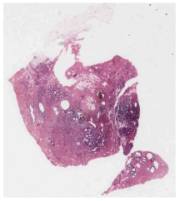
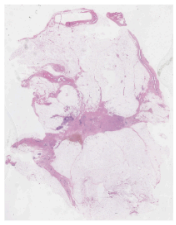
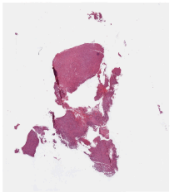
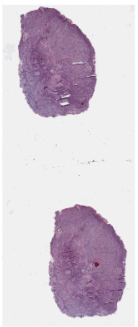
WSI	PRISM-generated Report	SlideFlame-generated Report	Preference
	High-grade spindle cell sarcoma in dermis and subcutis.	The slide shows a high-grade astrocytic neoplasm with marked cellularity and pleomorphism, including numerous small multinucleated tumor cells with brisk mitotic activity. Prominent microvascular proliferation and geographic necrosis with pseudopalisading of tumor cells are present.	Expert 1: SlideFlame Expert 2: SlideFlame Expert 3: SlideFlame
	Metastatic melanoma in a lymph node.	The slide shows invasive ductal carcinoma composed of predominantly solid nests and cords with only focal tubule formation, marked nuclear pleomorphism and hyperchromasia, and increased mitotic activity, consistent with high histologic grade. Foci of ductal carcinoma in situ with solid and cribriform architecture and intermediate-grade nuclei are present, with calcifications identified in benign ducts.	Expert 1: SlideFlame Expert 2: SlideFlame Expert 3: SlideFlame
	Pseudoangiomatous stromal hyperplasia in breast parenchyma.	Fibrocystic change with numerous dilated cystic spaces lined by a single layer of flattened to cuboidal epithelium and surrounded by dense collagenous stroma; the surrounding ducts and lobules show preserved architecture.	Expert 1: SlideFlame Expert 2: SlideFlame Expert 3: SlideFlame
	Invasive ductal carcinoma with perineural invasion in breast tissue.	Breast parenchyma is infiltrated by invasive ductal carcinoma composed of irregular nests and cords with focal tubule formation in a desmoplastic stroma, with moderate nuclear pleomorphism and low mitotic activity. Associated ductal carcinoma in situ is present with solid and cribriform architecture, intermediate-to-high nuclear grade, comedo-type necrosis, and calcifications.	Expert 1: SlideFlame Expert 2: SlideFlame Expert 3: SlideFlame
	High-grade B-cell lymphoma with plasmacytic differentiation in lung tissue.	The slide shows a hypercellular infiltrating glial neoplasm with marked nuclear pleomorphism, brisk mitotic activity, microvascular proliferation, and geographic necrosis.	Expert 1: SlideFlame Expert 2: SlideFlame Expert 3: SlideFlame
	High-grade invasive urothelial carcinoma with perineural invasion.	The slide shows an invasive squamous cell carcinoma composed of irregular nests and sheets of malignant polygonal cells with eosinophilic cytoplasm and intercellular bridges, with marked nuclear pleomorphism, hyperchromasia, and increased mitotic activity, consistent with a poorly differentiated (grade 3) squamous cell carcinoma.	Expert 1: SlideFlame Expert 2: SlideFlame Expert 3: SlideFlame

Figure S18 | Representative cases in which blinded pathologists preferred SlideFlame-generated reports over PRISM-generated reports.

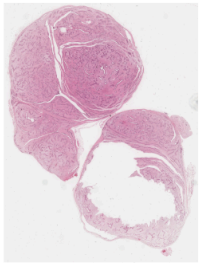
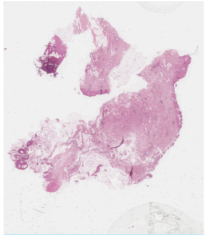
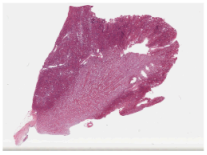
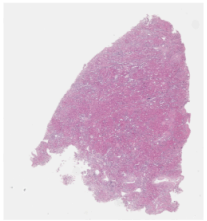
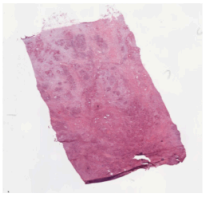
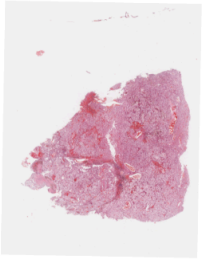
WSI	PRISM-generated Report	SlideFlame-generated Report	Preference
	Benign phyllodes tumor in breast parenchyma.	The slide shows a proliferation of tightly packed small acini and tubules lined by a single layer of relatively uniform cuboidal to columnar epithelial cells with minimal nuclear pleomorphism and low mitotic activity, embedded in a fibrous stroma.	Expert 1: SlideFlame Expert 2: PRISM Expert 3: SlideFlame
	Examined tissue: Benign breast parenchyma.	The slide shows dense collagenous fibrosis forming broad eosinophilic bands that replace the normal lobular architecture; scattered residual ducts and lobular units are present, lined by a single layer of cuboidal to flattened epithelium and containing scant eosinophilic secretions, and the fibrous stroma is interspersed with scattered chronic inflammatory cells.	Expert 1: SlideFlame Expert 2: SlideFlame Expert 3: PRISM
	Examined tissue: Benign pancreatic tissue.	The renal cortex demonstrates prominent vascular congestion with dilated capillaries and small veins containing erythrocytes; the interstitium shows moderate fibrosis composed of coarse collagen bundles that separate the acinar units and alter the lobular architecture.	Expert 1: SlideFlame Expert 2: None Expert 3: SlideFlame
	Examined tissue: Benign parathyroid gland.	The slide shows tightly packed renal cortical parenchyma with numerous glomeruli, many displaying global sclerosis with obliterated capillary loops and hyalinized matrix; the interstitium contains a moderate infiltrate of small lymphocytes and plasma cells with focal stromal fibrosis.	Expert 1: SlideFlame Expert 2: None Expert 3: SlideFlame
	Moderately to poorly differentiated invasive adenocarcinoma	The slide shows an infiltrative high - grade pancreatic ductal adenocarcinoma composed of irregular, angulated and fused glands with areas of solid growth in a desmoplastic stroma; tumor cells display marked nuclear pleomorphism, hyperchromasia, and brisk mitotic activity, with foci of tumor necrosis. Perineural invasion is present, and findings within endothelial - lined vascular / lymphatic spaces are suspicious for lymphovascular invasion.	Expert 1: SlideFlame Expert 2: PRISM Expert 3: SlideFlame
	Metastatic clear cell renal cell carcinoma in examined tissue.	The slide shows a renal neoplasm composed of nests, alveolar / acinar structures, and solid sheets of polygonal cells separated by a delicate branching capillary network. Tumor cells have abundant optically clear cytoplasm with distinct cell borders and enlarged irregular nuclei with conspicuous nucleoli, consistent with Fuhrman nuclear grade 3. Tumor cells are identified within vascular lumina, suspicious for lymphovascular invasion. Diagnosis: Clear cell renal cell carcinoma, Fuhrman grade 3, with tumor present within vascular lumina.	Expert 1: SlideFlame Expert 2: SlideFlame Expert 3: None

Figure S19 | Representative cases with discordant blinded pathologist preference between SlideFlame-generated and PRISM-generated reports.

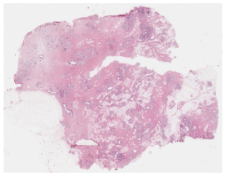
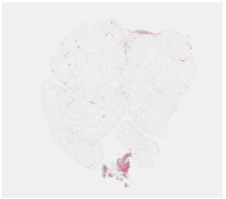
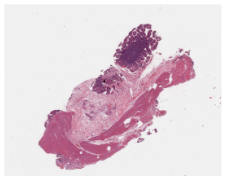
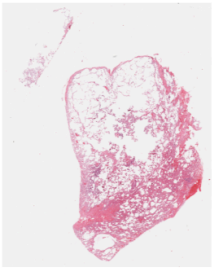
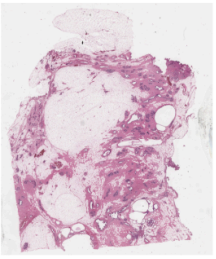
WSI	PRISM-generated Report	SlideFlame-generated Report	Preference
	Atypical ductal hyperplasia in breast parenchyma.	The slide shows multiple cystic spaces within a fibrous stroma; each cyst is lined by a single layer of flattened to cuboidal epithelial cells and the surrounding stroma is expanded by dense collagenous fibrosis.	Expert 1: PRISM Expert 2: PRISM Expert 3: PRISM
	Benign adipose tissue, no tumor present.	The omental tissue shows a collagenous fibrosis composed of thick eosinophilic bundles forming a compact stroma, with scattered spindle - shaped fibroblastic cells interspersed among the collagen.	Expert 1: PRISM Expert 2: PRISM Expert 3: PRISM
	Examined tissue: Benign small bowel mucosa	Pancreatic parenchyma shows tightly packed acini with basophilic granular cytoplasm and centrally placed round nuclei, interspersed with small ducts lined by a single layer of cuboidal to low columnar epithelium; the surrounding stroma contains a loose fibrous matrix with scattered fibroblasts.	Expert 1: PRISM Expert 2: PRISM Expert 3: PRISM
	Examined tissue: Benign fibroadipose without malignancy.	Dense collagenous fibrosis replaces normal breast architecture, forming thick eosinophilic bands that encircle and separate residual lobular units; the remaining ductal and acinar structures are compressed within the fibrotic stroma.	Expert 1: PRISM Expert 2: PRISM Expert 3: PRISM
	Benign lung parenchyma, no tumor present.	Lung parenchyma shows an infiltrative epithelial neoplasm composed of irregular nests, sheets, and duct - like / glandular structures within a desmoplastic stroma, with cytologic atypia including enlarged hyperchromatic pleomorphic nuclei and increased mitotic activity; focal tumor necrosis is present.	Expert 1: PRISM Expert 2: None Expert 3: PRISM
	Examined tissue: Benign breast parenchyma.	The slide shows focal basophilic, granular calcific deposits within the fibrous stroma in a sclerotic pattern, associated with dense collagenous fibrosis and scattered atrophic ductal and lobular structures exhibiting a sclerotic, distorted architecture.	Expert 1: PRISM Expert 2: None Expert 3: PRISM

Figure S20 | Representative cases in which blinded pathologists preferred PRISM-generated reports over SlideFlame-generated reports.

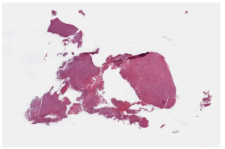
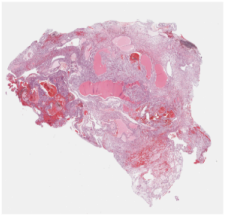
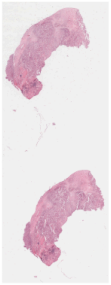
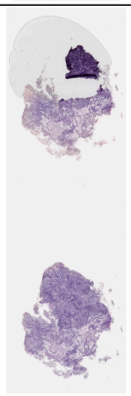
WSI	PRISM-generated Report	SlideFlame-generated Report	Majority Preference	Comments
	High-grade B-cell lymphoma with plasmacytic differentiation in lung tissue.	The slide shows a hypercellular infiltrating glial neoplasm with marked nuclear pleomorphism, brisk mitotic activity, microvascular proliferation, and geographic necrosis.	SlideFlame	While the specimen is probably fragmented due to necrosis, it is not clearly seen in this specimen.
	High-grade B-cell lymphoma with plasmacytic differentiation in lung tissue.	Clear cell renal cell carcinoma composed of nests and alveolar / acinar groups of polygonal cells with optically clear cytoplasm and delicate branching vasculature; nuclei show conspicuous atypia with prominent nucleoli consistent with Fuhrman nuclear grade 3. Background renal parenchyma shows arterionephrosclerosis.	SlideFlame	Not high nuclear grade.
	Malignant epithelioid neoplasm, likely renal cell carcinoma, PAX8 positive.	Clear cell renal cell carcinoma composed of nests and alveolar sheets of polygonal cells with optically clear cytoplasm and delicate branching vasculature, showing marked nuclear pleomorphism with prominent nucleoli consistent with high nuclear grade. Tumor is present within spaces suspicious for lymphovascular invasion.	SlideFlame	Unsure if lymphovascular invasion.
	Adenocarcinoma with mucinous features in lung tissue.	The slide shows a malignant epithelial neoplasm composed predominantly of papillary architecture with fibrovascular cores and associated solid nests and sheets of tumor cells. The neoplastic cells display cytologic atypia with enlarged, pleomorphic, hyperchromatic nuclei and conspicuous nucleoli, consistent with high nuclear grade. Foci of tumor necrosis are present. Diagnosis: Lung adenocarcinoma, papillary type, high nuclear grade.	PRISM	This is adenocarcinoma and there is mucous but there's not enough detail in the report.
	Prostatic adenocarcinoma identified, Gleason score 9, with perineural invasion.	Prostatic tissue shows numerous tightly packed glands lined by tall columnar to cuboidal epithelial cells with basal nuclei and moderate nuclear pleomorphism; the glands are forming irregular, variably sized glandular units, some containing fused or poorly formed glands, and the intervening stroma is desmoplastic with scattered inflammatory cells.	PRISM	Describe tumour and Gleason score 10.

Figure S21 | Representative edge cases in which pathologist preference reflected relative superiority rather than full clinical acceptability.

Representative paired PRISM- and SlideFlame-generated reports are shown alongside the corresponding whole-slide image, majority pathologist preference, and reviewer comments. These examples highlight cases in which one model output was preferred over the other,

although the selected report remained imperfect and required additional correction or clarification before reaching a clinically acceptable form.

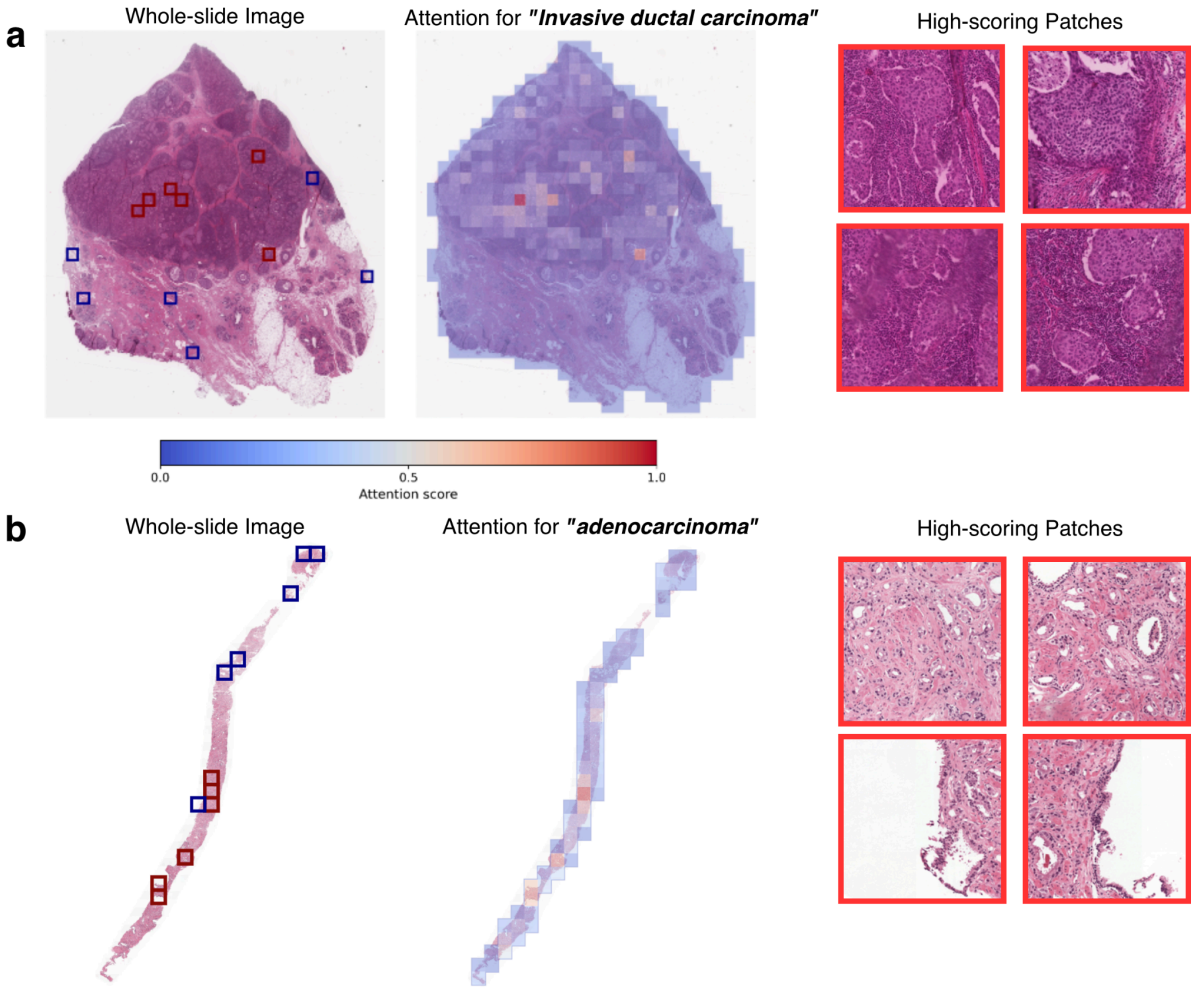


Figure S22 | Representative SlideFlame attention heatmaps for diagnostically aligned breast and prostate cases.

Whole-slide H&E images and corresponding attention heatmaps are shown for representative diagnostically aligned breast and prostate cases. Higher attention scores (patches within red boxes) localise to tissue regions morphologically associated with the predicted diagnosis.

Upon pathologist's review, high-attention regions in the breast example highlight tumor cells with minimal to absent glandular formation. In the prostate example, top-ranked tiles demonstrate small infiltrating malignant glands lacking basal cells, a characteristic feature of prostatic adenocarcinoma.

References

1. Shaikovski, G. *et al.* PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology. Preprint at https://doi.org/10.48550/arXiv.2405.10254 (2024).
2. Shaikovski, G. *et al.* PRISM2: Unlocking Multi-Modal General Pathology AI with Clinical Dialogue. https://doi.org/10.48550/ARXIV.2506.13063 (2025) doi:10.48550/ARXIV.2506.13063.
3. Tran, M. *et al.* Generating dermatopathology reports from gigapixel whole slide images with HistoGPT. *Nat. Commun.* **16**, 4886 (2025).
4. Sengupta, S. & Brown, D. E. Automatic Report Generation for Histopathology images using pre-trained Vision Transformers and BERT. Preprint at https://doi.org/10.48550/arXiv.2312.01435 (2024).
5. Tan, J. W. *et al.* Clinical-grade Multi-Organ Pathology Report Generation for Multi-scale Whole Slide Images via a Semantically Guided Medical Text Foundation Model. in (arXiv, 2024). doi:10.48550/ARXIV.2409.15574.
6. Sun, Y. *et al.* CPath-Omni: A Unified Multimodal Foundation Model for Patch and Whole Slide Image Analysis in Computational Pathology.
7. Moonemans, S. *et al.* Democratizing Pathology Co-Pilots: An Open Pipeline and Dataset for Whole-Slide Vision-Language Modelling. Preprint at https://doi.org/10.48550/arXiv.2512.17326 (2025).
8. Kim, K. A., Hong, S., Yoo, S., Kang, Y. & Shim, H. S. Enhancing Structured Pathology Report Generation With Foundation Model and Modular Design. *IEEE Access* **13**, 121290–121299 (2025).
9. Hao, G., Zhengrui and Ma, Jiabo and Xu, Yingxue and Wang, Yihui and Wang, Liansheng and Chen. HistGen: Histopathology Report Generation via Local-Global Feature Encoding and Cross-modal Context Interaction. *MICCAI 2024 - Open Access* https://papers.miccai.org/miccai-2024/387-Paper0796 (2024).

10. Hu, D. *et al.* Pathology report generation from whole slide images with knowledge retrieval and multi-level regional feature selection. *Comput. Methods Programs Biomed.* **263**, 108677 (2025).
11. Ahmed, F. *et al.* PolyPath: Adapting a Large Multimodal Model for Multi-slide Pathology Report Generation. Preprint at https://doi.org/10.48550/arXiv.2502.10536 (2025).
12. Chen, Y. *et al.* SlideChat: A Large Vision-Language Assistant for Whole-Slide Pathology Image Understanding.
13. Liang, Y. *et al.* WSI-LLaVA: A Multimodal Large Language Model for Whole Slide Image. Preprint at https://doi.org/10.48550/arXiv.2412.02141 (2025).
14. Chen, P., Zhu, C., Zheng, S., Li, H. & Yang, L. WSI-VQA: Interpreting Whole Slide Images by Generative Visual Question Answering. Preprint at https://doi.org/10.48550/arXiv.2407.05603 (2024).
15. Chen, P. *et al.* WsiCaption: Multiple Instance Generation of Pathology Reports for Gigapixel Whole-Slide Images. Preprint at https://doi.org/10.48550/arXiv.2311.16480 (2024).
16. Gao, Z. *et al.* ALPaCA: Adapting Llama for Pathology Context Analysis to enable slide-level question answering. 2025.04.22.25326190 Preprint at https://doi.org/10.1101/2025.04.22.25326190 (2025).
17. Huang, W. *et al.* Act Like a Pathologist: Tissue-Aware Whole Slide Image Reasoning. Preprint at https://doi.org/10.48550/arXiv.2603.00667 (2026).
18. Jiang, S., Liu, F., Wang, Z., Cai, L. & Zhang, Y. PathReasoner-R1: Instilling Structured Reasoning into Pathology Vision-Language Model via Knowledge-Guided Policy Optimization. Preprint at https://doi.org/10.48550/arXiv.2601.21617 (2026).