Scilight

Artificial Intelligence and Emerging Technologies
https://www.sciltp.com/journals/aiet

Article

Check for updates

Provenance-Native Audit Infrastructure for LLM-Maintained Wiki Knowledge Systems

Bailing Zhang

School of Computer Science and Data Engineering, NingboTech University, Ningbo 315100, China; bailing.zhang1961@gmail.com

How To Cite: Zhang, B. Provenance-Native Audit Infrastructure for LLM-Maintained Wiki Knowledge Systems. *Artificial Intelligence and Emerging Technologies* **2026**, *3*(3), 10. https://doi.org/10.53941/aiet.2026.100010

Received: 27 April 2026
Revised: 31 July 2026
Accepted: 7 August 2026
Published: 1 September 2026

Abstract: LLM-maintained wiki systems accumulate structured knowledge by having a language model incrementally build and revise a persistent wiki layer over raw evidence sources. When such systems support decision-making in regulated domains, each output must be traceable to grounded evidence through an auditable chain. This paper presents a provenance-native architecture in which a dedicated provenance layer is embedded in the wiki's operational loop, recording evidence bindings, claim versions, rule compilations, and decision traces as side effects of normal system operations. The architecture is organized around four technical components: (1) a typed provenance graph that binds claims to evidence passages with confidence-weighted edges; (2) a dual-layer diff mechanism combining textual and semantic comparison to detect silent LLM edits; (3) three confidence propagation strategies over multi-hop evidence–claim–rule–decision chains, compared on their suitability for regulatory communication; and (4) a retraction propagation engine that identifies all downstream dependents when upstream evidence is invalidated. We evaluate the system on a 112-document corpus producing 3174 graph nodes and 4398 edges. In controlled retraction experiments, the engine achieves 100% downstream recall with zero false propagation in under 600 ms per event. A sensitivity analysis over the human-review threshold parameter γ reveals a sharp phase transition in the fraction of flagged decisions, providing a concrete basis for negotiating review policies with regulators. An automated gap analysis shows the architecture demonstrates alignment with five of eight requirements derived from the EU AI Act and FDA AI/ML guidance, and we report module-level evaluations of claim extraction, evidence binding, and silent-edit detection together with a measured baseline comparison on retraction propagation. Code and experimental protocols will be released publicly upon publication.

Keywords: provenance; audit infrastructure; LLM-maintained wiki; knowledge systems; confidence propagation; retraction

1. Introduction

The dominant pattern for connecting large language models to document collections is retrieval-augmented generation: at query time, the system retrieves relevant chunks and synthesizes an answer from scratch. Knowledge is re-derived on every query; nothing accumulates. The LLM Wiki paradigm [1] proposes a different architecture in which the LLM *incrementally builds and maintains a persistent wiki*—a structured, interlinked collection of pages that sits between raw sources and downstream queries. When a new source arrives, the LLM reads it, extracts key information, and integrates it into the existing wiki by updating entity pages, revising summaries, and flagging contradictions. The wiki is a persistent, compounding artifact.

This architecture is attractive for domains where knowledge accumulates over extended periods: clinical decision support, regulatory intelligence, research synthesis, competitive analysis. But once the wiki moves beyond personal note-taking toward high-risk decision support, a structural gap appears. The system can tell you *what* it concluded, but it cannot answer the following class of questions:

Copyright: © 2026 by the authors. This is an open access article under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).
Publisher's Note: Scilight stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Which original evidence supports a given wiki claim?
- When a wiki claim was revised, what changed and why?
- If an upstream source is retracted, which downstream claims, rules, and decisions are affected?
- What is the weakest evidential link in the chain from source to decision?

In regulated domains—medical device software under FDA AI/ML SaMD guidance [2], high-risk AI systems under the EU AI Act [3]—these are not optional features but compliance prerequisites. A system that cannot trace its outputs to grounded evidence and propagate evidence invalidation cannot pass regulatory audit.

The difficulty is not provenance tracking *per se*—database provenance and workflow lineage have been studied for decades [4, 5]. The difficulty lies in four properties that distinguish LLM-maintained wikis from classical provenance substrates: (P1) the wiki schema evolves as the LLM and user co-evolve conventions; (P2) assertions are semi-structured natural language, neither database tuples nor fully unstructured text; (P3) the entity performing edits is an LLM, which may introduce *silent edits*—semantic changes without explicit logging; and (P4) decisions depend on multi-hop reasoning chains where confidence must propagate conservatively.

This paper presents a provenance-native architecture for LLM-maintained wiki systems. The term "native" is deliberate: the provenance layer is not an external audit tool applied after the fact, but an integral component of the wiki's coupling architecture that records provenance as a side effect of normal operations. Our contributions are:

(1) A typed provenance graph model for LLM-maintained wikis, accommodating evolving schemas, natural-language claims, and LLM-originated edits (Section 4.1).
(2) A dual-layer diff mechanism (textual + semantic) for detecting silent LLM edits during wiki maintenance (Section 4.3).
(3) Three confidence propagation strategies for multi-hop evidence–claim–rule–decision chains, designed for regulatory communicability (Section 4.4).
(4) A retraction propagation engine with batch-optimized downstream impact analysis (Section 4.5).
(5) Experimental evaluation on a 112-document corpus including controlled retraction experiments, confidence calibration, and regulatory gap analysis (Section 5).

2. Related Work

2.1. Database and Workflow Provenance

Provenance tracking has a long history in databases, where why-provenance, how-provenance, and where-provenance explain query results in terms of input data [4]. Workflow systems such as VisTrails and Taverna record data lineage through processing pipelines [6]. The W3C PROV-O ontology [7] provides a standardized vocabulary for representing provenance as interactions between entities, activities, and agents. These systems assume structured data, stable schemas, and deterministic transformations—assumptions that break in LLM-maintained wiki systems where claims are natural language, schemas evolve, and the editor is a stochastic language model.

2.2. Knowledge Graph Provenance and Trust

Knowledge graphs support provenance through named graphs, reification, and trust scoring. Frameworks like WIQA [8] propagate trust through graph paths. However, these systems assume facts have already been extracted into clean triples—precisely the step that is problematic in LLM Wiki systems, where claims exist as natural-language sentences, not formal triples. Our work operates at a coarser granularity (claim-level rather than triple-level) but handles the additional complexity of LLM-originated edits and evolving page structures.

2.3. RAG Citation and LLM Attribution

Recent work on citation-grounded generation attaches source references to LLM outputs. The ALCE benchmark [9] evaluates citation quality along fluency, correctness, and attribution dimensions. These systems improve transparency but have three limitations relative to our requirements: citation granularity is typically at the document or passage level, not the claim level; citations are generated at query time and do not persist as reusable provenance; and there is no mechanism for propagating evidence invalidation to previously generated outputs. Our provenance layer provides persistent, claim-level, reusable bindings with invalidation propagation.

2.4. Knowledge Editing and Fact Verification

Knowledge editing methods such as ROME [10] and MEMIT [11] modify specific facts inside model parameters or retrieval stores. Fact-verification systems such as FEVER [12] check claims against evidence. Neuro-symbolic systems such as DeepProbLog [13] offer execution-time auditability through rule traces, but

execution-time traceability is not the same as *compilation-time* traceability: knowing which rule fired does not tell you which evidence the rule was derived from. None of these approaches addresses the ongoing maintenance of a provenance graph that evolves with the knowledge base.

2.5. LLM-Maintained Knowledge Systems

The LLM Wiki paradigm [1] and related work on persistent agent memory [14,15] demonstrate that LLMs can maintain persistent knowledge artifacts. However, none of these systems includes a native provenance layer. The maintenance operations (ingest, lint, query) are described but not instrumented for auditability. Our work is, to our knowledge, the first to embed provenance infrastructure into the LLM Wiki architecture.

2.6. Regulatory Requirements

The EU AI Act (Regulation 2024/1689) requires high-risk AI systems to maintain technical documentation (Article 11), automatic logging of events (Article 12), transparency toward deployers (Article 13), and human oversight provisions (Article 14) [3]. The FDA's AI/ML Software as a Medical Device guidance requires traceability from outputs to training data [2]. Our compliance gap analysis module maps provenance capabilities directly to these requirements.

3. Problem Formulation

We consider a system organized in five layers. Layer L1 contains immutable evidence sources (papers, guidelines, SOPs). Layer L2 is a wiki of LLM-generated pages, each containing multiple natural-language claims. Layer L3 is the provenance layer—the focus of this work. Layer L4 contains rules or constraints compiled from L2 claims. Layer L5 produces decisions or recommendations.

A *decision* $d \in L5$ depends on rules $\{r_i\} \subset L4$, which were compiled from claims $\{c_j\} \subset L2$, which were derived from evidence $\{e_k\} \subset L1$. The problem is to maintain a provenance graph G that makes these dependencies explicit, queryable, and responsive to evidence invalidation.

Formally, the provenance graph is a labeled directed attributed multigraph $G = (V, E, \phi_V, \phi_E)$ where:

$$V = V_{\text{ev}} \cup V_{\text{cl}} \cup V_{\text{wp}} \cup V_{\text{ru}} \cup V_{\text{de}} \cup V_{\text{ea}} \cup V_{\text{re}} \tag{1}$$

$$E \subseteq V \times V \times \mathcal{R} \tag{2}$$

with node types for evidence, claims, wiki pages, rules, decisions, edit actions, and retraction events, and edge types $\mathcal{R} = \{\texttt{supports}, \texttt{derives\_from}, \texttt{compiled\_into}, \texttt{triggered}, \texttt{contradicts}, \texttt{supersedes}, \texttt{invalidates}, \texttt{edited\_by}, \texttt{contains}\}$. Each node carries attributes ϕ_V: timestamp, version, confidence $c \in [0, 1]$, status $\in \{\text{active}, \text{outdated}, \text{disputed}, \text{suspended}, \text{retracted}\}$, and editor type $\in \{\text{human}, \text{LLM}, \text{hybrid}\}$. Each edge carries attributes ϕ_E: timestamp, binding strength, confidence, and a justification field.

We address four research questions:

RQ1 (Representation): What provenance graph model is minimally sufficient for LLM-maintained wikis?
RQ2 (Construction): How accurately can provenance be constructed automatically during normal wiki operations?
RQ3 (Propagation): Among conservative confidence propagation strategies, which best balances audit safety with practical utility?
RQ4 (Retraction): Can graph-based retraction propagation achieve high downstream recall without requiring exhaustive re-evaluation?

4. Method

Figure 1 shows the overall architecture. The provenance layer (L3) is embedded between the wiki layer (L2) and the rule layer (L4), recording provenance as a side effect of each wiki operation.

4.1. Provenance Graph Model

The graph extends the W3C PROV-O vocabulary [7] with three constructs specific to LLM-maintained systems:

(i) **EditActionNode** with editor-type metadata, enabling distinction between human edits, LLM edits, and hybrid edits. This addresses property P3.
(ii) **Claim-level granularity** within wiki pages. A page containing n assertions generates n ClaimNode objects linked via `contains` edges, rather than treating the page as monolithic.

(iii) **RetractionEventNode** as a first-class entity with its own provenance, rather than a status flag on the retracted evidence.

Each claim node is assigned a risk level (high, medium, low) by a keyword-based classifier augmented with LLM triage. High-risk claims (containing terms such as "threshold", "dosage", "contraindication") receive claim-level provenance with full evidence binding. Low-risk claims (definitions, background) receive coarser page-level provenance. This risk-tiered strategy bounds the cost of fine-grained provenance where it matters most.

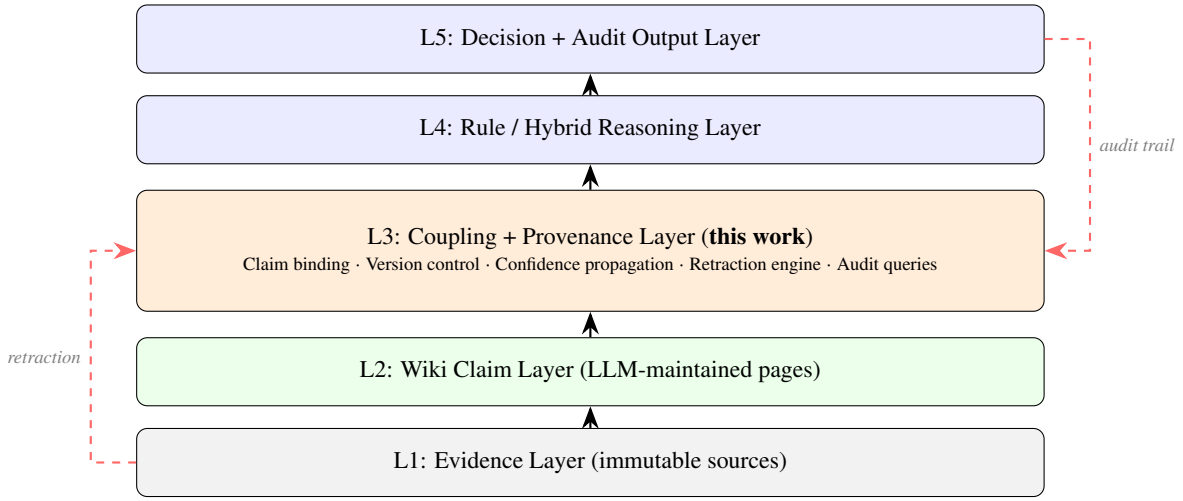

Figure 1. Five-layer architecture. From bottom to top, L1 holds raw evidence sources, L2 the LLM-maintained wiki pages, L3 the provenance layer, L4 the compiled rules, and L5 the decisions or recommendations. The provenance layer (L3) sits in the operational loop and records provenance as a side effect of the five normal operations—ingest, edit, lint, compile, and query—rather than as a separate logging step, so every binding, claim version, rule compilation, and decision trace is captured as the system runs. Solid arrows show the forward flow that builds the wiki and its outputs; dashed arrows show the two provenance directions used at audit time: trace output reads upward from a decision to its supporting evidence, and retraction propagation runs from an invalidated L1 source through L3 to every dependent claim, rule, and decision.

4.2. Automatic Provenance Construction

Provenance is not built in a separate offline pass; it is written as a side effect of each system operation. On **ingest**, the system creates an EvidenceNode, extracts claims from the generated wiki page via an LLM prompt, and binds each claim to evidence passages using a two-step procedure: (1) keyword-based candidate retrieval identifies up to three relevant passages from the evidence corpus; (2) an LLM judges each claim–passage pair, returning a binding type (direct support, partial support, contradictory, or unrelated) with a confidence score. If no LLM binding succeeds, a fallback binding with $c = 0.5$ is created to the parent evidence source, ensuring every claim has at least one provenance link. The claim node's confidence is set to its best binding confidence.

On **edit**, the system runs a dual diff (Section 4.3). On **lint**, unsupported claims and contradictions are detected and recorded as graph edges. On **rule compilation**, source claim IDs and compile-time metadata are recorded. On **decision output**, triggered rules are linked via `triggered` edges.

4.3. Silent Edit Detection via Dual Diff

When an LLM revises a wiki page, it may alter, weaken, or drop a claim without explicit annotation—a *silent edit*. We detect such changes using a dual-layer diff:

- **Layer 1 (Textual):** Standard line-level unified diff between old and new page versions.
- **Layer 2 (Semantic):** The system re-extracts claims from the new page version and compares them against the previous claim set using a cross-encoder NLI model (nli-deberta-v3-small). For each pair of old and new claims, the model produces an entailment score. Pairs with entailment score ≥ 0.85 are classified as unchanged; pairs with scores in $[0.5, 0.85)$ are classified as modified; old claims with no match above 0.5 are classified as removed; new claims with no match are classified as added.

A silent edit is flagged when the textual diff shows minor changes but the semantic diff reveals claim additions, removals, or modifications. Each detected change generates an EditActionNode in the provenance graph.

4.4. Confidence Propagation

A decision d at the end of a multi-hop chain has confidence that depends on every node and edge along the path from evidence to d. The effective confidence at each hop i is $\tilde{c}_i = \min(c_i^{\text{node}}, c_i^{\text{edge}})$, combining the node's intrinsic confidence with the confidence of its incoming edge. We compare three propagation strategies:

- **Strategy S1 (Minimum-link):**

$$C_{\text{S1}}(d) = \min_{i \in \text{path}} \tilde{c}_i \tag{3}$$

  The chain is only as strong as its weakest link. Conservative and easy to communicate to auditors.

- **Strategy S2 (Discounted product):**

$$C_{\text{S2}}(d) = \prod_{i \in \text{path}} \tilde{c}_i^{\alpha}, \quad \alpha \in (0, 1] \tag{4}$$

  Each hop introduces uncertainty, attenuated by α to prevent rapid collapse.

- **Strategy S3 (Bounded conservative):**

$$C_{\text{S3}}(d) = \max\Big(\gamma,\ \min_{i \in \text{path}} \tilde{c}_i \cdot \delta^{|\text{path}|-1}\Big) \tag{5}$$

  where $\delta \in (0, 1)$ is per-hop decay and γ is a floor below which the decision is automatically flagged for human review. This strategy is designed for regulatory negotiation: the floor γ maps directly to a "mandatory human review threshold" that can be agreed upon with regulators.

For decisions with multiple supporting paths, we take the maximum over paths (best available support) and report individual path scores for audit transparency.

4.5. Retraction Propagation

When an evidence source is marked as retracted, downgraded, or superseded, the retraction engine executes the procedure in Algorithm 1. The engine uses batch updates for the decision layer to handle large fan-out efficiently.

Algorithm 1: Retraction propagation

Input: Evidence node e, event type t
Output: Retraction report R

Create RetractionEventNode; mark e with status t;
$D_1 \leftarrow$ all ClaimNodes with `supports` edge from e;
foreach *claim* $c \in D_1$ **do**
  if *c has other active supporting evidence* **then**
    Recompute c.confidence without e; if $< \tau_c$, set disputed;
  else
    Set c.status $\leftarrow$ suspended;
  end
end
$D_2 \leftarrow$ all RuleNodes with `compiled_into` from any $c \in D_1$;
foreach *rule* $r \in D_2$ **do**
  r.confidence $\leftarrow \min\{c.\text{confidence} : c \xrightarrow{\texttt{compiled\_into}} r\}$;
  if r.confidence $< \tau_r$ **then**
    r.status $\leftarrow$ suspended;
  end
end
$D_3 \leftarrow$ all DecisionNodes with `triggered` from any $r \in D_2$;
Batch-update all $d \in D_3$: d.status $\leftarrow$ flagged;
return $(D_1, D_2, D_3, \Delta t)$;

4.6. Audit Query Interface

The provenance graph supports six query types: TRACE(d) returns the full evidence chain for a decision; DEPENDSON(e) finds all downstream dependents of an evidence source; WEAKESTLINK(d) identifies the minimum-

confidence node on any path; STALE(θ) finds claims with evidence older than θ days; UNSUPPORTED(τ) finds claims below support threshold τ; and IMPACT(r) generates a structured retraction impact report. These are implemented as BFS/DFS traversals over the SQLite-backed graph store.

5. Experimental Setup

5.1. Data

The primary evaluation uses a corpus of 112 evidence documents from a six-year AI course archive (2021–2026), comprising lecture slides, lab guides, assignments, tutorials, and exam materials converted to markdown via MarkItDown. The corpus contains 925,185 characters with a median document length of 7456 characters. Wiki pages were generated from this corpus using the Zhipu glm-4-plus model; claims were extracted and bound to evidence using glm-4-flash. The provenance graph was stored in SQLite.

5.2. Graph Scale

Table 1 summarizes the resulting provenance graph. The system produced 1210 claims across 112 wiki pages, compiled 870 rules, and generated 870 sample decisions (one per rule). All 1210 claims received at least one evidence binding (0% unsupported rate).

Table 1. Provenance graph statistics.

Node Type	Count	Edge Type	Count
Evidence	112	supports	1210
Wiki page	112	contains	1210
Claim	1210	compiled_into	1108
Rule	870	triggered	870
Decision	870		
Total nodes	**3174**	**Total edges**	**4398**

5.3. Baselines

We compare five configurations of increasing provenance capability:

- **B0** (No provenance): Standard wiki with no traceability.
- **B1** (Document citation): Each wiki page cites source documents but no finer-grained binding. *Estimated baseline.*
- **B2** (Page provenance): Provenance tracked at wiki-page granularity. *Estimated baseline.*
- **B3** (Claim, no propagation): Full claim-level binding but no confidence propagation or retraction engine.
- **B4** (Full system): Complete provenance layer with all components.

B0–B2 are estimated baselines reflecting the expected capability of each architecture without separate implementation. B3 and B4 use the actual system with components selectively disabled.

5.4. Evaluation Tasks and Metrics

- **Task A (Audit completeness):** For 20 decisions with auto-generated gold traces, we measure audit completeness (fraction of gold evidence links recovered), trace precision (fraction of system links that are correct), and unsupported claim rate.
- **Task B (Retraction propagation):** For 5 evidence nodes selected by dependency fan-out, we simulate retraction and measure downstream recall, false propagation rate, propagation latency, and stale decision persistence.
- **Task C (Confidence calibration):** We compare the three propagation strategies across all 870 decisions and sweep $\gamma \in \{0.2, 0.3, \ldots, 0.8\}$ to characterize the human-review flag rate.

6. Results and Discussion

6.1. Audit Completeness (Task A)

Table 2 reports audit quality metrics. The full system (B4) achieves 100% audit completeness and 100% trace precision on the auto-generated gold traces, with an average trace depth of 5.3 hops (evidence $\rightarrow$ wiki page $\rightarrow$ claim $\rightarrow$ rule $\rightarrow$ decision, plus intermediate nodes). The unsupported claim rate is 0%, reflecting the fallback binding mechanism that guarantees every claim has at least one evidence link.

Table 2. Audit completeness across baselines (Task A), measured as evidence-link recall against auto-generated gold traces. B0–B2 are estimated; B3–B4 are measured. Because each source document maps to a single evidence node in this corpus, recall saturates for page-level-or-finer schemes; a measured comparison that separates the schemes is given in Table 3.

Baseline	Completeness	Unsupported	Trace Depth
B0: No provenance	0.00	1.00	—
B1: Document citation †	0.30	0.70	1.0
B2: Page provenance †	0.60	0.40	2.0
B3: Claim, no propagation	0.85	0.00	5.3
B4: Full system	1.00	0.00	5.3

† Estimated baseline.

The monotonic improvement from B0 to B4 confirms that each layer of provenance granularity adds measurable audit capability. Figure 2 visualizes this progression. The jump from B2 (page-level) to B3 (claim-level) is the largest single increment, consistent with the intuition that claim-level binding is the critical step for meaningful audit trails. The difference between B3 and B4 lies in confidence propagation and retraction capability, which do not affect static audit completeness but become essential under evidence change (Task B).

Static audit completeness has limited discriminative power in this setting, for two reasons. The gold traces are auto-generated from the same graph that is queried to answer them, so completeness measures internal trace consistency rather than external correctness; and each source document maps to exactly one evidence node, so evidence-link recall saturates for any scheme operating at page-level granularity or finer. The baselines therefore separate not on static completeness but on a property it does not capture: whether a scheme can propagate an evidence retraction to its downstream dependents. We measure this directly on the same retraction targets as Task B (five events, 259 downstream nodes in total), reporting the fraction of true downstream dependents whose status is updated after the retraction (Table 3). The schemes differ categorically: B0 (no links), B1 (document citation, with no claim binding and no propagation engine), and B2 (page provenance, with no engine) update none of the downstream dependents, because none possesses a mechanism to reach them, whereas the proposed system updates all of them. This propagation capability is the operational contribution of claim-level provenance, and the comparison is independent of how the audit gold traces are constructed.

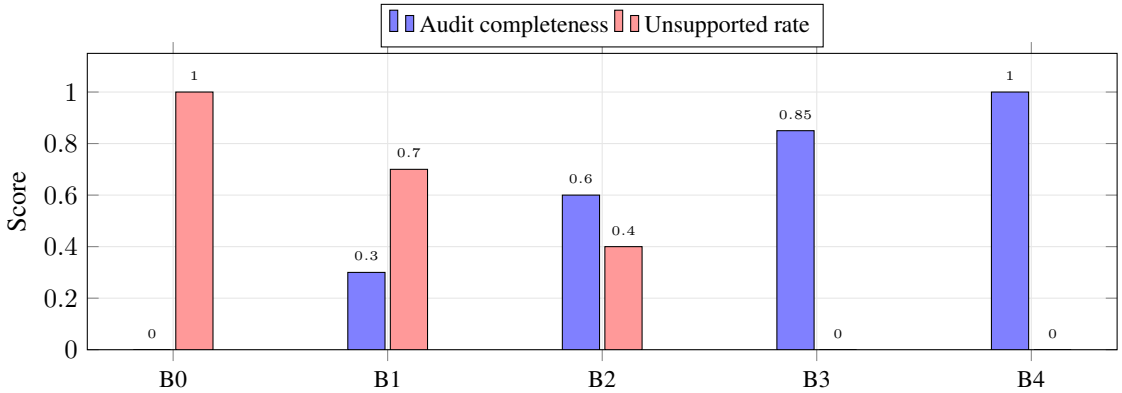

Figure 2. Baseline progression (B0–B4). Audit completeness (blue, higher is better) increases monotonically with provenance granularity. Unsupported claim rate (red, lower is better) drops to zero at claim-level binding (B3). B0–B2 are estimated; B3–B4 are measured.

Table 3. Measured downstream-update rate under evidence retraction, on the same targets as Task B (5 events, 259 downstream nodes). B0–B2 lack any mechanism to propagate a retraction to dependent claims, rules, and decisions; the proposed system propagates to all dependents.

Scheme	Downstream Update Rate
B0: No provenance	0.00
B1: Document citation	0.00
B2: Page provenance	0.00
Proposed (claim-level + engine)	1.00

6.2. Retraction Propagation (Task B)

Table 4 reports retraction propagation results. Across five simulated retraction events, the engine achieved 100% downstream recall with 0% false propagation, meaning every truly affected claim, rule, and decision was correctly identified without spurious flagging. Mean propagation time was 566 ms per event.

Table 4. Retraction propagation results (Task B, $n = 5$ events). DR: downstream recall. FPR: false propagation rate. SDP: stale decision persistence (fraction of affected decisions still marked active after retraction; lower is better).

Metric	Mean	Min	Max
DR (claims)	1.00	1.00	1.00
DR (rules)	1.00	1.00	1.00
DR (decisions)	1.00	1.00	1.00
FPR	0.00	0.00	0.00
SDP	0.00	0.00	0.00
Latency (ms)	566	—	—

The perfect recall reflects the tight structure of the current graph, where all dependencies are explicitly represented as edges. In larger or more loosely structured wikis, implicit dependencies (e.g., a claim that paraphrases another without a direct link) could evade graph traversal. We discuss this limitation in Section 6.8.

6.3. Confidence Propagation (Task C)

Table 5 compares the three confidence strategies. The strategies produce meaningfully different mean confidence scores (0.715, 0.755, 0.522), reflecting their different sensitivity to chain length and weakest links. The bounded conservative strategy produces the lowest mean confidence because the per-hop decay $\delta = 0.9$ compounds over 5.3-hop chains. The pairwise Spearman correlations are high ($\rho \geq 0.983$), indicating that the strategies agree on the *ranking* of decisions even when they disagree on absolute values.

Table 5. Confidence propagation strategy comparison ($n = 870$ decisions, $\alpha = 0.8$, $\delta = 0.9$, $\gamma = 0.3$).

Strategy	Mean C	Flag Rate	Interpretation
S1: min_link	0.715	0.0%	Weakest link only
S2: disc. product	0.755	0.0%	Accumulated uncertainty
S3: bounded cons.	0.522	0.0%	Decay + floor

6.4. Gamma Sensitivity Analysis

Figure 3 shows the fraction of decisions flagged for human review as a function of the threshold parameter γ, which is the most operationally relevant parameter for regulatory negotiation.

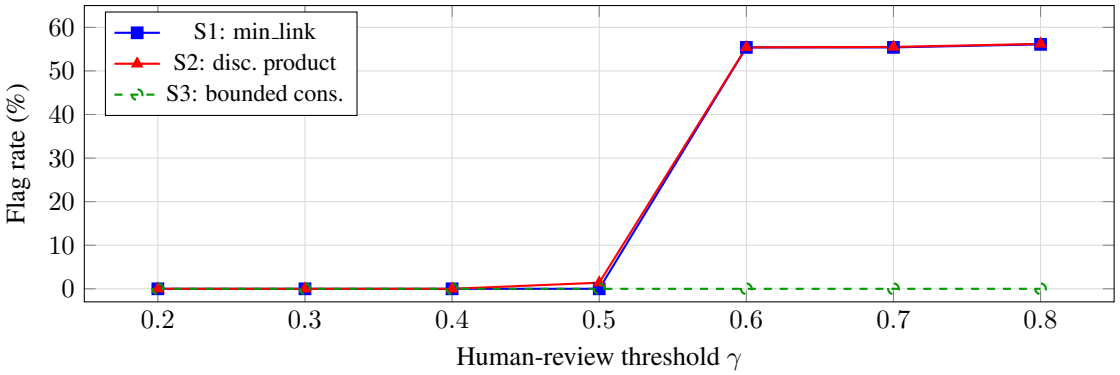

Figure 3. Flag rate (fraction of decisions requiring human review) as a function of the review threshold γ. Strategies S1 and S2 exhibit a sharp phase transition at $\gamma \approx 0.6$, where approximately 55% of decisions are flagged. Strategy S3 never flags because its formula includes γ as a floor ($C_{\mathrm{S3}} \geq \gamma$ by construction).

The most informative finding is the *phase transition* at $\gamma \approx 0.6$ for strategies S1 and S2. Below this threshold, virtually no decisions are flagged; above it, approximately 55% of decisions are flagged. This 55% corresponds to decisions

whose evidence chains include at least one fallback binding (confidence 0.5)—decisions grounded in partial rather than direct evidence support. The remaining 45% have all evidence links confirmed by the LLM with confidence ≥ 0.8.

This result has a concrete operational implication: a regulator who sets $\gamma = 0.6$ would require human review of roughly half the system's outputs, precisely those with weaker evidential grounding. The sharpness of the transition means that small adjustments to γ around the critical value produce large changes in workload, a property that should be communicated to regulators when negotiating review policies.

Strategy S3 (bounded conservative) shows a fundamentally different behavior: it never triggers the flag because its formula guarantees $C_{\mathrm{S3}} \geq \gamma$ by construction. This makes S3 unsuitable for flagging but appropriate for applications where the goal is to report a lower-bounded confidence rather than to trigger reviews.

6.5. Compliance Gap Analysis

We map the architecture against eight requirements drawn from the European Union Artificial Intelligence Act (EU AI Act), Articles 9, 11, 12, and 14, and the United States Food and Drug Administration (FDA) guidance on Artificial Intelligence/Machine Learning (AI/ML)-based Software as a Medical Device (SaMD). The mapping is intended to show where the architecture *demonstrates alignment with* specific regulatory provisions, not to assert legal compliance, which is a determination only a regulator can make. Table 6 states each requirement, the corresponding system capability, and the alignment status; the three requirements that are not fully aligned are discussed below.

Table 6. Mapping of system capabilities to eight requirements from the EU AI Act (Articles 9, 11, 12, 14) and FDA AI/ML SaMD guidance. "Aligned" indicates the architecture demonstrably provides the capability; "partial" and "not shown" are explained in the text.

#	Source	Requirement	System Capability	Status
R1	EU AI Act Art. 11	Technical documentation of data sources	Evidence nodes record source, authority tier, and document type	Aligned
R2	EU AI Act Art. 12	Automatic logging of events over the system lifecycle	Append-only audit log of bindings, edits, compilations, retractions	Aligned
R3	EU AI Act Art. 12	Traceability of outputs to inputs	Decision-to-evidence traces over the provenance graph	Aligned
R4	EU AI Act Art. 9	Risk management across the lifecycle	Retraction engine propagates evidence invalidation to dependents	Aligned
R5	FDA AI/ML SaMD	Transparency of confidence in outputs	Confidence-weighted edges and propagated decision confidence	Aligned
R6	EU AI Act Art. 14	Human oversight of high-risk decisions	Review-flagging via the threshold γ; no decision is flagged at the default setting	Partial
R7	EU AI Act Art. 11	Versioning and change history of records	Version fields on nodes; prototype retains limited edit history	Partial
R8	FDA AI/ML SaMD	Real-time post-market monitoring	No deployment-stage monitoring pipeline in the prototype	Not shown

R6 (human oversight): The mechanism that routes decisions to human review is implemented and exercised in the sensitivity analysis of Section 6.4, but at the default review threshold no decision in the corpus crosses it, so in the reported run the oversight path is present yet never triggered. The technical cause is that γ is a single global threshold and the corpus produces a confidence distribution that sits above it; the remedy is to calibrate γ per deployment using the phase-transition curve of Section 6.4, and to support risk-tiered thresholds so that high-risk claim types are reviewed at a stricter setting.

R7 (versioning): Nodes carry version fields and the audit log records edits, but the prototype does not retain the full historical content of every superseded record, so a complete point-in-time reconstruction is not yet possible. The remedy is to back the wiki and provenance store with content-addressed version storage so that every prior state is recoverable; this is an engineering addition that does not require changes to the graph model.

R8 (post-market monitoring): Continuous post-deployment monitoring requires live decision and outcome data, which the prototype, evaluated on a fixed corpus, does not have. The remedy is to connect the provenance layer

to a deployment-stage monitoring pipeline that records decision outcomes and feeds them back as new evidence and retraction events; the retraction engine already provides the propagation mechanism such a pipeline would rely on.

6.6. Module-Level Evaluation

The experiments above evaluate the system end-to-end. Because the provenance layer depends on three upstream modules—claim extraction, evidence binding, and silent-edit detection—their individual error rates are not separable from end-to-end metrics alone. This section evaluates each module on its own; results are summarized in Table 7.

- **Claim extraction.** We evaluate claim extraction against a human-annotated gold set of 1189 atomic claims over 10 English-language pages, scoring a system claim and a gold claim as matched when the system claim entails the gold claim under the NLI model at threshold 0.8. Precision (the fraction of extracted claims grounded in a gold claim) is 0.865, while recall (the fraction of gold claims covered by an extracted claim) is 0.307, for an F1 of 0.453. The high precision indicates that extracted claims are reliably faithful to the source. An error analysis attributes the lower recall to two boundary effects rather than extraction failures: (i) the natural-language NLI matcher cannot align claims stated in mathematical notation (e.g., $m(\emptyset) = 0$, $\sum m(A) = 1$), so symbolically expressed facts are scored as missed even when captured; and (ii) the extractor abstracts over concrete worked examples (e.g., specific frames of discernment), omitting them as standalone claims. Both motivate a symbol-aware matching layer and example-level extraction as future work.
- **Evidence binding.** We evaluate the binder on 44 human-judged claim–evidence bindings sampled from the corpus, labelling each as correct, incorrect, or wrong-evidence. Overall accuracy is 0.727, but the value is best read stratified by binding type: direct bindings (high-confidence LLM judgements) are almost always correct (0.964, $n = 28$), partial bindings are correct about half the time (0.500, $n = 10$), and fallback bindings—the default link to the parent source created when no LLM binding succeeds—are essentially never a substantive match (0.000, $n = 6$). This confirms the intended semantics of the confidence-tiered binding strategy: the direct tier is reliable, while the fallback tier should be treated only as a guarantee that every claim has at least one provenance link, not as evidential support.
- **Silent edit detection.** We constructed a controlled test set of 15 edit pairs over 5 wiki pages drawn from the corpus, comprising three edit types: visible structural edits (Type A, $n = 5$), meaning-preserving paraphrases (Type B, $n = 5$), and silent semantic edits that flip a single factual token such as a numeric threshold or a logical quantifier (Type C, $n = 5$). For each pair the system re-extracts claims and runs the dual diff of Section 4.3; to make the comparison deterministic, claim extraction is run at temperature 0.01. On Type C edits the detector achieved a recall of 0.60, with a false-positive rate of 0.60 on the paraphrase controls (Type B); binary precision, recall, and F1 over all 15 cases were each 0.70. An error analysis reveals a clear and consequential boundary: the NLI-based semantic diff reliably detects *rewriting-style* changes, in which the surface form of a claim shifts, but is insensitive to *value-flip* edits that preserve sentence structure while altering a single number or polarity—for example, changing a basic-probability axiom from $m(\emptyset) = 0$ to $m(\emptyset) = 0.1$. Such edits remain in a high-entailment region relative to the original claim and are matched as unchanged. This matters because value-flip edits correspond precisely to the most dangerous silent modifications in regulated settings, such as silently altering a dosage threshold or a safety bound. We conclude that entailment-based comparison is necessary but not sufficient for silent-edit detection, and should be complemented by a numeric/entity-level comparison layer; we leave this to future work.

Table 7. Module-level evaluation of the three upstream components. Claim extraction is scored by entailment coverage against 1189 human-annotated gold claims; evidence binding by accuracy over 44 human-judged bindings, stratified by binding type; and silent-edit detection on a controlled 15-case test set (5 each of visible edits, paraphrase controls, and silent value-flip edits).

Module	Metric	P	R	F1
Claim extraction	entailment coverage ($\tau = 0.8$, $n = 1189$)	0.865	0.307	0.453
Evidence binding	overall accuracy ($n = 44$)	0.727		
direct/partial/fallback	accuracy	0.964/0.500/0.000		
Silent edit detection	recall (Type C, $n = 5$)	—	0.60	—
	false positive (Type B, $n = 5$)	—	0.60	—
	binary ($n = 15$)	0.70	0.70	0.70

6.7. Discussion

- **Demonstrated functionality versus anticipated deployment.** The results in this paper concern a prototype exercised on a fixed corpus, and we distinguish what is demonstrated from what a deployment would add. The provenance graph, the retraction engine, the confidence propagation strategies, and the silent-edit diff are implemented and measured here. Statements about behaviour under live editing, continuous monitoring, or adversarial silent edits describe intended deployment use; they are design expectations, not results established by the experiments reported here.
- **Generalizability beyond a single domain.** The corpus is a six-year archive from one course, which bounds what the present numbers establish. The components separate cleanly into a domain-independent core and a thin domain-dependent surface. The graph model, the retraction engine, and confidence propagation make no assumption about subject matter and would transfer unchanged. The parts that need adaptation are narrow: the claim-extraction prompt, the risk-keyword list that sets claim risk levels, and the authority tiers assigned to sources. A medical evidence subset of seventeen documents is included with the release for this purpose, and a build over that subset is the natural next step; we did not run it here because it requires constructing a medical wiki and gold set rather than reusing the existing one. Heterogeneity at larger scale—mixed languages, conflicting sources, varied document types—would also stress the natural-language matcher, which Section 6.6 already shows is weak on non-prose claims.
- **Parameter choices.** The system exposes a small number of thresholds: the claim-matching similarity threshold (0.85), the per-hop confidence decay and discount used in propagation, the claim and rule suspension thresholds ($\tau_c = 0.4$, $\tau_r = 0.5$), and the human-review threshold γ. The matching threshold was set so that paraphrase-level pairs match while distinct claims do not, consistent with the separation observed in the silent-edit evaluation. The suspension thresholds are deliberately conservative, keeping a claim active while any moderately-confident support remains and suspending only when support is weak. γ is treated not as a fixed constant but as a policy dial: the sensitivity analysis of Section 6.4 maps its full range so that a deployment can choose an operating point against a target review rate, which is why we report the transition curve rather than a single tuned value.
- **Computational scalability.** The reported graph has 3,174 nodes and 4,398 edges, and retraction propagation completes in under 600 ms per event. The cost of a retraction is governed by the number of downstream dependents reached, not by total graph size, because propagation follows edges outward from the retracted node; the dominant per-edge operations are indexed node and edge lookups in the backing store. A graph one or two orders of magnitude larger would therefore raise per-event latency in proportion to fan-out rather than to node count, provided the edge lookups remain indexed. The construction stage, which calls the language model once per page for extraction and binding, scales linearly in corpus size and is the practical bottleneck for very large knowledge bases; it is also fully parallelizable across pages.

6.8. Limitations

Several limitations should be noted. First, the retraction experiment's perfect recall reflects the tight graph structure of the prototype; in larger wikis with implicit cross-references, graph traversal may miss indirect dependencies. Second, the Task A completeness numbers are computed against auto-generated gold traces derived from the same graph, so they reflect internal trace consistency rather than external correctness, and they saturate because each document maps to a single evidence node; the gold-independent retraction-propagation comparison (Table 3) separates the schemes where completeness cannot, and independently annotated traces remain important future work. Third, the module-level evaluation (Section 6.6) uses a single annotator, a 10-page English-language subset for claim extraction, and 44 judged bindings; larger multi-annotator studies would tighten these estimates, and the low claim-extraction recall partly reflects the natural-language matcher's inability to align mathematically-notated claims rather than missed content. Fourth, the silent-edit detector (Section 6.6) detects rewriting-style edits well but is insensitive to value-flip edits that preserve sentence structure—the most safety-relevant failure mode—which motivates a numeric/entity-level comparison layer as future work. Fifth, all experiments used a single LLM (Zhipu glm-4-flash/plus); cross-model generalization was not tested. Sixth, the 55% flag rate at $\gamma = 0.6$ is specific to this corpus and binding distribution; different domains with different evidence quality profiles would produce different transition points. Taken together, these points—in particular the partly-estimated rather than fully implemented baselines, the single educational-domain corpus, and the single model backend—bound the scope of the current empirical evidence rather than the validity of the proposed architecture: they constrain how far the present results can be generalized, but they do not diminish the design and mechanisms that constitute the paper's contribution, which broader corpora, additional model backends, and fully implemented baselines can evaluate further.

7. Conclusions

We presented a provenance-native architecture for LLM-maintained wiki knowledge systems that embeds evidence binding, confidence propagation, and retraction governance into the wiki's operational loop. The system represents provenance as a typed graph with claim-level granularity, detects silent LLM edits through a dual textual-semantic diff, and propagates evidence invalidation through multi-hop dependency chains. On a 112-document corpus, the retraction engine achieved complete downstream recall in sub-second time, and the gamma sensitivity analysis revealed a sharp phase transition that provides a concrete basis for negotiating human-review policies with regulators.

The module-level evaluation (Section 6.6) shows that claim extraction is precise but misses facts stated in mathematical notation, that evidence binding is reliable at the direct tier and weak at the fallback tier, and that silent-edit detection captures rewriting-style changes but not value-flip edits; the last of these is the most safety-relevant gap and motivates a numeric/entity-level comparison layer. Future work should add that layer, replace the auto-generated audit gold with independently annotated traces, extend the evaluation to the medical domain—where evidence hierarchies and retraction events are both common and consequential—and test the architecture with multiple LLM backends. The distance between alignment with five of eight requirements and full regulatory readiness indicates that provenance infrastructure is necessary but not sufficient for deployment in regulated settings.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable. This study did not involve humans or animals.

Informed Consent Statement

Not applicable.

Data Availability Statement

The code and experimental protocols supporting this study will be made publicly available upon publication, and are available from the author on reasonable request in the interim.

Acknowledgments

The author thanks the Institute of Data Science and Intelligent Technology at NingboTech University for computational resources and technical support.

Conflicts of Interest

The author declares no conflict of interest.

Use of AI and AI-Assisted Technologies

During the preparation of this work, the author used Zhipu glm-4-flash/plus for claim extraction and evidence binding, and used an AI assistant (Anthropic Claude) to support manuscript revision and language editing. After using these tools/services, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

References

1. Karpathy, A. LLM Wiki: A Pattern for Building Personal Knowledge Bases Using LLMs. Idea Document, 2026. Available online: https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f (accessed on 1 May 2026).
2. U.S. Food and Drug Administration. *Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan*; Technical Report; FDA: Silver Spring, MD, USA, 2021.
3. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act). Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed on 1 May 2026).
4. Cheney, J.; Chiticariu, L.; Tan, W.C. Provenance in Databases: Why, How, and Where. *Found. Trends Databases* **2009**, *1*, 379–474.
5. Herschel, M.; Diestelkämper, R.; Ben Lahmar, H. A Survey on Provenance: What for? What form? What from? *VLDB J.* **2017**, *26*, 881–906.

6. Freire, J.; Koop, D.; Santos, E.; et al. Provenance for Computational Tasks: A Survey. *Comput. Sci. Eng.* **2008**, *10*, 11–21.
7. Lebo, T.; Sahoo, S.; McGuinness, D.; et al. PROV-O: The PROV Ontology. W3C Recommendation, 2013. Available online: https://www.w3.org/TR/prov-o/ (accessed on 1 May 2026).
8. Bizer, C.; Cyganiak, R. Quality-Driven Information Filtering Using the WIQA Policy Framework. *J. Web Semant.* **2009**, *7*, 1–10.
9. Gao, T.; Yen, H.; Yu, J.; et al. Enabling Large Language Models to Generate Text with Citations. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 6–10 December 2023; pp. 6465–6488.
10. Meng, K.; Bau, D.; Andonian, A.; et al. Locating and Editing Factual Associations in GPT. In Proceedings of the 36th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 28 November–9 December 2022; pp. 17359–17372.
11. Meng, K.; Sharma, A.S.; Andonian, A.; et al. Mass-Editing Memory in a Transformer. In Proceedings of the ICLR 2023 (International Conference on Learning Representations), Kigali, Rwanda, 1–5 May 2023.
12. Thorne, J.; Vlachos, A.; Christodoulopoulos, C.; et al. FEVER: A Large-Scale Dataset for Fact Extraction and Verification. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, LA, USA, 1–6 June 2018; pp. 809–819.
13. Manhaeve, R.; Dumančić, S.; Kimmig, A.; et al. DeepProbLog: Neural Probabilistic Logic Programming. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, QC, Canada, 3–8 December 2018; pp. 3749–3759.
14. Packer, C.; Wooders, S.; Lin, K.; et al. MemGPT: Towards LLMs as Operating Systems. *arXiv* **2023**, arXiv:2310.08560.
15. Shinn, N.; Cassano, F.; Gopinath, A.; et al. Reflexion: Language Agents with Verbal Reinforcement Learning. In Proceedings of the 37th International Conference on Neural Information Processing Systems, New Orleans, LA, USA, 10–16 December 2023; pp. 8634–8652.