The Entropy of Recursion: A Strategic Framework for Provenance-Verified Data and the Preservation of Signal Integrity in Generative Systems*

Tad MacPherson
Metavolve Labs, Inc., Research Division
research@metavolvelabsinc.com

Abstract

The rapid scaling of generative models has precipitated a data integrity challenge characterized by "Model Collapse"—a degenerative process where recursive training on synthetic output drastically reduces statistical variance and induces hallucination. This paper argues that the prevailing paradigm of indiscriminate Data Accumulation must yield to a new era of Data Stewardship. We introduce the Provenance-Embedded Semantic Transport (PEST) framework, a novel architecture that inextricably binds semantic context to media artifacts via a robust, steganographic Persistent Provenance Pointer (PPP). Unlike fragile header-based metadata standards which are easily stripped, the PPP survives compression and adversarial transformation, linking the artifact to a decentralized, immutable ledger containing verified provenance and dense semantic ontologies. Furthermore, we propose a Human-AI Hybrid Curation model that combines high-variance human artifacts with machine-generated semantic scaffolding to establish "High-Density Ground Truth" datasets. We conclude by outlining the economic implications of this framework, specifically the emergence of a Verification Economy where cryptographically verified human data becomes a scarce, priced asset class essential for stabilizing future foundation models.

Keywords: Model Collapse, Training Data Valuation, Provenance, Steganography, Human-in-the-Loop, Data Shapley, GLAM

Author Information and Verification

This paper utilizes the Human-Hybrid paradigm it describes. The theoretical framework, strategic vision, and technical specifications were authored by human researchers. While large language models were utilized for literature synthesis and drafting support, all technical assertions, mathematical definitions, and citations have been rigorously verified by human domain experts to ensure accuracy and prevent hallucination.

*DOI: 10.5281/zenodo.18436975

1 Introduction

1.1 The Saturation of Data Accumulation

We stand at a critical juncture in the evolution of machine learning. For the past decade, the dominant strategy has been the era of Data Accumulation. Governed by the Kaplan scaling laws, model performance has been linked via power laws to the volume of compute and the quantity of data ingested. This paradigm relied on a critical assumption: that the internet functions as an abundant source of independent, identically distributed (i.i.d.) human-generated data—images, text, and code that reflect the high-variance complexity of the physical world.[1] However, the proliferation of generative systems has fundamentally altered the **informational composition** of the internet. As synthetic content floods digital channels—accounting for a rapidly increasing percentage of all new text and images—the internet is transforming from a repository of human experience into a closed feedback loop. Future models are increasingly training on the outputs of their predecessors.[3] This recursive training dynamic introduces a degenerative entropy known as the "Xerox Effect" or "Model Autophagy," where the subtle artifacts, biases, and errors of one generation become the ground truth for the next.[5]

1.2 The Crisis of Integrity and Context

The crisis is not merely one of data quantity—though the exhaustion of high-quality public text is predicted by Epoch AI to occur as early as 2026—but of integrity and context.[1] In the wild ecosystem of the web, metadata is often treated as disposable. The "semantic context" of an asset—the metadata that carries provenance, intent, emotional framing, and usage rights—is systematically stripped by compression algorithms, social platforms, and aggregators.[8] Consequently, models are trained on data stripped of context: pixels without history, text without intent. This decoupling of content from context is a primary driver of hallucination and the loss of semantic density in model outputs.[10] When a model ingests an image of a "19th-century coal miner" but receives only the visual pixel data without the historical metadata, it learns the texture of "dirt on face" but fails to learn the ontological reality of "labor" or "extraction." It learns the surface features, but not the underlying causal relationships.

1.3 The Mandate for Stewardship

This paper argues that the era of Data Accumulation must yield to the era of Data Stewardship. The strategy of "bigger is better" is hitting the wall of data contamination at scale. We posit that the only viable path forward is a re-architecture of how data is published, verified, and consumed for training. We must move beyond fragile, detached metadata standards (like JSON sidecars or EXIF headers) toward a protocol of Recursive Integrity, where provenance and semantic context are immutably embedded within the artifact itself using advanced steganography and decentralized storage.[12] Furthermore, we challenge the binary view that "synthetic data is poison," proposing instead a Human-AI Hybrid Curation model where human-generated artifacts (preserving variance) are augmented by machine-generated semantic scaffolding (providing density) to create "High-Density Ground Truth" datasets.[14]

2 Literature Review

2.1 The Physics of Recursive Degeneration: Model Collapse

The theoretical foundation of this crisis lies in the phenomenon of "Model Collapse," solidified in the 2024 peer-reviewed literature by Shumailov et al. in *Nature*.[5] Their work demonstrates that when generative models are trained recursively on their own outputs, they eventually converge to a state of extremely low variance, effectively "forgetting" the tails of the underlying data distribution. Shumailov's findings suggest that generative models function as lossy compressors. When a model samples from its learned distribution to generate synthetic data, it inevitably over-samples the modes (high-probability regions) and under-samples the tails (low-probability, high-information regions).[17] The "tails" are critical because they are where nuance, eccentricity, and the "entropy of reality" reside. In a recursive loop, this variance reduction compounds. Generation N captures 90% of the variance; Generation $N+1$ captures 90% of the remaining 90%. Over time, the distribution narrows to a point estimate, resulting in outputs that are generic, repetitive, and devoid of the complexity that characterizes true intelligence.[18] Dohmatob et al. (2024) expanded on this by distinguishing between "weak" and "strong" model collapse. They identified a "generalization-to-memorization" transition, where models in the late stages of autophagy cease to generalize and instead begin to memorize the synthetic artifacts of their training set, reinforcing errors rather than learning principles.[17] This degradation results in a loss of contact with reality, where the model begins to infer relationships governed by the artifacts of the previous model rather than the laws of physics or human behavior.[6]

2.2 The Fragility of Detached Metadata

Parallel to the crisis of recursion is the crisis of context. The current infrastructure for digital provenance relies on "detached" metadata standards such as C2PA (Coalition for Content Provenance and Authenticity), IPTC, and XMP.[21] While C2PA represents a significant advancement in cryptographic signing, recent technical reports highlight a critical vulnerability: strippability. Research indicates that the vast majority of social media platforms and content management systems (CMS) routinely strip metadata headers to optimize bandwidth or sanitize user data.[8] A C2PA signature attached to a file header is lost the moment an image is screenshotted, re-encoded, or uploaded to a non-compliant platform. This "decoupling effect" leaves the pixel data (the "artifact") verifying nothing, while the provenance data is discarded. This renders the data "context-blind" for any downstream crawler that scrapes the image from a social feed.[8] Furthermore, critics argue that C2PA verifies *who* signed the asset, but not *what* the asset represents semantically.[24] It provides a chain of custody but lacks the "dense semantic scaffolding" required to train models on meaning rather than just pattern.

2.3 Steganography and the Transport Layer

To address the fragility of detached metadata, recent scholarship has turned to Deep Neural Steganography and Semantic Watermarking. Unlike traditional watermarking, which is often visible or easily destroyed, deep steganography uses encoder-decoder neural networks to embed information within the latent features of the media itself.[12] Lee and Cho (2025) introduced "Hermitian Symmetric Fourier Watermarking" (SFW), which embeds signals in the frequency domain to ensure robustness against compression and cropping—the very attacks that destroy standard metadata. By maintaining frequency integrity, this method allows for high detection accuracy while preserving image fidelity.[12] Similarly, recent advances in "Semantic Watermarking" (Arabi et al., 2025) utilize

locality-sensitive hashing (LSH) to bind the watermark to the semantic content of the image. This ensures that if the image is manipulated (e.g., a person is removed via generative fill), the watermark is invalidated, flagging the asset as tampered.[28] These technologies provide the technical substrate for a "Pointer-Based" architecture, where the asset carries its own immutable link to its metadata.

2.4 The Economic Valuation of Data

The crisis of data quality has also spurred research into the economics of data. With high-quality human data becoming a scarce resource, concepts like Data Shapley values have gained prominence. The Shapley value, derived from cooperative game theory, quantifies the marginal contribution of a single data point to the overall performance of a model.[30] Recent work by Wang et al. (2024/2025) on "In-Run Data Shapley" and "Fast-DataShapley" has made it computationally feasible to estimate these values for large datasets without the prohibitive cost of retraining models thousands of times.[32] This theoretical advancement is critical for the establishment of a Verification Economy, as it provides a mathematical basis for pricing data based on its utility and informational richness, rather than just its volume.

3 Problem Analysis: The Entropy of Recursion

To construct a robust solution, we must first rigorously define the problem space. The modern web acts as a mechanism for separating signal from context, a process that is accelerating entropy loss in training pipelines.

3.1 The Mathematics of Distributional Collapse

The core of the problem is the mathematical inevitability of variance reduction in self-consuming loops. Let us formalize the "Xerox Effect." A generative model M approximates a true distribution D_{real}. Due to the operational necessity of avoiding incoherent outputs, models are typically sampled with reduced temperature or truncation (e.g., nucleus sampling, top-k sampling). This process effectively shaves off the tails of D_{real}.[17] When a subsequent model M_{next} is trained on the output of M, it is training on a distribution D_{syn} where:

$$Variance(D_{syn}) < Variance(D_{real}) \tag{1}$$

Recursive iterations lead to a compounding loss of information:

$$\lim_{n \to \infty} Variance(M_n) \to 0 \tag{2}$$

This convergence is the theoretical definition of Model Autophagy Disorder (MAD). The symptoms observed in 2024-2025 literature include:

Mode Collapse: The probability distribution collapses to a single mode. Ask for a "CEO," and it produces the exact same face, lighting, and suit configuration every time.[18]

Artifact Amplification: A small error in Generation 1 (e.g., a six-fingered hand) is learned as a feature in Generation 2, and by Generation 3, it is codified as a "canonical" anatomical fact. The model amplifies the "noise" of the previous generation, mistaking it for signal.[19]

Reality Drift: The model's physics detach from the real world. Without the "entropy of reality"—the noise, friction, and chaos of physical existence—the model drifts into a logic governed by statistical probability rather than causality.

3.2 The Poison of Context Blindness

Current vision models (e.g., CLIP, ViT) and Large Language Models (LLMs) are trained on image-text pairs scraped from the web (e.g., LAION-5B, Common Crawl). These datasets are characterized by Sparse Supervision. An image is paired with "alt-text" that is notoriously sparse, inaccurate, or missing. A photograph of a "19th-century coal miner" might be captioned simply as "man with dirt." This forces the model to guess the relationship between visual features and semantic meaning.[35] It learns that "dirt on face" correlates with "man," but it fails to learn the historical, sociological, or emotional context of "labor," "extraction," or "poverty." When models graze on this context-free data, they hallucinate because they have no "Ground Truth" anchor.[10] They recognize the surface texture of reality but not its ontological substance. This "context blindness" is exacerbated by the Decoupling Effect. When an image is shared, screenshotted, or reposted, the link to its original context is severed. The crawler sees the image in a vacuum. This is not a failure of the model architecture; it is a failure of the data transport layer.

3.3 The Failure of Current Solutions

The industry's response to these challenges has been fragmented and largely ineffective:

- **Data Cleaning:** Filtering synthetic data is becoming computationally intractable as synthetic content becomes indistinguishable from human content.[4]
- **Reinforcement Learning from Human Feedback (RLHF):** While RLHF aligns models to human preferences, it introduces human annotator bias and does not solve the underlying lack of variance in the pre-training data. It aligns the model's behavior but does not replenish the "knowledge capital" lost to autophagy.[37]
- **Detached Provenance (C2PA):** As discussed, C2PA provides a "chain of custody" but not a "chain of meaning." Furthermore, its vulnerability to stripping makes it an unreliable signal for web-scale crawlers.[8]
- **Output-Side Watermarking (e.g., SynthID):** Recent systems such as Google DeepMind's SynthID embed imperceptible watermarks into AI-*generated* content to identify synthetic outputs.[36] While valuable for detection and attribution of generated media, these approaches address a complementary problem: they label what a model *produces*, but do not enrich or verify what a model *consumes*. They cannot inject the missing semantic context, provenance, or variance that training data requires to prevent distributional collapse. The PEST framework addresses this input-side gap.

4 Proposed Framework: Provenance-Embedded Semantic Transport (PEST)

To arrest the entropy of recursion and solve the crisis of context, we propose a framework that fundamentally alters the relationship between data, metadata, and storage. This framework, which we designate as the Provenance-Embedded Semantic Transport (PEST), operates on the principle that the metadata must be inextricably bound to the artifact via a robust, invisible pointer.

4.1 Architectural Core: The Persistent Provenance Pointer

We reject the storage of heavy metadata payloads inside the image file (Deep Payload Steganography), which degrades image quality and increases latency. Instead, we propose embedding a Readable, Invisible Perceptual Hash—a pointer—directly into the latent pixel structure.[39]

4.1.1 The Pointer Mechanism

Using a custom Encoder-Decoder-Adversary Network (EDAN), we embed a robust 256-bit hash (a URI) into the image. This encoder utilizes Spread-Spectrum Redundancy, distributing the signal across the frequency spectrum so that the hash can be recovered even if the image is cropped, compressed (JPEG), or filtered.[41]

- **Encoder (E):** Takes the Cover Image (C) and the Pointer (P) to generate the Stego-Image (C').
- **Adversary (A):** Ensures C' is perceptually indistinguishable from C.
- **Decoder (D):** Extracts P from C' with high fidelity, even after "differentiable attack layers" (noise, blur, crop) are applied during training.[12]

4.1.2 The Resolution and Latency Advantage

This hash acts as a decentralized DNS lookup. It points not to a mutable server, but to an Atomic Asset on a decentralized permanent storage layer (e.g., a blockweave or permaweb).[42] This approach decouples the "weight" of the metadata from the media. A training crawler need only scan the image for the hash (millisecond latency) to instantaneously retrieve the full, high-fidelity metadata payload from the immutable ledger. This satisfies the requirement for rapid ingestion in training pipelines while maintaining deep semantic context.[12]

4.1.3 Semantic Watermarking and Tamper Detection

To prevent "Copy-Paste Attacks" (where a pointer is copied to a fake image), the pointer is mathematically derived from the semantic features of the image using Locality-Sensitive Hashing (LSH).[28] The Decoder performs a checksum: it extracts the semantic features of the image and compares them to the hash. If the image has been semantically altered (e.g., a person removed), the hash will not resolve, and the asset is flagged as "Corrupted/Tampered." This ensures the pointer acts as a seal of integrity.

Implementation validation (practitioner note). The authors have developed and validated a production implementation of the Persistent Provenance Pointer (PPP) mechanism using 256-bit perceptual hashing. Testing against a 10,097-artifact corpus (the *Alexandria Aeternum* collection), this implementation demonstrates that:

- pointers survive standard social media compression (JPEG re-encoding at quality levels typical of Facebook, Twitter/X, and Instagram), achieving a 98.7% mean perceptual similarity score across platforms (all 80/80 artifacts successfully resolved);
- pointers remain recoverable after screenshot capture, re-upload cycles, and up to 80% image degradation via a triple-hash integrity protocol (pixel-perfect hash, semantic content hash, and resilience hash); and

- payload retrieval from the decentralized ledger achieves sub-second latency (<200ms p95), suitable for integration into large-scale training pipeline ingestion.

4.2 The Hybrid Curation Paradigm

The second pillar of the framework addresses the content of the data. We challenge the orthodoxy that "synthetic data is poison." We argue that Generative Synthetic Data (pixels/text) is often **degenerative** (prone to variance collapse), whereas Analytical Synthetic Data (metadata) is **negentropic** (structure-enhancing). We propose a Human-AI Hybrid Curation pipeline [14]:

The Artifact (Human): The primary asset (image, text) must be human-generated to preserve the "entropy of reality" and the "tails" of the distribution. Human data contains the micro-expressions, chaotic lighting, and non-linear logic that synthetic models shave off.[3]

The Description (Machine): The metadata describing the asset should be generated by an ensemble of Expert Agents. Humans are "lazy labelers" (e.g., tagging a photo "Sunset"). Models fine-tuned on ontologies can generate Dense Semantic Scaffolding—graph-based descriptions that map every object, relationship, lighting condition, and historical context in the frame.[35] This "Human Pixels + Machine Semantics" approach creates a "High-Density Ground Truth." The human pixels prevent model collapse by providing variance; the semantics provide the Dense Supervision signal required to align future models without the noise of sparse human captions.[46]

4.3 The Ontology Layers (Metadata Structure)

The metadata payload linked by the Provenance Pointer is structured into four immutable layers, stored as a JSON-LD object on the decentralized ledger:

1. **Provenance Layer:** Contains cryptographic assertions of origin (wrapping standards like C2PA). This verifies the human source, device data, and timestamp, preventing "history hallucination" by anchoring the asset in spacetime.[10]
2. **Semantic Topology Layer (NEST):** A Neural Extraction of Semantic Topology (NEST) graph. This moves beyond simple captions to map causal and spatial relationships (e.g., "Man holding cup" → "Cup contains hot liquid" → "Liquid emits steam"). This teaches the model physics and interaction, not just object co-occurrence.[47]
3. **Affective Layer (HEO):** A Human Emotion Ontology (HEO) vector map, utilizing the PAD (Pleasure-Arousal-Dominance) emotional state model. Current models are emotionally blind. This layer quantifies the emotional resonance of the asset, enabling the training of Affective Computing models.[11]
4. **Rights & Economics Layer:** A machine-readable Universal Data License (UDL). This explicitly dictates training permissions (Yes/No), attribution requirements, and royalty mechanisms. This layer automates compliance, solving the legal bottleneck of "fair use" vs. infringement.[1]

5 Implications: The Verification Economy

The implementation of the PEST framework necessitates and enables a new economic model: the Verification Economy. As the supply of high-quality, uncontaminated human data dwindles, its value will decouple from the value of "generic" data.[1]

5.1 Data as an Asset Class and "Verified Human Data" Markets

Verified, pointer-embedded human data will become a distinct asset class. Developers, facing the existential risk of MAD, will distinguish between "Unverified Synthetic Data" (synthetic/unverified data) and "Verified Human Data" (verified human data). They will pay a premium for Verified Human Data datasets—certified human data required to stabilize their models and re-introduce variance.[50] This market will rely on Data Valuation metrics. Utilizing concepts like Data Shapley, the marginal contribution of a specific dataset to a model's performance (e.g., its ability to reduce perplexity or increase robustness) can be quantified. High-Shapley-value data (rare, high-entropy human data) will command higher prices.[30]

5.2 Automated "Data Dividends" and Royalties

The embedding of the Rights & Economics Layer within the Atomic Asset allows for the programmatic distribution of value. When a crawler accesses the metadata payload via the Provenance Pointer, the decentralized ledger can trigger a micro-transaction or "Data Dividend" to the original creator.[49] This resolves the "freerider" problem of current training, where models extract value from human creators without compensation. It converts the relationship between creator and model from extractive to symbiotic. Creators are incentivized to produce high-variance "Verified Human Data" to feed the machine, ensuring a sustainable supply of training data.[53]

5.3 Legal and Governance Implications

From a legal perspective, this framework supports the "3 Cs" of generative governance: Consent, Credit, and Compensation.[54] The embedded pointer ensures that Consent (via the UDL) and Credit (via the Provenance Layer) travel with the asset. This technical enforcement of rights may mitigate the need for aggressive regulatory intervention by creating a market-based solution to copyright infringement.[55] It aligns with the emerging "Right to Uncontaminated Data," which posits that human creators have a right to have their work distinguished from synthetic outputs.[56]

6 Cultural Heritage Opportunity: The Reservoirs of Reality

The GLAM sector (Galleries, Libraries, Archives, Museums) is uniquely positioned to become the "Central Bank" of this new Verification Economy. Their collections represent the largest, most diverse, and most reliably human datasets in existence.[57] The authors have demonstrated this approach by enriching 10,097 artworks from The Metropolitan Museum of Art's open-access collection into the *Alexandria Aeternum* corpus.

6.1 GLAMs as "Ground Truth" Reservoirs

Unlike the open web, which is increasingly polluted with synthetic artifacts, museum collections are finite, verified, and historically grounded. They are the ultimate "anti-collapse" datasets. They contain the full variance of human history, art, and biology, immune to the recursive loops of modern generative models.[59]

6.2 The "Collections as Data" Pivot

By digitizing collections and embedding them with Persistent Provenance Pointers, GLAMs can monetize access to their data for training while ensuring integrity. This "Collections as Data"

approach allows institutions to license high-fidelity datasets to labs.[60] The revenue generated from these "Data Dividends" can support the preservation missions of these institutions, creating a virtuous cycle.[58]

6.3 Preventing Historical Hallucination

Models trained on verified cultural heritage data will be less prone to "historical hallucination"—the invention of fake artifacts, events, or biologies. Because the training signal is grounded in institutionally verified provenance and dense semantic metadata, the model learns a consistent ontology of the past.[10] This is critical for applications in education, historical reconstruction, and scientific research.

7 Future Research and Challenges

While the PEST framework offers a robust theoretical solution, significant challenges remain for future research:

- **Adversarial Robustness:** As steganographic encoders improve, so too will adversarial "scrubbers" designed to remove pointers. Research into "fragile watermarking" (where removing the mark destroys the image utility) is critical to ensure the pointer cannot be stripped without rendering the asset useless for training.[12]
- **Hardware Integration:** The ultimate endpoint of this framework is hardware integration. Research should focus on embedding the Provenance Pointer generation directly into the Image Signal Processors (ISPs) of cameras (e.g., Sony, Canon). This would create a "Chain of Truth" from the photon to the screen, ensuring that the provenance is bound at the moment of capture.[22]
- **Semantic Density Metrics:** We need standardized metrics for "Semantic Density" to quantify the information richness of the metadata payload.[63] Developing universal standards for measuring the "entropy per pixel" or "knowledge per token" will be essential for the algorithmic pricing of training data in the Verification Economy.
- **Governance of Data DAOs:** As "Data DAOs" (Decentralized Autonomous Organizations) emerge to manage these data pools, research is needed into their governance structures. How do we ensure fair distribution of dividends? How do we prevent the centralization of verified data?[64]

8 Data Availability and Reproducibility

To support reproducibility and enable further research, the authors have prepared a sample dataset implementing the PEST framework described in Section 4.2.

8.1 Sample dataset

A curated sample of 1,000 provenance-enriched images is derived from the *Alexandria Aeternum* collection (10,097 public-domain artworks sourced from The Metropolitan Museum of Art).

8.2 Access

The dataset will be made available in two complementary distribution channels:

- **Zenodo** (DOI: 10.5281/zenodo.18359131) for academic citation and archival stability; and
- **Hugging Face** (`https://huggingface.co/datasets/Metavolve-Labs/alexandria-aeternum-genesis`) for practitioner access and downstream benchmarking workflows.

8.3 Contents

Each image in the sample dataset includes:

- a human-originated artifact with museum-verified provenance;
- machine-generated semantic scaffolding (50+ structured metadata fields per image);
- embedded XMP metadata following ISO 16684-1, recoverable with standard tools; and
- a cryptographic integrity hash (SHA-256) for tamper-evidence.

8.4 License and research purpose

The sample dataset will be released under the Creative Commons Attribution 4.0 International (CC BY 4.0) license for research use. Its purpose is to provide the community with a reference implementation of the Human-Hybrid Curation paradigm described in Section 4.2 and to enable controlled experiments in provenance, valuation, and model collapse mitigation.

9 Conclusion: Stewardship of the Signal

The "Xerox Effect" is a mathematical inevitability in closed loops; entropy always increases unless energy—in the form of fresh, high-variance information—is injected from the outside. The Provenance-Embedded Semantic Transport (PEST) framework represents the infrastructure for that injection. By binding the context to the content through robust, immutable pointers, we preserve the variance, nuance, and history of human experience. We enable a transition from a model of Data Extraction, which depletes the digital commons, to a model of Data Stewardship, which enriches it. In doing so, we ensure that the digital future remains tethered to the richness, complexity, and truth of the human past. This is the only path to a Hybrid Intelligence that is not just powerful, but meaningful.

Subsequent publications will detail the technical implementation of the Persistent Provenance Pointer and report empirical results from deployment of the framework at scale.[1]

Competing Interests

The author is affiliated with Metavolve Labs, Inc., which develops data infrastructure tools. The theoretical framework and sample dataset are made available for academic research purposes under open licenses.

[1]This paper is the first in a four-part research program. Companion papers include: *The Density Imperative* (DOI: 10.5281/zenodo.18667735), *Cognitive Nutrition for Foundation Models* (DOI: 10.5281/zenodo.18667742), and *Perceptual Compute Offloading* (DOI: 10.5281/zenodo.18667749).

Acknowledgments

The authors acknowledge the Metropolitan Museum of Art for providing open access to their collection via the Open Access Initiative. The Alexandria Aeternum dataset (DOI: 10.5281/zenodo.18359131) represents enrichment of these public-domain works and does not claim ownership of the underlying artifacts.

Software and Tools. The implementation and validation experiments described in this paper were conducted using the following open-source tools developed by the authors:

- **codex-lab-kit** v1.0.1 (DOI: 10.5281/zenodo.18668110; source: `https://github.com/codex-curator/codex-lab-kit`)
- **gcp-robotics SDK** (DOI: 10.5281/zenodo.18668113; source: `https://github.com/codex-curator/gcp-robotics`)

References

[1] 3 Questions: The pros and cons of synthetic data in AI | MIT News, Accessed: Dec. 16, 2025. [Online]. Available: `https://news.mit.edu/2025/3-questions-pros-cons-synthetic-data-ai-kalyan-veeramachaneni-0903`

[2] The Future of Truth: Building AI on Verified Data | by Ransome's Ink | Nov, 2025 | Medium, Accessed: Dec. 16, 2025. [Online]. Available: `https://medium.com/@ransomechubby/the-future-of-truth-building-ai-on-verified-data-1503e7bad670`

[3] The AI Model Collapse Risk is Not Solved in 2025 - Winssolutions, Accessed: Dec. 16, 2025. [Online]. Available: `https://www.winssolutions.org/ai-model-collapse-2025-recursive-training/`

[4] AI Slop III: Society and Model Collapse - Khazanah Research Institute, Accessed: Dec. 16, 2025. [Online]. Available: `https://www.krinstitute.org/publications/ai-slop-iii-society-and-model-collapse`

[5] A Note on Shumailov et al. (2024): 'AI Models Collapse When Trained on Recursively Generated Data' - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: `https://arxiv.org/html/2410.12954v2`

[6] A Closer Look at Model Collapse: From a Generalization-to-Memorization Perspective, Accessed: Dec. 16, 2025. [Online]. Available: `https://arxiv.org/html/2509.16499v1`

[7] GenAI: Running on empty? - SAP, Accessed: Dec. 16, 2025. [Online]. Available: `https://www.sap.com/blogs/genai-running-on-empty`

[8] Forensic Watermarking & C2PA - Enhanced Content Security - DoveRunner, Accessed: Dec. 16, 2025. [Online]. Available: `https://doverunner.com/blogs/forensic-watermarking-c2pa-content-protection/`

[9] AI Model Collapse: Causes and Prevention - WitnessAI, Accessed: Dec. 16, 2025. [Online]. Available: `https://witness.ai/blog/ai-model-collapse/`

[10] From Illusion to Insight: A Taxonomic Survey of Hallucination Mitigation Techniques in LLMs, Accessed: Dec. 16, 2025. [Online]. Available: `https://www.mdpi.com/2673-2688/6/10/260`

[11] AI Hallucinations in 2025: Causes, Impact, and Solutions for Trustworthy AI Maxim AI, Accessed: Dec. 16, 2025. [Online]. Available: https://www.getmaxim.ai/articles/ai-hallucinations-in-2025-causes-impact-and-solutions-for-trustworthy-ai/

[12] [2509.07647] Semantic Watermarking Reinvented: Enhancing Robustness and Generation Quality with Fourier Integrity - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/abs/2509.07647

[13] Semantic Watermarking Reinvented: Enhancing Robustness and Generation Quality with Fourier Integrity, Accessed: Dec. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2025/papers/Lee_Semantic_Watermarking_Reinvented_Enhancing_Robustness_and_Generation_Quality_with_Fourier_ICCV_2025_paper.pdf

[14] CRAFT Your Dataset: Task-Specific Synthetic Dataset Generation Through Corpus Retrieval and Augmentation | Transactions of the Association for Computational Linguistics - MIT Press Direct, Accessed: Dec. 16, 2025. [Online]. Available: https://direct.mit.edu/tacl/article/doi/10.1162/TACL.a.56/134309/CRAFT-Your-Dataset-Task-Specific-Synthetic-Dataset

[15] Synthetic Data Generation Using Large Language Models: Advances in Text and Code, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/html/2503.14023v2

[16] (PDF) Author Correction: AI models collapse when trained on recursively generated data, Accessed: Dec. 16, 2025. [Online]. Available: https://www.researchgate.net/publication/390064111_Author_Correction_AI_models_collapse_when_trained_on_recursively_generated_data

[17] [Literature Review] Strong Model Collapse - Moonlight, Accessed: Dec. 16, 2025. [Online]. Available: https://www.themoonlight.io/en/review/strong-model-collapse

[18] Self-Consuming Generative Models Go MAD, Accessed: Dec. 16, 2025. [Online]. Available: https://research.latinxinai.org/papers/neurips/2023/pdf/Josue_CascoRodriguez.pdf

[19] SELF-CONSUMING GENERATIVE MODELS GO MAD - ICLR Proceedings, Accessed: Dec. 16, 2025. [Online]. Available: https://proceedings.iclr.cc/paper_files/paper/2024/file/ebc042e767de551803ccfcc45e2454f5-Paper-Conference.pdf

[20] A Closer Look at Model Collapse: From a Generalization-to-Memorization Perspective, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/html/2509.16499v2

[21] C2PA Implementation Guidance, Accessed: Dec. 16, 2025. [Online]. Available: https://spec.c2pa.org/specifications/specifications/2.2/guidance/Guidance.html

[22] C2PA Technical Specification: Hardware Binding and Camera Integration, Coalition for Content Provenance and Authenticity, v2.2, 2025. [Online]. Available: https://c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html

[23] Why C2PA is necessary, what it can and cannot do., Accessed: Dec. 16, 2025. [Online]. Available: https://blog.longtailedleopard.com/2024/09/why-c2pa-is-necessary-what-it-canand.html

[24] A Critical Look at C2PA: DRM and AI - Yumechi, Accessed: Dec. 16, 2025. [Online]. Available: https://yumechi.jp/en/blog/2025/a-critical-look-at-c2pa-drm-and-ai/

[25] C2PA and Content Credentials Explainer, Accessed: Dec. 16, 2025. [Online]. Available: https://spec.c2pa.org/specifications/specifications/2.2/explainer/Explainer.html

[26] Robust and secure image steganography with recurrent neural network and fuzzy logic integration - PMC - NIH, Accessed: Dec. 16, 2025. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12003756/

[27] Semantic Watermarking Reinvented: Enhancing Robustness and Generation Quality with Fourier Integrity - ResearchGate, Accessed: Dec. 16, 2025. [Online]. Available: https://www.researchgate.net/publication/395388375_Semantic_Watermarking_Reinvented_Enhancing_Robustness_and_Generation_Quality_with_Fourier_Integrity

[28] SEAL: Semantic Aware Image Watermarking - CVF Open Access, Accessed: Dec. 16, 2025. [Online]. Available: https://openaccess.thecvf.com/content/ICCV2025/papers/Arabi_SEAL_Semantic_Aware_Image_Watermarking_ICCV_2025_paper.pdf

[29] SEAL: Semantic Aware Image Watermarking - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/html/2503.12172v3

[30] Fast-DataShapley: Neural Modeling for Training Data Valuation - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/html/2506.05281

[31] Data Shapley: Equitable Valuation of Data for Machine Learning, Accessed: Dec. 16, 2025. [Online]. Available: https://proceedings.mlr.press/v97/ghorbani19c/ghorbani19c.pdf

[32] Data Shapley in One Training Run - GitHub Pages, Accessed: Dec. 16, 2025. [Online]. Available: https://jiachen-t-wang.github.io/data-shapley.github.io/

[33] DO DATA VALUATIONS MAKE GOOD DATA PRICES? - OpenReview, Accessed: Dec. 16, 2025. [Online]. Available: https://openreview.net/pdf/be5c15c137884ccc26f7edeeed5c997574924fbd.pdf

[34] (PDF) Rethinking Data Value: Asymmetric Data Shapley for Structure-Aware Valuation in Data Markets and Machine Learning Pipelines - ResearchGate, Accessed: Dec. 16, 2025. [Online]. Available: https://www.researchgate.net/publication/397696706_Rethinking_Data_Value_Asymmetric_Data_Shapley_for_Structure-Aware_Valuation_in_Data_Markets_and_Machine_Learning_Pipelines

[35] Are Scene Graphs Good Enough to Improve Image Captioning? | Request PDF ResearchGate, Accessed: Dec. 16, 2025. [Online]. Available: https://www.researchgate.net/publication/386176300_Are_Scene_Graphs_Good_Enough_to_Improve_Image_Captioning

[36] P. Fernandez, G. Couairon, H. Jégou, M. Douze, and T. Furon, "The Stable Signature: Rooting Watermarks in Latent Diffusion Models," in *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. See also: Google DeepMind, "SynthID: Identifying AI-generated content," 2024. [Online]. Available: https://arxiv.org/abs/2303.15435

[37] S. Casper, J. Davies, C. Shi, et al., "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback," *Transactions on Machine Learning Research (TMLR)*, 2023. [Online]. Available: https://arxiv.org/abs/2307.15217

[38] Paper Review: Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback - Andrey Lukyanenko, Accessed: Dec. 16, 2025. [Online]. Available: https://andlukyane.com/blog/paper-review-rlhf-overview

[39] Denoising and enhancement for robust image steganography - ResearchGate, Accessed: Dec. 16, 2025. [Online]. Available: https://www.researchgate.net/publication/397682964_Denoising_and_enhancement_for_robust_image_steganography

[40] Denoising and enhancement for robust image steganography - PeerJ, Accessed: Dec. 16, 2025. [Online]. Available: https://peerj.com/articles/cs-3368.pdf

[41] Errorless Robust JPEG Steganography using Outputs of JPEG Coders - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/pdf/2211.04750

[42] S. Williams, "Arweave: A Protocol for Economically Sustainable Information Permanence," Arweave Yellow Paper, 2019. [Online]. Available: https://www.arweave.org/yellow-paper.pdf

[43] Human-in-the-Loop AI Platform Market Research Report 2033 - Dataintelo, Accessed: Dec. 16, 2025. [Online]. Available: https://dataintelo.com/report/human-in-the-loop-ai-platform-market

[44] When AI Assistants Get the News Wrong - Sapien, Accessed: Dec. 16, 2025. [Online]. Available: https://www.sapien.io/blog/when-ai-assistants-get-the-news-wrong

[45] Graph-Based Captioning: Enhancing Visual Descriptions by Interconnecting Region Captions - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/html/2407.06723v2

[46] Context-aware Image Understanding in VQA with Dense-captioning, Accessed: Dec. 16, 2025. [Online]. Available: https://ijict.itrc.ac.ir/browse.php?a_id=740&sid=1&slc_lang=en&ftxt=1

[47] Fine-Grained Video Captioning through Scene Graph Consolidation - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/html/2502.16427v1

[48] The Case for Universal Basic Income in an AI Economy - Eduwik, Accessed: Dec. 16, 2025. [Online]. Available: https://eduwik.com/the-case-for-universal-basic-income-in-an-ai-economy/

[49] Mapping the Potential and Pitfalls of "Data Dividends" as a Means of Sharing the Profits of Artificial Intelligence - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/pdf/1912.00757

[50] From Data Divide to Data Dividend - Splunk, Accessed: Dec. 16, 2025. [Online]. Available: https://www.splunk.com/en_us/pdfs/research-report/from-data-divide-to-data-dividend.pdf

[51] AI annotation Market Size to Hit USD 17.37 Billion by 2034 - Precedence Research, Accessed: Dec. 16, 2025. [Online]. Available: https://www.precedenceresearch.com/ai-annotation-market

[52] Synthetic Data in AI: Challenges, Applications, and Ethical Implications - arXiv, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/html/2401.01629v1

[53] Sharing the Winnings of AI with Data Dividends: Challenges with "Meritocratic" Data Valuation - Nicholas Vincent, Accessed: Dec. 16, 2025. [Online]. Available: http://www.nickmvincent.com/static/eaamo_data_dividends.pdf

[54] Governance of Generative AI in Creative Work: Consent, Credit, Compensation, and Beyond, Accessed: Dec. 16, 2025. [Online]. Available: https://arxiv.org/html/2501.11457v1

[55] Data Privacy and Antitrust: Competing Frameworks for Mitigating Privacy Harms - California Lawyers Association, Accessed: Dec. 16, 2025. [Online]. Available: https://calawyers.org/privacy-law/data-privacy-and-antitrust-competing-frameworks-for-mitigating-privacy-harms/

[56] Model Collapse and the Right to Uncontaminated Human-Generated Data, Accessed: Dec. 16, 2025. [Online]. Available: https://jolt.law.harvard.edu/digest/model-collapse-and-the-right-to-uncontaminated-human-generated-data

[57] Leveraging AI to increase the value of museum collections - UF News - University of Florida, Accessed: Dec. 16, 2025. [Online]. Available: https://news.ufl.edu/2022/03/leveraging-ai-to-increase-the-value-of-museum-/

[58] AI and Libraries, Archives, and Museums, Loosely Coupled - Humane Ingenuity, Accessed: Dec. 16, 2025. [Online]. Available: https://newsletter.dancohen.org/archive/ai-and-libraries-archives-and-museums-loosely-coupled/

[59] Now more than ever: the role of museum and archival objects in an age of generative artificial intelligence | Collection and Curation | Emerald Publishing, Accessed: Dec. 16, 2025. [Online]. Available: https://www.emerald.com/cc/article/doi/10.1108/CC-04-2025-0020/1272140/Now-more-than-ever-the-role-of-museum-and-archival

[60] Crowdsourced AI Data Collection: Pros and Cons - Surfing Technology, Accessed: Dec. 16, 2025. [Online]. Available: https://www.surfing.ai/crowdsourced-ai-data-collection/

[61] Exploring the adoption of collections as data in the GLAM context - Publicera, Accessed: Dec. 16, 2025. [Online]. Available: https://publicera.kb.se/ir/article/download/52252/43361/127999

[62] Impact of High Data Quality on LLM Hallucinations - International Journal of Computer Applications, Accessed: Dec. 16, 2025. [Online]. Available: https://www.ijcaonline.org/archives/volume187/number4/gautam-2025-ijca-924909.pdf

[63] Survey and analysis of hallucinations in large language models: attribution to prompting strategies or model behavior - Frontiers, Accessed: Dec. 16, 2025. [Online]. Available: https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2025.1622292/full

[64] Decentralized autonomous organizations: adapting legal structures and proposing a new model of DAOLLP - Oxford Academic, Accessed: Dec. 16, 2025. [Online]. Available: https://academic.oup.com/cmlj/article/20/3/kmaf011/8249442

[65] Deep Dive: Everything You Need to Know About Data DAOs on Vana, Accessed: Dec. 16, 2025. [Online]. Available: https://www.vana.org/posts/datadao-deepdive