AI Git: Unified Knowledge Provenance Framework for Safe Scaling

(Preventing Model Collapse via Anti-Self-Training and Federated Provenance Controls)

1. Core Architecture

A. Tiered Semantic Fingerprint Engine (Temporal + Clustered)

Scope Control:
Fingerprint only long-form completions, high-complexity reasoning, and multi-source synthesis, excluding trivial or ephemeral queries (e.g., short prompts, boilerplate).

Dual Embedding Spaces:
Current model embeddings for present relevance.

Frozen historical anchors to detect older outputs across model generations.

Embedding Timewarp Functions (Novel Invention): universal translation layer enabling cross-generational semantic fingerprint comparison, preserving detectability of older outputs as embedding models evolve.

Compression & Storage:
Quantized centroids + deltas for efficient representation (>90% reduction).

Fingerprints infrequently accessed are demoted to cost-efficient cold storage, retaining accessibility without bloating active memory.

B. Provenance & Attribution Registry (Privacy-Preserved)

Auto-tag every output:
Session pseudonym (hashed tokenization for GDPR/CCPA compliance).

Context metadata: abstracted topic labels (e.g., "children's literature," "legal memo").

Opt-in user attribution for named credit.

Cryptographic Provenance Stamping (Novel Invention): Invisible watermark tokens persisting across paraphrase/stripping.

Metadata stored separately from fingerprints in a secure enclave.

C. AI Git Version Tree (Authorship Lineage)

Commit Nodes:
Every output → commit node (AI root or human edit).

Branches track deltas:
Semantic similarity score.

Token-level edit distance.

Lineage tracking logs delta score plus external corroboration score, weighting human-added citations tied to independent factual sources.

Preservation of human enrichment:
E.g., AI-assisted draft later revised and published with new references remains trainable.

D. Delta Tracking Engine

Dual metrics:
Cosine similarity (semantic) for meaning preservation.

Token edit distance to block trivial rewrites.

Thresholds:
Identical (≥0.98): Suppress.

Paraphrase (0.85–0.98): Down-weight.

Novel (<0.85): Retain as enriched/human-guided.

2. Training Data Filter Pipeline

Exact Hash Deduplication (Epochal)

Remove verbatim matches using hash comparison against fingerprint DB.

Hash deduplication is performed in epochal batch passes (rather than token-by-token) to reduce compute overhead.

ANN Vector Comparison + Neural Novelty Estimator

ANN vector comparisons also run in offline batch epochs via centroid pruning, separate from runtime inference.

Domain-specific thresholds (legal, code, literature).

Neural Novelty Estimator (Novel Invention): small LLM that distinguishes trivial paraphrases from real enrichment.

Citation & Contextual Enrichment Scoring

Detect citations, DOIs, quotes.

Enrichment scoring explicitly cross-validates against human-authored reference corpora to anchor provenance to verified factual sources.

Score factual additions and unique entities.

Provenance Weighting

High Weight: Human-enriched, provenance-tagged, citations present.

Medium Weight: Novel but unprovenanced outputs.

Low/Archive: Clearly marked AI quotes are archived at low weight rather than discarded, preserving lineage without polluting high-weight training data.

Synthetic Cross-Check (Multi-Signal)

Burstiness detection, token entropy, watermarking to identify disguised AI text.

Suppress unless provenance verifies human-guided editing.

3. Governance & Federation

Rollout: Initial internal rollout for closed-loop filtering during next-generation model pretraining.

Neutral Standards Body: IEEE/W3C-like provenance registry for cross-lab alignment.

Zero-Knowledge Proof API: ZK-proof API supports cross-lab queries—'has this fingerprint appeared?'—without exposing raw data.

Model Certificate Authority (Novel Invention): Introduce an SSL-style Model Certificate Authority to cryptographically verify provenance across labs.

Federated Provenance Exchange: Signed fingerprint queries prevent cross-model contamination.

4. Rare Knowledge Preservation

Frequency-weighted fallback: Override suppression heuristics for rare knowledge items flagged near-duplicate to prevent factual attrition.

Knowledge Importance Estimator: Scores factual rarity + corroboration.

5. Human Review & Automation

Auto-escalation: Flag cases within a 45–55% classifier confidence band for human audit via moderation dashboards.

Cluster moderation: One representative per cluster sampled, not every output.

AI Moderation Dashboards (Novel Invention): LLMs pre-summarize flagged clusters for lightweight human approval.

6. Vulnerability Mitigations

Scalability:

ANN vector DB (FAISS/Milvus) with centroid pruning.

Tiered storage reduces infra cost.

Adversarial Evasion:
Combined token edit + semantic scores block trivial paraphrase bots.

High-authority injection audits: Synthetic-looking text from prestigious sources triggers review.

Privacy & Compliance:
Pseudonymized session tokens.

Metadata separation and secure enclave storage.

Explicit opt-in attribution controls.

Ephemeral Content:
Maintain explicit whitelists for boilerplate/low-semantic text (e.g., greetings) to prevent over-aggressive suppression.

7. Implementation Roadmap

- Phase 1: Fingerprint DB + hash deduplication.

- Phase 2: Semantic diff + citation/enrichment parser.

- Phase 3: Provenance-tagged AI/human enrichment weighting.

- Phase 4: Federated provenance API with zero-knowledge proofs + Model Certificate Authority.

- Phase 5: Embedding timewarp + cryptographic watermarking rollout.

8. Resulting Benefits

Scalable: Storage compression + clustering.

Safe: Prevents recursive self-ingestion and model collapse.

Interoperable: Built for cross-lab standards and API federation.

Knowledge-preserving: Protects rare, factual outputs.

Human-guided: Weighting ensures AI remains human-anchored.

Human-in-the-loop: Ambiguous or borderline cases routed to lightweight human moderation dashboards ensure oversight at scale.