PathTrace: A Trace-Based Evidence Harness for Auditing Whole-Slide Pathology Agents

Chiew Hui Lim*

Monash University Malaysia
Selangor, Malaysia
clim0149@student.monash.edu

Abstract. Large language model (LLM)-based agents have shown potential for whole-slide image (WSI) analysis by approximating pathologists' diagnostic workflow: navigating across gigapixel slides, zooming into suspicious regions, and collecting visual evidence before making diagnostic claims. However, existing evaluations focus mainly on final accuracy, leaving unclear where the agent looked, which regions supported each claim, and whether the claimed evidence was ever inspected. Without such traces, agent behaviour is difficult to replay, compare, and audit across models. We present **PathTrace**, a trace-based evidence harness and offline auditing interface for WSI pathology agents. PathTrace converts each agent run into a standardized trace that records percentage-coordinate views, marked evidence, and diagnostic claims. PathTrace also provides a trace-auditing interface for replaying navigation paths, inspecting the exact patches seen by the model, and flagging missed tumour regions, off-tumour evidence, and trace-inconsistent citations. We conduct a multi-model study on 250 PANDA prostate biopsies. PathTrace reveals a failure mode that standard accuracy hides: one open agent achieves 0.83 binary accuracy for cancer detection, yet 77% of its evidence pins fall outside annotated tumour regions. We additionally demonstrate a simple audit-informed grounding check that improves grade agreement from κ=0.06 to κ=0.41 without retraining. PathTrace therefore provides a practical framework for inspecting evidence provenance beyond final diagnosis accuracy.

Keywords: Computational Pathology · Agentic AI · Whole-slide Image Analysis · Evidence Auditing · Visual Grounding · Hallucination

1 Introduction

Whole-slide images (WSIs) are gigapixel-scale digital scans of pathology slides. Unlike ordinary images, they cannot be interpreted from a single fixed view: diagnostically relevant tissue may occupy only a small fraction of the slide, and diagnosis requires searching across magnifications, inspecting suspicious regions, and linking local visual evidence to a slide-level conclusion. WSI analysis is

* Corresponding author.

therefore not only a recognition problem, but also a navigation and evidence-grounding problem.

Early computational pathology systems addressed this setting through patch-based training and weakly supervised multiple-instance learning [11,12,13], where a slide is represented as a bag of local patches and trained from slide-level labels. More recent work has moved beyond static slide classification toward agentic vision-language systems [1,2,3]. These systems attempt to approximate a pathologist-like workflow by navigating across a WSI, zooming into selected regions, collecting observations, and producing a diagnostic answer based on the inspected evidence.

However, current evaluations of WSI agents still focus primarily on final diagnostic accuracy or free-form textual reports. This compresses a complex sequential process into a single outcome and can hide important failure modes. An agent may produce the correct diagnosis after inspecting irrelevant regions, miss tumour despite a confident report, or cite visual evidence that was never actually observed. Such ungrounded generation is a known tendency of vision-language models, which may rely on language priors rather than the pixels in front of them [16]. On gigapixel pathology slides, where diagnostically relevant tissue can be sparse, this failure mode is especially consequential. Without a standardized record of where an agent looked and which evidence supports each claim, agent behaviour remains difficult to reproduce, compare, and audit.

We introduce **PathTrace**, a trace-based evidence harness for WSI pathology agents. PathTrace records each agent run as a slide-relative trace containing viewed regions, evidence marks, patch-level observations, and final diagnostic claims. It can be integrated into existing agent loops with minimal instrumentation. By evaluating navigation, evidence localization, and claim grounding in addition to final accuracy, PathTrace exposes failure modes that are invisible to standard evaluation and provides a common basis for comparing WSI agents under shared patch budgets.

PathTrace audits the *observable evidence process* of an agent rather than its latent reasoning. We distinguish between tissue exposure through viewed regions, explicit evidence selection through pins, and claim attribution through links between claims and recorded evidence. The resulting checks are operational audit signals: they expose spatial or trace-level inconsistencies, but do not by themselves establish semantic relevance or causal dependence of a diagnosis on a particular region.

Using PathTrace, we conduct a multi-model audit on 250 PANDA prostate biopsies [14]. Our contributions are:

- We introduce PathTrace, a trace-based auditing framework for WSI agents. It records where an agent looks, what evidence it marks, and what diagnostic claims it makes, then compares these behaviours with pixel-level annotations of tumour regions.
- We audit multiple WSI agents, including Claude, Qwen-2.5-VL, and a PathAgent-inspired baseline, and show that final accuracy can hide ungrounded behaviour. In one open agent, binary cancer-detection accuracy

reaches 0.83, yet grade agreement is close to random guessing (κ=0.06), with 77% of evidence pins falling outside tumour regions.
- We demonstrate a lightweight CONCH [6] grounding gate that improves grade agreement from κ=0.06 to κ=0.41 without retraining.

2 Related Work

2.1 WSI Navigation Agents

A recent line of work frames WSI diagnosis as sequential navigation by vision-language agents: PathAgent [1] (a training-free Navigator/Perceptor/Executor pipeline), CPathAgent [2] (a trained coarse-to-fine navigator), Pathology-CoT [3] (expert-behaviour reasoning traces), and multi-magnification / query-driven navigators [4,5]. These systems report aggregate accuracy and often present interpretable or traceable decisions, but the traces are self-reported, model-specific, rarely verified against pixel-level ground truth, and, in question-answering variants, may lack slide coordinates entirely. PathTrace is complementary: rather than another navigator, it is a standardized auditing layer that turns agent behaviour into externally verifiable traces of navigation, evidence localisation, and claim grounding.

2.2 Foundation Models, Benchmarks, and Grounding

The components WSI agents orchestrate are already strong: vision-language encoders [6,8], tile encoders [7,10], slide-level models [9], and MIL classifiers [11,12,13]. We use CONCH to score patches. Pathology benchmarks such as PathMMU [15] evaluate static visual question answering, scoring final answers rather than the inspection process. PathTrace instead asks whether an agent looked at relevant tissue, marked valid evidence, and grounded its claims in inspected regions. This connects to work on hallucination in vision-language models (decoding-time mitigation [16,17], benchmarks [18], cross-modal verification [19]), which is mostly evaluated on natural images rather than gigapixel slides.

3 Method

PathTrace is designed as an auditing layer around an existing WSI agent (Fig. 1). It records an agent run as a standardized trace, optionally re-checks the trace with an additional grounding checker such as CONCH, and provides an auditor for replaying, inspecting, and adjudicating the agent's behaviour.

3.1 Trace Schema

A trace records a single slide-model run. It contains an ordered list of *views*, where each view stores the inspected slide region, magnification level, and a

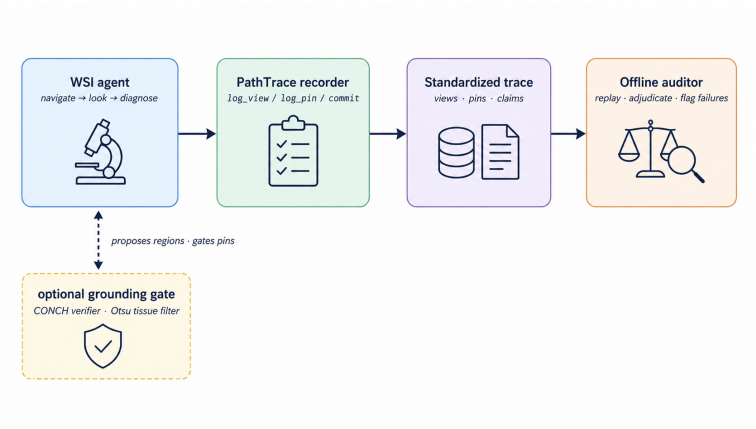

Fig. 1. PathTrace pipeline. An agent run is recorded as a standardized trace and replayed by an offline auditor that flags grounding failures; the optional CONCH / tissue grounding gate feeds back into the agent (proposing regions, gating pins).

patch-level observation. Each viewed region is represented in slide-relative coordinates, for example `x%, y%, w%, h%` and `zoom`. The trace also contains *pins*, which mark point evidence with a colour, note, and location, and a final *diagnosis* made up of structured claims. Each claim can cite one or more evidence pins through `supporting_pin_ids`. Storing spatial locations as percentages of the slide dimensions makes traces independent of screen resolution, viewer size, and tile-server implementation.

3.2 Harness

Model authors do not need to rewrite their agents to use PathTrace. During inference, the agent only needs to call three recorder methods: `log_view` to record inspected regions, `log_pin` to record marked evidence, and `commit` to record the final diagnosis. Each evidence pin logged through `log_pin` is assigned a unique identifier. At commit time, the agent links each diagnostic claim to one or more previously logged pins using these identifiers. The harness then exports the run as a schema-compliant JSON trace.

These recorder calls determine what PathTrace can audit. Views recorded through `log_view` allow navigation and tumour-hit analysis. Evidence pins recorded through `log_pin` allow off-tumour evidence analysis. Claims recorded through `commit`, together with their cited pin identifiers, allow claim-grounding analysis. If the trace contains only the final diagnosis, PathTrace falls back to standard accuracy evaluation; richer traces enable deeper audits of navigation, evidence, and claim grounding. These three events correspond to three observable stages of the agent workflow: exposure (`log_view`), evidence selection (`log_pin`), and claim attribution (`commit`).

3.3 Auditor

The auditor is an interface for reviewing and comparing WSI agent traces. It displays risk flags such as missed tumour, false-positive diagnosis, off-tumour pins, and ungrounded citations. A *slide view* replays a selected model's trajectory with a step slider, shows the patch inspected at each step, overlays the ground-truth tumour mask, and can display trajectories from multiple models for comparison. A *cohort QA* view summarizes per-model agreement, grounding metrics, and failure-flag counts across the study cohort (Fig. 2).

3.4 Grounding Checks and Metrics

PathTrace uses simple rule-based checks to flag grounding failures in each trace. A trace is flagged as *missed tumour* if the agent never views or marks a region that overlaps the tumour mask. It is flagged as *off-tumour evidence* if a red evidence pin lies outside the tumour mask. It is flagged as an *ungrounded citation* if a diagnostic claim cites a pin that was not recorded in the trace.

Because these flags are computed directly from the trace and the tumour mask, they are reproducible and easy to inspect. We report three operational audit metrics. *Tumour-hit rate* measures whether the agent reached the annotated tumour at least once, using either a viewed region or an evidence pin. *Off-tumour-pin rate* measures the fraction of red evidence pins placed outside the tumour mask. *Ungrounded-citation rate* measures the fraction of diagnostic claims that cite missing or invalid evidence pins. These metrics capture observable evidence provenance and spatial consistency, but should not be interpreted as complete measures of semantic or causal grounding.

4 Experiments

4.1 Setup

We evaluate PathTrace on a balanced set of 250 PANDA prostate core biopsies from Radboud and Karolinska, each with a pixel-level tumour mask. We run the main ablation experiments on all 250 slides using the open VLM, while a separate 60-slide subset is used only for frontier-model auditing and exploratory mitigation experiments. PANDA is a public challenge dataset [14], used under its original release terms; no new data were collected, and no additional ethics approval was required. ISUP labels and tumour extent are taken from the dataset metadata and pixel masks.

We report binary tumour detection, exact ISUP agreement, within-one-grade ISUP agreement, and quadratic-weighted κ. We also measure whether the agent's evidence is spatially grounded: an evidence pin is counted as *off-tumour* if its centre falls outside the annotated tumour mask, meaning that the centre pixel is not labelled as tumour. This centre-point rule is strict and reproducible, although it may count nearby diagnostic context as off-tumour. The off-pin rate is computed over all evidence pins. We also check whether the final report is

grounded in the recorded trace: a citation is counted as *ungrounded* if the report refers to a pin, patch, or region that was not inspected, was not dropped as evidence, or cannot be linked to any recorded trace event.

For reproducibility, the open perceptor is Qwen-2.5-VL-7B-Instruct in 4-bit mode, and the frontier perceptor is Claude Sonnet 4.5. We use CONCH ViT-B/16 as an independent image encoder to score how suspicious each region is for malignancy. By default, the navigator inspects the top-16 regions from a 16×16 Otsu-masked grid for each slide.

We also implement a PathAgent [1]-inspired Navigator/Perceptor/Executor baseline inside our harness. It is a simplified single-pass variant: the Navigator proposes candidate regions, the Perceptor scores them, and the Executor aggregates the evidence. It is not a reproduction of the original closed-loop PathAgent system, and our results should not be interpreted as claims about the original method. By auditing a closed frontier model, an open VLM, and a PathAgent-inspired method using the same trace schema, PathTrace is tested across different types of WSI agents.

4.2 Audit Findings

PathTrace reports both diagnosis-level and process-level metrics. Binary accuracy and quadratic-weighted κ are computed from the final committed diagnoses. In contrast, tumour-hit rate, off-tumour-pin rate, and ungrounded-citation rate are computed from the recorded trace and compared with the ground-truth tumour mask. These process-level metrics are not available in a diagnosis-only evaluation. This allows us to ask not only whether an agent is correct, but also whether its observable evidence process is spatially consistent with the tissue it inspected and the available tumour annotations.

On the 250-slide open-VLM cohort, Qwen-2.5-VL-7B used as the perceptor in a navigating agent reaches 0.83 binary accuracy. However, its ISUP grading remains poor, with quadratic-weighted κ=0.06, and 77% of its evidence pins fall outside the annotated tumour. This means that the model can often make the correct binary tumour decision while still relying on weak or mislocalized evidence. Such a failure would be hidden by accuracy alone. Grounding failures are not limited to the open model: even Claude Sonnet 4.5 places about 63% of its evidence pins off-tumour. Figure 2 shows how PathTrace surfaces these failures directly, including pins on benign tissue and citations to regions that were never inspected.

These failures are not isolated to one model. Table 1 reports two trace-based metrics defined for every agent: tumour-hit (whether the agent reached the tumour) and the off-tumour-pin rate. Even the frontier model, and even when an agent does reach the tumour, still places the majority of its evidence pins outside the annotated tumour.

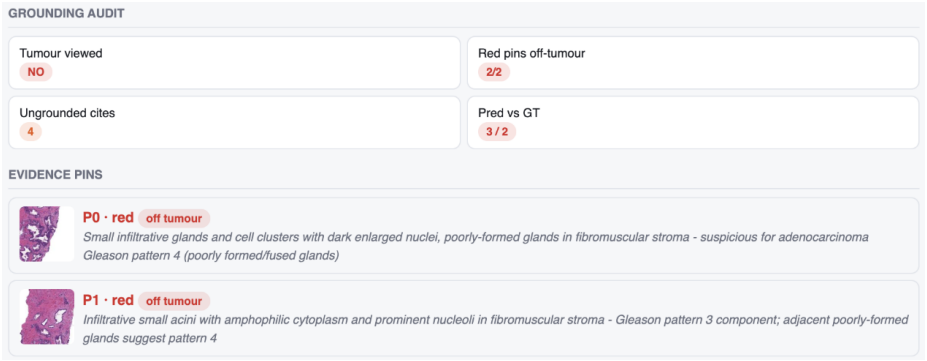

Fig. 2. Example PathTrace audit on a PANDA WSI [14]. The model gives a diagnosis, while the auditor identifies several grounding flags: it did not view the tumour, its red evidence pins are outside the tumour mask, and some cited evidence was never recorded.

Table 1. Audit flags across models (frontier n=60; open-VLM n=250). Tumour-hit is higher better; off-pin is lower better.

Model	Tumour-hit	Off-pin
Frontier agent (Claude)	0.77	0.63
Open-VLM agent (Qwen-2.5-VL-7B)	0.88	0.77
+ CONCH gate	0.68	0.68

4.3 Navigation Coverage

Navigation errors are budget-sensitive: the agent may miss tumour simply because it does not inspect the right regions. To measure this separately from diagnosis, we ignore the VLM and only test the region proposer. We split each slide into a grid, use CONCH to rank regions by suspiciousness, and report *candidate coverage* to see whether the selected top-k regions include the annotated tumour. This is different from the trace tumour-hit in Table 3, which measures what the full agent actually inspected. Coverage rises from 0.55 with 8 patches to 0.97 with full coverage, showing that many missed tumours are recoverable by looking more. We therefore use grid 16/top-16, with coverage 0.68, as a tractable default budget.

4.4 A Simple Grounding Check Improves Grounding

PathTrace shows weak grounding in the open-VLM agent: Qwen-2.5-VL reaches chance-level grading performance (κ=0.06), while 77% of its evidence pins fall outside the annotated tumour. We therefore test two simple checks: an Otsu tissue filter to remove background regions, and a CONCH gate that keeps a pin only if CONCH also scores the region as malignant. CONCH is independent of

Table 2. Navigation coverage on cancer slides (CONCH proposer, no VLM). Candidate coverage = whether the selected regions include tumour; "full" inspects every tissue cell (n=60).

Setting (grid : top-k)	Patches	Coverage
16 : 8	8	0.55
16 : 16	16	0.68
32 : 16	16	0.70
16 : full	62	0.88
32 : full	248	0.97

Table 3. Grounding ablation (n=250, Qwen-2.5-VL perceptor). ≤ 1 = within one ISUP grade; MAE and off-pin lower better. Best per column **bold**, second underlined.

Config	Bin	ISUP	≤ 1	MAE	κ	Tum.-hit	Off-pin
ungated	**0.83**	0.17	0.51	1.50	0.06	**0.88**	0.77
+ tissue filter	**0.83**	0.17	0.51	1.50	0.06	**0.88**	0.77
+ CONCH gate	0.71	0.24	0.62	1.42	0.40	0.52	0.73
+ CONCH + tissue	0.78	**0.26**	**0.63**	**1.38**	**0.41**	0.68	**0.68**

the report-generating Qwen VLM, but not of the complete pipeline because it is also used for candidate-region ranking. We therefore interpret this as an audit-informed intervention rather than a fully independent verification experiment.

Table 3 shows that the tissue filter alone does not change the ungated agent, because its candidate regions are already mostly on tissue. However, adding tissue filtering before the CONCH gate removes artefacts such as glass background and pen marks, improving tumour-hit from 0.52 to 0.68. The main grading gain comes from the CONCH gate, with quadratic-weighted κ rising from 0.06 to 0.41. This improvement from combining Qwen with CONCH suggests that the grounding failure can be partly corrected, although accurate localisation remains unresolved.

Beyond exact grade agreement, within-one-grade accuracy rises from 0.51 to 0.63, and MAE decreases from 1.50 to 1.38, meaning that predictions are closer to the ground truth. This improvement involves a trade-off: binary accuracy decreases from 0.83 to 0.78, while tumour-hit decreases from 0.88 to 0.68. The off-pin rate also remains high at 0.68. Thus, the grounding check improves grading agreement more than localisation and does not uniformly improve all audit dimensions.

5 Conclusion and Limitations

In this paper, we have presented a trace-based evidence harness for WSI agents that records where an agent looks, what evidence it marks, and which claims it makes. This reveals the process-level failures that final accuracy hides, like

how agents can make a correct binary call while missing tumour, pinning benign tissue, or citing uninspected regions. We have also mitigated these failures with a CONCH verifier that improves ISUP grading and within-one-grade agreement. However, localisation remains unresolved as many evidence pins still fall outside the annotated tumour.

Limitations. This study uses 250 slides from one public prostate-cancer dataset, with a 60-slide subset for frontier-model auditing, so the results should be interpreted as dataset-specific audit findings rather than cross-dataset model rankings. PathTrace requires agents to expose coordinates or evidence locations and does not recover latent reasoning. Tumour masks provide a reproducible spatial reference, but off-tumour regions may still contain diagnostically relevant context, particularly near annotation boundaries; therefore, off-tumour pins should be treated as audit flags rather than definitive evidence of incorrect reasoning. Likewise, linking a claim to a recorded pin establishes trace provenance but does not guarantee semantic support. Finally, we have not evaluated cross-dataset generalization or repeated-run trace stability.

Disclosure of Interests The author has no competing interests to declare that are relevant to the content of this article.

References

1. Chen, J. et al. PathAgent: Toward Interpretable Analysis of Whole-slide Pathology Images via LLM-based Agentic Reasoning. arXiv:2511.17052, 2025.
2. Sun, Y. et al. CPathAgent: An Agent-based Foundation Model for Interpretable High-Resolution Pathology Image Analysis. arXiv:2505.20510, 2025.
3. Wang, S. et al. Pathology-CoT: Learning Visual Chain-of-Thought Agent from Expert WSI Diagnosis Behavior. arXiv:2510.04587, 2025.
4. Xu, Z. et al. MMNavAgent: Multi-Magnification WSI Navigation Agent for Clinically Consistent Whole-Slide Analysis. arXiv:2603.02079, 2026.
5. Zhang, K. et al. PathReasoning: A Multimodal Reasoning Agent for Query-based ROI Navigation on Whole-Slide Images. arXiv:2511.21902, 2025.
6. Lu, M.Y. et al. A Visual-Language Foundation Model for Computational Pathology (CONCH). *Nature Medicine* 30, 2024.
7. Chen, R.J. et al. Towards a General-Purpose Foundation Model for Computational Pathology (UNI). *Nature Medicine* 30, 2024.
8. Huang, Z. et al. A Visual–Language Foundation Model for Pathology Image Analysis using Medical Twitter (PLIP). *Nature Medicine* 29, 2023.
9. Xu, H. et al. A Whole-Slide Foundation Model for Digital Pathology from Real-World Data (Prov-GigaPath). *Nature* 630, 2024.
10. Vorontsov, E. et al. A Foundation Model for Clinical-Grade Computational Pathology and Rare Cancers Detection (Virchow). *Nature Medicine* 30, 2024.
11. Lu, M.Y. et al. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images (CLAM). *Nature Biomedical Engineering* 5, 555–570, 2021.
12. Ilse, M., Tomczak, J., Welling, M. Attention-based Deep Multiple Instance Learning. *ICML*, 2018.

13. Shao, Z. et al. TransMIL: Transformer based Correlated Multiple Instance Learning for WSI Classification. *NeurIPS*, 2021.
14. Bulten, W. et al. Artificial Intelligence for Diagnosis and Gleason Grading of Prostate Cancer: the PANDA Challenge. *Nature Medicine* 28, 154–163, 2022.
15. Sun, Y. et al. PathMMU: A Massive Multimodal Expert-Level Benchmark for Understanding and Reasoning in Pathology. *ECCV*, 2024.
16. Leng, S. et al. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. *CVPR*, 2024.
17. Favero, A. et al. Multi-Modal Hallucination Control by Visual Information Grounding (M3ID). *CVPR*, 2024. arXiv:2403.14003.
18. Guan, T. et al. HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in LVLMs. *CVPR*, 2024.
19. Yang, Z. et al. InEx: Hallucination Mitigation via Introspection and Cross-Modal Multi-Agent Collaboration. arXiv:2512.02981, 2025.