Cognitive Nutrition: Architecture of a Provenance-Embedded Data Refinery

Tad MacPherson

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

Abstract

We present the architecture, operational metrics, and theoretical foundations of a production metadata factory that transforms raw media artifacts into provenance-verified, semantically dense training assets—what we term *Cognitive Nutrition* (CN) for foundation models. The system operationalizes the Provenance-Embedded Semantic Transport (PEST) framework introduced in *The Entropy of Recursion* [**?**], processing 10,097 cultural heritage artifacts from The Metropolitan Museum of Art into "CN Assets," each carrying 2,000–6,000 tokens of structured context across 111+ fields organized into four ontology layers. Central to our approach is the *Translation Paradigm*: enrichment constitutes machine-assisted articulation of existing human knowledge—not synthetic generation—a distinction with fundamental implications for how enriched datasets should be classified and trusted.

The factory employs six specialized agents orchestrated through a Firestore-based workflow engine, with the Nova agent performing multi-model *Orchestrated Multi-Model Inference*—an orchestrated pipeline where distinct frontier models contribute complementary cognitive capabilities to produce enrichment that no single model achieves alone. A triple-hash integrity protocol (content hash, perceptual hash, payload integrity hash) provides defense-in-depth verification, achieving 100% recovery (80/80) under four social media transformation classes. Processing cost is approximately $0.10–$0.12 per artifact at current API pricing, placing million-scale enrichment within institutional reach.

We formalize the mathematical foundations of Cognitive Nutrition using topological data analysis (persistent homology for distributional drift detection) and causal disentanglement (back-door adjustment for separating causal from spurious features). We introduce the *Amendability Framework*—a living curation model where originating institutions retain governance over enriched collections through versioned corrections and trust-level field tagging. Downstream validation via controlled ablation [**?**] demonstrates a 24.1-point CogBench cognitive swing between sparse and dense fine-tuning conditions, confirming that the semantic density this system produces measurably transforms model capability.

Keywords: data refinery, provenance, semantic density, metadata enrichment, perceptual hashing, GLAM, topological data analysis, causal inference, data stewardship, multi-model synthesis

[0]DOI: 10.5281/zenodo.18667742

1. Introduction

1.1 From Theory to Factory

In *The Entropy of Recursion* [**?**], we established that the era of indiscriminate Data Accumulation must yield to Data Stewardship. We identified two converging crises:

1. **Model Collapse** (the Xerox Effect): recursive training on synthetic output causes distributional tails to vanish, compounding variance loss across generations until models converge to low-entropy point estimates [**?**, **?**].
2. **Context Blindness**: the systematic stripping of metadata during web redistribution produces "lobotomized" training artifacts—pixels without provenance, intent, or semantic grounding [**?**].

We proposed the PEST framework, anchored by a Persistent Provenance Pointer (PPP) embedded via perceptual hashing and linked to a four-layer metadata ontology. In *The Density Imperative* [**?**], we validated the core hypothesis empirically: fine-tuning on PEST-enriched data produces a 24.1-point CogBench cognitive swing over sparse data, with 159.6% semantic coverage improvement, while sparse fine-tuning *actively destroys* pre-trained capabilities.

What neither paper presented was *the system itself.* This paper fills that gap. We describe a production metadata factory that has been operational since late 2025, processing museum-grade cultural heritage artifacts through a multi-agent pipeline of AI enrichment, metadata infusion, cryptographic hashing, and immutable registration. We report quantitative results from the *Alexandria Aeternum* 10K collection—the first large-scale corpus of Cognitive Nutrition assets—formalize the mathematical underpinnings that were only sketched in [**?**], and introduce the Amendability Framework for living curation.

1.2 The Urgency Has Intensified

Since the publication of [**?**], the evidence base for both model collapse and data scarcity has strengthened materially:

- Epoch AI's updated analysis estimates that high-quality, non-duplicated public text will be effectively exhausted under plausible scaling policies by 2026–2028 [**?**]. We are now within that window.
- Shumailov et al.'s model collapse findings [**?**] have been replicated across modalities, and Dohmatob et al. [**?**] have identified a "generalization-to-memorization" transition in late-stage autophagy where models cease learning principles and begin memorizing artifacts of previous generations.
- The data-centric AI community has demonstrated that dataset design—not just scale—drives model performance. DataComp [**?**] showed that data curation alone, at fixed compute, yields material downstream improvements. RefinedWeb [**?**] demonstrated that filtered web data can outperform curated mixtures. DoReMi [**?**] showed that optimizing data mixture weights improves training efficiency.
- Our own controlled ablation [**?**] has now demonstrated that the sign of fine-tuning data quality is not merely theoretical: sparse metadata *destroys* pre-trained capabilities (-54.0% CogBench, $+330\%$ hallucination rate), while dense PEST-enriched metadata *enhances* them ($+25.5\%$ visual perception, $+159.6\%$ semantic coverage).

These developments validate the central premise of [**?**]: the bottleneck is no longer compute or architecture but *the quality, provenance, and semantic density of the training signal itself.*

1.3 Contributions

1. **System architecture**: A production metadata factory employing six specialized agents and multi-model Orchestrated Multi-Model Inference that operationalizes the PEST framework at scale (Section 3).
2. **The Translation Paradigm**: A formal argument that NEST enrichment constitutes machine-assisted articulation of human knowledge, not synthetic generation—a distinction with fundamental implications for dataset classification and trust (Section 2.4).
3. **Empirical results**: Quantitative statistics from the Alexandria Aeternum 10K corpus (Section 4), including metadata density analysis, processing economics, and provenance integrity metrics.
4. **Mathematical formalization**: Topological data analysis (persistent homology) and causal disentanglement as theoretical foundations for measuring Cognitive Nutrition (Section 5).
5. **The Amendability Framework**: A living curation model enabling institutional governance, versioned corrections, and trust-level field tagging (Section 7).
6. **The Verification Economy**: An economic model for provenance-verified training data as an asset class (Section 8).

2. The Cognitive Nutrition Framework

2.1 Definitions

We retain and sharpen the definitions from [**?**]:

Provenance-Embedded Semantic Transport (PEST) is a protocol for binding provenance and semantic context to media artifacts via a durable, transformation-resistant pointer mechanism, rather than fragile file-header metadata.

Persistent Provenance Pointer (PPP) is a 256-bit perceptual hash serving as a content-addressable identifier that resolves to the artifact's full metadata record in an immutable registry. The PPP survives image transformations because it is derived from the perceptual content of the image itself, not stored in a strippable header.

Neural Extraction of Semantic Topology (NEST) is a deep-inference enrichment process that produces graph-structured semantic descriptions of an artifact—mapping entities, causal relationships, spatial compositions, and affective resonance—across 111+ structured fields producing 2,000–6,000 tokens per image (median ∼2,100 for batch-processed records, with curated records exceeding 5,000), consuming 45–60 seconds of frontier model inference per artifact.

Cognitive Nutrition (CN) is the measurable bundle of properties that distinguishes CN Assets from raw training data:

Property	Definition	Measurable Signal
Provenance integrity	PPP resolves reliably; assertions verifiable; tampering detectable	Extraction success rate, resolver latency, tamper detection accuracy
Semantic topology density	NEST provides graph-structured entity/causal/spatial context	Nodes/edges per asset, token density, graph entropy
Affective grounding	Structured emotional vectors (PAD model) and poetic interpretation	PAD coverage, calibration against human judgments
Rights computability	Machine-readable training permissions and obligations	License coverage, policy-evaluability rate
High-entropy coverage	Long-tail variance preserved from human-grounded sources	Embedding-space coverage, tail frequency mass

2.2 The Data State Taxonomy

We classify training data into four states based on their information-theoretic properties:

Data State	Variance	Semantic Density	Provenance	Training Impact
Human Ground Truth	High	Low (sparse captions)	Variable	High variance, low alignment signal
Stripped Pixels	Medium	Zero (no metadata)	None	Hallucination risk, context blindness
Synthetic Slop	Low (recursive decay)	Low (auto-generated)	None	Model collapse, artifact amplification
CN Assets (CN)	High (human-sourced)	High (NEST-enriched)	Verified (PPP)	Variance preservation + dense supervision

The Cognitive Nutrition hypothesis—now empirically validated [**?**]—is that CN Assets occupy a unique quadrant: *human variance* (resisting collapse) combined with *machine-articulated semantic density* (providing alignment signal). This is the "Human Pixels + Machine Articulation" paradigm.

2.3 The Four-Layer Ontology

The metadata payload resolved by the PPP is structured into four immutable layers:

1. **Provenance Layer**: Cryptographic assertions of origin (wrapping C2PA where available), device metadata, geolocation, timestamp, institutional attribution. This anchors the artifact in spacetime and prevents "history hallucination" [**?**].
2. **Semantic Topology Layer (NEST Graph)**: A property graph mapping entities, attributes, and causal/spatial relationships. Beyond enumerating objects, this layer encodes *interactions*: "steam rises from locomotive" implies thermodynamics; "light refracts through stained glass" implies optics. This teaches models physics and causality, not merely object co-occurrence.
3. **Affective Layer**: Pleasure-Arousal-Dominance (PAD) vectors quantifying the emotional resonance of the artifact, complemented by poetic interpretation ("Affective Interpretation")

that captures the artwork's meaning in human-resonant language. This layer enables affective computing and alignment—a model that understands "solemn reverence" processes a cathedral differently than one that sees "large building with colored windows."

4. **Rights and Economics Layer**: Machine-readable licensing (Universal Data License / SPDX expressions) specifying training permissions, attribution obligations, and compensation mechanisms. This automates compliance and enables the Verification Economy described in Section 8.

2.4 The Translation Paradigm: Human Knowledge, Machine Articulation

A critical distinction must be made between *synthesis* and *translation*. The NEST enrichment process does not fabricate information from nothing. It operates on a foundation of human-authored ground truth:

The input is human. Every artifact in Alexandria Aeternum carries institutionally verified curatorial metadata: artist attribution, date, medium, provenance, exhibition history, and conservation records. This data originates from generations of human scholarship.

The enrichment is articulation, not invention. When the Orchestrated Multi-Model Inference process describes a Baroque painting as employing "chiaroscuro technique influenced by Caravaggio's tenebrism," it is not generating novel claims. It is articulating what any trained art historian would recognize upon inspection—drawing on the museum's own records, the visual evidence of the artifact, and domain knowledge encoded in model weights derived from millions of verified art-historical texts.

The analogy is translation, not authorship. Consider a curator who knows, from years of training, that a particular shade of green in a 17th-century Dutch painting indicates the use of copper resinate pigment, and that the compositional balance follows Golden Ratio principles. This knowledge exists but has never been written down for *this specific painting* because the curator's time is allocated to the hundred most important works, not the ten thousand. The Orchestrated Multi-Model Inference performs the labor of articulation—rendering implicit human knowledge into explicit, structured, machine-readable form—for every artifact in the collection.

This distinction has fundamental implications for how enriched datasets should be classified:

- The pixel data is **human-created** (museum artifact).
- The baseline metadata is **human-authored** (curatorial records).
- The enriched semantic topology is **machine-articulated human knowledge** (structured translation of existing expertise).
- The affective layer (Affective Interpretation) is a **machine-authored interpretation** guided by human-grounded artistic context.

These are categorically different from synthetic data generated by a model operating on its own outputs. The entropy originates from humans; the density originates from systematic articulation. This is not the "Unverified synthetic data" of synthetic slop—it is the structured expression of what humanity already knows but has never had the means to write down at scale.

2.5 Triple-Hash Integrity Protocol

Each CN Asset is secured by a three-layer integrity system, providing defense in depth against different classes of transformation and tampering:

Hash	Algorithm	Purpose	Tolerance
Content Integrity	SHA-256 of encoded file	Exact file verification; detects any modification	Zero tolerance—any change invalidates
Perceptual (PPP)	pHash, 256-bit	Identity resolution after transformation; survives compression, screenshot, resize	Tolerates lossy compression, resize, screenshot; degrades under major crop
Payload (Payload Integrity Hash)	SHA-256 of canonical JSON	Detects any alteration to embedded metadata	Zero tolerance—any metadata change invalidates

Together, these three hashes create a complete provenance state machine:

Content Hash	PPP	Payload Integrity Hash	State
Match	Match	Match	**Pristine**: exact original with verified metadata
No match	Match	Match	**Transformed**: same artwork, re-encoded
No match	Match	No match	**Tampered metadata**: artwork identified, metad
No match	No match	—	**Unresolvable**: degraded beyond threshold

The Locality-Sensitive Hashing (LSH) decomposition of the PPP (16 bands of 4 hex characters) enables sub-linear similarity search for near-duplicate detection.

2.6 The Output-Side / Input-Side Distinction

A key distinction introduced in [**?**] and underappreciated in the broader literature is between *output-side* and *input-side* interventions:

- **Output-side** systems (e.g., Google DeepMind's SynthID [**?**]) watermark AI-generated content to identify synthetic outputs. They label what a model *produces*.

- **Input-side** systems (PEST) enrich and verify what a model *consumes*. They inject the missing semantic context, provenance, and variance that training data requires.

These are complementary, not competing. SynthID helps detect synthetic contamination; PEST ensures that non-synthetic data carries sufficient information density to stabilize models against collapse. The present work focuses exclusively on the input side.

3. System Architecture: The Metadata Factory

3.1 Overview

The metadata factory processes artifacts through a multi-agent pipeline orchestrated by a Firestore-based workflow engine ("Thalos") with pull-based job assignment, atomic status transitions, and immutable provenance logging. The system comprises six specialized agents:

Agent	Function	Technology
Aurora	Intake, archiving, sequential ID reservation	Cloud Run, GCS, Firestore atomic counters
Nova	AI enrichment (NEST extraction via Orchestrated Multi-Model Inference)	Vertex AI, multi-model orchestration, 45–60s deep inference
Flux	Resolution enhancement (4× upscaling)	ESRGAN on NVIDIA L4 GPU (∼1.15s/image)
Atlas	Metadata infusion (dual-layer XMP embedding)	ExifTool with custom XMP namespace
Archivus	Permanent storage on Arweave blockweave	Native AR SDK (direct L1 transactions), WebP compression
Mintra	On-chain registration and NFT minting	Polygon/Ethereum smart contracts

Each agent operates as an independent microservice, pulling jobs from a shared Firestore queue. This decoupled architecture enables horizontal scaling: additional Nova instances can be provisioned independently of Flux or Atlas capacity. The Thalos orchestrator ensures exactly-once processing through atomic status transitions with optimistic locking.

3.2 The Nova Agent: Multi-Model Orchestrated Multi-Model Inference

The enrichment process is the computational and epistemological core of the factory. It employs *Orchestrated Multi-Model Inference*—an orchestrated pipeline where distinct frontier models contribute complementary cognitive capabilities:

- **Deep visual-semantic translation**: A frontier vision-language model serves as the primary workhorse, performing 45–60 seconds of structured inference per artifact. This is not simple image captioning. The model executes a deep cognitive interpretation of visual information—recognizing artistic techniques, mapping compositional structure, identifying historical and cultural context, and rendering the physics of light, material, and space into structured semantic form. The output is closer to what a PhD art historian produces in an hour with a painting than to what an auto-captioner generates in a second.
- **Artistic interpretation and affective depth**: A secondary model, selected for poetic and art-historical sensibility, layers affective interpretation over the semantic foundation.

This produces the Affective Interpretation—a resonant, human-meaningful articulation of what the artwork *means*, not merely what it *depicts*.

- **Architectural reasoning and systematic integration**: A third model contributes to prompt engineering, schema design, and structural coherence—ensuring that outputs of the primary and secondary models are integrated into a unified, schema-compliant semantic graph that meets the provenance schema specification.
- **Human-directed orchestration**: All three models operate under human-authored prompt frameworks refined through thousands of iterations. The domain expertise—what questions to ask, what structure to impose, what relationships to extract—is human intellectual capital encoded into the orchestration layer.

The result is a *process*, not a tool. Applying the provenance schema to a single model's output does not replicate Orchestrated Multi-Model Inference. The value lies in the orchestration, the multi-perspective integration, and the human-refined prompt engineering iterated over a two-year development period.

3.3 The Atlas Agent: Dual-Layer Metadata Infusion

A critical design requirement is that metadata must be *recoverable from the artifact itself* without requiring registry access—defense in depth against registry unavailability. Atlas implements this via dual-layer XMP infusion:

Layer 1 (Standard XMP/IPTC): Title, creator, description, keywords, rights statement, and source identifiers written to standard metadata fields readable by any XMP-compliant tool.

Layer 2 (Full Codex Payload): The complete structured provenance JSON (4,000+ tokens) is gzipped, base64-encoded, and written to a custom XMP namespace (`http://ns.goldencodex.io/schema/1.0/`) as a `CodexPayload` field. Any tool with ExifTool and the namespace configuration can extract and decompress the full semantic payload.

This dual-layer approach ensures that even if the PPP registry is unavailable, the artifact carries its own complete metadata—a "self-sovereign" data object.

3.4 Processing Pipeline: Hybrid Runner

For the Alexandria Aeternum 10K, we developed a hybrid processing pipeline that operates on pre-existing Orchestrated Multi-Model Inference outputs (avoiding redundant AI inference) while performing schema transformation, PNG conversion, ExifTool infusion, perceptual hash computation, payload integrity hash generation, and GCS upload in parallel across 8 worker processes.

The pipeline implements checkpoint/resume via a progress tracker that serializes completed item IDs to disk, enabling crash recovery without reprocessing. Items are processed through multiprocessing `imap_unordered` for maximum throughput.

3.5 Infrastructure Economics

A key practical contribution is the demonstration that NEST enrichment is economically feasible at scale. The factory was optimized for Google Cloud NVIDIA L4 instances (24GB VRAM), achieving an 87% cost reduction compared to A100-class GPU clusters.

Component	Per-Image Cost	Per-Image Time
Nova (Orchestrated Multi-Model Inference via Vertex AI)	∼$0.065	45–60s
Flux (ESRGAN 4× upscale)	∼$0.005	∼1.15s
Atlas (ExifTool infusion + hashing)	<$0.002	∼1s
GCS storage + upload + overhead	∼$0.003	∼2s
Total per artifact (conservative)	**∼$0.10–$0.12**	**∼50–65s**

At this cost structure, processing 1 million artifacts costs approximately $100,000–$120,000 in compute—within reach of institutional or grant-funded research programs. Processing 10 million artifacts, sufficient to constitute a meaningful fraction of a frontier model's training mix, costs approximately $1 million.

Scale	Estimated Cost	Timeline (8 workers)
10K (pilot)	∼$950	∼3 hours
100K	∼$9,500	∼30 hours
1M	∼$95,000	∼13 days
10M	∼$950,000	∼130 days

4. Empirical Results: Alexandria Aeternum 10K

4.1 Dataset Overview

The Alexandria Aeternum collection consists of 10,097 public-domain artworks sourced from The Metropolitan Museum of Art's Open Access Initiative. The collection spans:

- **8 curatorial departments**: American Paintings, Asian Art, Drawings & Prints, European Paintings, Islamic Art, Medieval Art, Photographs, and Decorative Arts
- **Temporal range**: Antiquity through 20th century
- **Media types**: Paintings, drawings, prints, photographs, textiles, metalwork, ceramics
- **Geographic coverage**: Global (Americas, Europe, Asia, Middle East, Africa)

This corpus was chosen specifically because GLAM collections represent *verified human ground truth*: every artifact has institutional provenance, verified attribution, and known material properties. There is zero risk of synthetic contamination.

4.2 Processing Statistics

Metric	Value
Total source artifacts	10,097
Enrichment success rate	99.93% (10,090/10,097; 7 failures)
Final PNGs produced	10,000 (99.04% of sources)
Total processing time (8 workers)	181.3 min (∼3 hours)
Average output image size	2.8 MB (PNG, lossless)
Maximum dimension	2,560 px (long edge)
structured provenance JSON size	5–15 KB per artifact (57.2 MB aggregate)
Schema version	v1.1-GCS-CORE
Perceptual hash algorithm	pHash, 256-bit (hash_size=16)
Payload Integrity Hash integrity hash	SHA-256 of canonical JSON
Aggregate processing rate	0.93 items/sec (8 workers)

4.3 Metadata Density Analysis

Each artifact carries a structured provenance JSON document containing 2,000–6,000 tokens of structured context (median ∼2,100 for the 10K batch schema; curated records with expanded rights and archival sections exceed 5,000 tokens). The full nested schema contains 111+ fields across the four ontology layers:

Layer	Key Fields	Source	Coverage
Provenance	Creator, institution, date, medium, accession number	Museum API (human-authored)	100%
Semantic Topology	Title, description, artistic analysis, composition, cultural context, keywords	Museum baseline + Orchestrated Multi-Model Inference	100%
Affective	Affective Interpretation, mood descriptors, thematic resonance	Orchestrated Multi-Model Inference (machine-interpreted)	100%
Rights	Copyright holder, license, commercial use flag	Museum license terms (human-defined)	100%

A note on attribution (see Section 2.4): The "Source" column reflects the epistemic origin of each layer. Provenance and rights data derive directly from the museum's curatorial records—these are human-authored facts. The semantic topology layer begins with the museum's own metadata and extends it through Orchestrated Multi-Model Inference, which articulates latent knowledge that the curatorial record implies but does not spell out. The affective layer is the most interpretive, where the multi-model synthesis generates poetic and emotional readings grounded in the visual and contextual evidence. None of these layers constitute synthetic fabrication; they represent varying degrees of machine-assisted articulation of human knowledge and visual reality.

Comparison to standard web-scraped training data (web-scrape baseline drawn from LAION-5B [**?**] and DataComp [**?**] dataset statistics):

Metric	Typical Web Scrape	Alexandria Aeternum (CN)
Tokens per image	5–20 (alt text) [**?**]	2,000–6,000 (111+ structured fields)
Provenance available	<5% (stripped)	100% (PPP-bound)
Semantic structure	None (flat text)	Graph-based (entities, relations)
Affective signal	None	PAD vectors + poetic interpretation
Rights clarity	Ambiguous	Machine-readable (UDL)
Integrity verification	None	Triple hash (pixel + perceptual + payload integrity hash)

4.4 PPP Transformation Survival

To validate PPP robustness under realistic social media conditions, we subjected 20 randomly selected Alexandria Aeternum artifacts to five transformation classes ($N = 100$ total test cases):

Transform	Mean Sim.	Min Sim.	Recovery (N=20)
JPEG Q75 (Twitter/Instagram-like)	99.5%	98.4%	20/20 (100%)
Resize 50% then 100% (thumbnail bounce)	99.6%	99.2%	20/20 (100%)
JPEG Q30 (worst-case compression)	99.1%	97.7%	20/20 (100%)
Screenshot simulation (1080p + JPEG 85)	99.6%	99.2%	20/20 (100%)
Center crop to 80% (20% per edge)	57.1%	44.5%	0/20 (0%)

Recovery is defined as similarity $\geq \tau$, where τ is an application-specific acceptance threshold. For the results above, $\tau = 85\%$ (the operational LSH matching threshold for provenance resolution). Under the four transforms representative of social media redistribution (compression, resize, and screenshot capture), PPP recovery is **100% (80/80)** with mean similarity 99.5% and minimum observed similarity 97.7%.

The 80% center crop produces 0% recovery, which is the *intended cryptographic behavior*: the crop removes approximately 36% of the total pixel area, fundamentally destroying the artifact's frequency-domain topology. The pHash correctly identifies that the perceptual structure of the artifact has been irreversibly altered, preventing a severely mutilated image from falsely inheriting the provenance of the pristine original. This is a security feature, not a limitation.

Adaptive threshold architecture. A critical design property of the triple-hash system is that the acceptance threshold τ is not fixed but *tunable per application*. The three hashes operate at different tolerance levels by design: the content hash (SHA-256) breaks if a single bit changes, enforcing exact-match verification; the payload integrity hash (SHA-256 of canonical JSON) enforces zero-tolerance metadata integrity; and the perceptual hash (PPP) accepts a configurable degree of visual transformation. This separation enables application-specific threshold policies:

- **Provenance recovery for orphan web images**: An image scraped from the web with no metadata can be hashed and matched against the registry at a permissive threshold (e.g., $\tau = 80\%$), recovering its full semantic payload and institutional provenance from the registry as source of truth. Preliminary testing suggests operational thresholds will converge near $\tau = 95\%$ for high-confidence resolution.
- **Archival verification**: Institutional use cases requiring exact provenance confirmation can demand $\tau \geq 99\%$, accepting only compression-level transformations.
- **Embodied AI applications**: In robotics compute offloading (Section 10), the same perceptual hash architecture indexes pre-computed action schemas. Here, the threshold τ must be tuned per task: a pick-and-place operation in a known factory environment may

tolerate $\tau = 75\%$ to accommodate lighting variation, while a surgical instrument handoff demands $\tau \geq 98\%$ to prevent misidentification.

The ability to customize τ per deployment context—without modifying the underlying hash infrastructure—makes the triple-hash protocol a general-purpose identity resolution system that adapts from archival provenance to real-time robotic perception. Social media platforms typically apply minimal crops (aspect ratio adjustment of 2–5%), well within recovery tolerance at any reasonable τ.

4.5 Dual-Layer Embedding Verification

To verify metadata survival, we extracted and validated the dual-layer embedding on a random sample of 50 output artifacts using ExifTool:

Field	Example	Success
XMP-dc:Title	"Washington Crossing the Delaware"	50/50
XMP-dc:Creator	"Emanuel Leutze"	50/50
XMP-dc:Rights	"(c) 2026 Public Domain..."	50/50
XMP-gc:CodexPayload	gzip+base64 (decompresses to full JSON)	50/50
IPTC:Source	"Provenance Schema v1.1 | Alexandria Aeternum"	50/50

The CodexPayload field on each verified artifact decompressed to a valid JSON document matching the corresponding `_meta.json` file, confirming lossless round-trip encoding.

4.6 Downstream Validation: The Density Imperative

The definitive test of any data enrichment system is whether its output measurably improves downstream model performance. In a companion paper [**?**], we report a controlled ablation study on Llama 3.2 11B Vision-Instruct using 9,081 Alexandria Aeternum images under three conditions: no fine-tuning (Base), sparse captions (Group A), and dense NEST metadata (Group B). The key findings:

- **24.1-point CogBench cognitive swing**: Group A (0.174) vs. Group B (0.415), with Base at 0.378. All eight CogBench dimensions significant at $p < .001$ (Friedman test, $n = 100$).
- **159.6% semantic coverage improvement**: Group B produces structured, multi-dimensional analyses covering composition, color, technique, emotion, symbolism, and narrative—while using 15% fewer tokens than Base.
- **Sparse data actively destroys**: Group A reduces CogBench by 54.0%, increases hallucination rates by 330%, and collapses counterfactual reasoning by 51.3% relative to the un-fine-tuned base model.
- **The metacognitive scaffolding effect**: Dense NEST training teaches the model *how to think about* a domain—producing structured analytical outputs on images never seen during training—not merely what to say about specific examples.

These results validate the core thesis: the semantic density this system produces is not merely richer metadata—it is a *signed* intervention that measurably transforms model capability.

5. Mathematical Framework

The preceding sections describe what the factory *produces*. This section formalizes *how to measure* whether that production improves training outcomes. The tools below—persistent homology and causal disentanglement—define evaluation metrics for future experimental tracks beyond the ablation study reported in [**?**].

5.1 Topological Data Analysis for Distributional Drift

To move beyond informal descriptions of model collapse ("tails disappear"), we employ Topological Data Analysis (TDA) to formalize the geometry of distributional degradation [**?**, **?**].

Persistent Homology studies topological invariants that persist across multiple scales of resolution. Given a dataset $X \subset \mathbb{R}^d$, we construct a nested family of simplicial complexes (a filtration) parameterized by a scale parameter ϵ:

$$K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n \tag{1}$$

As ϵ increases, topological features (connected components, loops, voids) are *born* and *die*. Features that persist across a wide range of ϵ values represent genuine structural properties of the data; short-lived features typically represent noise [**?**].

Application to model collapse detection: We propose computing persistence diagrams for the embedding-space representations of training data across generational iterations $M_0, M_1, \ldots, M_n$. Model collapse manifests as:

1. **Loss of persistent features**: Long-lived topological features (representing distributional tails and modes) progressively vanish across generations.
2. **Concentration of birth-death pairs**: The persistence diagram converges toward the diagonal, indicating that all features become short-lived (noise-like).
3. **Reduction in total persistence**: The sum of all feature lifetimes decreases monotonically, quantifying progressive information loss.

This provides a *mathematically rigorous early-warning system* for autophagy: if the total persistence of a training corpus drops below a threshold, the dataset is losing the structural complexity required for robust generalization.

The CN hypothesis in TDA terms: CN-enriched datasets should exhibit *higher total persistence* and *more long-lived topological features* than equivalent-size datasets of stripped pixels or synthetic data. The NEST semantic graph, by providing dense causal and relational structure, acts as a "topological anchor" that preserves the high-dimensional structure of the data manifold even when mixed with lower-quality samples.

5.2 Spectral Methods for High-Dimensional Persistence

In high-dimensional embedding spaces, standard distance metrics become unreliable (the "curse of dimensionality"). Following Huguet et al. [**?**], we recommend computing persistence using spectral distances on the k-nearest-neighbor graph:

- **Diffusion distance**: Captures multi-scale geometry by integrating over random walks on the data manifold.
- **Effective resistance**: Provides noise-robust distance estimates that allow correct topology detection even in the presence of high-dimensional noise.

These methods are critical for practical deployment, since modern embedding spaces (CLIP, ViT) operate in 512–2048 dimensions.

5.3 Causal Disentanglement

Standard machine learning models capture statistical correlations, which are brittle under distribution shift. A model trained on correlated features (e.g., "dirt on face" correlates with "miner") fails when the correlation breaks. This is the core failure mode that NEST enrichment addresses.

We formalize this using the structural causal modeling (SCM) framework [**?**]:

Let X be the input, $Z = (Z_c, Z_s)$ the latent representation decomposed into causal features Z_c (task-relevant) and spurious features Z_s (environment-specific), and Y the target. The environment E acts as a confounder creating a backdoor path $X \leftarrow E \rightarrow Z_s \rightarrow Y$.

Back-door adjustment blocks this spurious path by conditioning on E:

$$P(Y \mid \text{do}(X)) = \sum_e P(Y \mid X, E = e) \cdot P(E = e) \tag{2}$$

In the NEST pipeline, the dense semantic topology provides the explicit causal structure that enables this disentanglement. When NEST labels an image not just as "man with dirt on face" but as "19th-century coal miner, context: extraction labor, environment: mine shaft, causality: coal dust from manual labor," it provides the model with the causal variables (Z_c) necessary to distinguish genuine semantic content from environmental confounders (Z_s).

The Causal Feature Extractor (CFE) operates in the frequency domain [**?**], assuming that domain-specific (non-causal) signals reside in extreme low and high frequency bands, while mid-frequency components preserve structure-relevant information. A content-aware mask retains mid-frequency causal components while applying perturbations to non-causal bands, forcing the model to attend to causally relevant structure.

This is precisely what NEST enrichment provides at the *data* level: explicit causal structure that would otherwise need to be learned (unreliably) from statistical co-occurrence. Critically, this causal disentanglement is not limited to digital media. In embodied AI applications (Section 10), the same mathematical extraction of causal structure (Z_c) from environmental noise (Z_s) is what enables a perceptual hash to reliably retrieve a deterministic physical action schema regardless of spurious variations in factory lighting, camera angle, or surface wear—the hash resolves to the causal identity of the scene, not its incidental appearance.

5.4 Semantic Density as Information-Theoretic Quantity

We define **Semantic Density** ρ_s of a training sample as:

$$\rho_s(x) = \frac{H(\text{NEST}(x))}{|\text{pixels}(x)|} \tag{3}$$

where $H(\text{NEST}(x))$ is the Shannon entropy of the semantic topology graph (computed over the distribution of entity types, relation types, and structural features), and $|\text{pixels}(x)|$ is the pixel count.

For Alexandria Aeternum assets, we observe ρ_s values approximately 100–300× higher than typical web-scraped image-text pairs (2,000–6,000 structured tokens across 111+ fields vs. 10–50

alt-text tokens, normalized by similar pixel counts). This quantity provides a standardized basis for comparing the "nutritional value" of different training corpora.

The empirical validation of this metric is provided in [**?**], where semantic density coverage scores showed 159.6% improvement for NEST-enriched Group B over Base across all six measured dimensions (composition, color harmony, technique, emotional resonance, symbolic meaning, narrative context), all significant at $p < .001$.

6. Cultural Heritage as Ground Truth Reservoirs

6.1 The GLAM Advantage

The Galleries, Libraries, Archives, and Museums (GLAM) sector holds a unique position in the data ecosystem [**?**]. Their collections represent:

- **Verified human authorship**: Every artifact has institutional attribution and curatorial verification. Zero synthetic contamination.
- **Maximal temporal and cultural variance**: Collections span millennia and every human civilization. This is the ultimate anti-collapse dataset—the full variance of human creative expression.
- **Established open-access frameworks**: Major institutions (Met, Rijksmuseum, Smithsonian, Chicago Art Institute) provide open-access digital collections under CC0 or equivalent licenses.
- **Inherent provenance infrastructure**: Museums maintain conservation records, exhibition histories, and material analyses that map naturally to the PEST provenance layer.

6.2 From Sharing to Stewardship

Museums currently view digital platforms as a means to share collections with the public. We argue this framing is incomplete. In the era of AI training, museum collections are not merely cultural resources to be *displayed*—they are epistemological anchors to be *stewarded*.

The PEST framework offers museums a role beyond passive sharing: active participation in the data stewardship movement. By enriching their collections with NEST semantic topology and embedding PPP identifiers, institutions become *partners* in the training data supply chain, with governance rights over how their collections are represented, used, and compensated.

This represents a "Collections as Data" pivot [**?**]: the production of high-entropy, provenance-verified training assets that command economic value in the emerging Verification Economy.

7. The Amendability Framework: Living Curation

7.1 The Problem of Static Datasets

Current AI training datasets are frozen snapshots. Once published, they are never corrected. If a label is wrong, a provenance claim is inaccurate, or new scholarship revises an attribution, the error persists in every model trained on that dataset. This is antithetical to how human knowledge actually works—scholarship is iterative, and understanding evolves over time.

Prior work has proposed static documentation standards—Datasheets for Datasets [**?**] and Dataset Nutrition Labels [**?**]—that help consumers evaluate a dataset's fitness for use. These frameworks are valuable but fundamentally *descriptive and retrospective*: they characterize data

after the fact without altering its intrinsic properties or enabling correction over time. Cognitive Nutrition datasets should be *living* artifacts that improve over time through institutional curation and scholarly review, extending the concept from static documentation to active, versioned institutional governance.

7.2 Versioned Amendments, Not Overwrites

We propose an amendment chain model inspired by version control systems:

- Each CN Asset has an **immutable base record** (the original mint, permanently preserved).
- **Amendments** are new records that reference the base via a provenance pointer. The original is never overwritten.
- The registry resolves to the **latest curated version** by default, but the full history is traversable.
- Each amendment carries a cryptographic signature from the amending institution or reviewer.

This is analogous to how scholarly publishing works: a correction does not erase the original paper; it adds a corrigendum that points to the original and specifies what changed and why.

7.3 Institutional Governance Gates

Originating institutions retain curatorial authority over their enriched collections:

- **The Met curates Met works.** The Rijksmuseum curates Rijksmuseum works. Each institution has gated access to review and amend *their* artifacts within the global dataset.
- A curator at the Met who identifies that a dress in painting GCX-AA-00042 is sage green, not grey (a model hallucination), can submit a correction. This correction is minted as an amendment, linked to the base record, and signed with the Met's institutional credential.
- **The reviewer's duty is comprehensive.** When a human reviewer corrects a single field, they are expected to review the entire artifact's metadata—not just the flagged issue. This ensures that human verification, when it occurs, is thorough.

7.4 Trust-Level Field Tagging

To enable downstream consumers to filter by epistemic source, we propose a trust-level tagging taxonomy for each metadata field:

Tag	Meaning	Set By
`verified:institution`	Confirmed by originating museum/archive records	Automated
`verified:human`	Reviewed and confirmed by a human curator	First reviewer
`verified:amended`	Originally AI-inferred, subsequently corrected	Institution
`inferred:model`	AI-generated interpretation via Orchestrated Multi-Model Inference	Automated
`inferred:cross-reference`	Derived from external sources or model weights	Automated

This taxonomy enables a critical capability: **filterable trust levels for downstream consumers.**

- A lab requiring *only* human-verified facts can filter to `verified:*` tags, receiving a dataset where every field has been confirmed by a human expert.
- A lab comfortable with machine-articulated enrichment uses all tags, receiving the full 4,000+ token semantic payload.
- A lab studying the reliability of AI enrichment can compare `inferred:model` fields against `verified:human` corrections, producing empirical accuracy metrics for the Orchestrated Multi-Model Inference process itself.

No existing training dataset offers this capability. Trust-level tagging transforms a static dataset into a *configurable epistemological instrument*.

7.5 The Living Dataset Vision

The full Amendability Framework creates a virtuous cycle:

1. **Factory produces** enriched CN Assets with trust-tagged metadata.
2. **Institutions review** their collections, correcting errors and confirming accurate enrichment.
3. **Corrections accumulate** as versioned amendments, improving dataset quality over time.
4. **Labs consume** datasets filtered to their trust requirements, with transparent provenance for every field.
5. **Accuracy metrics** emerge from the ratio of `verified:human` confirmations to `inferred:model` corrections, empirically measuring Orchestrated Multi-Model Inference quality.
6. **The dataset gets better over time**, unlike static training corpora that degrade through uncorrected errors.

8. The Verification Economy

8.1 Data as Asset Class

As forecast in [**?**], the exhaustion of high-quality human data creates a market bifurcation:

- **"Unverified synthetic data"**: Unverified, possibly synthetic, context-stripped data. Abundant, cheap, and increasingly dangerous for training.
- **"Verified Human Data"**: Verified human data with provenance, semantic density, and rights clarity. Scarce, valuable, and essential for model stability.

Foundation model developers facing the existential risk of autophagy will pay premiums for Verified Human Data datasets. The economic question is not *whether* this market emerges, but *how* it is structured.

8.2 Algorithmic Pricing via Data Shapley

The Shapley value from cooperative game theory quantifies the marginal contribution of a data point to model performance [**?**]. Recent advances in Fast-DataShapley [**?**] provide one-pass estimation, making it computationally feasible to assign per-artifact values in large datasets.

The PEST registry enables this by maintaining the mapping from artifact (via PPP) to provenance (via metadata bundle) to training usage (via ingestion logs). When a model lab uses CN-enriched data, the registry can compute:

1. Which artifacts were used (PPP tracking).
2. How much each contributed to downstream performance (Shapley estimation).
3. What compensation is owed (UDL terms).

This closes the loop between data creation, data consumption, and data compensation—the "3 Cs" of generative governance: Consent, Credit, and Compensation [**?**].

9. Limitations and Open Questions

1. **PPP is currently hash-based, not steganographic.** The current implementation uses perceptual hashing (pHash) for content-addressable identification, not true steganographic embedding as described in the theoretical PEST framework [**?**]. The hash survives transformations because it is *recomputed* from the transformed image, not *extracted* from a hidden channel. A full EDAN steganographic encoder remains future work.
2. **Affective layer validation.** The affective layer (Affective Interpretation) is generated by multi-model Orchestrated Multi-Model Inference and has not been validated against human judgments. Whether these vectors improve training outcomes is an open empirical question.
3. **Single domain.** Alexandria Aeternum covers fine art and decorative arts. Whether CN generalizes to other domains (scientific imagery, satellite data, medical imaging) requires investigation.
4. **Adversarial robustness.** We have not tested PPP survival against purpose-built adversarial scrubbers. As steganographic techniques improve, so will adversarial attacks.
5. **Scale validation.** 10K artifacts demonstrate feasibility but not scale behavior. Performance characteristics at 1M or 50M artifacts may differ.
6. **Trust-level tagging is proposed, not implemented.** The Amendability Framework (Section 7) is architectural design; the v1.2-GCS-TRUST schema is in development.
7. **Downstream validation is domain-specific.** The ablation study in [**?**] evaluates cultural heritage imagery. Whether the density imperative generalizes to other domains with PEST-enriched data requires future work.

10. Future Work

Multi-institutional scaling. Extend Alexandria Aeternum from Met-only to Rijksmuseum, Smithsonian, Chicago Art Institute, and other open-access collections to test cross-institutional generalization.

Domain expansion. Apply the PEST factory to scientific imagery, satellite data, and medical imaging to test whether Cognitive Nutrition is domain-general.

Recursive contamination stress tests. Train successive model generations with increasing fractions of synthetic data, comparing CN-anchored versus unanchored conditions to measure collapse resistance.

TDA implementation. Compute persistent homology diagrams for embedding-space representations of CN-enriched versus web-scraped datasets to empirically validate the topological anchor hypothesis.

Amendability Framework deployment. Implement the v1.2-GCS-TRUST schema with trust-level field tagging and institutional governance gates for pilot testing with museum partners.

Perceptual Compute Offloading for robotics. The same principle underlying PEST—dense structured metadata indexed by perceptual hash—has preliminary applications in embodied AI, where hash-based lookup of pre-computed action schemas replaces expensive VLA inference for recognized environments [**?**].

11. Conclusion

The Entropy of Recursion [**?**] argued that Data Stewardship must replace Data Accumulation. *The Density Imperative* [**?**] proved empirically that the quality of the supervision signal is signed—sparse data destroys, dense data enhances. This paper demonstrates that stewardship is not merely a philosophical position but an *engineering practice* with concrete implementation, measurable outputs, and economically viable scaling characteristics.

We have shown that:

- A production metadata factory employing multi-model Orchestrated Multi-Model Inference can process 10,097 artifacts with 99.93% success rate, producing dual-layer provenance-embedded assets at approximately \$0.10–\$0.12 per artifact (varying with API pricing and image complexity).
- Each asset carries 2,000–6,000 tokens of structured semantic context across 111+ fields—approximately 100–300× the information density of typical web-scraped training data—rooted in the Translation Paradigm, where machine enrichment articulates existing human knowledge rather than synthesizing from nothing.
- A triple-hash integrity protocol (pixel, perceptual, payload integrity hash) provides defense-in-depth verification, achieving 100% PPP recovery under four social media transformation classes.
- The mathematical foundations of Cognitive Nutrition can be formalized through topological data analysis (persistent homology for drift detection) and causal disentanglement (back-door adjustment for separating causal from spurious features).
- The Amendability Framework enables living curation: institutional governance, versioned corrections, and trust-level field tagging that lets downstream consumers select their epistemic confidence level.
- Controlled ablation [**?**] confirms that this system's output produces a 24.1-point CogBench cognitive swing, 159.6% semantic coverage improvement, and the metacognitive scaffolding effect—dense structured training teaches models how to think, not merely what to say.

The infrastructure of memory for the AI age is not a metaphor. It is a system: agents, schemas, hashes, registries, and immutable records. We have built it, and we have proved that the signal it produces measurably transforms the models that consume it.

The era of indiscriminate Data Accumulation is over, halted by the exhaustion of human-generated text and the thermodynamic limits of brute-force scaling. The future of foundation models—both in digital reasoning and embodied robotics—depends entirely on the fidelity of the

signal they consume. Cognitive Nutrition is the architecture of that signal. It is mathematically sound, empirically validated, and economically deployable at scale. The factory is open.

12. Data Availability

The Alexandria Aeternum 10K dataset [?] (DOI: 10.5281/zenodo.18359131) and all experimental materials are available for research use:

- **Dataset**: https://huggingface.co/datasets/Metavolve-Labs/alexandria-aeternum-genesis
- **Dataset (codex-curator)**: https://huggingface.co/datasets/codex-curator/alexandria-aetern
- **Experiment package**: https://huggingface.co/Metavolve-Labs/cognitive-nutrition-experime (LoRA adapters, evaluation scripts, raw outputs, interactive comparison tool)
- **Interactive evaluation explorer**: https://iaeternum.ai/evaluation
- **Source code**: `codex-lab-kit` [?] (https://github.com/codex-curator/codex-lab-kit); `gcp-robotics` [?] (https://github.com/codex-curator/gcp-robotics)
- **License**: CC BY 4.0 for enrichment metadata; underlying artworks are public domain (Met Museum Open Access).

Competing Interests

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

Author Contributions

System architecture, factory engineering, pipeline development, dataset production, and paper authorship by Tad MacPherson. Multiple AI systems were utilized for enrichment execution, literature synthesis, and drafting support under human direction. All technical assertions and empirical claims have been verified by the lead author.

Acknowledgments

The Metropolitan Museum of Art for providing open access to their collection via the Open Access Initiative. The Alexandria Aeternum dataset [?] represents enrichment of public-domain works and does not claim ownership of the underlying artifacts.

Software. The factory pipeline is implemented using the `codex-lab-kit` toolkit [?] and the `gcp-robotics` SDK [?].

References

[1] T. MacPherson, "The Entropy of Recursion: A Strategic Framework for Provenance-Verified Data and the Preservation of Signal Integrity in Generative Systems," Metavolve Labs, 2025. DOI: 10.5281/zenodo.18436975.

[2] T. MacPherson, "The Density Imperative: How Semantic Curation Depth Determines Vision-Language Model Capability," Metavolve Labs, 2026. DOI: 10.5281/zenodo.18667735.

[3] I. Shumailov, Z. Shumaylov, Y. Zhao, Y. Gal, N. Papernot, and R. Anderson, "AI models collapse when trained on recursively generated data," *Nature*, vol. 631, pp. 755–759, 2024.

[4] E. Dohmatob, Y. Feng, and J. Yang, "A Closer Look at Model Collapse: From a Generalization-to-Memorization Perspective," arXiv:2509.16499, 2025.

[5] A. Alemohammad, J. Casco-Rodriguez, et al., "Self-Consuming Generative Models Go MAD," in *Proc. ICLR*, 2024.

[6] P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho, "Will we run out of data? Limits of LLM scaling based on human-generated data," Epoch AI, 2024.

[7] S. Gadre, G. Ilharco, A. Fang, et al., "DataComp: In Search of the Next Generation of Multimodal Datasets," in *Proc. NeurIPS Datasets and Benchmarks*, 2023.

[8] G. Penedo, Q. Malartic, D. Hesslow, et al., "The RefinedWeb Dataset for Falcon LLM: Outperforming Curated Corpora with Web Data Alone," in *Proc. NeurIPS*, 2023.

[9] S. Xie, H. Pham, X. Dong, N. Du, H. Liu, and Y. Lu, "DoReMi: Optimizing Data Mixtures Speeds Up Language Model Pretraining," in *Proc. NeurIPS*, 2023.

[10] C. Schuhmann, R. Beaumont, R. Vencu, et al., "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models," in *Proc. NeurIPS Datasets and Benchmarks*, 2022.

[11] P. Fernandez, G. Couairon, H. Jegou, M. Douze, and T. Furon, "The Stable Signature: Rooting Watermarks in Latent Diffusion Models," in *Proc. ICCV*, 2023. See also: Google DeepMind, "SynthID," 2024.

[12] F. Chazal and B. Michel, "An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists," *Frontiers in AI*, 2021.

[13] H. Adams, T. Emerson, M. Kirby, et al., "Persistence Images: A Stable Vector Representation of Persistent Homology," *JMLR*, vol. 18, 2017.

[14] H. Edelsbrunner and J. Harer, "Persistent Homology – a Survey," *Contemporary Mathematics*, 2008.

[15] G. Huguet, A. Tong, et al., "Persistent Homology for High-dimensional Data Based on Spectral Methods," in *Proc. NeurIPS*, 2024.

[16] J. Pearl, *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge University Press, 2009.

[17] Q. Zhang, et al., "Causal Visual Generalization via Learning Causal Features and Geometric Topology," preprint, OpenReview, 2024.

[18] "Now more than ever: the role of museum and archival objects in an age of generative artificial intelligence," *Collection and Curation*, Emerald, 2025.

[19] D. Cohen, "AI and Libraries, Archives, and Museums, Loosely Coupled," *Humane Ingenuity*, 2024.

[20] A. Ghorbani and J. Zou, "Data Shapley: Equitable Valuation of Data for Machine Learning," in *Proc. ICML*, 2019.

[21] J. Wang, et al., "Fast-DataShapley: Neural Modeling for Training Data Valuation," arXiv:2506.05281, 2025.

[22] "Governance of Generative AI in Creative Work: Consent, Credit, Compensation, and Beyond," arXiv:2501.11457, 2025.

[23] C. Zhou, P. Liu, P. Xu, et al., "LIMA: Less Is More for Alignment," in *Proc. NeurIPS*, 2023.

[24] HuggingFace, "FineWeb: A Large-Scale Open Web Dataset," Technical Report, 2024.

[25] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski, "The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards," arXiv:1805.03677, 2020.

[26] T. Gebru, J. Morgenstern, B. Vecchione, et al., "Datasheets for Datasets," *Communications of the ACM*, vol. 64, no. 12, 2021.

[27] T. MacPherson, "Perceptual Compute Offloading: Hash-Indexed Action Schemas for Embodied AI," Metavolve Labs, 2026. DOI: 10.5281/zenodo.18667749.

[28] T. MacPherson, "Alexandria Aeternum Dataset," Metavolve Labs, 2025–2026. DOI: 10.5281/zenodo.18359131. Available: https://huggingface.co/datasets/codex-curator/alexandria-aeternum.

[29] T. MacPherson, "codex-lab-kit v1.0.1," Metavolve Labs, 2026. DOI: 10.5281/zenodo.18668110. Available: https://github.com/codex-curator/codex-lab-kit.

[30] T. MacPherson, "gcp-robotics SDK," Metavolve Labs, 2026. DOI: 10.5281/zenodo.18668113. Available: https://github.com/codex-curator/gcp-robotics.