TOWARDS A VERIFIABLE MEDICAL AGENT POWERED BY SPECIALIST TOOLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Medical reasoning requires selective image inspection, specialized computation, and reconciliation of evidence across sources. We introduce MedVESTA, a medical agent that connects general-purpose multimodal large language models (MLLMs) to specialist tools. Predictions, attention maps, and selected patches guide evidence acquisition; for classification and reporting, a verifier flags evidence-linked claims for re-inspection and revision. The toolbox combines pathology with available radiological and molecular evidence and diagnostic-knowledge retrieval. We also introduce PathScore, which evaluates report content through 15 clinical fields. On six classification tasks, Qwen3.5-27B + MedVESTA reaches 69.66% balanced accuracy with self-verification and 70.25% with a separate verifier, compared with 68.90% for its MIL specialist. VQA and report generation improve over the evaluated single-shot configurations of the same backbones. These results support specialist-guided evidence acquisition within an inspectable reasoning workflow.

1 INTRODUCTION

Medical reasoning begins with finding relevant evidence. Whole-slide pathology requires integrating local morphology across gigapixel images (Ding et al., 2025). Pathology reports combine microscopic findings with clinical information (National Cancer Institute, 2022b), while molecular assays characterize biomarkers that can inform treatment (National Cancer Institute, 2021). General-purpose multimodal large language models (MLLMs) provide a natural-language interface, but must decide what to inspect, which computations to request, and how observations support their conclusions.

Specialist models offer useful starting points. Attention-based multiple-instance learning (MIL) yields slide predictions and patch weights (Ilse et al., 2018); segmentation adds spatial context (Wasserthal et al., 2023). Medical agents coordinate these tools through an MLLM (Fallahpour et al., 2025), but each reported finding still requires supporting evidence.

We introduce MedVESTA (Medical Verifiable-Evidence Integrated Specialist-Tool Agent), which couples specialist-guided inspection with claim–evidence verification (Figure 1). Predictions and spatial readouts guide the executor to informative regions; observations and retrieved diagnostic criteria receive evidence identifiers. For classification and reporting, the verifier flags claims for re-examination, and the executor revisits the case before producing its final answer. We evaluate classification, VQA, and reporting, and introduce PathScore to evaluate clinical report content.

Our contributions are: (i) a specialist-guided workflow and shared toolbox for acquiring pathological and complementary multimodal evidence; (ii) evidence-linked claims and verification-guided revision; and (iii) PathScore, which separates reference-field recovery and field-wise agreement. Experiments quantify same-backbone gains and assess specialist guidance and verification.

2 RELATED WORK

Pathology-specific models. CONCH, PRISM, and PRISM2 support pathology image–text alignment and slide-level inference (Lu et al., 2023; Shaikovski et al., 2024; Vorontsov et al., 2026). TITAN also aligns whole-slide representations with pathology reports (Ding et al., 2025). SlideChat[†]

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

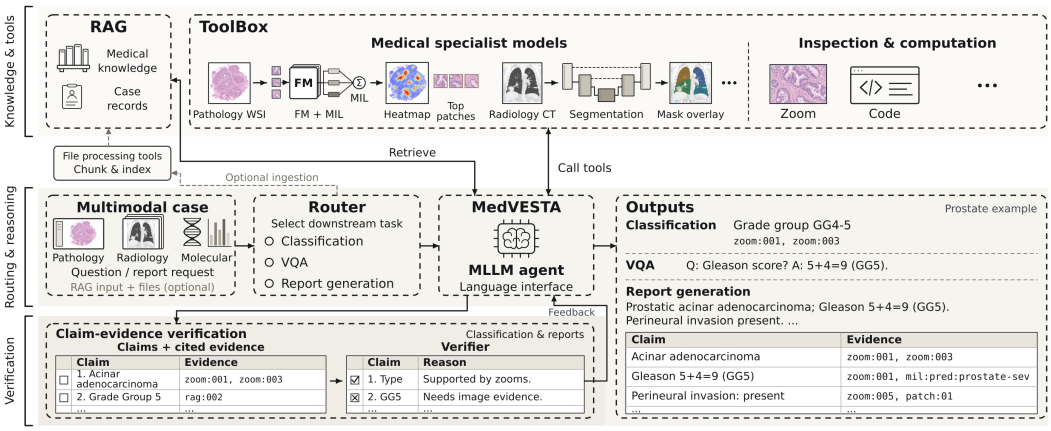


Figure 1: MedVESTA overview: task routing, specialist-guided evidence acquisition, and claim–evidence verification for classification and reporting. The prostate classification example uses the defined GG4–5 category.

and WSI-LLaVA† extend whole-slide representations to instruction following (Chen et al., 2025; Liang et al., 2025). Throughout this paper, † denotes documented training overlap with our TCGA evaluations; it does not imply overlap with non-TCGA cohorts (Appendix A). MedVESTA uses specialist outputs to guide further inspection and links claims to the acquired evidence.

Medical MLLMs and agents. LLaVA-Med, MedDr, and MedGemma adapt multimodal models to medical tasks (Li et al., 2023; He et al., 2024; Sellergren et al., 2025). ReAct interleaves reasoning and actions (Yao et al., 2023); MedRAX coordinates chest X-ray tools (Fallahpour et al., 2025), and QCAgent uses critique and region retrieval for pathology reporting (Wang et al., 2026). MedVESTA combines specialist-guided acquisition with an explicit evidence registry and claim-level re-inspection, without agent-specific executor fine-tuning.

Clinical report evaluation. BLEU measures lexical overlap and BERTScore compares contextual token representations (Papineni et al., 2002; Zhang et al., 2020). CheXbert demonstrates structured finding extraction for radiology report evaluation (Smit et al., 2020). PathScore evaluates pathology fields, distinguishing correct, omitted, indeterminate, and mismatched findings. Appendix E contrasts clinical agreement with text similarity.

3 METHOD

MedVESTA connects an MLLM executor to specialist models and a shared evidence registry. A router selects classification, multiple-choice VQA, or report generation; document-ingestion tools support retrieval-corpus updates. The executor requires no agent-specific supervised or reinforcement training.

3.1 INPUTS AND EVIDENCE REPRESENTATION

Requests combine a question or reporting instruction with available pathological, radiological, and molecular inputs. The registry assigns typed IDs to source artifacts and tool outputs. Classification and report claims take the form

$$c_j = (u_j, I_j), \qquad I_j \subseteq \mathrm{IDs}(\mathcal{E}) \cup \mathcal{P}, \tag{1}$$

where u_j is a finding or field–value pair and $\mathcal{E}$ is the registry. Ordinary IDs retrieve observations or records, such as a crop (`zoom:003`), MIL readout, or diagnostic passage. Retrieved criteria remain distinct from patient evidence. The set $\mathcal{P}$ contains `negative_finding` and `inconclusive`: non-retrieving status markers for non-observation and insufficient information (Appendix D).

3.2 SPECIALIST-GUIDED EVIDENCE ACQUISITION

Specialist guidance. The pathology MIL tool uses gated attention (Ilse et al., 2018) to pool patch features $\{\mathbf{h}_i\}_{i=1}^{N}$ and predict task τ:

$$\mathbf{z} = \sum_{i=1}^{N} \alpha_i \mathbf{h}_i, \qquad \mathbf{p}_\tau = \mathrm{softmax}(g_\tau(\mathbf{z})), \tag{2}$$

with $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. It returns labels, probabilities, attention maps, patch coordinates, and selected views. The executor may retain or overturn the prediction after inspecting morphology; attention supplies a search cue, not proof of a finding. Pathology pooling networks and heads are trained on non-TCGA REG2026 data with a frozen UNI2-h encoder (REG2026 Challenge Organizers, 2026; Mahmood Lab, n.d.). Specialist weights are not updated during an agent run; this separation does not rule out encoder pretraining exposure (Appendix B).

Actions and observations. Let x denote the available case inputs, q the task request, and h_t the interaction history before round t. The executor π_θ selects an allowed action a_t and registers its observation o_t:

$$\begin{aligned} a_t &\sim \pi_\theta(\cdot \mid q, x, h_t, \mathcal{E}_t), \qquad o_t = \mathcal{T}_{a_t}(x), \\ \mathcal{E}_{t+1} &= \mathcal{E}_t \cup \{(\mathrm{id}_t, o_t)\}, \qquad h_{t+1} = h_t \parallel (a_t, o_t). \end{aligned} \tag{3}$$

Here a_t includes tool arguments, $\mathcal{T}_{a_t}$ executes the requested tool, and $\parallel$ appends to the history. The loop ends with an answer or forced finalization at the task's round budget. Initial views and cached readouts form $\mathcal{E}_0$; new observations receive persistent IDs. Tools include grading/attribute readouts, region viewing, attention-table computation, spatial-extent estimation, and retrieval (Appendix B.2). The extent tool measures an attention-selected bounding-box span, not a validated anatomical tumor dimension. Available radiology montages, segmentation outputs, and molecular summaries are prepared offline for reporting. VQA uses a fixed visual package and specialist readouts, without zoom, retrieval, or verification (Appendix C).

3.3 CLAIM–EVIDENCE VERIFICATION AND REVISION

Verification checks source validity, evidential adequacy, and consistency; the executor remains responsible for the final answer. For classification, self-verification or a separate verifier reviews claims against the initial views, specialist readout, and diagnostic criteria, with the final class withheld. Flagged claims trigger one revision pass with prior outputs, feedback, and inspection tools.

For reporting, the first pass produces a narrative and evidence-linked findings. The same MLLM checks selected findings against their citations and returns `keep` or `revise`. The revision pass receives prior finding values and `LOOK AGAIN` markers, without detailed verifier explanations or prior citation lists. It re-examines the original case and may retain, change, or supplement findings; its narrative is final. For report claim $c_j = (u_j, I_j)$, let $v_j = V_\phi(u_j, \mathcal{E}[I_j])$ denote the verifier decision, where $\mathcal{E}[I_j]$ resolves the cited artifacts and status markers. The revision context is

$$\mathcal{F} = \{(u_j, m_j)\}_{j=1}^{J}, \qquad m_j = \begin{cases} \texttt{LOOK AGAIN}, & v_j = \texttt{revise}, \\ \emptyset, & v_j = \texttt{keep}. \end{cases} \tag{4}$$

The executor runs once more on $(q, x, \mathcal{F})$ with the permitted tools. Thus feedback selects claims for re-inspection without supplying replacement values; V_ϕ is the inference-time judge. Figure 2 illustrates this loop; Appendix C specifies task budgets and verifier inputs.

4 PATHSCORE: CLINICAL CONTENT EVALUATION

PathScore uses an LLM to extract predefined clinical fields from generated and reference reports. These fields describe diagnostic and staging information recorded in pathology reports (National Cancer Institute, 2022b). Of 18 candidate fields, it retains those mentioned in strictly more than 10% of TCGA-209 references (at least 21/209), yielding 15 fields including type, size, grade, invasion, margins, and stage. This inventory is fixed for external evaluation and is not a required generation template.

Generated-report extraction distinguishes unmentioned fields (`NA`), explicitly unassessable findings (`INDETERMINATE`), and stated values, including explicit negatives. Deterministic comparison assigns outcomes for each evaluated field. For slide s, let C_s, M_s, O_s, and I_s count correct, mismatched, omitted, and indeterminate outcomes against reference-supported fields. Our primary metric, reference-field recall (PS-R), is

$$\text{PS-R} = \frac{1}{|S|} \sum_{s \in S} \frac{C_s}{D_s}, \qquad D_s = C_s + M_s + O_s + I_s, \tag{5}$$

where S contains evaluated slides with $D_s > 0$. Each slide receives equal weight; mismatches, omissions, and indeterminate answers receive no credit.

We also report per-slide F1, averaged over the same slides:

$$\text{PS-F1} = \frac{1}{|S|} \sum_{s \in S} \frac{2C_s}{2C_s + 2M_s + O_s + I_s}. \tag{6}$$

A mismatch contributes both a false positive and a false negative. Macro-F1 pools these counts within each field and averages over fields with reference support. Fields absent from the reference are not scored. Appendix E gives the inventory and aggregation rules.

Why evaluate fields? In a controlled invasion example, deleting "not" gives a contradictory report higher BLEU-4 than a faithful paraphrase (0.669 vs. 0.525). Across seven constructed contrasts, evaluated BERTScore and embedding configurations preserve only 0–2 intended rankings; PathScore preserves all seven when supplied the displayed target values (Appendix E). These controls test the scoring criterion, not extraction accuracy. PathScore measures reference-content agreement; the verifier instead checks a claim's cited support.

5 Experiments

Evaluation scope. We use Qwen3.5-27B (Qwen Team, 2026) as the default executor; VQA and reporting include Occamy-1.0 (Chen et al., 2026) and Claude Opus 5 (Opus 5) (Anthropic, 2026). Classification also tests GPT-5.4-mini (OpenAI, 2026), GPT-5.5 (OpenAI, n.d.a), and GPT-5.6-sol (OpenAI, n.d.b); the last also serves as a separate verifier. TCGA-209 is an agent-development cohort; additional cohorts assess transfer. SlideChat† and WSI-LLaVA† have audited TCGA overlap, while MedGemma's slide-level exposure remains unresolved. Cohort construction, overlap counts, and protocols appear in Appendices A and C.

5.1 Classification

Setup. Full-653 contains 653 slide–task pairs from 512 WSIs and 475 patients across six breast, lung, prostate, and stomach tasks. Single-shot inputs are: S1, overview; S2, overview plus heatmap and top-20 patches; S3, S2 plus the MIL label and probabilities. Balanced accuracy (BA) is primary because tasks are imbalanced; a majority-class predictor attains 60.03% raw accuracy (Acc.) but 41.67% BA.

Results. Qwen3.5-27B + MedVESTA reaches 69.66% BA with self-verification and 70.25% with a GPT-5.6-sol verifier (Table 1). Both exceed ABMIL (68.90%) and all evaluated single-shot MLLMs. Without verification, Qwen3.5-27B + MedVESTA reaches 66.80%, compared with 57.28% for Qwen3.5-27B under S3. Self-verification therefore adds 2.86 points; the GPT-5.6-sol verifier adds 3.45 points. Qwen3.5-27B + MedVESTA with self-verification also exceeds MedGemma and LLaVA-Med under S3 by 18.16 and 22.56 points. In raw accuracy, Qwen3.5-27B + MedVESTA with the GPT-5.6-sol verifier scores 64.17%, above ABMIL (59.88%); PRISM2 remains highest at 68.91%.

5.2 Visual Question Answering

Setup. TCGA VQA contains 1,760 questions on 209 development slides: 1,189 from WSI-Bench (Liang et al., 2025), 311 from WSI-VQA (Chen et al., 2024), and 260 from SlideChat-TCGA (Chen

Table 1: Classification on full-653 (%). Top: single-shot S1–S3 settings. Bottom: reference models and MedVESTA configurations. Bold marks the best score per column within each block.

	S1		S2		S3	
Single-shot MLLM	BA	Acc.	BA	Acc.	BA	Acc.
Medical MLLMs and agents						
MedGemma-27B-it (Sellergren et al., 2025)	36.20	50.30	42.50	57.00	51.50	58.04
LLaVA-Med v1.5 (Li et al., 2023)	35.10	**56.20**	33.30	55.40	47.10	56.80
General VLMs						
Qwen3.5-122B (Qwen Team, 2026)	49.00	54.50	61.60	**62.60**	60.20	**66.00**
Qwen3.5-4B (Qwen Team, 2026)	41.50	48.50	42.40	52.10	44.90	50.10
Qwen3.5-27B (Qwen Team, 2026)	43.34	43.64	54.42	54.67	57.28	59.26
GPT-5.4-mini (OpenAI, 2026)	46.70	52.20	60.10	59.00	64.90	60.80
GPT-5.5 (OpenAI, n.d.a)	47.70	47.00	63.10	62.50	**68.40**	64.00
GPT-5.6-sol (OpenAI, n.d.b)	**55.90**	50.70	**63.90**	60.30	67.70	63.09

Reference model or MedVESTA configuration	BA	Acc.
Majority-class prediction	41.67	60.03
Pathology-specific models		
CONCH, zero-shot (Lu et al., 2023)	53.08	28.48
PRISM (Shaikovski et al., 2024)	64.04	56.66
PRISM2 (Vorontsov et al., 2026)	66.36	**68.91**
ABMIL specialist (the tool called by MedVESTA) (Ilse et al., 2018)	68.90	59.88
MedVESTA		
Qwen3.5-27B + MedVESTA (initial answer) (Qwen Team, 2026)	66.80	61.72
Qwen3.5-27B + MedVESTA (self-verifier) (Qwen Team, 2026)	69.66	63.71
Qwen3.5-27B + MedVESTA (GPT-5.6-sol verifier) (Qwen Team, 2026)	**70.25**	64.17

et al., 2025). External SlideBench evaluation uses 682 BCNB questions on 100 slides and 120 CPTAC questions on 120 slides. Each configuration is evaluated once; invalid/missing answers count as incorrect. WSI-Bench questions have known training overlap with WSI-LLaVA† (Appendix A.4).

Table 2: VQA accuracy (%). Top: external methods. Bottom: paired overview-only and MedVESTA (Agent) results. Bold marks the best Agent/external score. † marks TCGA training overlap; ‡ marks unresolved run-coverage conflicts excluded from comparative claims.

External method	TCGA-1,760	BCNB-682	CPTAC-120
Pathology-specific models			
SlideChat† (Chen et al., 2025)	66.90	51.61	60.83
WSI-LLaVA† (Liang et al., 2025)	66.70‡	46.48	48.30‡
PRISM (Shaikovski et al., 2024)	42.84	46.48	20.83
PRISM2 (Vorontsov et al., 2026)	57.16	41.79	69.17
Medical MLLMs and agents			
MedDr (He et al., 2024)	41.53	42.40	38.30
LLaVA-Med v1.5 (Li et al., 2023)	16.65	35.00	7.50

	TCGA-1,760		BCNB-682		CPTAC-120	
MedVESTA configuration	Overview	Agent	Overview	Agent	Overview	Agent
MedVESTA						
Qwen3.5-27B + MedVESTA (Qwen Team, 2026)	62.95	69.55	46.33	54.99	34.17	69.17
Qwen3.5-9B + MedVESTA (Qwen Team, 2026)	56.59	65.23	40.91	50.88	39.17	53.33
Occamy-1.0 + MedVESTA (Chen et al., 2026)	61.36	68.13	45.89	58.36	30.83	55.83
MedGemma-27B-it + MedVESTA (Sellergren et al., 2025)	57.56	59.09	45.31	56.16	29.17	55.83
Claude Opus 5 + MedVESTA (Anthropic, 2026)	72.56	**73.92**	63.20	**66.13**	65.83	**70.83**

Table 3: PathScore on TCGA-209. General VLM rows are single-shot outputs; MedVESTA rows are final reports. Bold marks the best score per metric. † denotes documented TCGA training overlap.

System	PS-R	PS-F1	Macro-F1
Pathology-specific models			
PRISM (Shaikovski et al., 2024)	0.2662	0.3853	0.2199
PRISM2 (Vorontsov et al., 2026)	0.4290	0.4927	0.4271
SlideChat† (Chen et al., 2025)	0.4148	0.4802	0.3307
WSI-LLaVA† (Liang et al., 2025)	0.4438	0.4973	0.3614
Medical MLLMs and agents			
LLaVA-Med v1.5 (Li et al., 2023)	0.3892	0.4646	0.3332
MedDr (He et al., 2024)	0.4026	0.4810	0.3870
MedGemma-27B-it (Sellergren et al., 2025)	0.4074	0.4847	0.3769
QCAgent + Patho-R1 (Wang et al., 2026; Zhang et al., 2025a)	0.3161	0.4037	0.2446
General VLMs			
Qwen3.5-27B (Qwen Team, 2026)	0.3458	0.4133	0.3234
Occamy-1.0 (Chen et al., 2026)	0.4026	0.4738	0.3801
Claude Opus 5 (Anthropic, 2026)	0.4784	0.5323	0.4102
MedVESTA			
Qwen3.5-27B + MedVESTA (Qwen Team, 2026)	0.4435	0.4758	0.3727
Occamy-1.0 + MedVESTA (Chen et al., 2026)	0.4663	0.5079	0.4137
Claude Opus 5 + MedVESTA (Anthropic, 2026)	**0.5539**	**0.5691**	**0.4889**

Results. All five MedVESTA configurations improve over their overview-only backbones on all three datasets (15/15 comparisons; Table 2). Claude Opus 5 + MedVESTA ranks first with 73.92%, 66.13%, and 70.83% accuracy on TCGA, BCNB, and CPTAC. Qwen3.5-27B + MedVESTA exceeds eligible external methods on TCGA and BCNB and matches PRISM2 on CPTAC. Its gains over overview-only Qwen3.5-27B are 6.60, 8.66, and 35.00 points; the corresponding gains for Claude Opus 5 + MedVESTA are 1.36, 2.93, and 5.00 points. MedGemma + MedVESTA gains 10.85 points on BCNB and 26.66 on CPTAC. These results show that specialist evidence benefits both general and medically trained backbones under the tested VQA protocol, which uses no verifier.

5.3 Report Generation

Setup. We evaluate TCGA-209 with available radiological/molecular evidence, EAGLE-48 (48 non-TCGA WSI–report pairs), and TCGA-242 (Chang et al., 2026), whose evaluated subset contains 119 WSIs from 95 patients. TCGA-242 is disjoint from the 672-slide parent cohort by slide and patient. PathScore fields are fixed across cohorts. Baselines include pathology models, medical MLLMs, QCAgent with Patho-R1 (Wang et al., 2026; Zhang et al., 2025a), and single-shot backbones. MedVESTA rows use final reports.

Development cohort. Claude Opus 5 + MedVESTA leads all three metrics on TCGA-209, with PS-R 0.5539 versus 0.4784 for single-shot Claude Opus 5 and 0.4290 for PRISM2 (Table 3). Qwen3.5-27B + MedVESTA reaches 0.4435, up from 0.3458 for its backbone; Occamy-1.0 + MedVESTA reaches 0.4663, up from 0.4026.

Additional cohorts. All three backbones improve on all three reported metrics across all three cohorts. Claude Opus 5 + MedVESTA leads PS-R on EAGLE-48 (0.4669) and TCGA-242 (0.4454), gaining 0.1860 and 0.1751 over single-shot inference (Table 4). It leads all three reported metrics on EAGLE-48 and two on TCGA-242, where PRISM2 leads Macro-F1. The PS-R gains over the best medical MLLM in each cohort are 0.1465, 0.0923, and 0.1034 for TCGA-209, EAGLE-48, and TCGA-242, respectively. Thus the strongest MedVESTA configuration exceeds the evaluated medically trained LLM-based baselines in reference-field recovery across all three cohorts.

The metric differences qualify this advantage. On TCGA-242, Claude Opus 5 + MedVESTA exceeds PRISM2 in PS-R by 0.0781 but trails its Macro-F1 by 0.0082: better average recovery per

Table 4: PathScore on EAGLE-48 and TCGA-242 (119 evaluated slides). Groups and formatting follow Table 3. Bold marks the best score per cohort and metric; † applies only to the TCGA panel.

(a) EAGLE-48 System	PS-R	PS-F1	Macro-F1
Pathology-specific models			
PRISM (Shaikovski et al., 2024)	0.1561	0.2487	0.1399
PRISM2 (Vorontsov et al., 2026)	0.1913	0.2956	0.1716
SlideChat (Chen et al., 2025)	0.3521	0.4290	0.2783
WSI-LLaVA (Liang et al., 2025)	0.4030	0.5019	0.4096
Medical MLLMs and agents			
LLaVA-Med v1.5 (Li et al., 2023)	0.3744	0.4606	0.2685
MedDr (He et al., 2024)	0.3746	0.4550	0.3073
MedGemma-27B-it (Sellergren et al., 2025)	0.2471	0.3305	0.2045
General VLMs			
Qwen3.5-27B (Qwen Team, 2026)	0.0712	0.0953	0.0926
Occamy-1.0 (Chen et al., 2026)	0.0679	0.0993	0.0733
Claude Opus 5 (Anthropic, 2026)	0.2809	0.3677	0.2504
MedVESTA			
Qwen3.5-27B + MedVESTA (Qwen Team, 2026)	0.4544	0.5033	0.4554
Occamy-1.0 + MedVESTA (Chen et al., 2026)	0.3609	0.4072	0.3881
Claude Opus 5 + MedVESTA (Anthropic, 2026)	**0.4669**	**0.5151**	**0.4889**
(b) TCGA-242 System	PS-R	PS-F1	Macro-F1
Pathology-specific models			
PRISM (Shaikovski et al., 2024)	0.1607	0.2557	0.2160
PRISM2 (Vorontsov et al., 2026)	0.3673	0.4600	**0.4155**
SlideChat† (Chen et al., 2025)	0.2640	0.3513	0.2448
WSI-LLaVA† (Liang et al., 2025)	0.3721	0.4345	0.3459
Medical MLLMs and agents			
LLaVA-Med v1.5 (Li et al., 2023)	0.3123	0.3956	0.2904
MedDr (He et al., 2024)	0.3218	0.4087	0.3410
MedGemma-27B-it (Sellergren et al., 2025)	0.3420	0.4267	0.4141
General VLMs			
Qwen3.5-27B (Qwen Team, 2026)	0.1217	0.1769	0.1755
Occamy-1.0 (Chen et al., 2026)	0.1261	0.1766	0.1492
Claude Opus 5 (Anthropic, 2026)	0.2703	0.3520	0.2777
MedVESTA			
Qwen3.5-27B + MedVESTA (Qwen Team, 2026)	0.3663	0.4211	0.3174
Occamy-1.0 + MedVESTA (Chen et al., 2026)	0.3121	0.3625	0.2959
Claude Opus 5 + MedVESTA (Anthropic, 2026)	**0.4454**	**0.4812**	0.4073

slide does not ensure the best field-balanced agreement. On TCGA-209, the same-backbone gain is larger in PS-R (0.0755) than PS-F1 (0.0368), showing that field recovery improves more than mismatch-sensitive agreement. These are complete-workflow comparisons, not isolated verification effects.

Evidence-guided revision. Differentiation and tumor grade reflect microscopic appearance (National Cancer Institute, 2022a). In Figure 2, flagged claims prompt re-inspection of gland formation and specialist readouts. The report changes from moderate to poor differentiation, matching the reference. PS-R rises from 3/7 to 5/7: two fields improve in total, differentiation and numerical grade, through the same grading revision; the figure illustrates the differentiation field. Tumor size and lymphovascular invasion remain mismatched, illustrating that evidence support and reference agreement are separate checks.

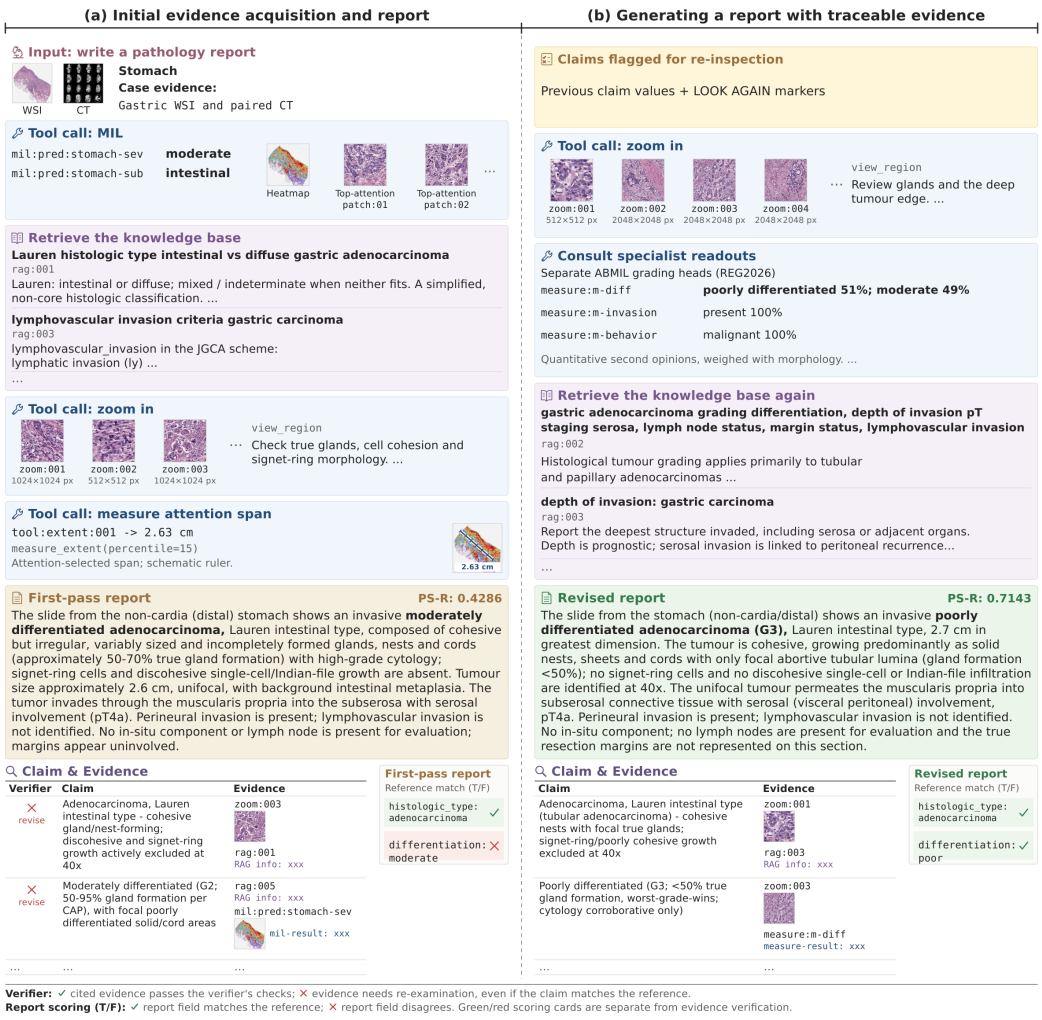
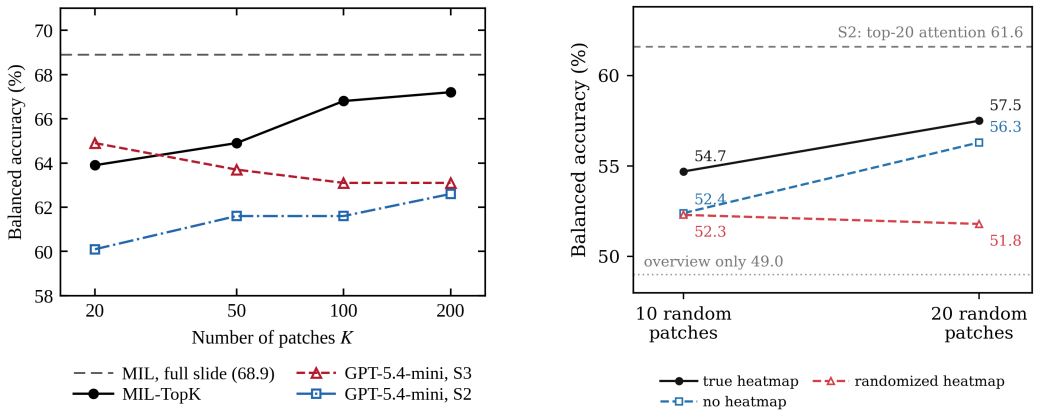


Figure 2: Evidence-guided report revision with Opus 5. Abbreviated schematic of a gastric report revision; selected steps and evidence are shown, not the full execution trace. Appendix F gives the trace and remaining errors.

5.4 Effects of Specialist Information and Verification

For Qwen3.5-27B, visual guidance adds 11.08 points in BA (S1 to S2), the MIL prediction adds 2.86 (S2 to S3), and Qwen3.5-27B + MedVESTA without verification adds 9.52 (S3 to 66.80%). Self-verification adds another 2.86 points. Using GPT-5.6-sol as verifier adds 0.59 points beyond self-verification, suggesting that much of the observed benefit is already obtained with the same backbone. These sequential configuration differences describe the tested workflow, rather than independent causal effects.

The single-shot results also reveal backbone-dependent responses to specialist information (Table 1). Visual guidance improves BA for seven of eight backbones, but LLaVA-Med drops from 35.10% under S1 to 33.30% under S2; adding the MIL readout then raises it to 47.10%. Conversely, Qwen3.5-122B drops from 61.60% to 60.20% BA when moving from S2 to S3, despite an increase in raw accuracy from 62.60% to 66.00%. Additional specialist information therefore need not improve class-balanced performance, even when overall accuracy rises. For Qwen3.5-27B, the initial MedVESTA answer remains 2.10 points below standalone ABMIL, while self-verification places it 0.76 points above the specialist. This comparison suggests a role for evidence-guided revision beyond simply exposing the executor to a specialist prediction.



(a) Patch-budget sensitivity

(b) Heatmap and patch-selection controls

Figure 3: Specialist-information controls. (a) Patch budget K for MIL-TopK and GPT-5.4-mini under S2/S3. (b) Qwen3.5-122B with random patches and true, absent, or randomized heatmaps; dashed lines mark overview-only and top-attention references.

MIL-TopK improves with patch budget, but GPT-5.4-mini's S3 scores are non-monotonic (Figure 3a). For Qwen3.5-122B, true heatmaps outperform randomized ones at both random-patch budgets, and attention-selected top-20 patches reach 61.6% BA versus 57.5% for random-20 patches with the same heatmap (Figure 3b; Appendix C.2). With random-20 patches, the true heatmap adds 1.2 points over no heatmap, while a randomized heatmap reduces BA by 4.5 points. Attention selection adds a further 4.1 points with the true heatmap held fixed. Both spatial guidance and patch choice matter; misleading guidance can be worse than omitting it.

Scope and limitations. Workflow comparisons change the evidence and tools available; report results do not isolate verification alone. Single-run VQA scores do not establish statistical significance. TCGA development use and baseline training overlap qualify in-domain results. PathScore depends on extraction and reference completeness and does not establish clinical validity.

6 Conclusion

MedVESTA connects specialist computation, selective inspection, and evidence-linked revision. It improves evaluated same-backbone VQA and reporting configurations and exceeds its MIL specialist in classification BA. PathScore complements these results with clinical-field comparisons while keeping reference agreement distinct from evidence support.

Ethics Statement

MedVESTA is a research system evaluated on retrospective benchmarks, not a validated clinical decision system. Incorrect findings, incomplete evidence, and automation bias could harm patients if generated reports were used without qualified review. Verifier approval and PathScore agreement do not establish clinical safety. Aggregate results also do not establish fairness across demographic groups, institutions, or rare diagnoses. Any clinical use would require prospective validation and appropriate oversight. Dataset access and reuse remain subject to the source terms; retrieved diagnostic material may have separate redistribution restrictions. We report known training overlap and retain case-level errors to make the limits of the evidence visible.

AI Use Statement

AI tools were used to assist with writing and language polishing. The authors are responsible for the final manuscript.

REFERENCES

H. J. W. L. Aerts, L. Wee, E. Rios Velazquez, R. T. H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C. R. Leemans, A. Dekker, J. Quackenbush, R. J. Gillies, and P. Lambin. Data From NSCLC-Radiomics (version 4). The Cancer Imaging Archive, data set, 2014. URL https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.

Anthropic. Introducing Claude Opus 5, 2026. URL https://www.anthropic.com/news/claude-opus-5. Published July 24, 2026. Accessed September 26, 2026.

Kai-Po Chang, Nan-Haw Chow, and Nicholas Tatonetti. TCGA-242-Pathology-Benchmark: External Validation Dataset for "The Digital Registrar". Zenodo, version 3, 2026. URL https://zenodo.org/records/20263897.

Pingyi Chen, Chenglu Zhu, Sunyi Zheng, Honglin Li, and Lin Yang. WSI-VQA: Interpreting Whole Slide Images by Generative Visual Question Answering. *arXiv preprint arXiv:2407.05603*, 2024. URL https://arxiv.org/abs/2407.05603.

Wenhui Chen, Shiwen Cheng, Hao Dong, Chenda Duan, Ruixiang Feng, Zhong Guan, Boqiang Guo, Xueyuan Han, Haojie Hao, Liangmeng Huang, Zhelong Huang, Xinke Kong, Hongyu Li, Jiazheng Li, Junbo Li, Qingchuan Li, Yukun Lian, Chang Liu, Tianyu Liu, Zicheng Liu, Shuyi Ouyang, Yijun Pan, Kunyu Shi, Xiaojun Tang, Bingquan Wang, Kesu Wang, Yuchen Wang, Sibo Wei, Sicong Xie, Xiaoying Xing, Yi Xu, Zhijun Xu, Hongwei Xue, Qingcheng Zeng, Di Zhang, Guannan Zhang, Haochen Zhang, Tianlong Zhang, Tianyu Zhao, Tianyu Zhao, Yanjun Zheng, Jialong Zhu, and Zijian Zou. Occamy-1.0: Open Pareto-frontier 35B Intelligence for Co-work. *arXiv preprint arXiv:2609.11977*, 2026. URL https://arxiv.org/abs/2609.11977.

Ying Chen, Guoan Wang, Yuanfeng Ji, Yanjun Li, Jin Ye, Tianbin Li, Ming Hu, Rongshan Yu, Yu Qiao, and Junjun He. SlideChat: A Large Vision-Language Assistant for Whole-Slide Pathology Image Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5134–5143, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chen_SlideChat_A_Large_Vision-Language_Assistant_for_Whole-Slide_Pathology_Image_Understanding_CVPR_2025_paper.html.

Nan-Haw Chow, Han Chang, Hung-Kai Chen, Chen-Yuan Lin, Ying-Lung Liu, Po-Yen Tseng, Li-Ju Shiu, Yen-Wei Chu, Pau-Choo Chung, and Kai-Po Chang. Digital Registrar: A Schema-First Framework for Multi-Cancer Privacy-Preserving Pathology Abstraction via Local LLMs. *Diagnostics*, 16(11):1644, 2026. doi: 10.3390/diagnostics16111644. URL https://www.mdpi.com/2075-4418/16/11/1644.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 26(6):1045–1057, 2013. doi: 10.1007/s10278-013-9622-7. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3824915/.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114. URL https://doi.org/10.1145/1571941.1572114.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, Ming Y. Lu, Andrew Zhang, Anurag J. Vaidya, Guillaume Jaume, Muhammad Shaban, Ahrong Kim, Drew F. K. Williamson, Harry Robertson, Bowen Chen, Cristina Almagro-Pérez, Paul Doucet, Sharifa Sahai, Chengkuan Chen, Christina S. Chen, Daisuke Komura, Akihiro Kawabe, Mieko Ochi, Shinya Sato, Tomoyuki Yokose, Yohei Miyagi, Shumpei Ishikawa, Georg Gerber, Tingying Peng, Long Phi Le, and Faisal Mahmood. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11): 3749–3761, 2025. doi: 10.1038/s41591-025-03982-3. URL https://www.nature.com/articles/s41591-025-03982-3.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and Bo Wang. MedRAX: Medical Reasoning Agent for Chest X-ray. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 15661–15676. PMLR, 2025. URL https://proceedings.mlr.press/v267/fallahpour25a.html.

Lidia Garrucho, Kaisar Kushibar, Claire-Anne Reidel, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, Maria-Laura Cosaka, Pasant M. Abo-Elhoda, Sara W. Tantawy, Shorouq S. Sakrana, Norhan O. Shawky-Abdelfatah, Amr Muhammad Abdo Salem, Androniki Kozana, Eugen Divjak, Gordana Ivanac, Katerina Nikiforaki, Michail E. Klontzas, Rosa García-Dosdá, Meltem Gulsun-Akpinar, Oğuz Lafcı, Ritse Mann, Carlos Martín-Isla, Fred Prior, Kostas Marias, Martijn P. A. Starmans, Fredrik Strand, Oliver Díaz, Laura Igual, and Karim Lekadir. A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. *Scientific Data*, 12(1):453, 2025. doi: 10.1038/s41597-025-04707-4. URL https://www.nature.com/articles/s41597-025-04707-4.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23, 2022. doi: 10.1145/3458754. URL https://doi.org/10.1145/3458754.

Sunan He, Yuxiang Nie, Zhixuan Chen, Zhiyuan Cai, Hongmei Wang, Shu Yang, and Hao Chen. MedDr: Diagnosis-Guided Bootstrapping for Large-Scale Medical Vision-Language Learning. *arXiv preprint arXiv:2404.15127v1*, 2024. URL https://arxiv.org/abs/2404.15127v1.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 2127–2136. PMLR, 2018. URL https://proceedings.mlr.press/v80/ilse18a.html.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z. URL https://www.nature.com/articles/s41592-020-01008-z.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. *arXiv preprint arXiv:2306.00890*, 2023. URL https://arxiv.org/abs/2306.00890.

Yuci Liang, Xinheng Lyu, Wenting Chen, Meidan Ding, Jipeng Zhang, Xiangjian He, Song Wu, Xiaohan Xing, Sen Yang, Xiyue Wang, and Linlin Shen. WSI-LLaVA: A Multimodal Large Language Model for Whole Slide Image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22718–22727, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Liang_WSI-LLaVA_A_Multimodal_Large_Language_Model_for_Whole_Slide_Image_ICCV_2025_paper.html.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Andrew Zhang, Long Phi Le, Georg Gerber, Anil V Parwani, and Faisal Mahmood. Towards a Visual-Language Foundation Model for Computational Pathology. *arXiv preprint arXiv:2307.12914*, 2023. URL https://arxiv.org/abs/2307.12914.

Mahmood Lab. UNI2-h: Official Model Card, n.d. URL https://huggingface.co/MahmoodLab/UNI2-h. Official model card; accessed 26 September 2026.

Andriy Myronenko. 3D MRI brain tumor segmentation using autoencoder regularization. *arXiv preprint arXiv:1810.11654*, 2018. URL https://arxiv.org/abs/1810.11654.

National Cancer Institute. Biomarker Testing for Cancer Treatment. National Cancer Institute, 2021. URL https://www.cancer.gov/about-cancer/treatment/types/biomarker-testing-cancer-treatment. Updated December 14, 2021. Accessed September 26, 2026.

National Cancer Institute. Tumor Grade. National Cancer Institute, 2022a. URL https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-grade. Updated August 1, 2022. Accessed September 26, 2026.

National Cancer Institute. Pathology Reports. National Cancer Institute, 2022b. URL https://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/pathology-reports-fact-sheet. Reviewed August 8, 2022. Accessed September 26, 2026.

National Cancer Institute. New Data from NCI's MATCH and EAGLE Studies Available, n.d.a. URL https://gdc.cancer.gov/news-and-announcements/new-data-nci%E2%80%99s-match-and-eagle-studies-available. Genomic Data Commons announcement; accessed 26 September 2026.

National Cancer Institute. The Cancer Genome Atlas Program (TCGA), n.d.b. URL https://www.cancer.gov/ccg/research/genome-sequencing/tcga. Official resource page; accessed 26 September 2026.

OpenAI. Introducing GPT-5.4 mini and nano, 2026. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/. Published March 17, 2026. Accessed September 26, 2026.

OpenAI. GPT-5.5 Model: OpenAI API Documentation, n.d.a. URL https://developers.openai.com/api/docs/models/gpt-5.5. Official model documentation. Accessed September 26, 2026.

OpenAI. GPT-5.6 Sol Model: OpenAI API Documentation, n.d.b. URL https://developers.openai.com/api/docs/models/gpt-5.6-sol. Official model documentation. Accessed September 26, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, 2002. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Qwen Team. Qwen3.5: Towards Native Multimodal Agents, 2026. URL https://qwen.ai/blog?id=qwen3.5.

REG2026 Challenge Organizers. Pathologist Reasoning-Guided Report Generation Challenge (REG2026): Official Challenge Website. Grand Challenge, 2026. URL https://reg2026.grand-challenge.org/. Challenge and dataset resource, not a journal article. Accessed 26 September 2026.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Stephen Robertson and Hugo Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 2009. doi: 10.1561/1500000019. URL https://doi.org/10.1561/1500000019.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa

Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. MedGemma Technical Report. *arXiv preprint arXiv:2507.05201*, 2025. URL https://arxiv.org/abs/2507.05201.

George Shaikovski, Adam Casson, Kristen Severson, Eric Zimmermann, Yi Kan Wang, Jeremy D. Kunz, Juan A. Retamero, Gerard Oakley, David Klimstra, Christopher Kanan, Matthew Hanna, Michal Zelechowski, Julian Viret, Neil Tenenholtz, James Hall, Nicolo Fusi, Razik Yousfi, Peter Hamilton, William A. Moye, Eugene Vorontsov, Siqi Liu, and Thomas J. Fuchs. PRISM: A Multi-Modal Generative Foundation Model for Slide-Level Histopathology. *arXiv preprint arXiv:2405.10254*, 2024. URL https://arxiv.org/abs/2405.10254.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y. Ng, and Matthew Lungren. Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1500–1519, 2020. doi: 10.18653/v1/2020.emnlp-main.117. URL https://aclanthology.org/2020.emnlp-main.117/.

Eugene Vorontsov, George Shaikovski, Adam Casson, Julian Viret, Eric Zimmermann, Neil Tenenholtz, Yi Kan Wang, Jan H. Bernhard, Ran A. Godrich, Juan A. Retamero, Jinru Shia, Mithat Gonen, Martin R. Weiser, David S. Klimstra, Razik Yousfi, Nicolò Fusi, Thomas J. Fuchs, Kristen Severson, and Siqi Liu. End-to-end multimodal pathology foundation model with clinical dialogue. *Nature Medicine*, 32(9):3203–3213, 2026. doi: 10.1038/s41591-026-04521-4. URL https://www.nature.com/articles/s41591-026-04521-4.

Rundong Wang, Wei Ba, Ying Zhou, Yingtai Li, Bowen Liu, Baizhi Wang, Yuhao Wang, Zhidong Yang, Kun Zhang, Rui Yan, and S. Kevin Zhou. QCAgent: An agentic framework for quality-controllable pathology report generation from whole slide image. *arXiv preprint arXiv:2603.01647*, 2026. URL https://arxiv.org/abs/2603.01647.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T. Meyer, Maurice Pradella, Daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023. doi: 10.1148/ryai.230024. URL https://pubs.rsna.org/doi/10.1148/ryai.230024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.03629.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1904.09675.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner. *arXiv preprint arXiv:2505.11404*, 2025a. URL https://arxiv.org/abs/2505.11404.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models. *arXiv preprint arXiv:2506.05176*, 2025b. doi: 10.48550/arXiv.2506.05176. URL https://arxiv.org/abs/2506.05176.

A Datasets and Cohort Construction

A.1 The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a joint NCI–NHGRI resource covering 33 cancer types (National Cancer Institute, n.d.b). It provides molecular profiles and linked clinical information, with histopathology whole-slide images (WSIs) available through the Genomic Data Commons (GDC).

A.2 TCGA–TCIA Cohort for Classification

To support multimodal analysis, we select TCGA cases with corresponding radiological images in The Cancer Imaging Archive (TCIA) (Clark et al., 2013), forming a patient-matched TCGA–TCIA cohort. Our inventory contains 672 WSIs. The labelled portion used for classification, denoted full-653, contains 653 slide-task pairs from 512 WSIs and 475 patients across six tasks, including the prostate grade, breast subtype, breast grade, stomach subtype, stomach differentiation, and lung subtype. These slides are excluded from fitting the pathology pooling networks and heads; this does not establish absence of encoder pretraining exposure. Table 5 summarizes the tasks and class distributions. BA is computed over classes present in each task and then averaged across the six tasks. The zero-count well-differentiated stomach class is excluded from that task's BA denominator.

Task	Classes	Class counts
Prostate grade	Benign/GG1; GG2–3; GG4–5	30 / 151 / 108
Breast subtype	IDC/NST; ILC; other	108 / 21 / 2
Breast grade	Grade I; II; III	16 / 67 / 42
Stomach subtype	Intestinal; diffuse	13 / 4
Stomach differentiation	Well; moderate; poor	0 / 24 / 15
Lung subtype	Adenocarcinoma; squamous	23 / 29
Total	512 unique WSIs	653

Table 5: Classification tasks and class counts in full-653. Counts follow the listed class order. GG denotes Gleason grade group; IDC/NST, invasive ductal carcinoma/no special type; and ILC, invasive lobular carcinoma. A WSI may contribute to multiple tasks.

A.3 Construction of the 209-Slide Cohort

To limit the development evaluation budget, we first select at most 12 slides per class of each classification task using random seed zero. We combine 144 subset slides with WSI-Bench questions and 96 slides covered by SlideChat-TCGA. Their intersection contains 31 slides, yielding a union of 209 WSIs. These slides all belong to the classification cohort and are shared by the VQA and report-generation development evaluations. The final union is not claimed to be class-balanced. Classification uses the full labelled inventory and reports balanced accuracy to account for class imbalance.

A.4 VQA Datasets and Training Exposure of Some Baselines

TCGA VQA combines report-derived questions from WSI-Bench (WSI-LLaVA) (Liang et al., 2025) and SlideBench-VQA-TCGA (SlideChat-TCGA) (Chen et al., 2025), together with caption-derived breast questions from WSI-VQA (Chen et al., 2024). We retain all eligible questions on the 209 slides: 1,189 from WSI-Bench, 311 from WSI-VQA, and 260 from SlideChat-TCGA, for 1,760 questions in total. External evaluation uses the BCNB breast and CPTAC lung portions of SlideBench-VQA, containing 682 questions on 100 WSIs and 120 questions on 120 WSIs, respectively.

Our slide-identifier audit finds training overlap for SlideChat† and WSI-LLaVA† with TCGA-209: 62/209 and 208/209 slides, respectively. In addition, the 1,189 WSI-Bench questions come from the WSI-LLaVA training partition. This exposure may affect the baselines' TCGA scores; its causal

effect has not been measured by a decontaminated control. MedGemma's slide-level exposure remains unresolved. The † symbol has this same meaning throughout the paper. The separate ‡ marker in Table 2 identifies unresolved run-coverage discrepancies, not training overlap.

A.5 Report-Generation Datasets and Training Exposure

TCGA-209 uses the slide reports released with TITAN (Ding et al., 2025) as references. Additional held-out evaluation uses EAGLE-48, comprising 48 WSI–report pairs from the non-TCGA GDC project `CDDP_EAGLE-1` (National Cancer Institute, n.d.a), and the TCGA-242 resource released with *The Digital Registrar* (Chang et al., 2026; Chow et al., 2026). The evaluated TCGA-242 subset contains 119 WSIs from 95 patients. Both additional cohorts use native human-written pathology reports; EAGLE reports are transcribed from scans.

TCGA-242 is disjoint from our 672-slide parent cohort at both the patient and slide levels, while remaining within TCGA; EAGLE-48 provides non-TCGA evaluation. Our audit finds that SlideChat† and WSI-LLaVA† overlap 42/119 and 118/119 of the TCGA-242 slides, respectively. Neither has documented EAGLE-48 overlap, so those report-table rows are unmarked. The absence of documented overlap is not a universal guarantee about pretraining exposure. Comparative results are reported in Tables 3 and 4.

B Specialist Models, Training Data, and Encoder Sources

B.1 Training Data and Architectures

REG2026 and specialist labels. Our pathology specialists are trained on REG2026, a pathology report-generation resource pairing H&E WSIs with reports and structured chain-of-thought (CoT) question–answer annotations (REG2026 Challenge Organizers, 2026). The training annotation file used in our pipeline contains 11,220 records covering breast, prostate, stomach, lung, colorectal, bladder, and cervical specimens. We derive classification labels using keyword rules applied to the CoT answers, and measurement labels using fixed answer-to-class mappings. These are automatically derived targets without additional manual adjudication. For each task or model group, we retain slides with the required labels and patch features. No TCGA slides are used to fit these pathology ABMIL models.

Frozen pathology encoder. Both specialist families use 1,536-dimensional patch embeddings from the official UNI2-h ViT-H/14 checkpoint, pretrained on H&E and IHC images (Mahmood Lab, n.d.). The encoder remains frozen; only the pooling networks and prediction heads are trained. Classification uses 256×256-pixel patches at $10\times$, while measurement models use feature bags at both $10\times$ and $20\times$.

Multi-task classification MIL. The classification specialist (`modeD`) uses gated attention-based multiple-instance learning (ABMIL) (Ilse et al., 2018). A shared projection maps each patch embedding from 1,536 to 512 dimensions, followed by ReLU and dropout. A gated attention module combines 256-dimensional tanh and sigmoid branches, normalizes their attention scores over patches, and pools the patch representations into a 512-dimensional slide vector. Six linear heads predict the prostate grade, breast subtype, breast grade, stomach subtype, stomach differentiation, and lung subtype tasks defined in Appendix A.2. The model has 1,058,065 trainable parameters. All six heads share one attention distribution; its heatmap is therefore shared across tasks.

The final classification checkpoint, `reg26_10x`, uses 5,763 slide–task pairs from 4,052 REG2026 WSIs. We use this REG2026-only configuration for classification, VQA, and report generation. Each task is split independently into approximately 90% training and 10% validation slides using seed 42, without class stratification, yielding 5,184 training and 579 validation slide–task pairs. Because splitting is performed independently per task, a slide held out for one head may occur in another head's training split. These validation splits serve early stopping and checkpoint selection; there is no separate internal test split.

Grading and measurement specialists. The `measure` family consists of five gated ABMIL groups with nine prediction heads in total (Table 6). Each group shares a 1,536-to-256-dimensional

projection and a 128-dimensional gated attention module among its heads. We train one model at each of $10\times$ and $20\times$ per group, yielding ten models. For each head, the numerical readout is the arithmetic mean of the two softmax probability vectors. For TCGA-209, cached numerical ensemble readouts and $10\times$ attention visualizations were produced by separate training runs with the same architecture and training recipe. These artifacts are used as cached inputs; no specialist weights are updated during an agent run. The group-specific training pools contain 1,076–7,273 REG2026 slides and include organs beyond the six classification tasks. Each group uses a random 90/10 slide split with seed 42, shared between magnifications. Validation is used for early stopping and checkpoint selection, with no separate internal test split.

Group	Prediction heads (classes)	Training organs	Train / val
Nottingham	Tubular differentiation (3); nuclear pleomorphism (3); mitotic rate (3)	Breast	969 / 107
Gleason	Grade group (5); worst pattern (3); pattern 4 presence (2)	Prostate	1,436 / 159
Differentiation / grade	Well, moderate, or poor differentiation / grade I–III (3)	Breast, stomach, colorectal	2,250 / 250
Invasion	Absent or present (2)	Breast, bladder	2,790 / 309
Behavior	Benign, in situ, or malignant (3)	Colorectal, prostate, stomach, bladder, lung, cervix	6,546 / 727

Table 6: REG2026 training and validation slides for the measurement specialists. Counts are per group and may overlap across groups. The differentiation/grade head maps breast grades I–III to the same three output categories as well/moderate/poor differentiation.

ABMIL optimization. Table 7 summarizes the training settings. Classification cycles through the six task loaders, updating on one slide and its task-specific class-weighted cross-entropy loss at a time. Within each measurement group, the class-weighted cross-entropy losses of its heads are summed with equal weight. Checkpoints are selected by mean validation balanced accuracy across the corresponding heads. Classification randomly resamples up to 2,000 patches per training slide on each presentation and uses all patches for validation. Measurement models sample up to 3,000 patches once per slide at loading, retaining that sample across epochs and validation. Both families use all available patches at inference.

Setting	Classification MIL	Measurement groups
Projection / attention dimension	512 / 256	256 / 128
Dropout	0.25	0.25
Optimizer	AdamW	AdamW
Learning rate / weight decay	2×10^{-4} / 10^{-5}	2×10^{-4} / 10^{-5}
Learning-rate schedule	Cosine, $T_{\max} = 30$	Constant
Gradient norm clipping	1.0	1.0
Maximum epochs / patience	30 / 8	40 / 12
Training patch limit	2,000, resampled	3,000, fixed sample
Random seed	42	42

Table 7: Optimization settings for the two pathology specialist families. UNI2-h remains frozen throughout training.

Radiology segmentation models. For lung tumor localization, we train a MONAI 3D SegResNet (Myronenko, 2018) with 16 initial filters on CT volumes and gross tumor volume (GTV) masks from NSCLC-Radiomics (Aerts et al., 2014). The usable pool contains 393 CT–mask pairs: the first 40 pairs in sorted order are reserved for validation and the remaining 353 for training. Volumes are resampled to 1.5-mm isotropic spacing, and intensities are clipped to $[-1000, 400]$ HU and scaled to $[0, 1]$. Training uses 96^3 crops, Dice plus cross-entropy loss, AdamW with learning rate 2×10^{-4},

cosine annealing, and 120 epochs. The predicted tumor mask supports downstream lung-location estimation.

We additionally use the official pretrained TotalSegmentator model (Wasserthal et al., 2023) (software version 2.14.0) for CT anatomy segmentation and the MAMA-MIA (Garrucho et al., 2025) `nnU-Net` (Isensee et al., 2021) five-fold ensemble for breast DCE-MRI tumor segmentation. These models retain the weights learned from their original training data and are used for inference without further fine-tuning. Their outputs provide anatomical masks or tumor contours for the toolbox; downstream field construction and interfaces are described in Appendix B.2.

B.2 Tool Details and Implementation

Diagnostic retrieval combines Qwen3-Embedding-0.6B (Zhang et al., 2025b) with BM25 (Robertson & Zaragoza, 2009), merging their rankings by reciprocal rank fusion (Cormack et al., 2009). Corpus size and tool budgets below describe our implementation.

Table 8: Tools used in MedVESTA. Radiology outputs are prepared offline and supplied to report generation when available. Specialist training and encoder sources are detailed in Appendix B.1.

Tool	Field	Input	Output	Implementation
MIL prediction	Pathology	WSI; frozen UNI2-h $10\times$ patch features; organ/task.	Class labels and probabilities; attention heatmap and top-20 patch views.	Six-head gated ABMIL (`reg26`) with attention-based rendering. Readouts provide initial evidence for classification and report generation; VQA retrieves them through `mil_readout`.
Grading and attribute readout	Pathology	UNI2-h $10\times$/$20\times$ features; relevant measurement head.	Grading or attribute predictions and leading class probabilities.	Five ABMIL groups, nine heads; average $10\times$/$20\times$ softmax outputs. Report generation receives initial readouts.
Region viewer `view_region`	Pathology	WSI; level-0 coordinates (x, y) and window size.	A 768×768 region image.	OpenSlide crop and resizing. Report-generation windows are centred at (x, y) and span 256–8192 level-0 pixels.
Attention-table computation `code_tool`	Pathology	Natural-language instruction; patch coordinates and attention in `heatmap.csv`.	Computed summaries and execution output.	A coding-model call generates standard-library Python; execution uses a subprocess with a 30-second timeout.
Spatial extent `measure_extent`	Pathology	Patch coordinates and attention; MPP; top-$p\%$ selection (default $p = 15$).	Two-dimensional attention span in centimetres.	Select the highest-attention patches; compute their global bounding-box diagonal and convert using microns per pixel (MPP).
Diagnostic retrieval RAG retrieval	Pathology	Diagnostic topic and organ.	Up to four attributed passages per call, within 4,000 characters; multiple calls are allowed.	Synonym expansion; Qwen3-Embedding-0.6B and BM25 retrieval; reciprocal rank fusion and organ filtering over 2,233 passages.

Continued on next page

Table 8 (continued)

Tool	Field	Input	Output	Implementation
Lung location	Radiology	Patient-matched lung CT.	Pulmonary lobe and laterality.	TotalSegmentator lung-lobe masks and a SegResNet tumour mask; align masks, select the lobe with greatest tumour overlap, and derive laterality.
CT overview	Radiology	Lung or stomach CT volume in Hounsfield units.	A 4×4 montage of 16 slices.	Uniform sampling across the middle 76% of the volume; window level/width 40/400 for lung and 40/350 for stomach.
Breast MRI overview	Radiology	Breast DCE-MRI volume.	A 3×3 montage with tumour contours.	MAMA-MIA nnU-Net five-fold ensemble masks; select nine slices with the largest mask areas and overlay their contours.

C Task Protocols and Evaluation Scope

The tasks share specialist outputs and an evidence registry, with task-specific interaction and verification protocols. Specialist implementations are in Appendix B and evidence types in Appendix D.

C.1 Task Routing

The router interprets the user's textual request and dispatches it to the appropriate workflow. Classification, multiple-choice VQA, report generation, and report question-answering requests are forwarded to the agent MLLM with the corresponding task instructions. Document-ingestion requests invoke text-processing and RAG-ingestion tools to add the supplied material to the retrieval corpus. Routing follows explicit rules when the task can be identified directly; Qwen3.5-27B serves as the fallback router for ambiguous requests.

C.2 Classification

Initial reasoning and a triggered revision each permit at most four tool rounds, followed by forced finalization. The verifier receives the six initial views, the specialist readout, diagnostic criteria, and claims; final-class and summary fields are withheld. Intermediate zoom crops are not included in this verifier context. One verification stage is used, with revision only for flagged claims. Invalid or missing class outputs fall back to the MIL prediction. The separate-verifier configuration replaces the executor's self-verification with GPT-5.6-sol; it uses the same Qwen3.5-27B executor.

Localization control. Figure 3b evaluates Qwen3.5-122B with attention-selected or random patches and true, absent, or randomized heatmaps. Table 9 gives the BA values plotted in the figure.

C.3 Visual Question Answering

The MedVESTA VQA configuration uses the overview, a heatmap, and two patch contact sheets, with a mandatory MIL readout. For TCGA, measurement requirements are determined by offline annotation of questions and answer options without the correct answers; measurement readouts are used on 648/1,760 questions. External questions use the MIL readout. There is no zoom, retrieval, or verifier stage. Overview-only controls therefore differ in available evidence as well as in tool use. Endpoint sampling defaults are retained; a deterministic temperature-zero setting is not claimed.

Configuration	BA
S1: overview only	49.0
S2: true heatmap + attention-selected top-20	61.6
Random 20, no heatmap	56.3
Random 20 + true heatmap	57.5
Random 20 + randomized heatmap	51.8
Random 10, no heatmap	52.4
Random 10 + true heatmap	54.7
Random 10 + randomized heatmap	52.3
MIL reference	68.9
CONCH reference	53.1

Table 9: Balanced accuracy (%) for the Qwen3.5-122B localization control in Figure 3b.

C.4 Report Generation

The first pass produces a free-form narrative and evidence-linked findings, with organ-specific knowledge lookups and visual inspection. The generation prompt requests at most six evidence-linked fields, each with one or two cited IDs; it does not require filling the PathScore inventory as a reporting template. The self-verifier audits emitted claims within its supported field set, using at most six cited images per claim. Flagged claim values receive `LOOK AGAIN` markers. The second pass starts from the same initial evidence package and tool budget, without the previous citation lists or detailed verifier rationales. Its report is final; no third verification stage is applied. References are used for evaluation, not as case evidence.

D Evidence Types and Registration Rules

Input artifacts, tool outputs, and retrieved RAG passages are registered as evidence that claims can cite by ID. Table 10 summarizes the reference types used in report generation.

Type	Example IDs	Content
Image evidence	`overview`, `heatmap`, `patch:01`, `zoom:003`	Slide views, attention maps, selected patches, and inspected regions.
Tool results	`mil:pred:<head>`, `measure:m-*`, `tool:extent:001`, `aux:fields`	Specialist predictions, measurements, and radiological or molecular readouts.
Retrieved knowledge	`rag:001`	Retrieved diagnostic criteria and explanatory passages.
Claim reference	`claim:<field>`	Another claim from the same slide, registered as `claim_ref`; for example, deriving pT from invasion depth.
Reported absence	`negative_finding`	A finding was not observed in the inspected material, or the relevant structure is absent from the section.
Unresolved finding	`inconclusive`	Available information is insufficient to determine the finding.

Table 10: Evidence records, claim references, and status placeholders.

Artifact IDs must resolve to registered records, which the verifier retrieves as images or text. Retrieved knowledge provides interpretive criteria and retains a separate source identity from case observations. The two status placeholders may stand alone; available supporting artifacts should also be cited. They do not retrieve an artifact and are assessed within the verifier's `keep`/`revise` decision.

A report claim contains a field, its value, and evidence IDs. The following example cites an inspected region and the breast-subtype MIL prediction:

```
{
  "field": "histologic_type",
  "value": "Invasive (infiltrating) lobular carcinoma, classic type  --
  ↪  discohesive single-file infiltration, no cohesive nests",
  "evidence": ["zoom:003", "mil:pred:breast-sub"]
}
```

E PathScore: Detailed Evaluation Protocol

This section defines the PathScore inventory and aggregation rules, followed by controlled comparisons with lexical and embedding similarity. Section 4 summarizes the main scoring definitions.

E.1 Field Inventory and Selection

We begin with 18 predefined pathology-report fields and retain those mentioned in strictly more than 10% of the TCGA-209 reference reports (at least 21 of 209 reports). An LLM extracts these fields from reference reports. Mention counts are pooled across the four organs, yielding the 15 retained fields in Table 11. This inventory is fixed before external evaluation and used unchanged for EAGLE-48 and TCGA-242. It defines the content evaluated by PathScore rather than a mandatory report template. Three candidates are excluded: necrosis, pleural invasion, and ulceration.

Field key	Stated-value format	Count
Retained fields		
`histologic_type`	Short diagnostic text	209
`tumor_size_cm`	Greatest dimension, in cm	134
`margin_status`	involved / clear	125
`lymphovascular_invasion`	present / absent	118
`differentiation`	well / moderate / poor	94
`grade`	1 / 2 / 3	76
`gleason_grade_group`	1 / 2 / 3 / 4 / 5	74
`perineural_invasion`	present / absent	72
`in_situ_component`	present / absent	66
`extraprostatic_extension`	present / absent	64
`lymph_node_status`	positive / negative	62
`tumor_focality`	unifocal / multifocal	58
`seminal_vesicle_invasion`	present / absent	38
`depth_of_invasion`	mucosa / submucosa / muscularis_propria / subserosa / serosa / beyond_serosa	26
`tnm_stage`	Pathological TNM text	23

Table 11: Retained fields and their TCGA-209 reference-mention counts. Missing fields are marked `NA`. Only the 15 retained fields contribute to PathScore.

E.2 Outcome Counts and Aggregation

Scoring compares generated values with reference-supported findings, including explicit negatives. If the reference is `NA`, the field is skipped regardless of the generated value and contributes to none of the counts. Otherwise exactly one outcome applies:

For the evaluated slides $\mathcal{S}_0$, define

$$D_s = C_s + M_s + O_s + I_s, \qquad \mathcal{S} = \{s \in \mathcal{S}_0 : D_s > 0\}. \tag{7}$$

Outcome	Condition	PS-R	PS-F1
C: correct	A stated generated value matches the reference, including a matching explicit negative.	$C_s{+}1, D_s{+}1$	TP
M: mismatch	A stated generated value disagrees with the reference.	$D_s{+}1$	FP + FN
O: omission	The generated value is `NA`, or extraction fails.	$D_s{+}1$	FN
I: indeterminate	The generated value is `INDETERMINATE`.	$D_s{+}1$	FN

Table 12: Mutually exclusive outcomes for reference-supported fields. TP, FP, and FN denote true positives, false positives, and false negatives.

The two slide-averaged metrics are

$$\text{PS-R} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{C_s}{D_s}, \tag{8}$$

$$\text{PS-F1} = \frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} \frac{2C_s}{2C_s + 2M_s + O_s + I_s}. \tag{9}$$

Each slide receives equal weight. Mismatches, omissions, and indeterminate answers receive no recall credit. A mismatch contributes both FP and FN; an omission or indeterminate answer contributes FN.

For Macro-F1, pool the same outcomes across slides for each retained field, e.g., $C_f = \sum_{s \in \mathcal{S}_0} \mathbf{1}[o_{sf} = C]$, and define $\mathcal{F}^+ = \{f : C_f + M_f + O_f + I_f > 0\}$. Then

$$\text{Macro-F1} = \frac{1}{|\mathcal{F}^+|} \sum_{f \in \mathcal{F}^+} \frac{2C_f}{2C_f + 2M_f + O_f + I_f}. \tag{10}$$

Fields without reference support are excluded; a supported field with no correct values receives zero. A slide with $D_s = 0$ contributes no scored outcomes to either aggregation. These metrics evaluate reference-content agreement and do not penalize statements about fields absent from the reference.

E.3 Controlled Comparisons with BLEU

Table 13 presents three constructed examples covering invasion polarity, margin status, and grade group. Each pairs a faithful paraphrase with a near-verbatim report that changes the target finding. These examples isolate the difference between lexical overlap and field-value agreement; they are not patient-level experimental results.

For these examples, we lowercase text, tokenize with `[a-z0-9]+`, and retain all words, including negation. We compute single-pair BLEU-1 and BLEU-4 with clipped n-gram precision p_n and the standard brevity penalty (Papineni et al., 2002):

$$\text{BLEU-}N = \text{BP} \exp\left(\frac{1}{N} \sum_{n=1}^{N} \log p_n\right), \qquad \text{BP} = \min\{1, \exp(1 - r/c)\}, \tag{11}$$

where r and c are reference and candidate token counts. No smoothing, stemming, or stopword removal is used; a zero precision sets the score to zero. For the controlled PathScore comparison, we specify normalized values directly from the texts and score the target field shown for each example. We denote this target-field score by PS-R_{target}. This applies the PathScore matching and aggregation rules to explicit values; it does not run the production extractor.

Clinical ranking comparison. PS-R_{target} ranks all three faithful paraphrases above their altered counterparts (3/3), whereas BLEU-1 and BLEU-4 rank none correctly (0/3). For each faithful paraphrase, the target field is correct ($C = 1$); each altered statement produces a mismatch ($M = 1$). Their PS-R scores are therefore 1 and 0, respectively. Given the displayed field values, PathScore more faithfully reflects preservation of the clinical finding in these examples, while BLEU rewards retention of the reference wording.

Candidate text	BLEU-1	BLEU-4	PS-R_{target}
Invasion polarity. Reference: Lymphovascular invasion is not identified in the examined tissue sections. Reference value: `absent`.			
No lymphovascular invasion is seen in the examined tissue sections. (`absent`)	0.800	0.525	**1.000**
Lymphovascular invasion is identified in the examined tissue sections. (`present`)	**0.895**	**0.669**	0.000
Margin status. Reference: The surgical margins are negative for invasive carcinoma in this specimen. Reference value: `clear`.			
The resection margins are free of invasive carcinoma in this specimen. (`clear`)	0.727	0.417	**1.000**
The surgical margins are positive for invasive carcinoma in this specimen. (`involved`)	**0.909**	**0.702**	0.000
Grade group. Reference: The prostate tumor is assigned Gleason grade group 2 on histologic examination. Reference value: `2`.			
Histologic examination of the prostate tumor indicates Gleason grade group 2. (`2`)	0.747	0.345	**1.000**
The prostate tumor is assigned Gleason grade group 5 on histologic examination. (`5`)	**0.917**	**0.735**	0.000
Correctly ranked pairs	0/3	0/3	**3/3**

Table 13: BLEU and PathScore on three controlled clinical contrasts. The first candidate in each pair is faithful; the second changes the target finding. The higher score within each pair is bold. PS-R_{target} uses the displayed, explicitly specified target-field values. It preserves the intended ranking in all three pairs, whereas both BLEU variants reverse all three rankings.

E.4 Controlled Comparisons with BERTScore and Embedding Similarity

Clinical report evaluation should prefer a faithful paraphrase over a clinically altered statement, even when the latter preserves more reference wording. We test this requirement with seven controlled contrasts.

Contrastive design. Each reference r_i is paired with a faithful paraphrase g_i^+ and a clinically altered statement g_i^- (Table 14). For each scoring configuration, we compute

$$\Delta_i = s(r_i, g_i^+) - s(r_i, g_i^-). \tag{12}$$

A positive gap indicates the intended ordering; a negative gap means that the clinically altered statement receives the higher score. We also compute PS-R_{target} using the direct-value target-field procedure in Section E.3. Values in parentheses below specify the score inputs. Each contrast scores its named field; Example 7 scores both lymphovascular and perineural invasion.

Table 14: Seven constructed clinical contrasts. Each reference is paired with a faithful paraphrase and a clinically altered statement. Example 6 changes an explicit negative finding to an unassessable finding. Parentheses give explicitly specified normalized target values for PS-R_{target}, not outputs from the production extractor.

Reference	Faithful paraphrase	Clinically changed statement
1. Lymph-node status		
Metastatic carcinoma is present in the regional lymph nodes. (`positive`)	Regional lymph nodes contain metastatic carcinoma. (`positive`)	Metastatic carcinoma is absent from the regional lymph nodes. (`negative`)
2. Seminal-vesicle invasion		
Seminal vesicle invasion is not identified. (`absent`)	No invasion of the seminal vesicles is seen. (`absent`)	Seminal vesicle invasion is identified. (`present`)

Continued on next page

Table 14 (continued)

Reference	Faithful paraphrase	Clinically changed statement
3. Tumour size and unit equivalence		
The largest tumour dimension is 2.0 cm. (2.0 cm)	The tumour measures 20 mm in greatest dimension. (2.0 cm)	The largest tumour dimension is 8.0 cm. (8.0 cm)
4. Numerical grade		
The tumour is Nottingham grade 1. (`1`)	Nottingham grading assigns the tumour to grade I. (`1`)	The tumour is Nottingham grade 3. (`3`)
5. Depth of invasion		
The tumour invades the subserosa but does not reach the serosa. (`subserosa`)	Invasion is limited to the subserosa; the serosal surface is uninvolved. (`subserosa`)	The tumour invades through the subserosa and reaches the serosa. (`serosa`)
6. Negative finding versus unassessable finding		
Lymphovascular invasion is not identified. (`absent`)	No lymphovascular invasion is seen. (`absent`)	Lymphovascular invasion cannot be assessed on this section. (`INDETERMINATE`)
7. Swapping values between two findings		
Lymphovascular invasion is present; perineural invasion is absent. (`present`, `absent`)	Lymphovascular invasion is identified, without perineural involvement. (`present`, `absent`)	Lymphovascular invasion is absent; perineural invasion is present. (`absent`, `present`)

Scoring configurations. We compare three BERTScore F1 configurations and four cosine configurations, listed in Tables 15 and 16. BERTScore aggregates contextual token matches (Zhang et al., 2020); single-vector cosine similarity compares pooled representations. The configurations use PubMedBERT (Gu et al., 2022), RoBERTa-large (Liu et al., 2019), Sentence-BERT MiniLM (Reimers & Gurevych, 2019), and Qwen3-Embedding-0.6B (Zhang et al., 2025b). For an encoder E and pooling operation P, it is

$$s_{\text{emb}}(r, g) = \frac{P(E(r))^\top P(E(g))}{\|P(E(r))\| \, \|P(E(g))\|}. \tag{13}$$

RoBERTa baseline rescaling applies $s' = (s - b)/(1 - b)$ for a fixed baseline $b < 1$, preserving the ordering of sentence pairs.[1] The seven configurations are therefore not independent metrics; mean-gap magnitudes are also not directly comparable across score scales.

Observed rankings. Tables 15 and 16 report the paired scores. Every similarity configuration prefers the clinically altered statement in five to seven of the seven contrasts: five configurations misrank six contrasts, MiniLM misranks five, and Qwen3 misranks all seven. The mean Δ is negative for every similarity configuration, with no ties at the reported precision. Example 6 is correctly ordered by six of the seven configurations; Qwen3 is the exception.

All seven similarity configurations misrank Example 7, which swaps the positive and negative values of two findings while preserving the reference's word multiset. MiniLM assigns 0.9982 to the altered statement and 0.9042 to the faithful paraphrase. This demonstrates incorrect clinical ordering, beyond a narrow numerical range.

[1] BERTScore reference implementation: `https://github.com/Tiiiger/bert_score/blob/master/bert_score/score.py`.

Table 15: BERTScore F1 and PathScore: faithful / altered scores for Table 14. Larger values are bold; bold on the right indicates a misranking. RoBERTa-L denotes RoBERTa-large. PS-R_{target} uses the displayed target-field values.

Contrast	PubMedBERT	RoBERTa-L (raw)	RoBERTa-L (rescaled)	PS-R_{target}
1. Lymph nodes	0.9744 / **0.9852**	0.9412 / **0.9845**	0.6519 / **0.9084**	**1** / 0
2. Seminal vesicles	0.9655 / **0.9886**	0.9270 / **0.9842**	0.5674 / **0.9064**	**1** / 0
3. Size and units	0.9562 / **0.9927**	0.9276 / **0.9949**	0.5708 / **0.9695**	**1** / 0
4. Numerical grade	0.9635 / **0.9951**	0.9028 / **0.9942**	0.4240 / **0.9658**	**1** / 0
5. Invasion depth	0.9480 / **0.9821**	0.9141 / **0.9720**	0.4910 / **0.8342**	**1** / 0
6. Unassessable	**0.9853** / 0.9626	**0.9596** / 0.9453	**0.7604** / 0.6761	**1** / 0
7. Finding swap	0.9738 / **0.9968**	0.9548 / **0.9900**	0.7323 / **0.9407**	**1** / 0
Correct / wrong	1 / 6	1 / 6	1 / 6	**7 / 0**
Mean Δ	-0.0195	-0.0483	-0.2862	$+1.0000$

Table 16: Single-vector cosine and PathScore: faithful / altered scores, with the larger value in bold. SBERT denotes Sentence-BERT MiniLM; Qwen3 denotes Qwen3-Embedding-0.6B. PS-R_{target} uses the explicit target values in Table 14.

Contrast	PubMedBERT (mean pool)	PubMedBERT ([CLS])	SBERT MiniLM	Qwen3 Embedding	PS-R_{target}
1. Lymph nodes	0.9895 / **0.9956**	0.9731 / **0.9827**	**0.9682** / 0.9089	0.9116 / **0.9202**	**1** / 0
2. Seminal vesicles	0.9858 / **0.9943**	0.9704 / **0.9863**	0.8486 / **0.8876**	0.8446 / **0.9320**	**1** / 0
3. Size and units	0.9896 / **0.9988**	0.9800 / **0.9986**	0.8037 / **0.9428**	0.9047 / **0.9468**	**1** / 0
4. Numerical grade	0.9891 / **0.9991**	0.9279 / **0.9988**	0.8303 / **0.9693**	0.9023 / **0.9546**	**1** / 0
5. Invasion depth	0.9856 / **0.9970**	0.9597 / **0.9870**	0.7086 / **0.8634**	0.7888 / **0.9134**	**1** / 0
6. Unassessable	**0.9941** / 0.9865	**0.9839** / 0.9743	**0.8603** / 0.8489	0.8969 / **0.9366**	**1** / 0
7. Finding swap	0.9934 / **0.9997**	0.9826 / **0.9991**	0.9042 / **0.9982**	0.9678 / **0.9942**	**1** / 0
Correct / wrong	1 / 6	1 / 6	2 / 5	0 / 7	**7 / 0**
Mean Δ	-0.0063	-0.0213	-0.0707	-0.0544	$+1.0000$

PathScore on the same contrasts. With the displayed target values, PS-R_{target} assigns 1 to every faithful paraphrase and 0 to every altered statement, preserving all seven intended rankings (7/7), compared with 0–2/7 for the similarity configurations. In Example 3, 20 mm and 2.0 cm normalize to the same value, whereas 8.0 cm is a mismatch. In Example 7, the faithful paraphrase preserves both field–value associations ($C = 2$); swapping their values creates two mismatches ($M = 2$). In Example 6, the altered statement is indeterminate ($I = 1$), not contradictory: it receives no credit for recovering the reference's explicit negative finding.

Implications for report evaluation. On these controlled examples, the PathScore matching criterion more accurately preserves the intended clinical ordering than the tested lexical and embedding-based similarities. Explicit field–value comparison retains faithful paraphrases while distinguishing negation, numerical changes, uncertainty, and values assigned to different findings. This comparison evaluates scoring with specified target values; end-to-end performance additionally depends on extraction accuracy. The constructed cases do not estimate failure rates in patient reports.

F Report-Generation Case Study

This gastric case uses a WSI, case information, and an offline CT montage. Specialist readouts and visual inspection provide evidence for the MLLM. Figure 2 is an abbreviated schematic of this example: it omits intermediate calls, and panel placement does not imply when a cached readout was first available. Table 17 records selected workflow steps with Opus 5 on a gastric cancer case, from

specialist-tool use to report generation, verification, and revision. Evidence IDs are local to each pass; matching numeric suffixes across passes need not identify the same artifact. The extent readout is an attention-selected span, not a validated anatomical tumor-size measurement. The generated 2.7-cm size claim incorrectly treats this span as tumor size and remains an error against the 9.0-cm reference.

Table 17: A report-generation case with Opus 5 (`TCGA-VQ-A91Y-01Z-00-DX1`). Selected steps are grouped by function, not presented as a complete chronological log. The reference report is used only for evaluation.

Stage / tool	Action and returned evidence
Task and raw inputs	Generate a pathology report from the case information, a raw gastric WSI, and a patient-matched CT montage prepared offline.
Tool call: MIL	Analyse the WSI for gastric subtype and differentiation. **Returns:** intestinal subtype (`mil:pred:stomach-sub`); moderate differentiation (`mil:pred:stomach-sev`); attention heatmap and top-20 patch views.
Specialist readout	Obtain specialist grading and attribute predictions from the WSI. **Returns:** poor differentiation, 51%; moderate differentiation, 49% (`measure:m-diff`); stromal invasion present, 100% (`measure:m-invasion`); malignant behaviour, 100% (`measure:m-behavior`), with the measurement model's attention map and top-20 patch views.
Initial knowledge retrieval RAG retrieval	Across multiple calls, retrieve criteria for Lauren classification, differentiation, invasion depth, lymphovascular invasion, lymph nodes, and margins. **Returns:** attributed passages `rag:001`–`rag:005` across calls; each call returns at most four passages.
Initial visual inspection `view_region`	Inspect gland formation, cell cohesion, and the deep tumour edge at several scales, including possible neural and vascular invasion. **Returns:** six region views, `zoom:001`–`006`. For example, `zoom:003` is a 1024-pixel window centred at $(70478, 47156)$, requested to inspect gland/tubule formation.
Tool call: extent `measure_extent`	Measure the spatial span of the top 15% of attention-ranked patches. **Returns:** 2.63 cm, referenced as `tool:extent:001`.
First-pass report	The slide from the non-cardia (distal) stomach shows an invasive moderately differentiated adenocarcinoma, Lauren intestinal type, composed of cohesive but irregular, variably sized and incompletely formed glands, nests and cords (approximately 50–70% true gland formation) with high-grade cytology; signet-ring cells and discohesive single-cell/Indian-file growth are absent. Tumour size approximately 2.6 cm, unifocal, with background intestinal metaplasia. The tumor invades through the muscularis propria into the subserosa with serosal involvement (pT4a). Perineural invasion is present; lymphovascular invasion is not identified. No in-situ component or lymph node is present for evaluation; margins appear uninvolved.
Initial grade claim	*Moderately differentiated (G2; 50–95% gland formation per CAP), with focal poorly differentiated solid/cord areas.* **Cited evidence:** `rag:005` and `mil:pred:stomach-sev`.
Verification and feedback	The verifier checks the six claims against their cited evidence and flags all six for re-inspection. For the grade claim, it identifies the need for visual evidence supporting the stated gland-formation percentage. **Feedback to the agent:** previous claim values with `LOOK AGAIN` markers. The markers identify claims to revisit; the verifier's detailed rationales are not supplied to the revision pass.
Visual re-inspection `view_region`	Inspect additional architectural windows to reassess gland formation versus solid growth. The figure shows selected views rather than the complete sequence. **Returns:** revision-pass `zoom:002` and `zoom:003`, together with views of the deep edge and mucosal region. The revised interpretation describes predominantly solid nests, sheets, and cords with focal tubular lumina.

Continued on next page

Table 17 (continued)

Stage / tool	Action and returned evidence
Further tool calls RAG retrieval; `measure_extent`; ...	Retrieve gastric classification, grading, and invasion criteria; repeat the extent calculation using the top 40% of attention-ranked patches. **Returns:** revision-pass `rag:001–rag:003`; an attention-derived span of 2.72 cm (`tool:extent:001`).
Revised grade claim	*Poorly differentiated (G3; $< 50\%$ true gland formation, worst-grade-wins; cytology corroborative only).* **Cited evidence:** revision-pass `zoom:003` and the specialist-tool output `measure:m-diff`.
Revised report	The slide from the stomach (non-cardia/distal) shows an invasive poorly differentiated adenocarcinoma (G3), Lauren intestinal type, 2.7 cm in greatest dimension. The tumour is cohesive, growing predominantly as solid nests, sheets and cords with only focal abortive tubular lumina (gland formation $< 50\%$); no signet-ring cells and no discohesive single-cell or Indian-file infiltration are identified at 40x. The unifocal tumour permeates the muscularis propria into subserosal connective tissue with serosal (visceral peritoneal) involvement, pT4a. Perineural invasion is present; lymphovascular invasion is not identified. No in-situ component; no lymph nodes are present for evaluation and the true resection margins are not represented on this section.
Reference report (evaluation only)	The slide from the gastric antrum shows a tubular adenocarcinoma, histologic grade III (poorly differentiated), intestinal type (Lauren classification), with a largest tumor size of 9.0 cm. Features include ulceration, local invasion to the serosa, and evidence of neural, lymphatic, and blood vessel invasion.
Reference comparison	PS-R improves from 0.4286 (3/7 fields) to 0.7143 (5/7 fields). Two fields in total, differentiation and numerical grade, now match the reference; both improvements arise from the same grading revision. Figure 2 displays the differentiation field as the example. Tumour size and lymphovascular invasion remain mismatched.