PATHOLOGY HARNESS: EVIDENCE-GUIDED DIAGNOSTIC FINALIZATION FOR PATHOLOGY AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in computational pathology have enabled systems that combine whole-slide image analysis with diagnostic reasoning. These systems aggregate regional features, generate morphological descriptions, or navigate tissue at multiple magnifications to formulate diagnostic proposals. However, producing a plausible proposal does not ensure that the collected observations document the morphological findings and differential checks needed to support it. This gap raises a practical question: when is the available evidence sufficient to finalize a diagnosis, and what should be examined when it is not? To address this problem, we introduce Pathology Harness, a framework that governs diagnostic finalization through explicit evidence requirements. The framework binds each fixed diagnostic proposal to an evidence contract, audits the recorded observations against its requirements, and determines whether to release the proposal, request further observation, or block finalization. To resolve recoverable evidence gaps, we develop Differential-Contract Evidence Acquisition (DCE-Scout), which selects complementary regions and magnifications under a limited observation budget. Newly acquired findings can support or contradict the proposal and are recorded for renewed audit, preserving the basis for each release decision. Experiments reveal that even correct diagnostic proposals can lack the recorded evidence required for finalization, highlighting a gap between diagnostic prediction and evidential support. Pathology Harness addresses this gap through targeted evidence acquisition and selective release, allowing more cases to meet explicit finalization criteria than initial-record auditing alone and lowering the error rate among released diagnoses compared with unconditional release. These findings demonstrate the value of coupling evidence acquisition with explicit finalization criteria and suggest a promising direction for pathology agents that produce traceable, evidence-supported diagnostic outputs.

1 INTRODUCTION

Histopathological examination is central to cancer diagnosis, making the analysis of tissue morphology an important application of artificial intelligence. Digitizing tissue slides as whole-slide images (WSIs) has enabled computational methods to assist this analysis at scale (Litjens et al., 2017; van der Laak et al., 2021). Recent pathology foundation models have further advanced tissue representation and visual understanding (Chen et al., 2024; Xu et al., 2024; Lu et al., 2024a). These developments create opportunities for diagnostic systems that integrate observations across tissue regions and magnifications to support their conclusions.

Computational pathology has progressed from slide-level prediction toward interactive evidence gathering. Foundation-model features and multiple-instance learning aggregate regional information into diagnostic predictions (Chen et al., 2024; Lu et al., 2021; Shao et al., 2021), while pathology vision–language models connect tissue images with morphological descriptions and diagnostic dialogue (Lu et al., 2024a;b). More recently, pathology agents have begun to organize the observation process itself: CPathAgent navigates slides across magnifications, and WSI-Agents incorporates verification of intermediate findings (Sun et al., 2025; Lyu et al., 2025). SlideSeek and BEACON further integrate decisions about continued observation into diagnostic reasoning (Weishaupt et al., 2025; Wong et al., 2026). Together, these advances provide a basis for systems that actively gather, interpret, and assess evidence while forming a diagnosis.

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.

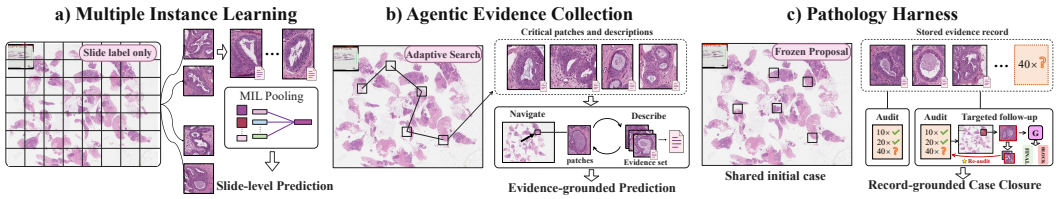


Figure 1: From slide-level prediction and agentic evidence collection to Pathology Harness's record-grounded case closure. Panel (c) separates proposal and record computations over shared images.

However, forming a plausible diagnosis does not by itself establish that the observations needed to support it have been collected. Those requirements depend on the proposed diagnosis and its alternatives, and relevant findings may occur in different regions or at different scales. For example, a low-magnification description may suggest a diagnosis while leaving a distinguishing tissue boundary unexamined at the required resolution. Deciding whether to finalize therefore requires connecting the proposed diagnosis to specific evidence requirements, identifying which are satisfied, and determining what remains to be observed. Figure 1 illustrates this distinction between producing a prediction, collecting evidence, and establishing a basis for finalization. This raises a practical question: when do the available observations meet the evidence requirements for a diagnosis, and what should be examined when they do not?

To address this challenge, we introduce Pathology Harness, a framework that makes diagnostic finalization depend on explicit evidence requirements. Our central idea is to connect the observations required for a proposed diagnosis with both the decision to finalize and the choice of further observations. We express these requirements as an evidence contract specifying the morphological findings and differential checks to be assessed. Comparing this contract with the stored observations establishes which requirements are satisfied and which remain unresolved, providing a common basis for finalization, further examination, or withholding the diagnosis.

Specifically, Pathology Harness receives an initial diagnostic proposal and observation record, binds the proposal to a compatible contract, and audits the recorded findings (Figure 2). For recoverable gaps, Differential-Contract Evidence Acquisition (DCE-Scout) selects complementary regions and magnifications under a limited observation budget. New findings are appended to the record and re-audited; they may support or contradict the proposal, which remains fixed throughout this process. The framework then determines whether to finalize, continue observation, or block finalization, retaining the basis for each decision. On BRACS and EBRAINS, this evidence-repair policy lowers the error rate among finalized diagnoses relative to unconditional release at reduced coverage; on BRACS, it also finalizes more cases than auditing the initial record alone. Our contributions are:

- **Evidence-guided diagnostic finalization.** We introduce a framework that connects proposal-specific evidence requirements with decisions to finalize, continue observation, or block finalization, while retaining the evidence and audit history.
- **Contract-guided evidence acquisition.** We develop DCE-Scout to translate missing findings and differential checks into budgeted, non-redundant observation requests that allow both supporting and contradictory evidence.
- **Matched policy evaluation.** We evaluate how evidence-based finalization and follow-up acquisition affect release coverage and held-out-label error, using shared, fixed proposals to separate these effects from diagnostic prediction.

2 Related Work

2.1 Representations and Regional Evidence in Pathology

WSI representations and multiple-instance learning derive slide proposals from regional features. Encoders such as CTransPath, UNI, Virchow, and Prov-GigaPath capture tissue appearance, while hierarchical and spatial models aggregate across scales (Wang et al., 2022; Chen et al., 2024; Vorontsov et al., 2024; Xu et al., 2024; Chen et al., 2022; Wang et al., 2024; Ding et al., 2025).

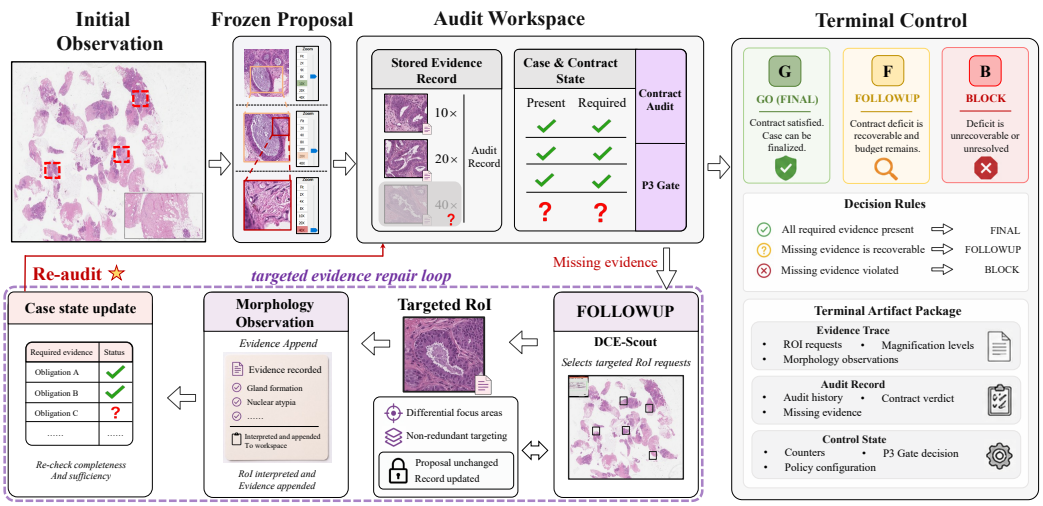


Figure 2: Pathology Harness workflow. Contract Auditor and P3 Gate govern finalization; recoverable gaps trigger budgeted DCE-Scout acquisition and re-audit.

Attention-based MIL, CLAM, TransMIL, and related methods link regional features to slide predictions and localization (Ilse et al., 2018; Li et al., 2021; Lu et al., 2021; Shao et al., 2021; Zhang et al., 2022; Chen et al., 2021). Pathology vision–language models such as CONCH, PLIP, and PathChat provide complementary morphological descriptions (Lu et al., 2024a; Huang et al., 2023; Lu et al., 2023; Ikezogwo et al., 2023; Seyfioglu et al., 2024; Lu et al., 2024b). The Harness uses these proposal and observation sources; its contracts govern stored observations' role in case closure.

2.2 Agentic Observation and Evidence Acquisition

Agentic pathology extends the reasoning–action loop to gigapixel images (Yao et al., 2023). CPathAgent performs multi-scale navigation, and PathFinder coordinates diagnostic decision making across specialist components (Sun et al., 2025; Ghezloo et al., 2025). WSI-Agents verifies claims through internal consistency and external knowledge checks, retaining verification logs for its final summary (Lyu et al., 2025). SlideSeek, built around PathChat+, updates exploration plans and diagnostic hypotheses until its supervisor judges the evidence sufficient (Weishaupt et al., 2025). These systems establish that evidence gathering, verification, and stopping are central to agentic WSI analysis.

BEACON selects observations using expected information gain over competing diagnostic hypotheses and controls whether to continue, inspect at higher resolution, or answer (Wong et al., 2026). DCE-Scout instead targets the missing obligations of a contract already bound to a fixed proposal. Its acquisition objective combines coverage of those obligations with differential relevance, diversity, and cost; the resulting observations return to a separate audit and authorization procedure. DCE-Scout thus links contract-targeted acquisition to audit-governed release authorization.

2.3 Selective Prediction and Runtime Authorization

Selective classification, error detection, conformal uncertainty, and learning to defer provide ways to withhold unreliable predictions (Geifman & El-Yaniv, 2017; 2019; Hendrycks & Gimpel, 2017; Angelopoulos & Bates, 2021; Madras et al., 2018; Mozannar & Sontag, 2020). These frameworks motivate joint risk–coverage evaluation. Our score-based comparison shares the Harness proposal generator, distinguishing initial-score selection from a policy that acquires further observations.

Shared-state systems externalize intermediate computation (Engelmore & Morgan, 1986), and AgentSpec enforces runtime predicates at action and completion events (Wang et al., 2025). Recent work also argues for requiring verifiable evidence of task completion within an agent harness (Ng et al., 2026). We build on this broader separation of agent generation and runtime control. The pathology-specific contribution is to bind regional evidence obligations to a fixed proposal and connect recoverable deficits to acquisition, re-audit, and terminal authorization within one lifecycle.

3 METHODOLOGY

3.1 PROBLEM FORMULATION AND LIFECYCLE OVERVIEW

Given a case $c = (I, \pi)$ comprising an anonymized whole-slide image I and an optional clinical prior π, Pathology Harness determines whether the acquired evidence supports releasing a diagnostic proposal $\hat{y} \in \mathcal{Y}$. We keep $\hat{y}$ fixed throughout case management; a revised diagnosis starts a new episode. Figure 2 shows the resulting lifecycle: bind the proposal to an evidence contract, audit the record, and either release the proposal, acquire further observations, or block finalization.

The initial Evidence Scout selects ROI–magnification requests before proposal generation. The crops feed two computations: features produce $\hat{y}$, while the Morphology Interpreter describes the crops and the Evidence Compiler constructs the initial record E_c^0. The runtime record is

$$T_c = (R_c, E_c, A_c, z_c),$$

where R_c stores observation requests, E_c stores evidence claims, A_c retains audits, and $z_c \in \{\text{FINAL}, \text{FOLLOWUP}, \text{BLOCK}\}$ records the control state. FOLLOWUP resumes observation; FINAL and BLOCK end the episode.

Correctness compares the proposal with a held-out label. *Readiness* concerns the observations, magnifications, and differential checks required by its contract. We define *unsupported finalization* as releasing a proposal whose stored record fails to satisfy its declared evidence obligations. *Authorization* determines whether the audited record permits release under the terminal policy. Matched policies share $(R_c^0, E_c^0, \hat{y})$ and the proposal–contract binding. Follow-up appends evidence without recomputing the proposal or its initial scores.

3.2 PATHOLOGY HARNESS AND CONTRACT-BOUND RECORDS

The Harness separates observation, record assessment, and release execution. We write

$$\mathcal{H} = (\mathcal{A}, \mathcal{W}, \mathcal{S}, \delta, g),$$

where $\mathcal{A}$ specifies role adapters and permissions, $\mathcal{W}$ is the append-only workspace, $\mathcal{S} = \{S_y\}$ is the contract library, δ governs case-state transitions, and g is the deterministic P3 Terminal Authorization Gate (P3 Gate). Adapters restrict each role to its artifact class: the Scout writes requests, the Interpreter supplies observations, the Compiler writes claims, and the Auditor assesses them. The Harness executes transitions authorized by the control policy. Appendix A.1 details the permitted reads and writes for each role.

Each proposal is bound to $S_{\phi(\hat{y})}$ through a fixed mapping $\phi : \mathcal{Y}_{\text{map}} \to \mathcal{Y}_{\text{contract}}$. A contract item identifies a required feature, acceptable values, minimum magnification, and any distinguishing rule for compatible competing hypotheses. Proposal, contract, and competitors must share the declared label granularity. Each evidence claim has a stable identifier, morphology statement, image pointer, coordinates, magnification, and producing role. These fields allow an audit to associate a contract requirement with the observations used to assess it (Appendix A.2).

For example, a contract may require assessment of a tissue boundary at a specified magnification. A stored description from an insufficient magnification leaves that requirement unresolved. The Auditor records the feature and required scale as a missing item; the subsequent request asks for a view in which the boundary can be assessed. Its observed presence, absence, or indeterminate status is then retained with the source image.

The workspace preserves earlier claims as new observations arrive. At each audit, the Contract Auditor returns unresolved items $M_c^{(t)}$ and a recoverability status. Recoverable deficits supply DCE-Scout's targets in Section 3.3; the authorization procedure in Section 3.4 determines whether acquisition may proceed.

3.3 DIFFERENTIAL-CONTRACT EVIDENCE ACQUISITION

DCE-Scout selects follow-up observations after a recoverable audit objection. At round t, it receives

$$(\hat{y}, E_c, M_c^{(t)}, \mathcal{Y}_c^-, \mathcal{P}_c, B),$$

where $\mathcal{Y}_c^-$ contains compatible competing hypotheses, $\mathcal{P}_c$ is the candidate ROI–magnification pool, and $B \geq 0$ is the remaining integer request-count cap. Competitors come from the Frozen Proposal Generator or the declared contract hierarchy at the proposal's label granularity.

Each missing item becomes an observability target specifying the feature, admissible outcomes, required magnification, and permitted region cues. In the boundary example, this target guides selection toward an assessable field. Observations may support or contradict the proposal, or leave the item unresolved; all enter the record. To account for competing diagnoses, DCE-Scout also forms

$$D_c^{(t)} = \bigcup_{y \in \mathcal{Y}_c^-} \mathrm{Diff}(\hat{y}, y),$$

where Diff returns contract features that distinguish the proposal from a compatible alternative. Missing-item coverage targets gaps in the current record, while differential relevance favors views that help assess these distinctions.

Nearby fields can cover the same feature with little additional information. We therefore select requests jointly under the remaining cap:

$$\mathcal{Q}_c^\star = \arg \max_{\substack{\mathcal{Q} \subseteq \mathcal{P}_c \\ |\mathcal{Q}| \leq B}} U(\mathcal{Q}),$$

where

$$\begin{aligned} U(\mathcal{Q}) = & \sum_{f \in M_c^{(t)}} w_f \max_{(r,m) \in \mathcal{Q}} o_f(r, m) \\ & + \alpha \sum_{d \in D_c^{(t)}} \lambda_d \max_{(r,m) \in \mathcal{Q}} q_d(r, m) \\ & + \beta \, \mathrm{Div}(\mathcal{Q}; E_c) - \gamma \, \mathrm{Cost}(\mathcal{Q}). \end{aligned}$$

Here, $o_f(r, m)$ measures observability of missing feature f in region r at magnification m, and $q_d(r, m)$ measures relevance to differential obligation d. The maxima credit the best selected view for each obligation, limiting redundant coverage. Diversity accounts for both stored observations and selected requests. Cost penalizes magnification and model-call expenditure, while B limits request count. All weights are nonnegative, and heterogeneous expert scores are normalized within each WSI before fusion. The maxima, diversity, and cost are zero for an empty selection.

Selected requests are appended to R_c. The Morphology Interpreter observes their crops, and the Evidence Compiler appends the resulting claims to E_c for renewed audit under the same contract. DCE-Scout writes requests and selection provenance; interpretation and authorization remain with their respective components. Appendix A.8 specifies target and trace fields.

3.4 Contract Audit and Terminal Authorization

Before release, the Contract Auditor assesses the stored claims against $S_{\phi(\hat{y})}$, citing the observations used and identifying missing features, insufficient magnification, unresolved differentials, or conflicting claims. Given $(\hat{y}, S_{\phi(\hat{y})}, E_c)$, it returns

$$A_c^{(t)} = (v_c^{(t)}, C_c^{(t)}, \tau_c^{(t)}, M_c^{(t)}, \kappa_c^{(t)}),$$

comprising a verdict, cited claim IDs, objection type, typed missing-item set, and recoverability status. The latter is recoverable, unrecoverable, or unspecified. An unspecified status permits follow-up only if a frozen deterministic policy rule resolves it as recoverable; below, $\bar{\kappa}_c^{(t)}$ denotes the status after this policy resolution.

The P3 Gate checks whether the retained artifacts authorize release. Its predicate groups four checks from Appendix A.9:

$$[g_c^{(t)} = \mathrm{pass}] \equiv G_{\mathrm{artifact}} \wedge G_{\mathrm{reference}} \wedge G_{\mathrm{audit}} \wedge G_{\mathrm{policy}}.$$

These check schemas, identifiers, and immutable binding; resolvable references with valid coordinates and magnifications; consistency of the verdict, cited claims, deficits, and blocking signals; and admissible transitions under the policy and budgets, respectively. Each predicate is evaluated

Table 1: Full case-management outcomes for Open Signout (M3) and Evidence Repair (M4-Full), which share frozen proposals and initial records within each dataset. Acc. denotes selective accuracy on finalized cases; Risk denotes the fraction of finalized proposals incorrect against held-out labels.

Dataset	Policy	Finalized	Cov.	Acc.	Wrong	Risk
BRACS	Open Signout (M3)	387/387	1.000	0.444	215	0.556
	Evidence Repair (M4-Full)	206/387	0.532	**0.811**	39	**0.189**
EBRAINS	Open Signout (M3)	2335/2335	1.000	0.719	657	0.281
	Evidence Repair (M4-Full)	936/2335	0.401	**0.890**	103	**0.110**

on stored artifacts and policy state. The Auditor judges claim content against the contract, while P3 checks the retained authorization basis. Release is controlled at the execution boundary, following AgentSpec's runtime-enforcement principle (Wang et al., 2025).

An incomplete record may still permit further observation. Let $h_c^{(t)}$ indicate that the artifacts and audit are structurally valid for follow-up consideration. With t completed rounds, b_c acquired follow-up requests, and fixed caps $R_{\max}, B_{\max}$, eligibility is

$$\chi_c^{(t)} \equiv [\bar{\kappa}_c^{(t)} = \text{recoverable}] \wedge [M_c^{(t)} \neq \emptyset] \wedge h_c^{(t)} \\ \wedge [t < R_{\max}] \wedge [b_c < B_{\max}].$$

The remaining cap passed to DCE-Scout is $B = B_{\max} - b_c$. Defining $F_c^{(t)} \equiv [v_c^{(t)} = \text{pass}] \wedge [g_c^{(t)} = \text{pass}]$, the Harness applies

$$z_c^{(t)} = \begin{cases} \textsc{Final}, & F_c^{(t)}, \\ \textsc{Followup}, & \neg F_c^{(t)} \wedge \chi_c^{(t)}, \\ \textsc{Block}, & \text{otherwise}. \end{cases}$$

Follow-up invokes DCE-Scout, appends observations, increments t by one and b_c by $|\mathcal{Q}_c|$, and repeats the audit and Gate checks. An empty selection blocks the case. Unrecoverable objections, blocking contradictions, or unresolved deficits without remaining budget also yield BLOCK. Appendix A.10 gives the loop and its no-follow-up specialization. The terminal package retains references, policy configuration, counters, and P3 results for deterministic Gate replay; repeating component inference additionally requires its models, prompts, and case artifacts.

4 EXPERIMENTS

Evaluation scope. We assess whether recorded observations support releasing a fixed diagnostic proposal and how further observation affects finalization counts and errors. Main policy comparisons use BRACS and EBRAINS. Upstream classification across six benchmarks (BRACS, CAMELYON16, CAMELYON17, EBRAINS, NSCLC, and RCC) provides context for the proposals (Figure 3; Appendix E); external results retain their original evaluation protocols. Appendix F compares thumbnail and multi-region evidence sources and evaluates initial navigation on CAMELYON16/17.

4.1 EXPERIMENTAL SETUP

We compare three policies. Direct release of the initial proposal is Open Signout (M3). Checking existing evidence before deciding whether to release is Evidence Audit (M4-NoFU). Acquiring observations to address missing evidence and then checking again is Evidence Repair (M4-Full). All three start from the same initial crops, observations, evidence records, and diagnostic proposals. Evidence Audit uses the Contract Auditor and P3 Gate to assess the initial record and authorize release. Evidence Repair adds DCE-Scout follow-up after recoverable objections. Proposal generation and evidence compilation follow the two paths defined in Section 3.1. Follow-up updates the evidence and release decision while keeping the proposal and initial confidence scores fixed.

The finalization evaluation uses 387 BRACS cases with three-family labels and 2,335 EBRAINS cases with 31-class diagnostic labels (Brancati et al., 2022; Roetzer-Pejrimovsky et al., 2022; Ding et al., 2025). Coverage is the finalized fraction of the evaluated cohort. Selective accuracy and selective prediction risk are the correct and incorrect fractions of finalized proposals, respectively, against

Table 2: BRACS policies sharing the initial record and frozen proposal. M4-Full alone may append evidence before re-audit. Risk is the held-out-label error fraction among finalized proposals. Gate-bypass counts finalizations outside the joint Contract Auditor and P3 Gate path.

Metric	Group	Open Signout (M3)	Conf.	Schema	Contract Auditor	Evidence Audit (M4-NoFU)	Evidence Repair (M4-Full)
Finalized (of cohort)	*Release set*	387/387	76/387	387/387	152/387	64/387	206/387
Coverage		1.000	0.196	1.000	0.393	0.165	0.532
Net increase in finalizations vs. M4-NoFU		-	-	-	-	-	142
Correct finalized	*Finalized-case outcomes*	172	54	172	132	64	167
Selective accuracy		0.444	0.711	0.444	0.868	1.000	**0.811**
Risk		0.556	0.289	0.556	0.132	0.000	**0.189**
Wrong kept	*Blocking behavior*	215	22	215	20	0	39
Correctly blocked wrong		0	193	0	195	215	176
Correct prediction blocked		0	118	0	40	108	5
Gate-bypass finalizations	*Release authorization*	387	76	387	152	0	0

held-out labels. We also count blocked incorrect and withheld correct proposals. Appendix B.3 defines cohorts and metrics; Appendix C.4 details the datasets (Ehteshami Bejnordi et al., 2017; Bandi et al., 2019; Coudray et al., 2018; The Cancer Genome Atlas Research Network, 2014; 2012; 2013; 2016; Davis et al., 2014; The Cancer Genome Atlas Network, 2012).

4.2 Case-Closure Outcomes

Table 1 shows that Evidence Repair releases fewer cases than Open Signout and reduces the number and fraction of incorrect finalized proposals on both datasets. On BRACS, it finalizes 206 of 387 cases, including 39 incorrect proposals versus 215 when all cases are released. Risk falls from 0.556 to 0.189. On EBRAINS, it finalizes 936 of 2,335 cases, including 103 incorrect proposals, compared with 657 under unconditional release; risk falls from 0.281 to 0.110.

Withholding differs between cohorts. On BRACS, Evidence Repair blocks 176 incorrect proposals and five correct proposals. On EBRAINS, it blocks 554 incorrect and 845 correct proposals. Thus, lower error rates among finalized cases come with substantially more correct proposals withheld on EBRAINS. Family-level BRACS results and EBRAINS details are in Appendices C.2 and C.4.

4.3 Authorization and Follow-up Acquisition

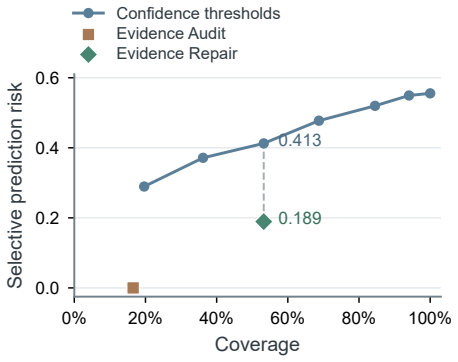
Fixed-record authorization. We first vary how the initial BRACS record is checked before release (Table 2). Schema validation checks record format and leaves all 387 proposals released. Contract Auditor review checks the recorded evidence against the contract and finalizes 152 cases, including 20 incorrect proposals. Adding the P3 Gate's structural and authorization checks gives Evidence Audit, which finalizes 64 cases (16.5% coverage). No held-out-label errors are observed among these 64 finalizations. This policy withholds all 215 incorrect proposals and 108 correct proposals, showing that even a correct proposal can lack a record that meets the release requirements.

Audit-guided acquisition. We next allow further observation with the same audit and authorization checks. Evidence Repair raises finalizations from 64 to 206, a net gain of 142. Coverage rises from 16.5% to 53.2%, and correct proposals withheld decrease from 108 to five. The larger release set also includes 39 incorrect proposals, compared with none among the 64 initial-record finalizations. Follow-up thus allows more cases to meet release requirements while increasing the observed error fraction among finalized cases. Appendix B.1 defines each comparison policy.

4.4 Evidence-Based Selection and Confidence

We compare selecting proposals by their evidence requirements with selecting them by initial confidence. Confidence selection uses initial prediction scores, whereas Evidence Repair combines contract audit with additional observations. At the same coverage of 53.2%, both policies finalize 206

Table 3: Risk–coverage comparison on the shared 387-case BRACS cohort. The line joins seven measured confidence thresholds; Audit and Repair are individual operating points. N denotes finalized cases, Cov. is $N/387$, and Risk is Wrong/N.



Policy	Thr.	N	Cov.	Wrong	Risk
Conf.	0.00	387	1.000	215	0.556
Conf.	0.55	364	0.941	200	0.549
Conf.	0.60	327	0.845	170	0.520
Conf.	0.65	266	0.687	127	0.477
Conf.	0.70	206	0.532	85	0.413
Conf.	0.75	140	0.362	52	0.371
Conf.	0.80	76	0.196	22	0.289
Audit	-	64	0.165	0	0.000
Repair	-	206	0.532	39	0.189

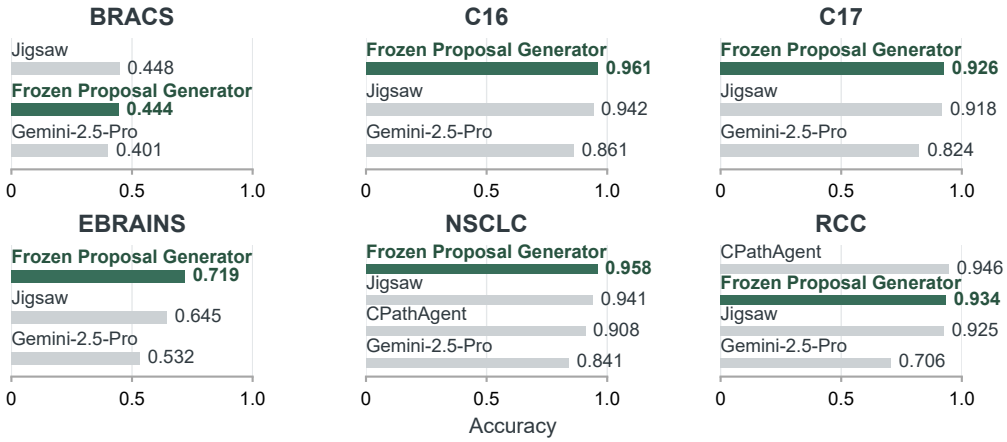


Figure 3: Upstream accuracy on six pathology benchmarks (descending per panel); Frozen Proposal Generator highlighted. External protocols unchanged. Full results: Appendix E, Table 8.

BRACS cases, but Evidence Repair releases 39 incorrect proposals compared with 85 at confidence threshold 0.70 (Table 3). Their risks are 0.189 and 0.413, respectively.

Across the seven reported confidence thresholds, raising the threshold from 0.00 to 0.80 reduces coverage from 100% to 19.6% and risk from 0.556 to 0.289. The source-specific M1 and M2 sweeps appear in Appendix C.3.

4.5 Family-Level Release Behavior

Evidence Repair releases malignant cases most frequently, with 76.1% coverage and risk 0.129, compared with 43.6% coverage and risk 0.262 for benign non-atypical cases (Table 4). BRACS groups follow Appendix B.3: benign non-atypical includes Normal, Pathological Benign, and UDH; atypical borderline includes FEA and ADH; malignant includes DCIS and Invasive Carcinoma.

The atypical borderline group has both the lowest coverage (22.7%) and the highest residual risk (0.353). In that group, 57 incorrect proposals and one correct proposal are blocked, while six incorrect proposals remain among the 17 finalizations.

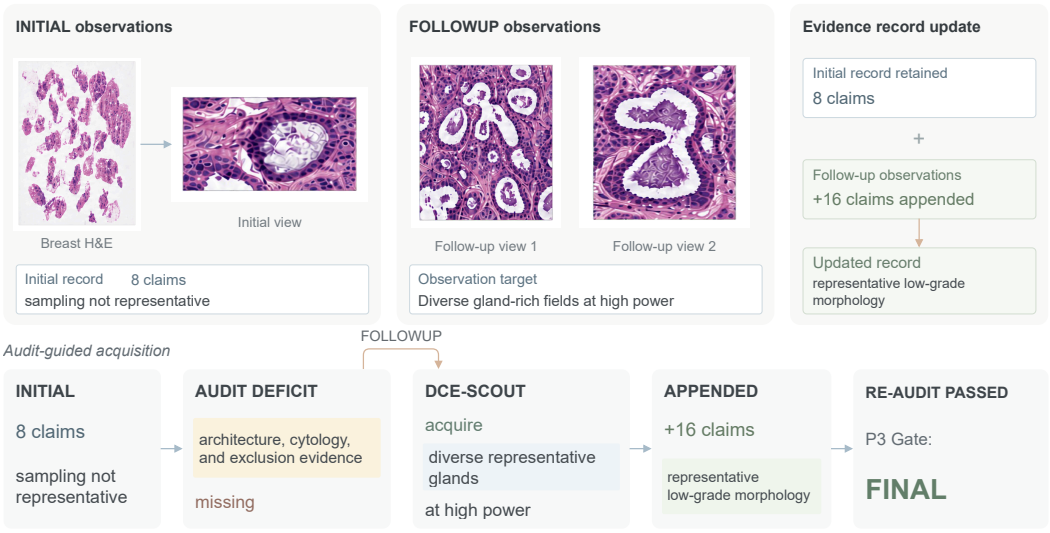


Figure 4: Audit-guided acquisition on a breast H&E case.

Table 4: Family-level Evidence Repair outcomes on BRACS. Cov. is N/Cases; Risk is the error fraction among finalized cases. Wrong block and Corr. block count blocked incorrect and correct proposals. Full counts appear in Appendix C.2.

BRACS family	Cases	N	Cov.	Risk	Wrong block	Corr. block
Benign non-atypical	149	65	0.436	0.262	82	2
Atypical borderline	75	17	0.227	0.353	57	1
Malignant	163	124	0.761	0.129	37	2
Overall	387	206	0.532	0.189	176	5

4.6 Qualitative Case Study

Figure 4 shows how further observation addresses missing evidence in a breast H&E case. The initial high-power view is dominated by a small, partial gland profile and does not establish representative architecture, cytology, or exclusion evidence. DCE-Scout requests diverse gland-rich fields at high power. These views broaden architectural sampling and document relatively uniform nuclei with largely preserved polarity, without conspicuous solid or cribriform growth or irregular infiltrative arrangement. On re-audit, the added observations satisfy the missing family-level requirements, allowing finalization of the unchanged atypical borderline proposal. Full panels and a complementary evidence-complete case appear in Appendix D.

5 Conclusion

In this work, we introduced Pathology Harness, a framework that helps pathology agents determine whether the observations they have collected are sufficient to finalize a proposed diagnosis. It identifies missing evidence, guides further inspection of relevant tissue regions at appropriate magnifications, and reassesses the findings before deciding whether to finalize the diagnosis. The system retains both supporting and contradictory observations, making the basis for each decision available for inspection. We conducted evaluations across multiple pathology datasets, covering diagnostic prediction, regional evidence collection, and diagnostic finalization. The results show that combining explicit evidence checks with targeted follow-up reduces errors among finalized diagnoses compared with accepting every prediction, at the cost of finalizing fewer cases. Further observation also enables more cases to satisfy the evidence requirements than checking the initial observations alone. These findings highlight the value of coordinating evidence gathering with diagnostic finalization and point to a new direction for pathology agents.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript to improve readability and to assist with LaTeX typesetting and formatting. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Peter Bandi, Oscar Geessink, Quirine Manson, Marcory van Dijk, Maschenka Balkenhol, Meyke Hermsen, et al. From detection of individual metastases to classification of lymph node status at the patient level: The CAMELYON17 challenge. *IEEE Transactions on Medical Imaging*, 38(2): 550–560, 2019. doi: 10.1109/TMI.2018.2867350.

Nadia Brancati, Anna Maria Anniciello, Pushpak Pati, Daniel Riccio, Giosuè Scognamiglio, Guillaume Jaume, Giuseppe De Pietro, Maurizio Di Bonito, Antonio Foncubierta, Gerardo Botti, Maria Gabrani, Florinda Feroce, and Maria Frucci. BRACS: A dataset for breast carcinoma subtyping in H&E histology images. *Database: The Journal of Biological Databases and Curation*, 2022:baac093, 2022. doi: 10.1093/database/baac093.

Richard J. Chen, Ming Y. Lu, Muhammad Shaban, Chengkuan Chen, Tiffany Y. Chen, Drew F. K. Williamson, and Faisal Mahmood. Whole slide images are 2d point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, volume 12908 of *Lecture Notes in Computer Science*, pp. 339–349. Springer, 2021. doi: 10.1007/978-3-030-87237-3_33.

Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16144–16155, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Chen_Scaling_Vision_Transformers_to_Gigapixel_Images_via_Hierarchical_Self-Supervised_Learning_CVPR_2022_paper.html.

Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, 2024. doi: 10.1038/s41591-024-02857-3.

Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L. Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018. doi: 10.1038/s41591-018-0177-5.

Caleb F. Davis, Christopher J. Ricketts, Min Wang, Lixing Yang, Andrew D. Cherniack, Hui Shen, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. *Cancer Cell*, 26(3): 319–330, 2014. doi: 10.1016/j.ccr.2014.07.014.

Tong Ding, Sophia J. Wagner, Andrew H. Song, Richard J. Chen, et al. A multimodal whole-slide foundation model for pathology. *Nature Medicine*, 31(11):3749–3761, 2025. doi: 10.1038/s41591-025-03982-3.

Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes van Diest, Bram van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen A. W. M. van der Laak, and the CAMELYON16 Consortium. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *Journal of the American Medical Association (JAMA)*, 318(22):2199–2210, 2017. doi: 10.1001/jama.2017.14585.

Joann G. Elmore, Gary M. Longton, Patricia A. Carney, Berta M. Geller, Tracy Onega, Anna N. A. Tosteson, Heidi D. Nelson, Margaret S. Pepe, Kimberly H. Allison, Stuart J. Schnitt, Frances P. O'Malley, and Donald L. Weaver. Diagnostic concordance among pathologists interpreting breast biopsy specimens. *Journal of the American Medical Association (JAMA)*, 313(11):1122–1132, 2015. doi: 10.1001/jama.2015.1405.

Robert Engelmore and Anthony Morgan (eds.). *Blackboard Systems*. Addison-Wesley, Reading, Mass., 1986.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, 2019.

Fatemeh Ghezloo, Mehmet Saygin Seyfioglu, Rustin Soraki, Wisdom O. Ikezogwo, Beibin Li, Tejoram Vivekanandan, Joann G. Elmore, Ranjay Krishna, and Linda Shapiro. PathFinder: A multi-modal multi-agent system for medical diagnostic decision-making applied to histopathology. *arXiv preprint arXiv:2502.08916*, 2025.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1610.02136.

Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J. Montine, and James Zou. A visual-language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29:2307–2316, 2023. doi: 10.1038/s41591-023-02504-3.

Wisdom Ikezogwo, Saygin Seyfioglu, Fatemeh Ghezloo, Dylan Geva, Fatwir Sheikh Mohammed, Pavan Kumar Anand, Ranjay Krishna, and Linda G. Shapiro. Quilt-1M: One million image-text pairs for histopathology. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/775ec578876fa6812c062644964b9870-Abstract-Datasets_and_Benchmarks.html.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 2127–2136. PMLR, 2018.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14318–14328, 2021.

Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen A. W. M. van der Laak, Bram van Ginneken, and Clara I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42: 60–88, 2017. doi: 10.1016/j.media.2017.07.005.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021. doi: 10.1038/s41551-020-00682-w.

Ming Y. Lu, Bowen Chen, Andrew Zhang, Drew F. K. Williamson, Richard J. Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19764–19775, 2023.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024a. doi: 10.1038/s41591-024-02856-4.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K. Chow, Kenji Ikemura, Ahrong Kim, Pouria Pouli, Ankush Patel, Aiman Soliman, Cynthia Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, and Faisal Mahmood. A multimodal generative AI copilot for human pathology. *Nature*, 634:466–473, 2024b. doi: 10.1038/s41586-024-07618-3.

Xinheng Lyu, Yuci Liang, Wenting Chen, Meidan Ding, Jiaqi Yang, Guolin Huang, Daokun Zhang, Xiangjian He, and Linlin Shen. WSI-Agents: A collaborative multi-agent system for multi-modal whole slide image analysis. *arXiv preprint arXiv:2507.14680*, 2025.

David Madras, Toni Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://papers.neurips.cc/paper_files/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087. PMLR, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Albus W. Ng, Yi Han, Jusheng Zhang, and Wenhao Wang. Agent Safety Should Be a Runtime Contract, 2026. URL https://arxiv.org/abs/2608.11274.

Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, Baran Atli, Clemens Christian Vogel, et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1):55, 2022. doi: 10.1038/s41597-022-01157-0.

Mehmet Saygin Seyfioglu, Wisdom O. Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-llava: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 2136–2147, 2021.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Kai Zhang, Zhongyi Shui, Bowen Ding, Tao Lin, and Lin Yang. CPathAgent: An agent-based foundation model for interpretable high-resolution pathology image analysis mimicking pathologists' diagnostic logic. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2505.20510.

The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature*, 489(7417):519–525, 2012. doi: 10.1038/nature11404.

The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of clear cell renal cell carcinoma. *Nature*, 499(7456):43–49, 2013. doi: 10.1038/nature12222.

The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–550, 2014. doi: 10.1038/nature13385.

The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of papillary renal-cell carcinoma. *New England Journal of Medicine*, 374(2):135–145, 2016. doi: 10.1056/NEJMoa1505917.

Jeroen van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: The path to the clinic. *Nature Medicine*, 27(5):775–784, 2021. doi: 10.1038/s41591-021-01343-4.

Eugene Vorontsov, Alican Bozkurt, Adam Casson, George Shaikovski, Michal Zelechowski, Kristen Severson, Eric Zimmermann, James Hall, Neil Tenenholtz, Nicolo Fusi, Ellen Yang, Philippe Mathieu, Alexander van Eck, Donghun Lee, Julian Viret, Eric Robert, Yi Kan Wang, Jeremy D. Kunz, Matthew C. H. Lee, Jan H. Bernhard, Ran A. Godrich, Gerard Oakley, Ewan Millar, Matthew Hanna, Juan Retamero, William A. Moye, Razik Yousfi, Christopher Kanan, David S. Klimstra, Brandon Rothrock, and Thomas J. Fuchs. A foundation model for clinical-grade computational pathology and rare cancer detection. *Nature Medicine*, 30(10):2924–2935, 2024. doi: 10.1038/s41591-024-03141-0.

Haoyu Wang, Christopher M. Poskitt, and Jun Sun. AgentSpec: Customizable Runtime Enforcement for Safe and Reliable LLM Agents, 2025. URL https://arxiv.org/abs/2503.18666.

Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

Xiyue Wang, Junhan Zhao, Eliana Marostica, Wei Yuan, et al. A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature*, 634(8035):970–978, 2024. doi: 10.1038/s41586-024-07894-z.

Luca L. Weishaupt, Chengkuan Chen, Drew F. K. Williamson, Richard J. Chen, Guillaume Jaume, Tong Ding, Bowen Chen, Anurag Vaidya, Long Phi Le, Ming Y. Lu, and Faisal Mahmood. Evidence-based diagnostic reasoning with multi-agent copilot for human pathology. *arXiv preprint arXiv:2506.20964*, 2025.

Bryan Wong, Xun Xu, Huazhu Fu, Nancy F. Chen, and Mun Yong Yi. Beyond Relevance: Bayesian Evidence Acquisition for Agentic Whole-Slide Image Reasoning, 2026. URL https://arxiv.org/abs/2608.05757.

Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, 2024. doi: 10.1038/s41586-024-07441-w.

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 (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E. Coupland, and Yalin Zheng. DTFD-MIL: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18780–18790, 2022. doi: 10.1109/CVPR52688.2022.01824.

APPENDIX OVERVIEW

Organization. The appendix provides supporting experiments and cases, policy definitions, evaluation protocols, implementation details, and additional limitations.

The appendix provides upstream classification context, initial spatial acquisition results, qualitative cases, and family-level terminal outcomes.

A Full Implementation Details . . . 15
- A.1 Harness Interfaces and Invariants . . . 15
- A.2 Skill Contract Interface and Label-Granularity Boundary . . . 16
- A.3 Initial Evidence Construction and Proposal Binding . . . 16
- A.4 Morphology Interpreter Implementation . . . 16
- A.5 Evidence Compiler Implementation . . . 17
- A.6 Frozen Proposal Generator Implementation . . . 17
- A.7 Contract Auditor Implementation . . . 17
- A.8 DCE-Scout Implementation . . . 17
- A.9 P3 Gate and Replay . . . 18
- A.10 Finite Harness State Machine . . . 19

B Evaluation Protocols and Matched Comparisons . . . 19
- B.1 Control Policies and Ablation Factors . . . 19
- B.2 Experimental Design Rationale . . . 20
- B.3 Protocol Details and Metric Definitions . . . 20
- B.4 Implementation Checks . . . 21
- B.5 Interpreting the Comparisons . . . 21

C Additional Terminal-Control and Dataset Results . . . 21
- C.1 Release Accounting . . . 21
- C.2 Family-Level Terminal Outcomes . . . 21
- C.3 Full Confidence Threshold Results . . . 21
- C.4 NSCLC, RCC, BRCA-PAM50, and EBRAINS Details . . . 23

D Qualitative Case Studies . . . 23

E Frozen Proposal Substrate . . . 26

F Spatial Evidence Acquisition . . . 28
- F.1 Navigation Quality and Protocol . . . 28
- F.2 CAMELYON17 Testing Details . . . 28

G Extended Discussion and Limitations . . . 29
- G.1 Extended Discussion . . . 29
- G.2 Scope and Limitations . . . 29
- G.3 Additional Limitations . . . 29

A Full Implementation Details

A.1 Harness Interfaces and Invariants

This section records the reusable Harness substrate. Its five functions are interface control, state externalization, proposal–contract execution, terminal-state transition, and artifact retention for deterministic P3 Gate reruns and paired policy evaluation. The append-only workspace stores the initial request ledger, observations and structured Evidence Chain claims, the frozen proposal–contract binding, audit history, follow-up selection trace, policy counters, and the terminal package.

Each role reads only its declared inputs and may write only its declared artifact class. The permissions follow the case workflow.

Role	Interface specification
Initial evidence construction.	
Evidence Scout	The Evidence Scout reads (I, π) and the frozen initial-scanning configuration and writes initial ROI–magnification requests.
Morphology Interpreter	The Morphology Interpreter reads approved requests and their WSI crops and writes morphology observations.
Evidence Compiler	The Evidence Compiler reads these observations and the active schema and writes structured evidence claims.
Frozen proposal generation.	
Frozen Proposal Generator	The evaluated Frozen Proposal Generator reads CONCH v1.5 features from the shared initially inspected regions and writes the frozen proposal and its provenance. This input specification describes the evaluated feature-based adapter (Appendix A.6).
Audit and follow-up acquisition.	
Contract Auditor	The Contract Auditor reads the fixed proposal, bound contract, and stored claims and writes an audit verdict and typed deficit fields.
DCE-Scout	DCE-Scout reads the fixed proposal, bound contract, audit deficit, stored metadata, candidate pool, and remaining budget and writes follow-up requests and selection provenance.
Terminal authorization and control.	
P3 Gate	The P3 Gate reads stored artifacts, audit output, and frozen policy configuration and writes a deterministic authorization result and failure codes.
Harness control	Harness control reads the policy state and outputs from declared roles and executes a FINAL, FOLLOWUP, or BLOCK transition.

The adapter boundary prevents a selector from creating morphology, the Contract Auditor from relabeling or finalizing, and the P3 Gate from interpreting morphology. The evaluation additionally enforces four invariants:

proposal invariance fixes $(R_c^0, E_c^0, \hat{y})$ before matched policies diverge;

record-grounded authorization restricts release to stored claims, the bound contract, audit output, and P3 Gate result;

append-only evidence preserves earlier events; and

role non-interference enforces the declared role permissions.

Open Signout (M3), Evidence Audit (M4-NoFU), and Evidence Repair (M4-Full) are management policies over this same substrate. M3 directly releases the fixed proposal, M4-NoFU audits and authorizes the fixed first-pass record, and M4-Full may append DCE-Scout-selected evidence after a recoverable objection.

A.2 Skill Contract Interface and Label-Granularity Boundary

The Skill interface binds each frozen proposal to evidence obligations at the same declared label granularity:

```
proposal_label: <frozen_proposal_label>
label_granularity: <family_or_diagnosis>
contract_id: <versioned_contract_id>
contract_version_or_hash: <frozen_version_or_hash>
required_features:
  - feature_id: <stable_feature_identifier>
    feature_description: <required_observation>
    min_magnification: <minimum_scale>
    acceptable_values: [<admissible_value>, ...]
    observability_target: <answer_neutral_observation_target>
    distinguishing_rule: <proposal_vs_hypothesis_rule>
    distinguishing_from: [<compatible_hypothesis>, ...]
    region_cue: <optional_case_local_region_cue>
```

Before a contract is selected, the proposal label, each competing hypothesis, and the selected Skill contract must be compatible at the declared granularity. DCE-Scout may use compatible differential hypotheses to choose where to inspect.

DCE-Scout repair trace. The request-trace schema and frozen run-manifest fields are specified in Appendix A.8.

A.3 Initial Evidence Construction and Proposal Binding

Initial evidence triage occurs before proposal generation. The Evidence Scout performs proposal-agnostic spatial triage over (I, π) and writes descriptors

$$r_i = (b_i, m_i, u_i)$$

to R_c^0, where b_i is the level-0 spatial support, m_i the requested magnification, and u_i an answer-neutral observation target. The current first-pass candidates may be generated by random tissue sampling at $10\times$ after Otsu thumbnail tissue detection, M2 CPathAgent-style global screening at $1.25\times$ followed by planning at $5\times$ and inspection at $10\times$ and $20\times$, or CONCH v1.5 Scout-A similarity against 128 region templates with the top five regions retained at $10\times$.

The Morphology Interpreter converts each requested crop into an observation. The Evidence Compiler then applies its schema-constrained projection and writes a structured claim to E_c^0. A claim records a stable identifier, morphology statement, image pointer, magnification, level-0 coordinates, and producing role. In a separate computation over the same initial crops, CONCH v1.5 image features feed a separately trained, frozen logistic-regression proposal generator to produce $\hat{y}$ in the instantiation described in Appendix A.6. The Harness then freezes the compatible proposal–contract binding. Every matched policy begins from the same R_c^0, E_c^0, and $\hat{y}$.

A.4 Morphology Interpreter Implementation

The Morphology Interpreter module uses Patho-R1, a pathology-fine-tuned 7B-parameter VLM based on a LLaVA-style architecture. In this paper, Patho-R1 is used in frozen inference mode. For each candidate region, the Morphology Interpreter crops the WSI at the specified magnification ($10\times$ for architectural assessment, $20\times$ for nuclear detail, $40\times$ for myoepithelial assessment and comedonecrosis detection) and queries the VLM with structured prompts. The shared initial pass uses proposal-agnostic observation targets, whereas post-objection follow-up prompts target features from the bound proposal-compatible contract. Each prompt asks for four fields: (a) presence/absence, (b) grade/severity if present, (c) confidence in the observation, and (d) a brief morphological description.

A.5 Evidence Compiler Implementation

The Evidence Compiler receives the Morphology Interpreter's free-text observations and a JSON schema. The shared initial pass uses a proposal-agnostic schema; after proposal binding, follow-up uses the bound proposal-compatible contract. It extracts structured claims through regex-based field extraction and LLM-based semantic parsing with schema-constrained output. Each claim is validated for field completeness, type correctness, and coordinate validity before insertion into the Evidence_Chain.

A.6 Frozen Proposal Generator Implementation

The Frozen Proposal Generator is a frozen logistic regression classifier trained on CONCH v1.5 image features extracted from the shared initially inspected regions. The initial proposal and the scores used for confidence-based selection remain fixed during terminal management; follow-up regions update the evidence record without being used to recompute either output. For BRACS, it maps to 3 families (benign_non_atypical, atypical_borderline, malignant). For CAMELYON, it maps to 2 classes (Normal, Tumor). The Frozen Proposal Generator is the only component in the Harness stack fit within our experimental setup. After fitting, the Frozen Proposal Generator is frozen before any selective-finalization experiments.

A.7 Contract Auditor Implementation

The Contract Auditor receives the proposal, bound contract, and stored Evidence_Chain, then emits a structured verdict. Its interface separates record assessment from image interpretation and terminal-state execution. A general verifier, rule-constrained LLM, or deterministic checker can implement this interface when compatible with the contract language.

The structured `Audit_Verdict` exposes the following fields to the state machine and follow-up selector:

Field	Meaning
`missing_feature`:	the contract feature identifier and, where needed, its required acceptable value;
`required_magnification`:	the minimum magnification still needed for that feature;
`differential_obligation`:	a compatible competing hypothesis or distinguishing rule that the next observation must address;
`recoverable`:	whether additional acquisition is permitted under the frozen policy and why;
`cited_claim_ids`:	stored claims considered during the verdict; and
`objection_type`:	one of `missing_observation`, `insufficient_magnification`, `unresolved_differential`, `conflicting_claims`, `unresolvable_reference`, or `unrecoverable`.

A recoverable typed objection supplies the missing-item targets for DCE-Scout. The audit concerns stored claims; morphology is supplied by the Interpreter and release is authorized through P3.

A.8 DCE-Scout Implementation

Repair-stage inputs and targets. DCE-Scout is invoked only after a typed, recoverable Contract Auditor objection. It receives the fixed proposal and bound contract identity/version, typed missing-feature and differential obligations, cited claims and stored region metadata, a case-local or pre-frozen candidate pool, and the remaining ROI–magnification request-count cap. Magnification and model-call expenditure are represented by the DCE-Scout cost term. Each missing item is rendered as an answer-neutral observability target:

```
target_type: observability | differential
feature_id: <stable_feature_identifier>
admissible_outcomes: [<permitted_outcome>, ...]
```

```
required_magnification: <minimum_scale>
permitted_region_cues: <optional_case_local_cues>
```

For example, a target may ask to find ductal structures in which the peripheral myoepithelial boundary can be reliably assessed at $40\times$. Each target permits supporting, contradictory, or unresolved outcomes.

Candidate pool, utilities, and selection. The candidate pool is case-local or frozen before evaluation. The selector combines contract-semantic observability, morphology or feature-space novelty, differential relevance, spatial or feature diversity, and magnification or model-call cost under the non-redundant set objective in Section 3.3. The remaining ROI–magnification request-count cap limits request cardinality, whereas magnification and model-call expenditure enter the cost term. Diversity is computed relative to both existing evidence and the selected set. Heterogeneous expert scores are normalized within each WSI before fusion. The solver, tie handling, budget, and hyperparameters are frozen policy configuration.

By convention, $\max_{(r,m)\in\emptyset} o_f(r,m) = \max_{(r,m)\in\emptyset} q_d(r,m) = 0$, $\mathrm{Div}(\emptyset; E_c) = 0$, and $\mathrm{Cost}(\emptyset) = 0$, so the optimization remains defined for an empty candidate pool or $B = 0$.

Trace and boundary. Each selected request appends the acquisition round, `roi_id`, level-0 coordinates, magnification, `missing_feature`, `differential_target`, component scores, marginal set gain, reason for selection, Morphology Interpreter output pointer, and resulting claim IDs. The frozen run manifest additionally binds these fields to case IDs, initial-record hashes, proposal hashes, contract hashes, initial audit IDs, candidate-pool identities, policy version, and budget state.

A.9 P3 Gate and Replay

Input boundary. The P3 Gate, implemented as the deterministic P3 predicate, operates only on retained terminal information. Its inputs comprise the proposal–contract binding, stored evidence, audit output, policy state and configuration, and follow-up trace. It applies the following checks:

- required artifact presence, schemas, unique identifiers, and immutable proposal–contract binding;
- resolution of claim, audit, ROI, and output references, including coordinate and magnification validity;
- consistency between the verdict, cited claims, missing-feature fields, differential obligations, recoverability, and blocking signals;
- policy-state consistency, including authorized follow-up caps, ROI and magnification budgets, and admissible terminal transitions; and
- a deterministic pass/block result with input digests and P3 Gate version recorded in the terminal package.

Failure information and terminal record. The resulting failure codes identify absent information, unresolved references, invalid coordinates or magnification, identifier or schema violations, audit inconsistencies, blocking signals, and unauthorized transitions or cap violations. The retained terminal information links each FINAL or BLOCK decision to the fixed proposal, evidence, audit output, policy state, P3 result, counters, and transition reason.

Gate replay. The retained terminal information and P3 version support deterministic Gate replay without repeating image-model inference.

Full inference replay. Full inference replay additionally requires the case artifacts, models, prompts, candidate regions, and configurations for each component.

A.10 FINITE HARNESS STATE MACHINE

Algorithm 1 gives executable pseudocode for Evidence Repair (M4-Full). Disabling the acquisition branch yields Evidence Audit (M4-NoFU); Open Signout (M3) uses the shared initial record and fixed proposal but directly releases the proposal.

Algorithm 1: Finite fixed-proposal state machine on Pathology Harness

Input: case c, policy ρ, and caps $R_{\max}, B_{\max}$
Output: terminal trace T_c

Fix initial requests and inspect crops; construct E_c^0 from morphology and separately generate $\hat{y}$ from initial image features; freeze the initial state and compatible proposal–contract binding; set $t, b_c \leftarrow 0$.
while $t \leq R_{\max}$ and $b_c \leq B_{\max}$ **do**
  Run Contract Auditor and P3 Gate to obtain $(A_c^{(t)}, M_c^{(t)}, \kappa_c^{(t)}, g_c^{(t)}, h_c^{(t)})$.
  if Contract Auditor passes and $g_c^{(t)} = \text{pass}$ **then**
    $z_c \leftarrow \text{FINAL}$; record the transition.
    break.
  else if $\chi_c^{(t)}$ **then**
    $z_c \leftarrow \text{FOLLOWUP}$; DCE-Scout selects $\mathcal{Q}_c$ under the remaining ROI–magnification request-count cap.
    if $\mathcal{Q}_c = \emptyset$ **then**
      $z_c \leftarrow \text{BLOCK}$; **break**.
    end if
    Append morphology-derived claims from $\mathcal{Q}_c$; $t \leftarrow t + 1$, $b_c \leftarrow b_c + |\mathcal{Q}_c|$.
  else
    $z_c \leftarrow \text{BLOCK}$; record the reason and **break**.
  end if
end while
if $z_c \notin \{\text{FINAL}, \text{BLOCK}\}$ **then**
  $z_c \leftarrow \text{BLOCK}$ because a finite cap was reached.
end if
return $T_c = (R_c, E_c, A_c, z_c)$.

B EVALUATION PROTOCOLS AND MATCHED COMPARISONS

This section defines the policy variants and the information shared by their matched comparisons.

B.1 CONTROL POLICIES AND ABLATION FACTORS

The policy notation separates data-plane evidence sources from management policies executed on the shared Harness core:

- **M0**: BRACS ROI feature-based baseline using CONCH v1.5 features and logistic regression. This provides ROI-scale context.
- **M1**: Single-shot baseline using one anonymized thumbnail and a structured diagnosis-confidence prompt.
- **M2**: Navigation-only baseline using CPathAgent-style slide screening and multi-scale region inspection. The Frozen Proposal Generator output is taken as final. No Harness terminal gate is applied.
- **Harness core**: Shared case state, role interfaces, append-only event persistence, proposal–contract binding, and terminal-state execution. It is the common substrate for Open Signout (M3), Evidence Audit (M4-NoFU), and Evidence Repair (M4-Full).
- **Open Signout (M3)**: Direct release of a Harness-compatible first-pass record without evidence authorization. It stores the fixed diagnostic proposal and record.
- **Evidence Audit (M4-NoFU)**: Authorization by the Contract Auditor and P3 Gate on exactly the fixed first-pass record and frozen proposal used by Open Signout (M3). It permits only `FINAL` or `BLOCK` and receives no new image observation.

- **Evidence Repair (M4-Full)**: The follow-up-enabled Harness management policy. It begins from the same first-pass record and frozen proposal, invokes DCE-Scout after a recoverable Contract Auditor objection, appends the resulting evidence, and reruns audit and authorization before terminal release.

M1 through M4 use the same Frozen Proposal Generator for the reported BRACS three-family and CAMELYON binary settings where applicable. Patient-disjoint train/validation/test partitions are fixed before terminal-control evaluation, and the Frozen Proposal Generator is trained before any release policy is scored. Open Signout (M3) versus Evidence Audit (M4-NoFU) is the authorization-only comparison. Evidence Audit (M4-NoFU) versus Evidence Repair (M4-Full) measures the added effect of DCE-Scout-guided follow-up acquisition.

The fixed-information ablations vary the predicate that converts the first-pass record and diagnostic proposal into a terminal state. **No-gate mapped** takes the Frozen Proposal Generator output as final. **Schema gate only** checks JSON schema compliance, required fields, and valid types without consulting the Contract Auditor. **Contract Auditor only** uses structured pass/block verdicts but does not enforce P3 Gate structural checks. **Contract Auditor + P3 Gate** combines contract audit with deterministic structural validation. Evidence Repair (M4-Full) is analyzed separately because it may append observations through DCE-Scout. The score-only policies use $s(c) = \max_y p(y \mid c)$ with thresholds 0.0, 0.55, 0.60, 0.65, 0.70, 0.75, and 0.80. The 0.0 threshold is equivalent to no-gate mapped finalization.

B.2 Experimental Design Rationale

Paired policies begin with identical initial requests, observations, compiled evidence, and proposal–contract bindings. This shared prefix separates upstream evidence construction from later management. NoFU uses the prefix alone; Full can extend it after a recoverable objection. Label-based outcomes are computed after terminal decisions, allowing the release set to change while every case retains its original proposal.

B.3 Protocol Details and Metric Definitions

BRACS and CAMELYON16 cohorts. BRACS comprises 387 WSIs from 189 patients and spans seven lesion classes grouped into three families: benign_non_atypical (Normal, Pathological Benign, and UDH), atypical_borderline (FEA and ADH), and malignant (DCIS and Invasive Carcinoma) (Brancati et al., 2022). It is used for selective-finalization and ROI analyses. CAMELYON16 contains 399 WSIs (159 normal and 240 tumor-containing); the primary paired navigation analysis uses 48 XML-annotated tumor slides and 80 normal slides from its testing split (Ehteshami Bejnordi et al., 2017).

The primary Harness-managed evaluation uses the 387-case BRACS ROI-bearing WSI cohort and the 2,335-WSI EBRAINS cohort. The shared Harness core persists the initial observations, `Evidence_Chain`, frozen diagnostic proposal, audit record, and terminal state. Open Signout (M3) directly signs out that Harness-compatible first-pass record. Evidence Audit (M4-NoFU) applies Contract Auditor and P3 Gate to the same fixed record. Evidence Repair (M4-Full) invokes DCE-Scout after a recoverable Contract Auditor objection, appends the resulting observations, and reruns Contract Auditor and P3 Gate on the updated record.

Selective-finalization metrics are defined as follows:

- **Coverage**: fraction of cases finalized under a stated management policy.
- **Selective accuracy**: accuracy on finalized cases only.
- **Selective prediction risk (Risk)**: incorrect finalized proposals divided by finalized cases, evaluated against held-out labels.
- **Wrong kept**: number of incorrect predictions that are finalized.
- **Correctly blocked**: number of wrong source predictions blocked by the gate.
- **Correct prediction blocked**: number of correct frozen proposals blocked under the management policy.

- **Gate-bypass finalization**: number of finalized cases released without terminal authorization from the full Contract Auditor and P3 Gate path.
- **Net increase in finalizations vs. M4-NoFU**: the number finalized under M4-Full minus the number finalized under M4-NoFU on the matched cohort.

B.4 Implementation Checks

Cohorts and paired inputs. Dataset-specific cohort manifests lock the splits and task labels to prevent cross-task denominator mixing. Frozen Proposal Generator inputs are shared across matched controls, keeping terminal-control comparisons paired.

Record validation. Required Harness fields are checked before terminal control so that case records support P3 Gate reruns. Structural consistency is checked before case release, separating artifact admissibility from subsequent disease-label scoring.

B.5 Interpreting the Comparisons

The fixed-record comparison identifies how the authorization rule changes release on an unchanged input. The follow-up comparison adds a source of observations and therefore evaluates acquisition together with re-audit. The confidence comparison holds the proposal generator and release count fixed, while allowing the Full policy additional evidence.

C Additional Terminal-Control and Dataset Results

C.1 Release Accounting

Table 2 contains the release endpoints for the shared 387-case BRACS cohort: M4-NoFU finalizes 64 and withholds 323; M4-Full finalizes 206 and blocks 181. Case-level acquisition and re-audit are illustrated in Figures 4 and 6.

C.2 Family-Level Terminal Outcomes

Table 5 reports terminal outcomes across the benign non-atypical, atypical borderline, and malignant BRACS families. Ungated wrong-kept proposals are concentrated in the benign non-atypical and atypical borderline families (99 and 63 cases, respectively), whereas the malignant family contributes 53. Evidence Repair finalizes 65, 17, and 124 cases in the three families.

Table 5: Per-family BRACS outcomes on the selective-finalization cohort. "Ungated wrong kept" counts incorrect family predictions that would have been finalized without control. "Gated final." counts cases finalized under Contract Auditor and P3 Gate. "Wrong block" counts incorrect ungated predictions blocked by the control layer. "Corr. block" counts correct ungated predictions blocked under the management policy.

Family	Cases	Ungated corr.	Ungated wrong kept	Gated final.	Wrong block	Corr. block
Benign non-atypical	149	50	99	65	82	2
Atypical borderline	75	12	63	17	57	1
Malignant	163	110	53	124	37	2
Overall	387	172	215	206	176	5

The family-level split records which correct and incorrect proposals the policy releases or withholds.

C.3 Full Confidence Threshold Results

Table 6 expands the confidence-threshold comparison into a full per-source, per-threshold sweep.

Table 6: Complete confidence-threshold sweep for the three source policies reported in the main paper. The threshold 0.0 is equivalent to no-gate mapped finalization. Risk is selective prediction risk: incorrect finalized proposals divided by finalized cases, using held-out labels.

Source	Metric	Confidence threshold						
		0.0	0.55	0.60	0.65	0.70	0.75	0.80
M1 single-shot	Finalized	285	264	203	178	128	81	32
	Wrong kept	108	96	64	49	25	11	2
	Risk	0.379	0.364	0.315	0.275	0.195	0.136	0.063
M2 navigation-only	Finalized	345	320	279	240	186	120	51
	Wrong kept	151	136	114	87	54	25	6
	Risk	0.438	0.425	0.409	0.363	0.290	0.208	0.118
Matched artifact source	Finalized	387	364	327	266	206	140	76
	Wrong kept	215	200	170	127	85	52	22
	Risk	0.556	0.549	0.520	0.477	0.413	0.371	0.289

C.4 NSCLC, RCC, BRCA-PAM50, AND EBRAINS DETAILS

NSCLC, RCC, and BRCA-PAM50 source cohorts: These analyses use task-specific diagnostic H&E WSI subsets. For morphology-caption alignment, Morphology Interpreter observations are compared against reference captions from pathology reports using BLEU, ROUGE-L, and BERTScore. This descriptive quality check complements the downstream classification analyses. For downstream classification, patient- or slide-level metadata define labels for NSCLC (LUAD versus LUSC), RCC (KICH versus KIRC versus KIRP), and BRCA-PAM50 tasks, while the image pipeline receives only diagnostic H&E WSIs and anonymized slide-level bags. NSCLC and RCC follow histologic subtype classification settings, whereas BRCA-PAM50 depends on the intersection of WSI availability, matched molecular-subtype labels, and task-specific filtering rules.

Additional BRCA-PAM50 benchmark. BRCA-PAM50 classification is included as an additional downstream MIL benchmark, distinct from the main NSCLC/RCC diagnostic-subtype result. The task uses diagnostic H&E WSIs with matched PAM50 labels. The retained label set is Luminal A, Luminal B, HER2-enriched, and Basal-like.

Table 7: Additional BRCA-PAM50 molecular-subtype proposal-substrate benchmark.

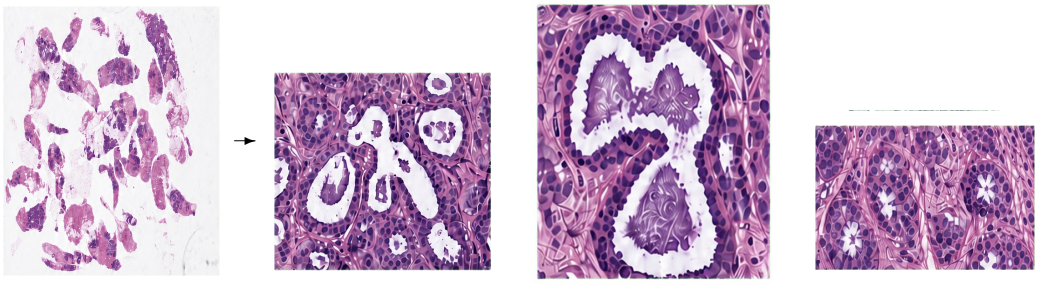
Method	Acc.	AUC
CLAM (Lu et al., 2021)	0.291	0.689
TransMIL (Shao et al., 2021)	0.254	0.631
MambaMIL	0.357	0.698
ACMIL	0.395	0.715
RPMIL	0.404	0.719
Jigsaw	0.418	0.725
Harness proposal substrate (Frozen Proposal Generator)	**0.496**	**0.732**

EBRAINS (2,335 WSIs): Human brain banking archive WSIs spanning neurodegenerative, vascular, and neoplastic conditions (Roetzer-Pejrimovsky et al., 2022). The EBRAINS experiment uses a 31-class diagnostic label space with WSI-level predictions aligned to ground truth for selective-finalization risk-coverage evaluation. Under Open Signout (M3), all 2,335 cases are finalized, with 1,678 correct finalizations, 657 wrong-kept diagnoses, accuracy 0.719, and selective prediction risk 0.281. Under Evidence Repair (M4-Full), 936/2,335 cases are finalized, with 833 correct finalizations, 103 wrong-kept diagnoses, selective accuracy 0.890, and selective prediction risk 0.110. The remaining 1,399 blocked cases include 554 wrong proposals and 845 correct predictions withheld under the management policy. All Evidence Repair finalizations pass the internal Contract Auditor and P3 Gate checks.

D QUALITATIVE CASE STUDIES

Figures 5 and 6 illustrate two complementary closure paths for the same frozen proposal class. In the evidence-complete case, observations distributed across multiple tissue fragments and gland-rich fields document the required architecture, cytology, representative sampling, and exclusion features. The stored record therefore passes contract audit and reaches a `FINAL` P3 Gate state without changing the proposal.

The repair case begins from a narrower gland-rich sample that does not establish representative architecture, cytology, or exclusion evidence. The Contract Auditor names these missing obligations, and DCE-Scout requests diverse high-power observations while keeping the proposal fixed. The appended evidence resolves the stated deficit on re-audit and permits `FINAL`.



WSI **E1** **E2** **E3**

Pathology VLM narrative evidence · 20 claims in 10 observations

[E01–E02] At low power, the specimen contains *multiple tissue fragments* with several *gland-rich foci.* The reviewed evidence is *distributed across more than one fragment* rather than a single isolated gland.

[E03–E04] At intermediate power, the sampled epithelial structures are *variably dilated glands and acini* with predominantly *round-to-oval contours. No markedly irregular outline* is apparent in the displayed fields.

[E05–E06] The glands *remain separated by intervening fibrous stroma* and do not form a confluent epithelial mass. Several lumina contain *conspicuous eosinophilic secretory material.*

[E07–E08] Focal *apical snouts* are visible along the luminal surface. The glands are lined by *cuboidal-to-columnar epithelial cells*, supporting a columnar-cell pattern in the sampled views.

[E09–E10] The epithelial lining is *generally one to a few cell layers thick. Mild crowding* is present focally, but *there is no broad multilayered proliferation* in the reviewed patches.

[E11–E12] The nuclei are *relatively uniform* from cell to cell with *only mild enlargement. Marked pleomorphism or striking variation in nuclear size* is not evident in these fields.

[E13–E15] Nuclear contours are *predominantly smooth,* with *fine-to-slightly coarse chromatin* and *inconspicuous nucleoli.* These cytologic findings *remain low grade* in the available high-power evidence.

[E16] Apical orientation and epithelial polarity are *largely preserved* around the glandular lumina. *Loss of organization is not conspicuous* in the sampled structures.

[E17–E19] *No conspicuous solid or cribriform epithelial growth* is identified. *Comedo-type necrosis* and *readily apparent mitotic activity* are also not seen in the reviewed fields.

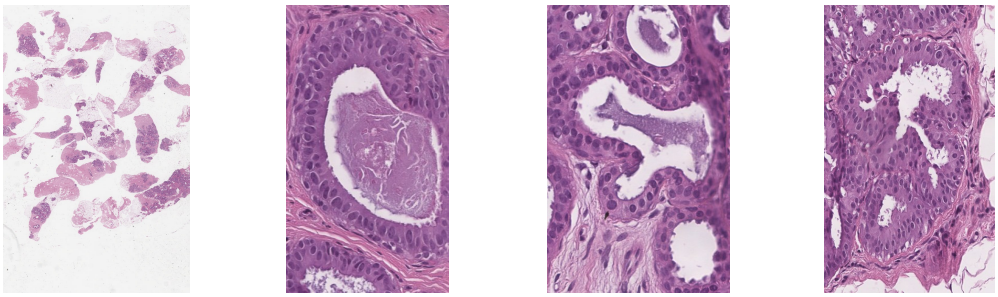
[E20] The sampled glands *do not show an irregular infiltrative arrangement* or a desmoplastic stromal response. *This exclusion is limited to the displayed and stored evidence.*

Frozen proposal: atypical_borderline (BRACS family; no FEA/ADH subclass selected).

Family audit: family-level architecture, cytology, sampling, and cross-family exclusions are documented.

P3 Gate: FINAL

Figure 5: Evidence-complete case study. Multiple stored observations satisfy the declared contract bound to the frozen `atypical_borderline` BRACS family proposal by documenting representative architecture, cytology, sampling, and cross-family exclusion evidence. The Contract Auditor accepts the record, and the P3 Gate authorizes `FINAL`.



WSI **Initial view** **Follow-up 1** **Follow-up 2**

Initial view $\xrightarrow{\text{FOLLOWUP}}$ Follow-up 1 / Follow-up 2

Pathology VLM narrative evidence · 24 claims in 12 observations

INITIAL RECORD · E01–E08

[E01–E02] At low power, the record contains multiple fragmented tissue cores, but only a *small fraction* of the *gland-rich tissue* has been reviewed. The sampling is therefore insufficient to establish the distribution of the epithelial change.

[E03–E04] The available high-power field is dominated by *one small gland profile* containing luminal *secretory material*. This finding is observable, but it represents only a *narrow portion* of the lesion.

[E05–E06] Epithelial lining cells and their nuclei are *visible*, although the gland profile is *partial* and the field is not representative. Nuclear *uniformity* and *stratification* cannot yet be assessed across multiple structures.

[E07–E08] The *extent* of the glandular change *remains uncertain* in the stored record. Higher-grade architectural patterns *cannot be excluded* without broader and more diverse sampling.

Frozen proposal: atypical_borderline (BRACS family; no FEA/ADH subclass selected).

Family-level audit deficit: representative architecture, cytology, and cross-family exclusions are incomplete.

DCE-Scout request: acquire diverse high-power epithelial regions; keep the family proposal fixed.

APPENDED AFTER FOLLOWUP · E09–E24

[E09–E10] Follow-up adds multiple *gland-rich fields* containing *variably dilated glands and acini*. The appended views substantially broaden the architectural sampling beyond the *initial isolated profile*.

[E11–E12] Most glands retain *round-to-oval contours* and remain *separated* by intervening fibrous stroma. *A confluent or irregular infiltrative arrangement is not apparent* in these views.

[E13–E14] Intraluminal *secretions* are confirmed across several glands, and *focal apical snouts* are visible. These findings are reproduced in more than one appended field.

[E15–E16] The glands are lined by *cuboidal-to-columnar epithelial cells*, generally *one to a few layers thick*. Broad multilayering or a *complex epithelial proliferation* is not identified.

[E17–E18] *Mild epithelial crowding* is present focally, while the nuclei remain *relatively uniform* with only mild enlargement. Marked *pleomorphism is not evident* in the follow-up evidence.

[E19–E20] Nuclear contours are predominantly *smooth*, nucleoli are inconspicuous, and *apical polarity is largely preserved*. The cytologic pattern remains *low grade* across the newly reviewed glands.

[E21–E23] *No conspicuous solid or cribriform growth* is identified in the expanded record. *Comedo-type necrosis* and readily apparent mitotic activity are also *not seen*.

[E24] The appended fields do not show *irregular infiltrative glands* or a desmoplastic stromal response. This exclusion *remains bounded* by the images stored in the case workspace.

Re-audit: the augmented record now satisfies the family-level architecture, cytology, and cross-family exclusion obligations.

P3 Gate: FINAL

Figure 6: Audit-guided evidence-repair case study. The initial record is too narrow to satisfy the declared contract bound to the frozen `atypical_borderline` BRACS family proposal. DCE-Scout requests targeted high-power follow-up; the appended observations address the family-level architecture, cytology, and cross-family exclusion obligations, after which re-audit and the P3 Gate return `FINAL`.

E FROZEN PROPOSAL SUBSTRATE

Table 8 places the proposal substrate alongside representative MIL, general multimodal, and pathology-specific systems. These external results follow their original protocols and serve as diagnostic context. The matched terminal-policy comparisons are reported in the main text. Additional molecular-subtype context appears in Table 7.

Table 8: Upstream classification context for the frozen proposal substrate. Acc. and AUC denote accuracy and area under the ROC curve.

Family	Method	Interface	BRACS		C16		C17	
			Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑
MIL	CLAM	Slide score + localization	0.432	0.704	0.928	0.944	0.887	0.928
	TransMIL		0.421	0.686	0.902	0.966	0.874	0.946
	MambaMIL		0.438	0.711	0.921	0.968	0.892	0.953
	ACMIL		0.440	0.716	0.934	0.972	0.906	0.954
	RPMIL		0.443	0.721	0.937	0.976	0.913	0.962
	Jigsaw		0.448	0.724	0.942	0.974	0.918	0.957
General MLLMs	InstructBLIP-FLAN-T5	ROI-level interpretation	0.318	0.582	0.621	0.701	0.588	0.676
	LLaVA-OneVision-72B		0.351	0.621	0.713	0.803	0.674	0.772
	GPT-4V		0.337	0.604	0.594	0.673	0.552	0.642
	Qwen2.5-VL-72B		0.366	0.638	0.742	0.829	0.708	0.804
	GPT-4.1-mini		0.359	0.629	0.729	0.817	0.692	0.789
	GPT-4.1		0.382	0.657	0.786	0.872	0.748	0.846
	Gemini-2.5-Flash		0.389	0.664	0.833	0.912	0.791	0.881
	Gemini-2.5-Pro		0.401	0.682	0.861	0.931	0.824	0.905
Pathology MLLMs	CPath-Omni	WSI evidence + proposal	-	-	-	-	-	-
	CPathAgent		-	-	-	-	-	-
Frozen substrate	**Frozen Proposal Generator**	Fixed proposal + initial record	0.444	-	0.961	-	0.926	-

Family	Method	Interface	EBRAINS		NSCLC		RCC		Average	
			Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑	Acc. ↑	AUC ↑
MIL	CLAM	Slide score + localization	0.665	0.975	0.928	0.978	0.907	0.976	0.791	0.918
	TransMIL		0.601	0.969	0.921	0.974	0.912	0.979	0.772	0.920
	MambaMIL		0.644	0.969	0.932	0.981	0.918	0.983	0.791	0.928
	ACMIL		0.641	0.962	0.934	0.982	0.915	0.981	0.795	0.928
	RPMIL		0.639	0.955	0.937	0.983	0.922	0.984	0.799	0.930
	Jigsaw		0.645	0.959	0.941	0.982	0.925	0.986	0.803	0.930
General MLLMs	InstructBLIP-FLAN-T5	ROI-level interpretation	0.268	0.641	0.553	0.711	0.438	0.623	0.464	0.656
	LLaVA-OneVision-72B		0.326	0.713	0.642	0.799	0.512	0.694	0.536	0.734
	GPT-4V		0.244	0.607	0.511	0.682	0.409	0.596	0.441	0.634
	Qwen2.5-VL-72B		0.358	0.742	0.654	0.814	0.529	0.719	0.560	0.758
	GPT-4.1-mini		0.341	0.728	0.651	0.803	0.526	0.707	0.550	0.746
	GPT-4.1		0.437	0.807	0.666	0.838	0.551	0.746	0.595	0.794
	Gemini-2.5-Flash		0.463	0.831	0.845	0.936	0.604	0.781	0.654	0.834
	Gemini-2.5-Pro		0.532	0.879	0.841	0.863	0.706	0.843	0.694	0.851
Pathology MLLMs	CPath-Omni	WSI evidence + proposal	-	-	0.888	-	0.940	-	-	-
	CPathAgent		-	-	0.908	-	0.946	-	-	-
Frozen substrate	**Frozen Proposal Generator**	Fixed proposal + initial record	0.719	-	0.958	-	0.934	-	0.824	-

F Spatial Evidence Acquisition

Tables 9 and 10 evaluate initial evidence acquisition before terminal policies diverge. The former compares thumbnail and multi-region sources; the latter measures geometric localization on CAMELYON16/17.

Table 9: Evidence-source comparison before terminal control. M1 uses one thumbnail per WSI. M2 uses multiple 10× ROI observations. "Mapped" reports Frozen Proposal Generator mappability.

Task	M1 mapped	M2 mapped	M1 mapped acc.	M2 mapped acc.	M2-only corr.	M1-only corr.	Interpretation
BRACS 3-family	0.742	0.893	0.644	0.523	32	36	Higher mappability, lower mapped accuracy
CAMELYON binary	0.761	0.732	0.868	0.863	112	138	Similar mapped accuracy, with more M1-only cases
EBRAINS 31-class	0.922	1.000	0.682	0.717	362	156	M2 improves coverage and accuracy
Cancer-site family	0.734	0.829	0.604	0.663	1,053	285	ROI evidence improves both
All labeled pairs	0.777	0.860	0.647	0.686	1,559	615	Net gain from ROI evidence

Multi-region observation changes both mappability and mapped accuracy, with task-dependent effects. The aggregate row summarizes the reported labeled pairs; the per-task rows retain these differences. The navigation analysis below measures whether inspected regions reach annotated lesions.

F.1 Navigation Quality and Protocol

Navigation quality is measured on CAMELYON16 (48 tumor slides with XML lesion annotations and 80 normal slides from the testing split) and CAMELYON17 (50 training XML slides). Hit@K uses nonzero overlap between top-K ROI fields and XML lesion polygons in level-0 pixel space. Center@5 records whether any top-5 ROI center falls inside a lesion polygon. Tissue efficiency is the fraction of viewed tissue area that overlaps lesion tissue, and lesion coverage is the fraction of total lesion area covered by the union of viewed ROIs.

F.2 CAMELYON17 Testing Details

CAMELYON17 consists of 1,000 WSIs (500 training, 500 testing) from 5 medical centers (Bandi et al., 2019). We report navigation quality on the local training XML cohort (50 slides with available lesion annotations), using the same Hit@K and tissue efficiency metrics as CAMELYON16.

Table 10: CAMELYON navigation quality at 10×. C16 is the paired testing cohort. C17 uses training XML cases only. Ctr. acc. is Center@5, defined as whether any top-5 ROI center falls inside a lesion polygon. Tiss. eff. is Mean tissue@5.

Method	C16 H@1	C16 H@3	C16 H@5	C16 Ctr. acc.	C16 Les. cov.	C16 Tiss. eff.	C17 H@5
Random tissue	0.133	0.262	0.343	0.153	0.025	0.359	0.240
M2 navigation	0.208	0.333	0.375	0.188	0.033	0.559	0.260
CONCH v1.5 Scout-A	0.250	0.417	0.500	0.292	0.051	0.839	-
Retained ROI subset	0.250	0.417	0.417	0.208	0.038	0.815	-
Annotation oracle (eval.-only)	0.542	0.542	0.542	0.208	0.071	0.547	0.480

G Extended Discussion and Limitations

G.1 Extended Discussion

The fixed-proposal setting separates two ways of changing a case outcome. A release rule can withhold a proposal on the current record, while follow-up can alter the evidence available to that rule. DCE-Scout supplies the observations requested after a recoverable objection; the Harness supplies the binding, history, and authorization pathway. A contradictory observation remains informative even when it prevents finalization, because it changes the basis for the terminal decision without rewriting the proposal.

G.2 Scope and Limitations

Pathology Harness applies to workflows in which regional observations can be retained, proposals can be bound to compatible evidence contracts, and release can be determined from stored artifacts. Its interfaces allow different evidence producers, but the reported evaluations use particular components, contracts, and cohorts. Their quality limits both the audit and the acquisition policy; correlated errors in interpretation and verification can survive the control pathway.

The evaluation characterizes retrospective case-closure decisions under fixed proposals and declared policies. Appendix G.3 details the application scope and replay requirements.

G.3 Additional Limitations

Contracts and interpretation. The contracts determine which observations are required at a given label granularity. Their adequacy and the correctness of the stored morphology are separate from structural compliance. Correlated Interpreter and Auditor errors and confirmation bias can persist. Independent expert review is particularly relevant at difficult diagnostic boundaries (Elmore et al., 2015).

Acquisition and scope. DCE-Scout depends on candidate-pool quality, observability estimates, and access to the requested tissue and magnification. Required evidence may be absent or remain ambiguous. The current NoFU–Full comparison evaluates the acquisition loop as a whole. Equal-budget selector comparisons and explicit acquisition costs are needed to assess its individual objective terms.

Evaluation context. Terminal-policy results use BRACS and EBRAINS; CAMELYON navigation and the NSCLC, RCC, BRCA-PAM50, and cancer-site analyses describe upstream evidence sources. External rows in Table 8 follow their original protocols. The M2 implementation is CPathAgent-style inference. The BRACS confidence comparison is specific to its proposal generator and reported thresholds.

Replay. The retained terminal package supports P3 replay; full component replay also needs released models, prompts, candidate pools, and configurations.