

ThatAIguy

AI · CREATIVITY · REAL RESULTS

thataiguy.org →

RESEARCH PAPER · VERSION 0.7 · NOVEMBER 27, 2026

Neuronal Sea.

Exact Mathematical Memory and the Limits of Further Compression

JOSEPH ANADY

THATAIGUY RESEARCH

01	02	03	04
32,228	655,116	5.876×	82.98%
canonical records, zero mismatches	shared structural nodes	active checkpoint ratio	reduction from canonical content

RESEARCH PAPER · NEURONAL SEA 0.7

Abstract

Mathematical memory must preserve more than the appearance of an equation. Its operators, argument order, domains, assumptions, examples, relationships and provenance determine what can be recalled and applied correctly. This paper describes Neuronal Sea 0.7, a native software prototype that represents canonical mathematical records as shared typed structures, compresses immutable pages, and interprets supported programs through generation-bound node references. The frozen checkpoint contains 32,228 records and 655,116 structural nodes. Its 73,419,411 bytes of canonical record content occupy 12,494,630 bytes in the active checkpoint, including its relation graph and snapshot metadata: a 5.876-fold ratio and 82.9818 percent reduction. All records passed a canonical hash comparison. We add a reproducible six-run compression study using identical exported inputs and three conventional codecs. A records-only XZ stream occupies 9,052,952 bytes, but omits graph and access indexes. Recompressing a complete active-checkpoint archive does not improve on the existing active file sum; the smallest tested wrapper is 14,699 bytes larger. We therefore distinguish measured storage savings from conditional opportunities in representation, grammar, indexing and page organization. Source-extracted diagrams show real shared expressions, concept relationships and native agent connectivity. Further figures explain why pattern overlap, learned decoders and wave transforms require explicit information accounting. The results establish exact canonical preservation and bounded execution on the supplied corpus. They do not establish a universal compression optimum, biological memory equivalence, general mathematical mastery or performance on an untested Apple device.

KEYWORDS

Mathematical knowledge representation · structural sharing · lossless compression · exact arithmetic · selective interpretation · reproducible systems research.

MAIN OBSERVATION

CANONICAL RECORDS AUDITED

32,228 with zero mismatches

ACTIVE CHECKPOINT

12,494,630 bytes

REDUCTION FROM CANONICAL RECORD CONTENT

82.9818 percent

FURTHER REDUCTION FROM COMPLETE-CHECKPOINT RECOMPRESSION

None in the three tested configurations



SECTION 01

Research question and contribution

The research question is whether one compact representation can preserve mathematical records, connect their concepts and support exact operations without reconstructing every stored object for each request. The intended use is a mathematical artificial narrow intelligence with an explicit, inspectable capability boundary. The present implementation combines storage, retrieval, finite mathematical kernels, a bounded program learner, and a native reference-transport mechanism. Figure 1 locates these components and their responsibilities. [A1, A4]

The contribution reported here is an implementation and measurement study of that combination. Structural sharing, content-addressed storage and compression-aware execution already have substantial prior art. The evidence does not support a claim that those general ideas were invented by this project. Instead, the paper makes the representation, its exactness contract, its current overheads, and its remaining limitations concrete enough to reproduce and improve. [3, 5, 6]

Three distinctions organize the analysis. A canonical mathematical record differs from the original document that supplied it. A compact representation differs from the complete state needed to use it. And a program that reads selected structures differs from an unrestricted intelligence capable of understanding every stored statement. Each distinction changes the denominator or the correctness claim in a compression experiment.

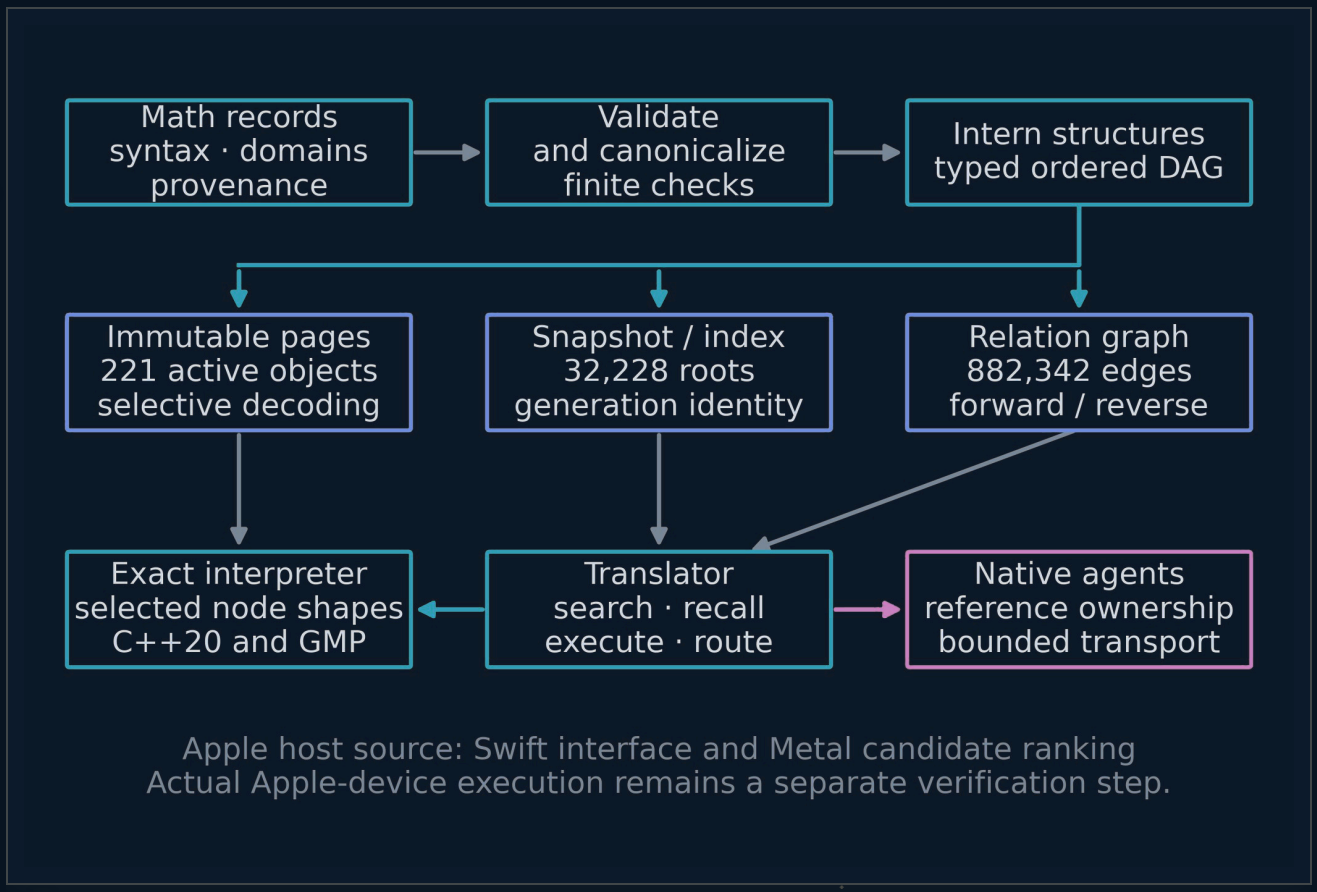
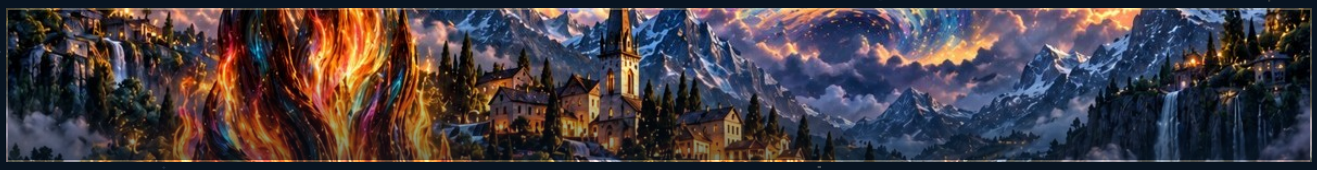


FIGURE 01

Implemented representation and control boundaries. The arrows summarize data dependencies, not physical wires or biological pathways. The shared Sea holds exact records; the relation graph connects records; the native agents transport references. Apple interface and ranking source is included in the release, while actual Apple-device execution remains a separate verification step. Sources: [A1, A4].



SECTION 02

Shared patterns as exact mathematical structure

An accepted record is encoded as a typed, ordered directed acyclic graph, or DAG. Primitive values and containers have distinct tags. Arrays preserve operand order and repeated children. Object fields have a deterministic order. When a newly encoded structure has exactly the same payload bytes as an existing one,

the writer reuses its identifier. The interning decision compares complete encoded structure; a hash is not treated as proof that two mathematical meanings are equivalent. [A1, A4]

This permits many records to share one stored term. The actual checkpoint supplies a useful example: node 639694 represents `x + y`. The mean program multiplies that term by one-half, its specification divides it by two, and the square-sum program raises it to the second power. Figure 2 shows these actual identifiers. Integer two is also shared, although its role differs between division and exponentiation. [A4]

Hash-consing is established prior art for identifying and sharing structurally equal values. The relevant question for this system is how its chosen term representation, compressed pages and access indexes interact on mathematical workloads. A small illustrative DAG explains the mechanism; only the complete byte accounting establishes the system's compression ratio. [3]

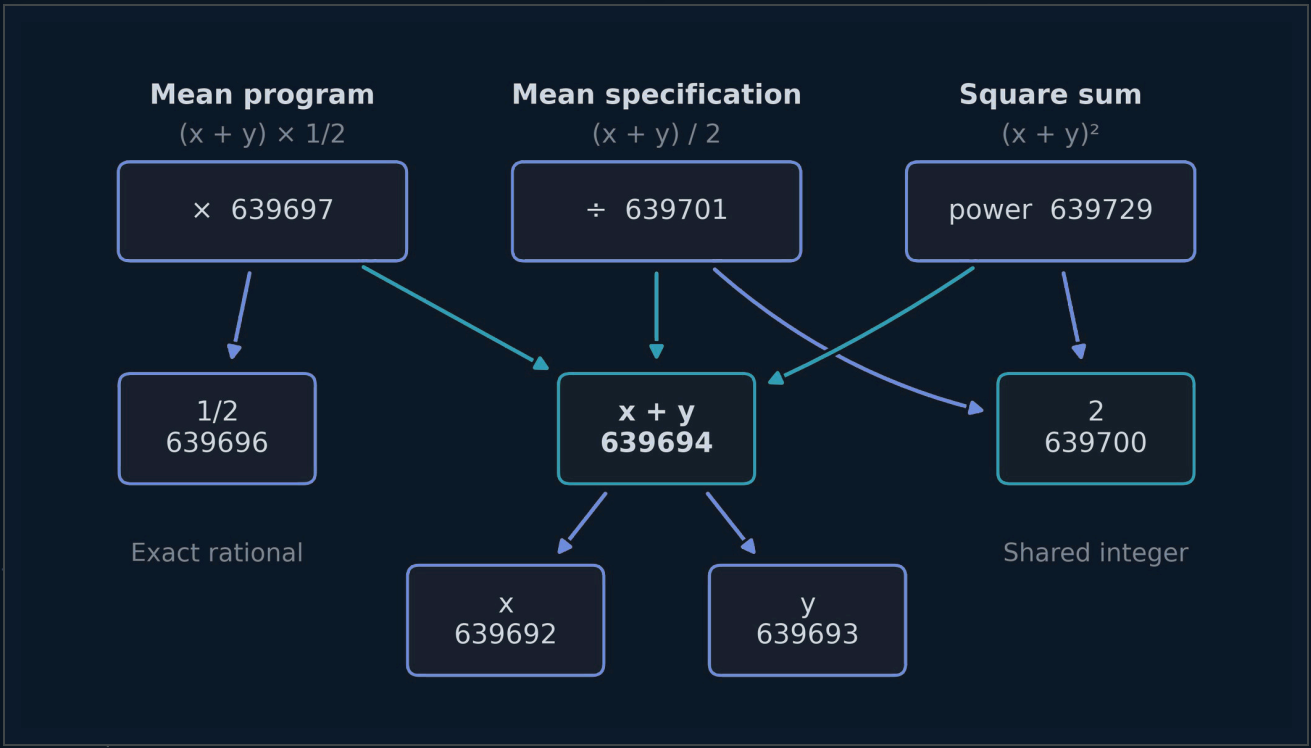
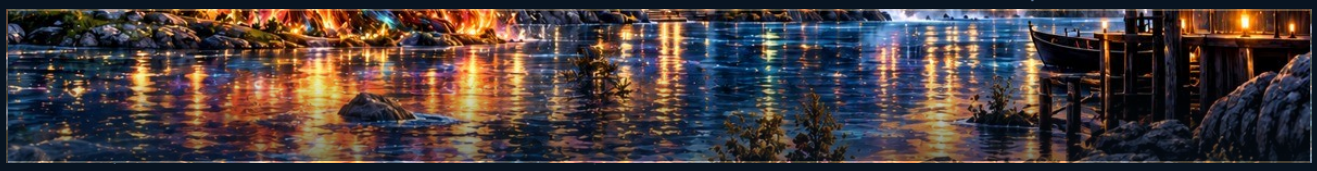


FIGURE 02

Source-extracted shared mathematical terms. Operator-tag strings and other record fields are collapsed for readability. Edge crossings are not junctions. With x = 1/2 and y = 1/3, native evaluation gives sum 5/6, mean 5/12 and square sum 25/36. These are fresh exact checks of retained programs, not a separate compression benchmark. Source: [A4].



SECTION 03

What the stored pattern must mean

A node's mathematical role depends on its type, its position and the interpreter's rules. The numeral two can be a denominator, an exponent, a matrix dimension or a count. Preserving the symbol alone does not preserve those distinctions. Neuronal Sea therefore retains mathematical structure and associated record fields, while the interpreter supplies a bounded grammar for applying supported terms. The translator does not need a separate complete copy of every pattern, but it does require an explicit language of operators and checks. [A1, A4]

The actual common-denominator lesson illustrates a fuller meaning record. It contains a definition, notation, nonzero-denominator conditions, prerequisites, a worked example, a nonexample, a source URL and verification scope. Its checked example is 5/12 + 7/18 = 29/36. This context is part of the stored data. A cue vector used to locate the lesson is a separate, replaceable retrieval aid. [A4]

ASPECT RETAINED	WHY IT MATTERS
Operator and ordered arguments	Distinguishes division, multiplication, power and operand roles
Domain and assumptions	Prevents using a rule where its original expression is undefined
Definitions and examples	Connects a formal expression to its educational interpretation
Prerequisites and related concepts	Supports directed navigation through mathematical dependencies
Provenance and verification scope	Distinguishes a source statement, a finite check and a proof-backed result

A domain example shows why this is necessary. Solving the following equation over the reals gives an apparent candidate x = 1 after cancellation, but the original denominator excludes that point. The retained native solver returned the empty solution set in the recorded inspection. [A4]

$$\frac{x^2-1}{x-1} = 2, \qquad x \neq 1$$

The system must preserve both the expression and the restriction. Similarly, structural sharing cannot automatically identify every semantically equivalent formula. Such identification is valid only when a specific equivalence procedure establishes it under the required assumptions. The current direct-program checker has

exact polynomial-equivalence logic within its admitted grammar; this does not establish equivalence for arbitrary mathematics.



SECTION 04

Translation and selective interpretation

The translator is a dispatcher with distinct modes: search, answer, execute, render, inspect, route, calculate and match. The mode determines what must be reconstructed and which checks are authoritative. A search ranks candidates using cue and lexical information. An explicit recall reconstructs requested content. A calculation parses a newly supplied expression and invokes a mathematical kernel. Direct execution follows an admitted stored program. These operations should not be conflated. [A1, A4]

For supported direct execution, the runtime verifies generation identity, program status, grammar, domain and exact agreement with the supplied specification. It decodes selected pages and reads one-hop node shapes. It then builds bounded program descriptors, coefficient maps and exact working values. It avoids constructing the entire stored JSON record and its nested JSON program or specification tree along that path. Figure 3 makes the page-decoding step explicit. [A2, A4]

Compression-aware execution has established precedent. Some encodings support operations directly; others require decompression before the relevant operator runs. In Neuronal Sea, avoiding whole-record materialization is the supported claim. Executing arbitrary arithmetic directly on an undecoded XZ byte stream is not the implemented mechanism. [5]

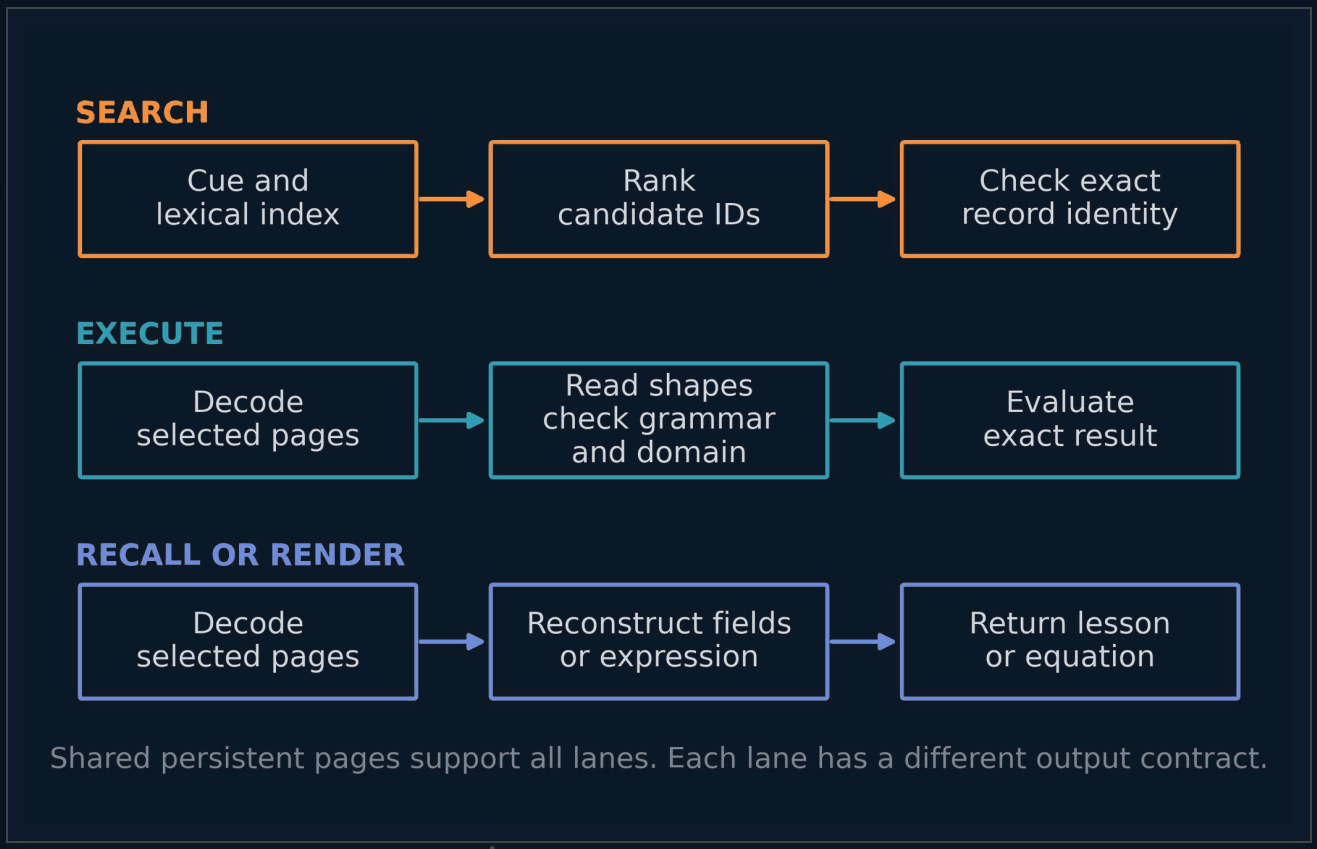
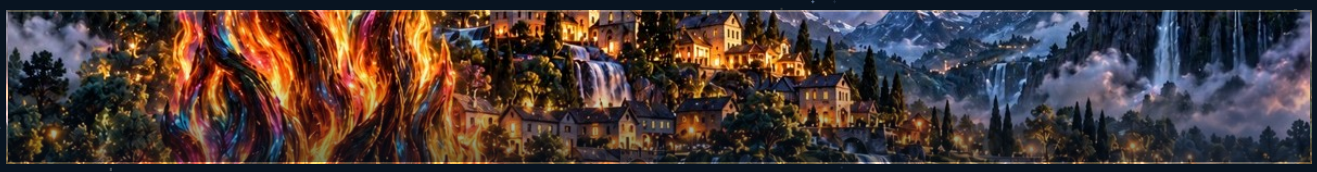


FIGURE 03

Three implemented paths through the representation. Candidate ranking proposes identities; exact record checks and mathematical rules determine supported results. Page decoding is present in the direct path. Recall and rendering intentionally construct the selected content needed by the caller. Sources: [A1, A2, A4].



SECTION 05

Connections between mathematical concepts

The relation graph connects concepts, formal assertions and learned-program records. Its active version has 32,226 graph nodes and 882,342 edges; these counts differ from the 655,116 structural nodes because they describe a different representation layer. Structural nodes encode values and subexpressions. Relation-graph nodes identify knowledge records and their declared associations. [A1]

A prerequisite edge is directed from the dependent concept to its prerequisite. Reverse adjacency allows the runtime to find material that depends on a known concept. Other edge kinds have different roles, including proof references, illustrative links and curated contextual relationships. Traversal uses explicit visit, edge and hop budgets. An exhausted budget means that exploration is incomplete; it does not prove that two concepts are disconnected. [A4]

Figure 4 is an induced subgraph of actual current records. It connects addition, integers, rational numbers, reduced fractions, common denominators, weighted averages and the mean program. It shows how a single shared mathematical store can support cross-topic navigation without partitioning exact content into isolated subject silos. External prerequisites are omitted from the figure for legibility.

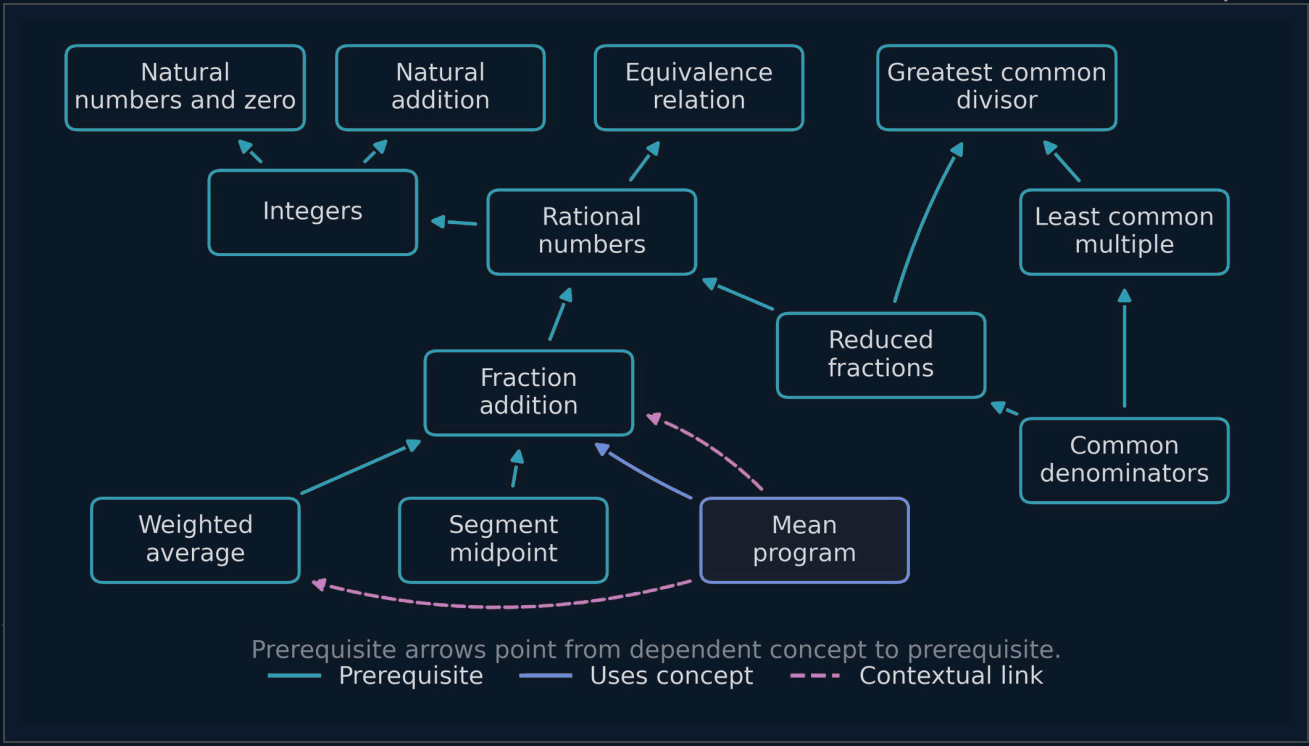


FIGURE 04

Thirteen actual nodes and fifteen actual stored edges. Solid teal prerequisite arrows point dependent to prerequisite. Blue and dashed magenta edges have distinct operational and contextual meanings. The layout and shortened labels are schematic. Connections encode context; they do not themselves establish theorem validity or learned mastery. Source: [A4].



SECTION 06

Native agents and the Gate 5000 boundary

The native transport component is separate from the shared structural store. Its retained configuration contains 88 agent positions, 44 pair bonds and 133 shards: one for each agent, one for each bond, and a substrate-directory shard. Longitudinal movement follows 16 configured family paths, with 72 reciprocal up/down links. The actual connectivity in Figure 5 is more specific than the general double-helix design metaphor. [A4]

The C++ bridge stores a bounded envelope containing a record identity, a generation-bound pattern reference and a requested operation. The full mathematical payload remains in the Sea. Recall checks the owner, digest, record identity and generation before exact C++ interpretation. This gives the agent or bond a verifiable reference to the knowledge associated with it. It does not make each native agent an independent learned mathematical model. The current bridge explicitly reports that native mathematical learning is absent. [A4]

Gate-5000 contains 5,000 gate-state entries and 20,000 edges for activation and scheduling. Its public interface excludes treating that activity as the exact knowledge store. Routing is bounded by permissible neighbors, repeated-visit checks and work limits. Reciprocal transport can contain graph cycles even when an individual job terminates and the structural DAG is acyclic.

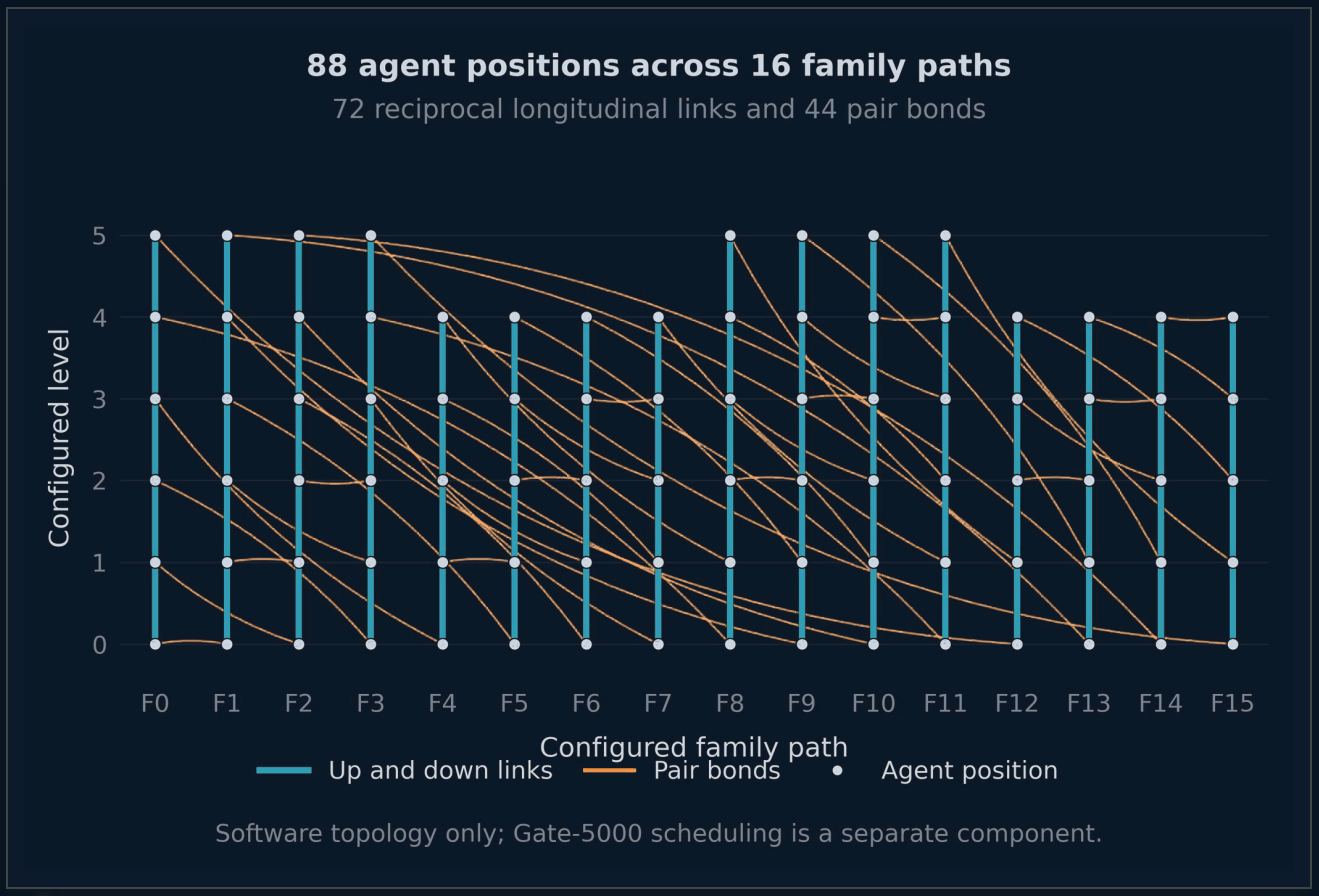


FIGURE 05

Retained software connectivity. Source family and level values determine plotted positions; they are not physical distances. Teal links show longitudinal neighbors and orange links show pair bonds. The diagram is not a biological neuron map, optical memory or literal two-strand physical helix. Source: [A4].



SECTION 07

Scope of the mathematical foundation

The installed catalog contains 691 educational concepts across 51 topic identifiers. It combines 221 legacy cards with 470 added curriculum cards. Topics span elementary numeration and arithmetic through algebra, calculus, probability, topology, measure theory, functional analysis and foundations. Topic names and counts measure what has been stored; they are not measurements of competence on every problem in those fields. [A1]

Within the new 470-card curriculum, 242 cards have 259 declared native finite checks, all of which passed. The remaining 228 are reference-only under the current checker. Legacy cards have separate verification history, so Figure 6 does not relabel them as newly checked or unverified by the same procedure. Keeping the categories separate prevents a coverage chart from becoming a false mastery chart. [A1]

The checkpoint also retains 31,052 formal assertions from a selected extract of the Metamath `set.mm` database, pinned to commit `c82282455b6ccfe3ab3eedf922705463fbe2bf7b`. This includes 29,356 proof-bearing assertions and 1,696 axioms, definitions and syntax entries. The retained provenance records the extraction policy and source assumptions. These proofs are inherited mathematical work. Their formal verification and storage identity are distinct from the finite examples attached to educational cards. [A1, 13]

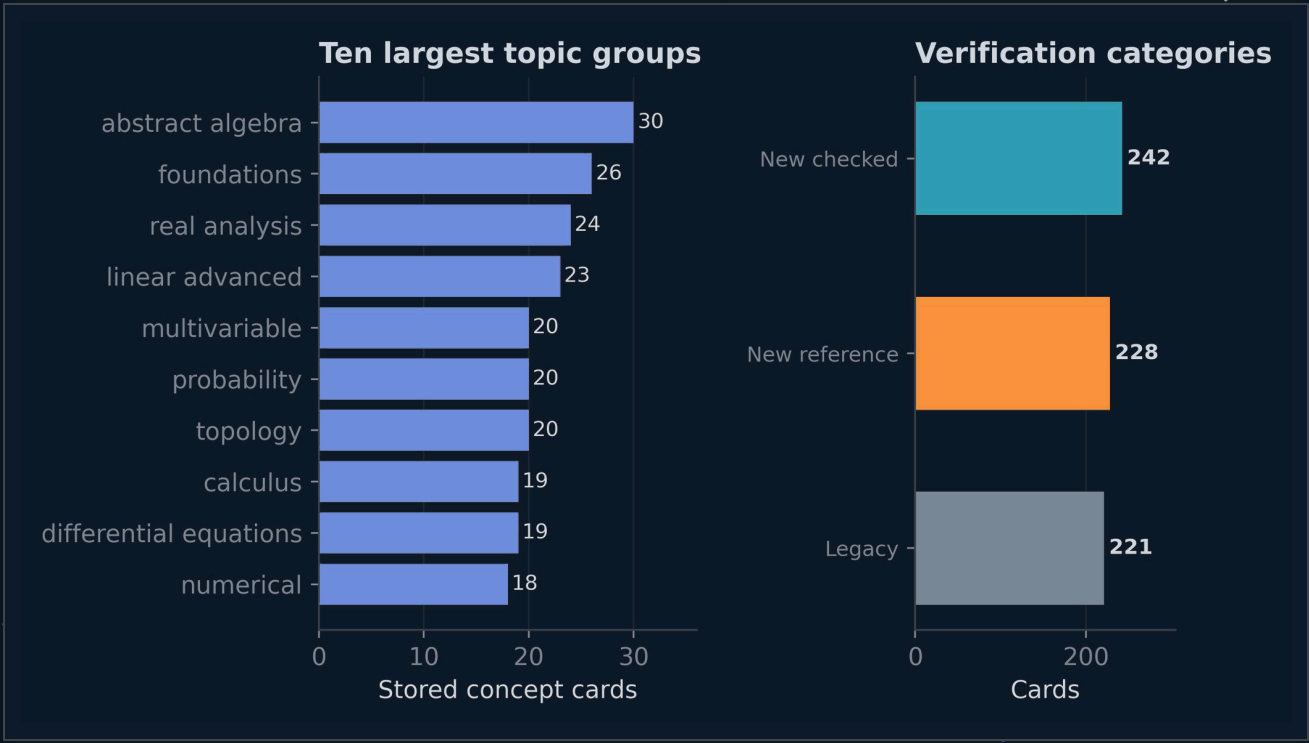


FIGURE 06

Stored curriculum coverage. The left panel shows the ten largest topic groups. The right panel separates the 242 new cards with finite checks, 228 new reference cards and 221 legacy cards. No bar represents the percentage of a mathematical field mastered. Source: [A1].



SECTION 08

Correctness contracts and their evidence

The primary storage contract is exact recovery of accepted canonical records. Let r be a record after the documented normalization and let E and D be the encoder and decoder. The intended property is:

$$D(E(r)) = r$$

This is a canonical-record guarantee. Whitespace, source-document typography, screenshots and original file packaging are outside that equality unless separately retained. Exact recovery also says nothing by itself about whether a mathematical assertion in r is true. A perfectly reproduced incorrect statement is still incorrect.

The structural design supports an inductive argument for recovery. Primitive tags carry their accepted values. Container nodes carry ordered references to previously created children. If those children decode correctly and their order and tags are preserved, the parent reconstructs correctly. Exact interning substitutes only an equal structure. Child identifiers preceding parents support acyclicity. This describes the representation's intended reasoning; it is not a machine-checked proof of the complete C++ implementation. [A4]

The frozen release's full audit compared all 32,228 reconstructed canonical records with expected SHA-256 values and found zero mismatches. The added study traversed the same checkpoint through the native API, checked each record before writing it, and began compression only after all 32,228 comparisons had passed. Both checks share parts of the production decoding path, which limits their independence. A separately implemented decoder would strengthen this evidence. [A1, A3]

The dated execution-path addendum exercised 2,102 fixture cases covering 419 active programs and seven inactive records. Throwing linker wrappers intercepted four full-materializer entry points, and all four positive controls were active. The exercised direct paths made zero calls to those wrapped entry points. This is evidence about those specific APIs and cases. It does not mean that no pages, shapes, coefficients, metadata or answer objects were constructed. [A2]

EVIDENCE LAYER	ESTABLISHED BY THE RECORDED CHECK	OUTSIDE THAT CHECK
Canonical record hashes	Equality of decoded canonical records	Original formatting and mathematical truth
Domain and exact kernel tests	Expected results for finite declared cases	All possible expressions and inputs
Formal proof verification	Validity under the retained proof system and assumptions	Autonomous theorem discovery
Materializer interception	Absence of calls to four APIs on exercised paths	Absence of all reconstruction or allocation



SECTION 09

Measured storage and its denominator

The active checkpoint occupies 12,494,630 logical bytes, compared with 73,419,411 bytes obtained by summing the canonical records. The ratio is 5.876077 and the reduction is 82.981844 percent. Figure 7 shows both the baseline and the components charged to the active representation. File sizes are logical byte lengths, not filesystem block allocation. [A1]

$$R = \frac{73{,}419{,}411}{12{,}494{,}630} \approx 5.876077 \qquad S = 1 - \frac{12{,}494{,}630}{73{,}419{,}411}$$

The compressed pages dominate the active total. The graph, root index, manifest and current-generation pointer are also required and are counted. The entire state directory is larger because it includes retained history and auxiliary files. Source copies, test data, executable code and linked libraries are excluded from the active-checkpoint ratio and must be considered separately for deployment. [A1, A3]

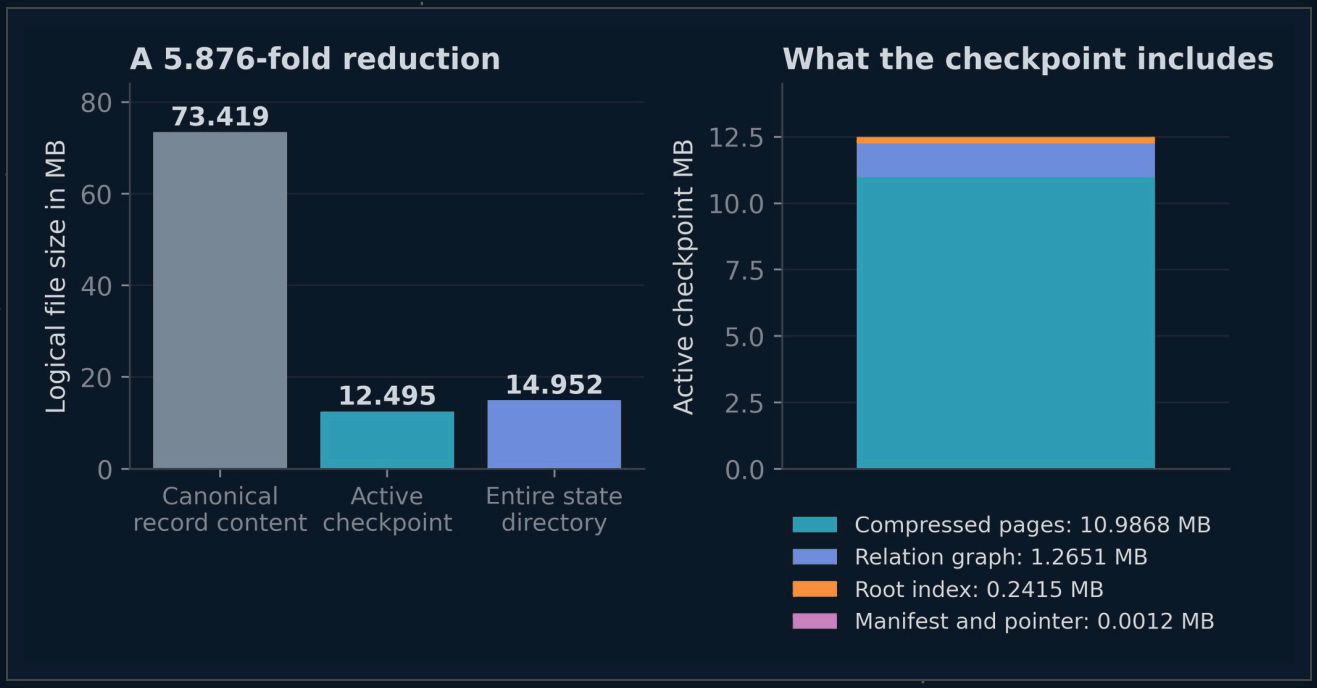
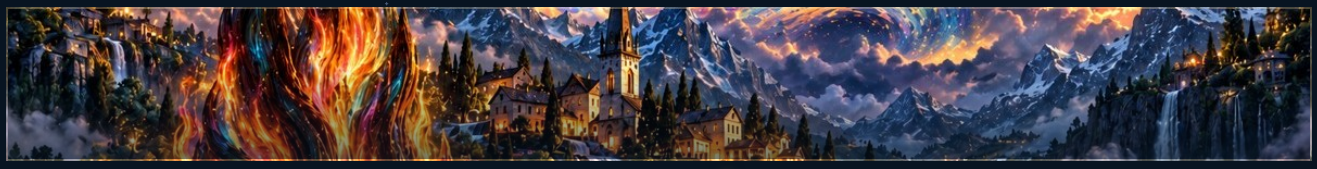


FIGURE 07

Current logical-file accounting in decimal MB, where one MB is 1,000,000 bytes. The active checkpoint includes 10,986,796 page bytes, 1,265,101 graph bytes, 241,511 index bytes, a 1,189-byte manifest and a 33-byte pointer. The full state directory occupies 14,951,719 bytes. Source: [A1].

The node store alone is 11,228,307 bytes, but quoting that smaller number as the whole system would omit the relation graph and snapshot metadata. Likewise, the 33,410,386-byte distribution ZIP is not the active-memory footprint: it includes source, reports and other release material. The appropriate denominator depends on the specific claim.



SECTION 10

A new controlled compression study

We performed a bounded additional study on the exact frozen checkpoint. A study-only C++ adapter used the existing native store API to enumerate records in root order, reconstruct each canonical value and verify its expected hash. The concatenated JSONL stream contains 73,451,639 bytes: the 73,419,411 canonical bytes plus exactly 32,228 newline separators. All three codecs received this identical byte stream. [A3]

A second stream contains the complete active checkpoint. The study copied all 225 referenced active files, whose logical sizes sum to 12,494,630 bytes, then created a deterministic tar archive with normalized metadata. The tar stream is 12,677,120 bytes, including container overhead. Each codec compressed that identical tar stream. Decoding every result reproduced its corresponding input SHA-256. The frozen release manifest and source-state hashes remained unchanged. [A3]

The settings were gzip 1.13 at level 9, XZ 5.8.1 at level 9, and Zstandard 1.5.7 at level 19. Jobs ran sequentially on Linux x86-64 during 2026-11-28 00:02:50–00:04:47 UTC, corresponding to November 27 in Chicago. Each configuration was measured once. No training corpus or external dictionary was supplied. [A3]

Codec wall times include process startup, file I/O and completion polling at nominal two-millisecond intervals. They exclude durable fsync; the operating-system cache was not flushed. Peak RSS is the Linux child-process high-water mark. The timings are descriptive observations, not statistically stable rankings or durable-storage throughput measurements. [A3]

INPUT AND CODEC	OUTPUT BYTES	ENCODE SECONDS	DECODE SECONDS
Records only with gzip 9	11,528,082	4.333	0.573
Records only with XZ 9	9,052,952	30.439	0.415
Records only with Zstandard 19	9,463,084	40.054	0.144
Complete active tar with gzip 9	12,530,613	0.362	0.067
Complete active tar with XZ 9	12,634,388	4.433	0.148
Complete active tar with Zstandard 19	12,509,329	3.468	0.019

TABLE 01

New same-input compression measurements. Records-only outputs exclude the graph and retrieval indexes. Complete-tar outputs include the active checkpoint but require restoration before ordinary queries. None is a replacement query engine. Source: [A3].



SECTION 11

What the new measurements establish

The XZ records-only stream is 9,052,952 bytes, or an 87.6749 percent reduction from its 73,451,639-byte JSONL input. This is a valid lossless archival result. It is not a 9 MB version of the complete queryable ANI. The native graph and access structures are absent, and an offset into the uncompressed stream does not by itself permit independent access inside a whole-stream compressed file. [A3]

The study also produced a 3,147,464-byte explicit record index containing offsets, root identifiers and hashes, and a 10,701-byte metadata file. These are useful reproducibility sidecars, not an optimized competing database design. Their existence makes the accounting issue visible: the cost of identifying, locating and validating records cannot disappear from a fair system comparison. Codec executables and libraries are also external dependencies, just as the native application's runtime code is external to its checkpoint. [A3]

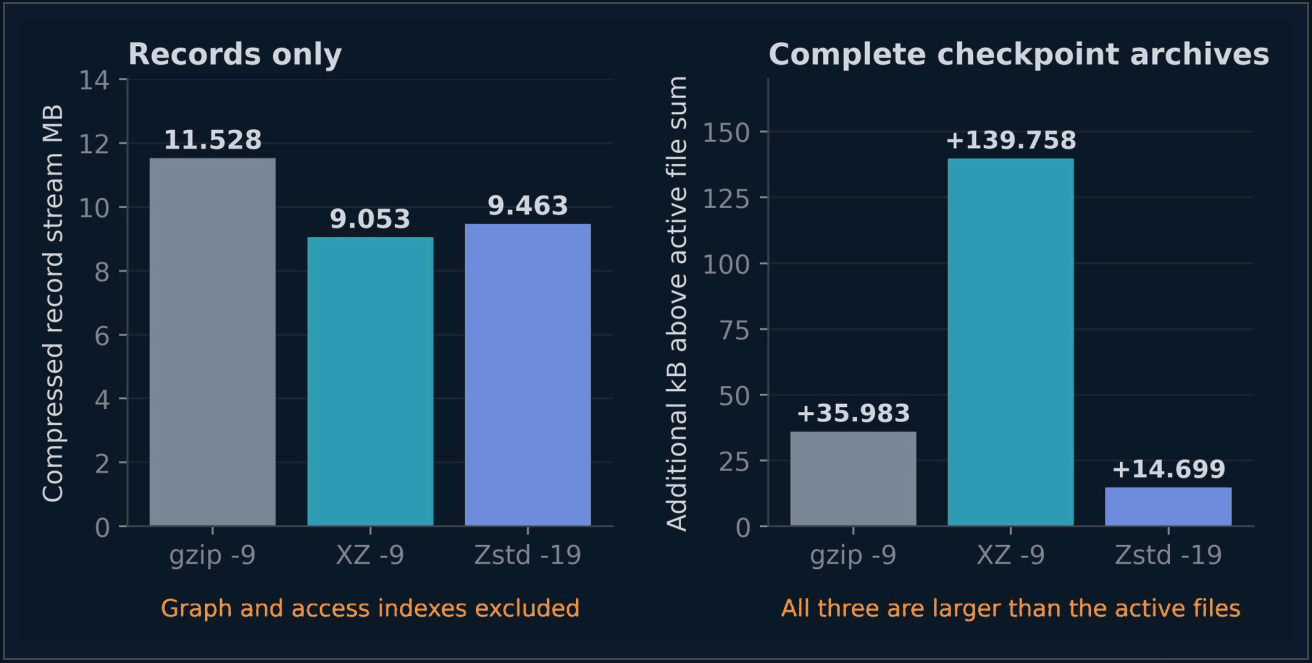


FIGURE 08

Two different questions with different denominators. Left: archival compression of the same canonical JSONL stream. Right: extra bytes above the current active file sum after compressing the complete deterministic tar. Even the smallest tested complete wrapper adds 14,699 bytes. The right panel uses a zero-based difference axis; positive bars mean larger files. Source: [A3].

The current pages are already compressed, so the negative recompression result is plausible without implying optimality. It rules out the specific tested strategy of wrapping these active bytes in those codec configurations as a space-saving improvement. It does not rule out a different page arrangement, a new mathematical grammar, a better index representation or another codec. Those changes require their own exactness and access measurements.



SECTION 12

Incremental knowledge growth

The curriculum extension added 470 concept cards and one curriculum-index record. Comparing the recorded pre-extension and current active footprints gives an increase from 12,341,855 to 12,494,630 bytes, a difference of 152,775 bytes. The concept count rose from 221 to 691. Figure 9 decomposes the reported storage increase. [A1, A3]

The comparison is historical, not a controlled ablation. It includes changes to the relation graph, index, manifest and a newly introduced 33-byte current-generation pointer. The added page-file total is 122,012 active bytes. This differs from the 140,204 bytes of newly written page objects because 18,192 old object bytes became unreferenced by the active generation. Retained history can still occupy space outside the active footprint. [A3]

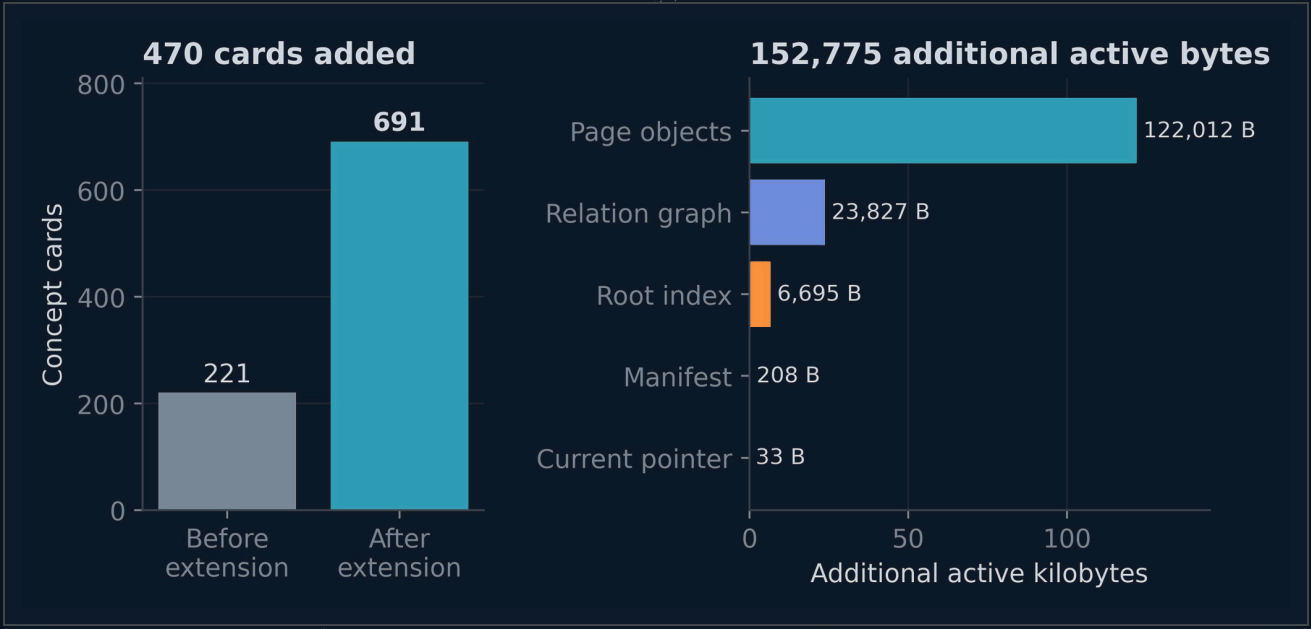
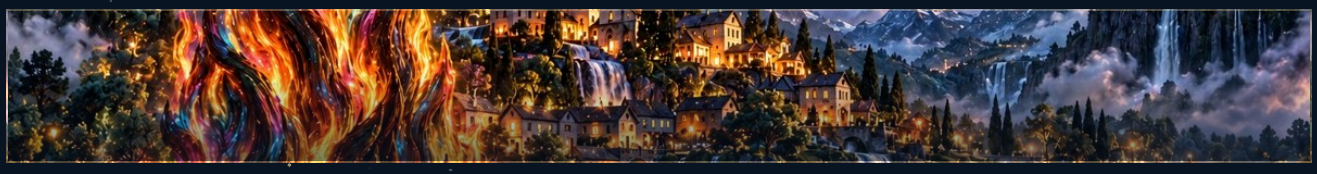


FIGURE 09

Historical active-footprint growth. The additional active bytes comprise 122,012 page bytes, 23,827 graph bytes, 6,695 index bytes, 208 manifest bytes and 33 pointer bytes. Source generations and their receipts are preserved in the study. This is a corpus-specific observation, not a storage-per-concept law. Sources: [A1, A3].

The arithmetic average is about 325 additional active bytes per added card, with the extra index record and graph effects included. It should not be extrapolated to all future mathematics. Highly repetitive lessons can reuse existing fields and terms; new proofs, unfamiliar symbols or long unique explanations can require much more new information. Growth experiments must vary novelty as well as record count.



SECTION 13

Execution speed and working memory

The retained native benchmark executed 1,000 distinct exact-mean inputs through one persistent engine. All 1,000 answers were checked using independent bounded integer arithmetic and greatest-common-divisor calculations. The first invocation took 14.018749 milliseconds. The 999 subsequent calls had a median of

0.164596 milliseconds and a 95th percentile of 0.211075 milliseconds. Figure 10 reports these summaries without inventing an unrecorded latency distribution. [A1]

Engine construction took 0.672543 seconds outside the individual invocation timers. The operating-system file cache was not flushed, and request serialization, transport and process creation were excluded from those timers. The result therefore describes a local supported operation under its recorded cache conditions. It does not establish end-to-end service latency, an Apple Silicon result, or a speed advantage over a matched conventional implementation. [A1]

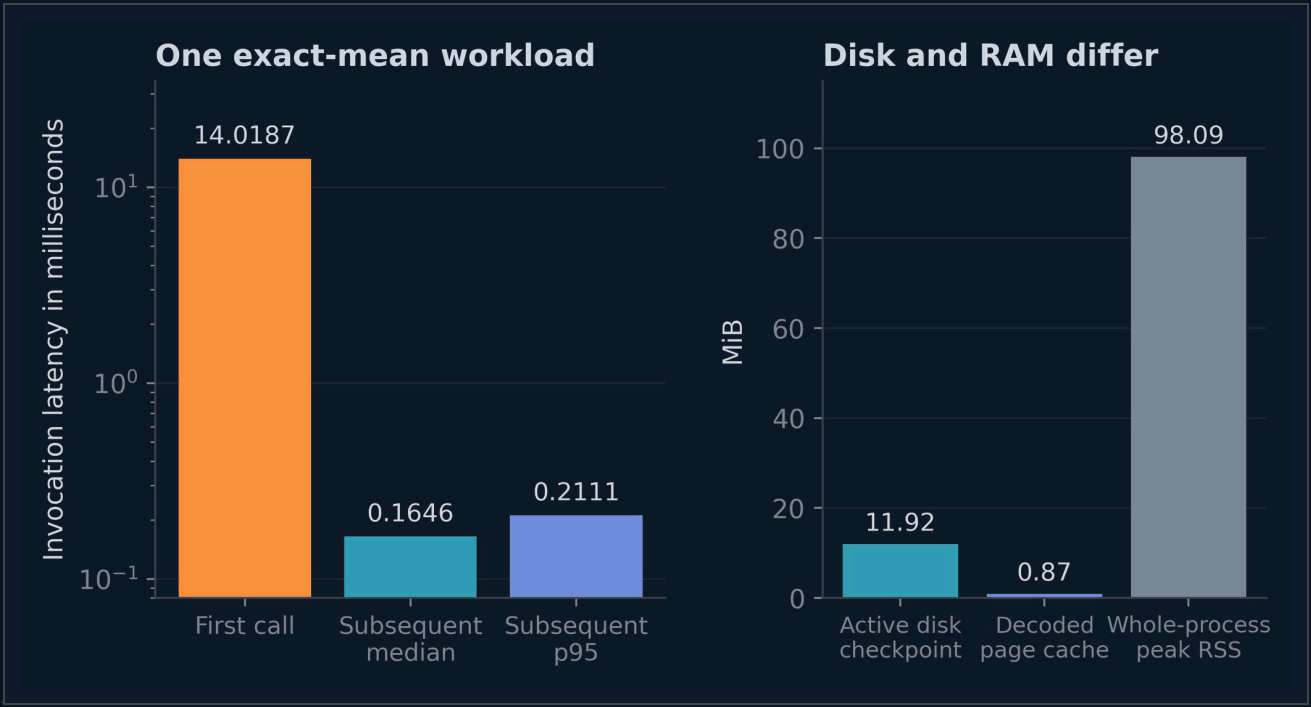


FIGURE 10

Left: retained exact-mean summary timings on a logarithmic latency axis. Right: storage and memory scopes in MiB, where one MiB is 1,048,576 bytes. Whole-process peak RSS is 98.09 MiB; the seven decoded pages occupy about 0.875 MiB. These are different accounting scopes, not additive bars. Source: [A1].

The process peak was 102,854,656 bytes, including libraries, graph metadata, indexes, caches, allocator overhead and measurement allocations. Candidate indexing, ingestion and helix activation were not exercised by that benchmark. A small persistent checkpoint cannot be equated with the total RAM requirement. In the new codec study, XZ's largest observed encoder peak was about 642.4 MiB, further illustrating that a small output may require a larger working set. [A1, A3]



SECTION 14

Finite stress testing and learning evidence

The retained foundation stress workload executed 101,804 tasks with 248,608 assertions. It covers rational arithmetic, polynomial derivatives, modular residues, matrix determinants, finite statistics and sets, probability, linear equations and systems, combinatorial identities, and invalid-domain rejection. Its deterministic seed and independent bounded oracles are recorded. Figure 11 shows the workload composition. Repeated cases are executions, not distinct mathematical discoveries. [A1]

The 5,000 invalid-domain cases are successful rejections rather than numerical answers. Treating them as ordinary correct calculations would obscure a necessary capability: refusing an operation when its declared domain is violated. The separate curriculum checks and full-record audits address different questions and should not be pooled into one universal intelligence score. [A1]

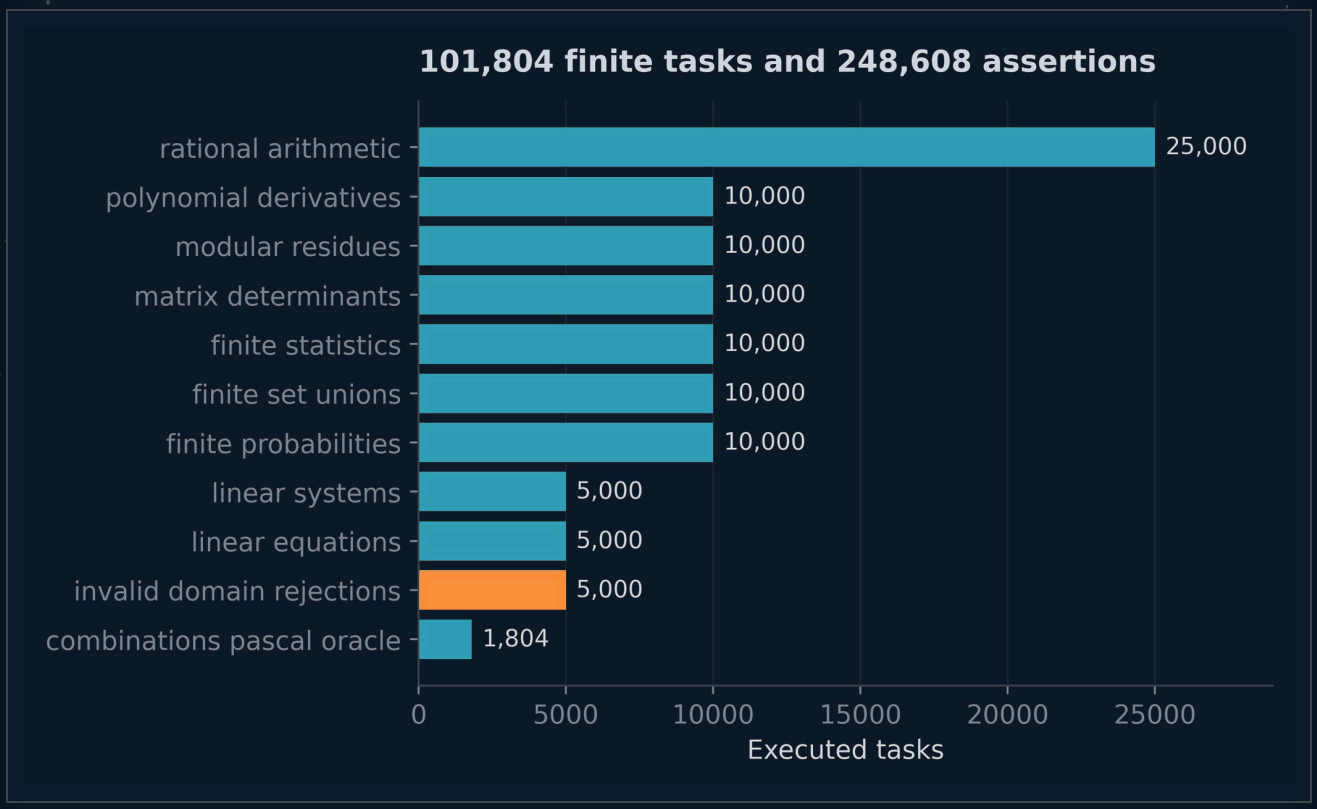


FIGURE 11

Finite regression workload. Orange identifies expected invalid-domain rejections. Other bars count supported task executions with their recorded oracles. Passing these tests narrows known implementation risks; it does not prove the absence of all bugs or establish general theorem-proving ability. Source: [A1].

The implemented learner searches a bounded rational-polynomial grammar. A supplied exact specification judges candidate identity, while training replay and distinct held-out inputs provide additional checks. An unresolved search remains unresolved or inactive. Reference-card admission is a separate operation and does not automatically create an executable rule. The release's 419 active programs therefore represent capability-specific evidence, not unrestricted learning from raw text. [A1, A4]



SECTION 15

Quantifying further compression opportunities

The largest measured target is the page payload: 10,986,796 bytes, or 87.9321 percent of the active checkpoint. The other components total 1,507,834 bytes. Suppose a future change reduces the page bytes by fraction s while preserving those other components and adding M bytes of model, dictionary or index overhead. Its resulting size would be:

$$B_{\text{new}} = 1{,}507{,}834 + (1 - s)\,10{,}986{,}796 + M$$

This is an accounting equation, not a prediction that any listed reduction is achievable. Figure 12 shows its consequences. A 10 percent reduction of page bytes produces only an 8.793 percent reduction of the whole checkpoint when M is zero. Additional model cost reduces the benefit. [A1; derived arithmetic]

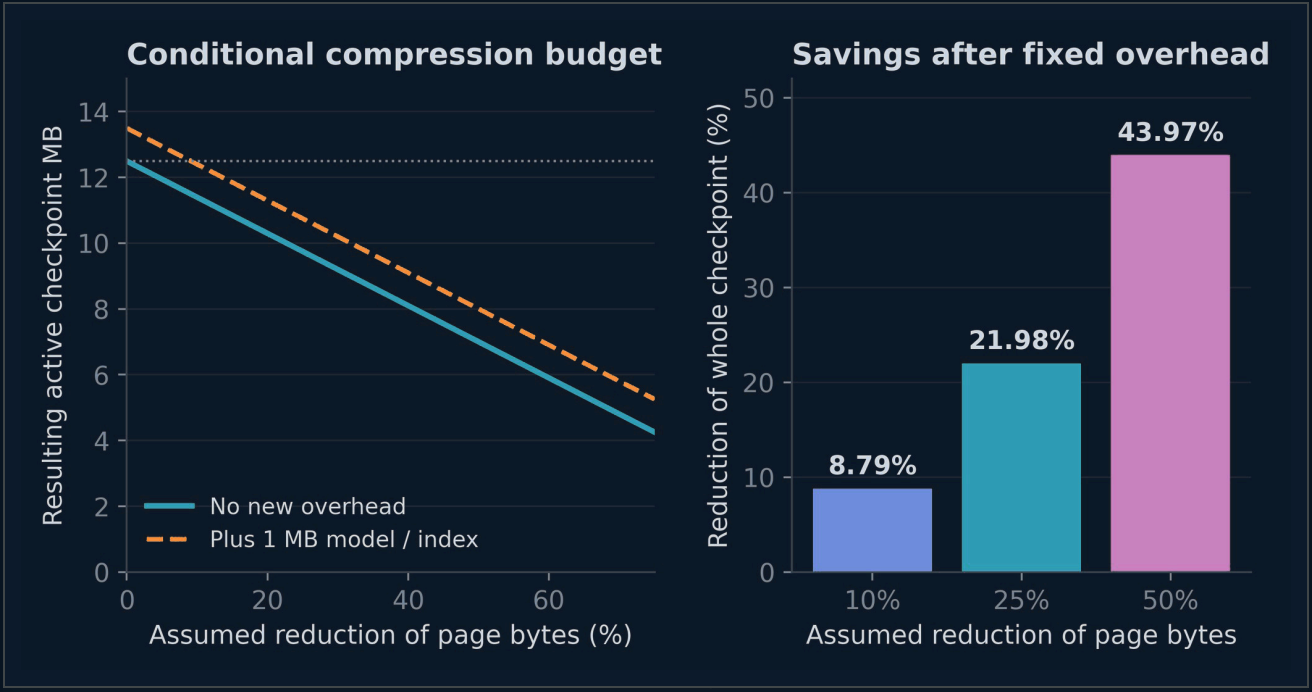


FIGURE 12

Conditional scenarios, not measurements. The left panel holds graph, index and metadata bytes fixed and compares zero extra overhead with a 1 MB addition. The right panel shows whole-checkpoint savings for three assumed page reductions. These curves are not entropy bounds or performance forecasts.

ASSUMED PAGE REDUCTION	RESULTING CHECKPOINT MB	ADDITIONAL WHOLE-CHECKPOINT REDUCTION
10 percent	11.396	8.793 percent
25 percent	9.748	21.983 percent
50 percent	7.001	43.966 percent

For the last scenario, the total reduction from the original canonical content would be 90.4641 percent. No tested implementation in this study achieved that queryable size. The table states a concrete engineering target and the required accounting assumptions.



SECTION 16

Pattern overlap and recoverable identity

Distinct patterns can overlap without damaging information when their identities, structure and associations remain recoverable. The shared `x + y` term in Figure 2 is an example: separate roots preserve which larger expression is being requested. Sharing one exact substructure does not erase the distinct multiplication, division or power operation above it.

Unlabelled aggregation is different. Figure 13 constructs two different pairs of binary patterns with the same elementwise sum. Observing that sum alone cannot identify which pair was stored. This is an elementary ambiguity in the representation, even though each original pattern is distinct. More elaborate binding schemes can change the interference behavior, but their keys, cleanup memory or residual information must be counted.

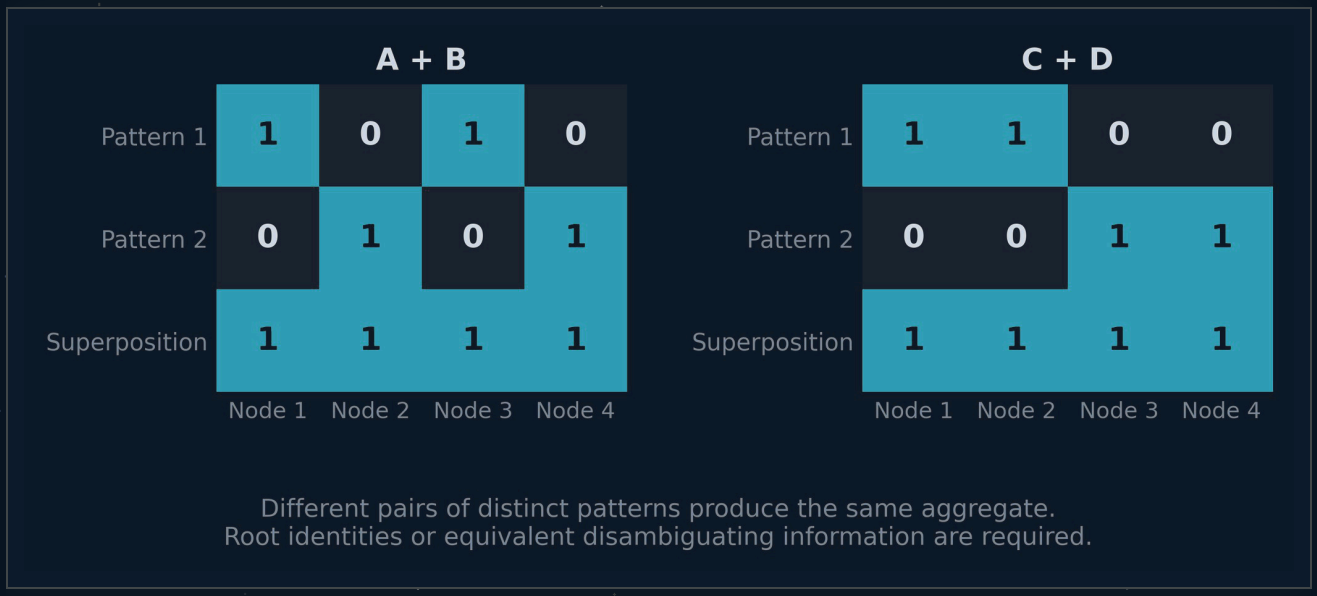


FIGURE 13

Constructed mathematical counterexample, not a measured Neuronal Sea failure. A = (1,0,1,0) and B = (0,1,0,1) have the same sum as C = (1,1,0,0) and D = (0,0,1,1). Both displayed pairs give (1,1,1,1). The exact Sea preserves separate roots and ordered references rather than relying on this aggregate alone.

Associative-memory research offers useful retrieval mechanisms. Hopfield's model retrieves a stored association from a sufficiently informative partial cue. Plate's holographic representations combine structured items but explicitly use cleanup memory for noisy reconstructions. Hyperdimensional methods support binding, superposition and similarity while retaining finite reliable capacity. These are relevant designs for a candidate-retrieval layer. They do not establish that a fixed-size vector can losslessly archive an unlimited number of arbitrary mathematical records. [7, 8, 9]

A useful future design can combine associative candidate selection with an exact store. The candidate mechanism helps decide where to look; exact identifiers, stored assumptions and mathematical checks determine what can be returned. The current cue-index design follows this separation at a limited scale. [A4]



SECTION 17

Waves transformations and physical information

An invertible change of basis can expose regularity that a practical compressor uses. It cannot, by itself, remove the information needed to invert it. Figure 14 uses an exact four-element Hadamard transform to make this distinction visible. With H chosen so that its transpose times H equals four times the identity, the coordinates satisfy:

$$c = Hx, \qquad x = \frac{H^{\mathsf{T}}c}{4}$$

A repeated input produces one nonzero coefficient, while the irregular example produces four. An encoder may exploit the sparse case if coefficient values, positions, length and decoding rules cost fewer bits than the original. The transform alone does not guarantee that outcome. Discarding nonzero coefficients or quantizing them changes fidelity unless an exact residual is retained.

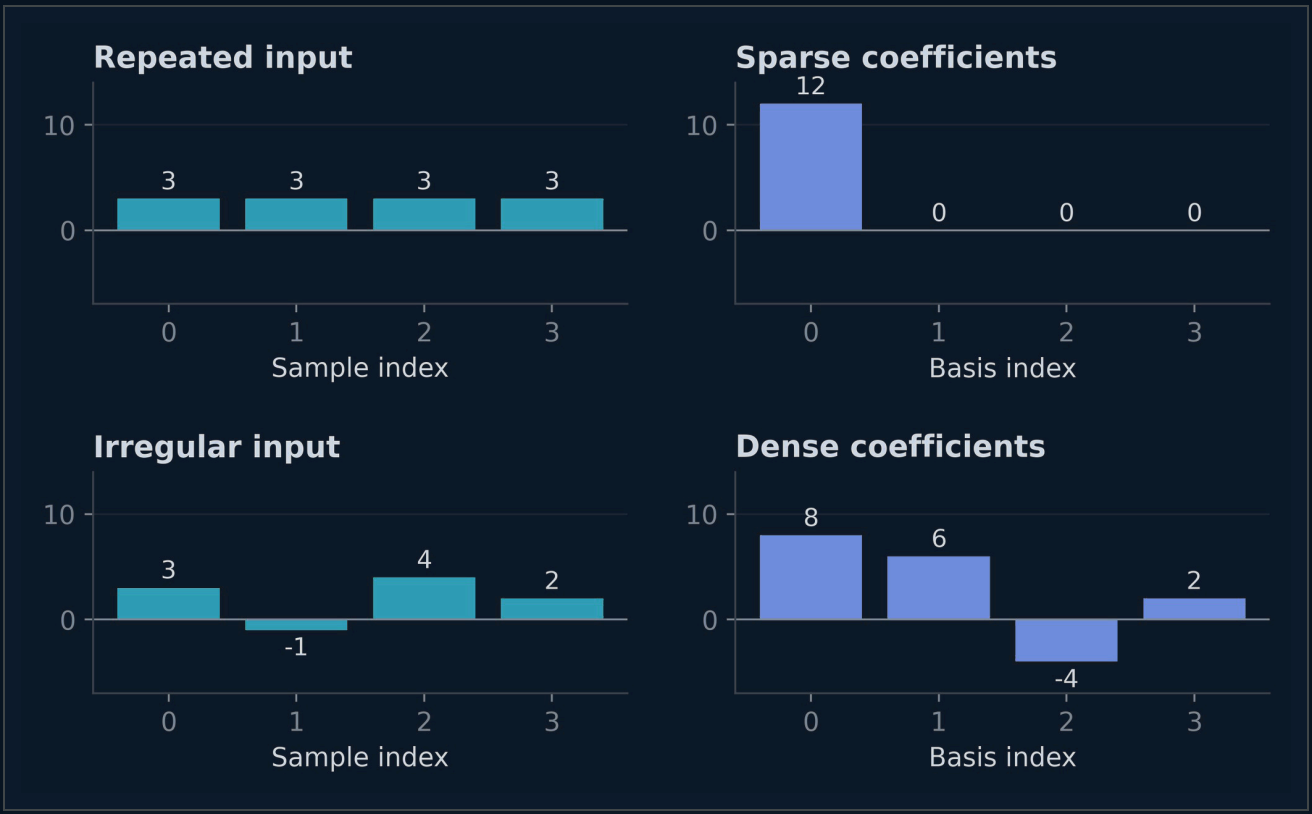
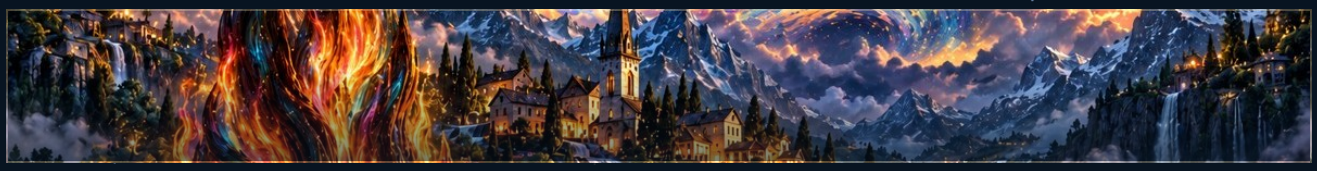


FIGURE 14

Exact constructed examples. The repeated vector (3,3,3,3) maps to (12,0,0,0); (3,-1,4,2) maps to (8,6,-4,2). The inverse recovers both inputs exactly. These panels illustrate representation choices; they are not observations of photon storage or biological waves.

Physical proposals require a separate capacity model. Distinguishable states depend on finite precision, noise, energy, bandwidth or available modes, and on reliable readout. Giovannetti and colleagues derive an energy-constrained capacity for a specific lossy bosonic communication channel. That result is not a measured storage limit for Neuronal Sea, but it shows why a photon claim needs explicit physical assumptions. [12]

No optical device, biological tissue or wave-based physical memory was used in this study. Neuronal Sea's demonstrated compression comes from digital structure sharing and page codecs. A future physical substrate would need measurements of retention, error rates, write/read cost, control state and the complete recovery process before its density could be compared with these software results.



SECTION 18

Keeping the translator and model costs honest

A general interpreter can apply a finite grammar to many stored programs without memorizing every complete expression. This is the useful form of reuse sought by the translator design. The grammar and execution code remain necessary shared state. If a future learned model replaces some explicit structures, its parameters and any exact residual must also remain in the accounting.

A simple hypothetical example gives the break-even condition. Suppose independent records cost 100 bytes each. A common model costs 1,000 bytes and each record then needs a four-byte reference, with no residual or other costs. The shared scheme is smaller only when:

$$1000 + 4N < 100N, \qquad N > \frac{1000}{96}$$

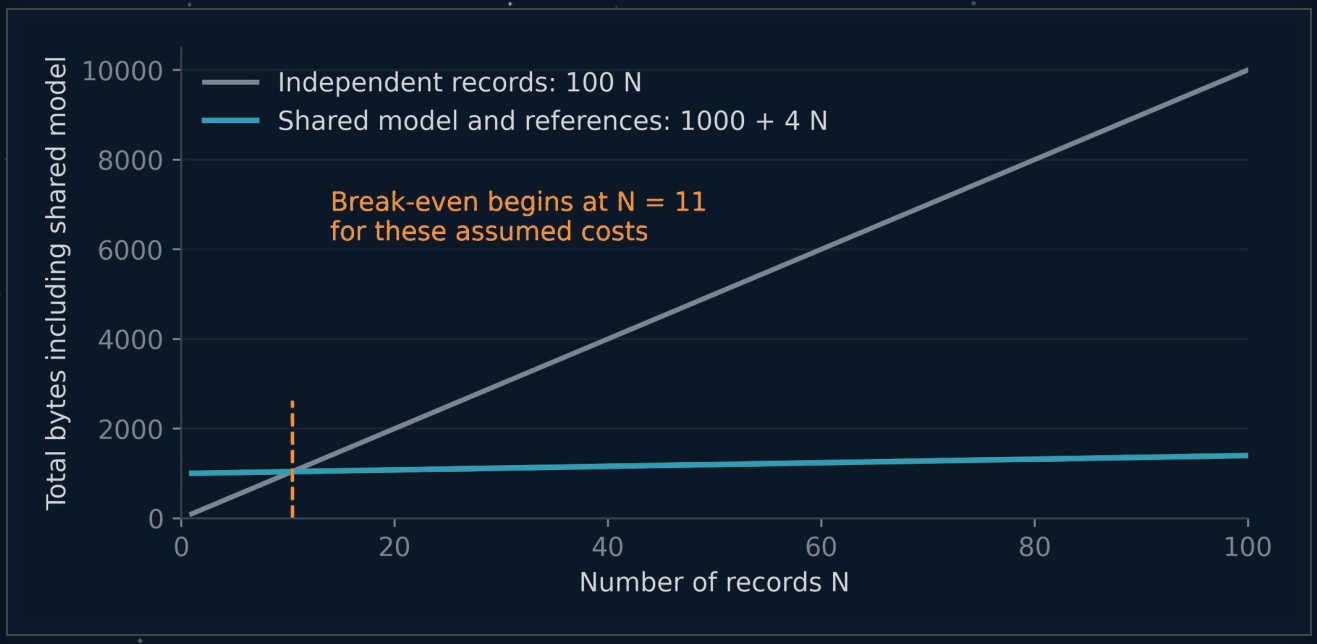
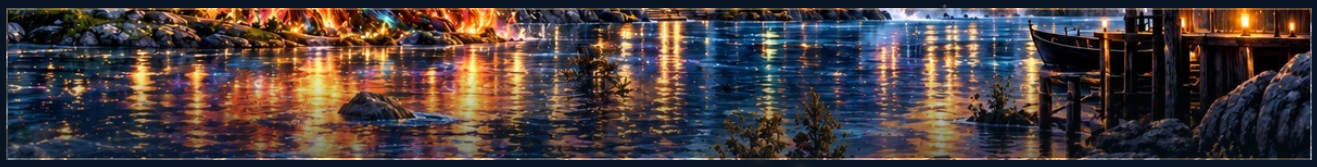


FIGURE 15

Explicit hypothetical costs. The shared representation becomes smaller at eleven records. Its assumed model is charged once, and every reference is charged. Real systems also need residuals, indexes, integrity metadata and working memory. This graph is a teaching example, not an experiment on the checkpoint.

Learned lossless compression is a legitimate research direction. BB-ANS demonstrates a latent-variable coding scheme and explicitly discusses the cost of communicating or storing model parameters. Those costs may amortize over a large corpus, but cannot be silently excluded from a self-contained artifact. Learned image compression also illustrates side-information accounting; exact coding of quantized latents does not make the original image reconstruction lossless. [10, 11]

The corresponding design rule for this project is to measure the complete reusable representation: model or grammar, record codes, residuals, graph, access indexes, metadata and decoder dependencies. A tiny pattern identifier plus an omitted large translator is not a small memory. A verified compact shared grammar with small exact record differences could be one, if the complete totals and access costs support it.



SECTION 19

Limits and meaningful storage targets

There is no established numerical minimum for this particular checkpoint. Its current ratio does not reveal the amount of remaining redundancy, and the failure of three wrappers does not prove optimality. Shannon's source-coding theory bounds expected code length under specified source and coding assumptions. It does not supply an oracle for the shortest exact encoding of this one file. [1]

An elementary counting argument prevents universal lossless shrinking. There are two to the power n possible n-bit inputs, but fewer binary strings shorter than n bits. A fixed decoder cannot map distinct inputs to the same encoded state and still recover both exactly:

$$\sum_{k=0}^{n-1} 2^k = 2^n - 1 < 2^n$$

Some inputs must therefore fail to shrink or expand. A known mathematical family can nevertheless have a short generative rule. A formula for a sequence can specify many values compactly because those values are constrained. It does not encode arbitrary independent labels, proofs, examples and exceptions for free. The distinction is especially important when describing storage of an entire mathematical foundation.

Rate-distortion theory addresses a different question: how small a representation may become when a defined amount of reconstruction error is allowed. Approximate retrieval cues may tolerate such errors, while exact integers, operator order, proof references and domain restrictions generally cannot use the same tolerance. Each representation needs its own fidelity contract. [2]

TOTAL REDUCTION TARGET FROM CANONICAL BYTES	TARGET STORED MB	FURTHER REDUCTION NEEDED FROM CURRENT CHECKPOINT
90 percent	7.342	41.239 percent
95 percent	3.671	70.620 percent
99 percent	0.734	94.124 percent

TABLE 02

Arithmetic targets only. These are not claimed achievable sizes. A 99 percent target is smaller than the current graph alone. It cannot be achieved while retaining all current non-page components unchanged. Such a target would require changing those representations as well, preserving exact recovery, and accounting for every new dependency.



SECTION 20

Representation changes worth testing

The next experiments should isolate the mechanism responsible for an improvement. Each alternative must ingest the same canonical records, preserve the same semantic fields, and satisfy the same record and relationship checks. A smaller file with missing provenance, weaker domains or a different recall contract does not answer the same research question.

Grammar compression beyond repeated subtrees. Current sharing recognizes complete equal structures. A grammar can potentially reuse recurring contexts and parameterized fragments even when a whole subtree is not identical. TreeRePair provides a concrete precedent for ordered-tree grammar compression. Its XML experiments also illustrate that smaller memory representations can have slower traversal. Its reported ratios are not forecasts for this mathematical corpus. [4]

Page organization and coding context. Current pages are independently compressed. Different record or node ordering, larger coding contexts, or grouped block formats might expose more repeated structure. These changes can increase the amount that must be decoded for one request. Measurements should report bytes decoded per query, cache misses, tail latency, peak memory and update amplification alongside final size.

Compact identifiers and relation indexes. The graph and snapshot index are measurable targets. Delta encodings, succinct adjacency structures and compact field dictionaries may reduce them, but validation and traversal still need efficient access. Any auxiliary navigation table must be counted. A scheme that saves persistent bytes while materializing a large runtime index has changed the disk/RAM tradeoff.

Certified mathematical normalization. Within a suitable domain, different expressions may have a common exact normal form. If original syntax must also be recoverable, the representation needs a reversible syntax map or residual. Domain restrictions and provenance remain attached. Extending normalization beyond a supported theory requires a new correctness argument, not a similarity threshold.

Retrieval and agent learning. Learned candidate selection could improve which records or pages are inspected first. Mobile maintenance agents could propose better placement or links. Their gains should be measured against a fixed routing policy with identical workloads and storage accounting. More bonds or higher activity are not themselves evidence of better recall, reasoning or compression.

EXPERIMENT	PRIMARY COMPARISON	REQUIRED SAFEGUARDS
Sharing ablation	Same codec with and without interning	Same records, ordering policy and complete metadata
Page-size study	Several bounded page and dictionary settings	Exact recovery, query latency, memory and update cost
Indexed-record baseline	Conventional blocks with equivalent graph and lookup support	Match access capabilities and side-information costs
Interpreter ablation	Full AST construction versus direct shapes	Same exact kernel, inputs and output checks
Learned retrieval	Fixed versus trained candidate policy	Held-out inputs, false candidates and complete model size



SECTION 21

Threats to validity and scope

This evaluation uses one supplied corpus and one development environment. Its mathematical records have particular patterns of repeated syntax, source proofs and educational fields. A highly repetitive synthetic corpus, a large text-heavy proof collection and a low-redundancy corpus can behave differently. No confidence interval or population-level advantage follows from the single-run codec timings. The observed exact-mean latency describes one operation and one engine lifetime. [A1, A3]

Several checks share implementation components. The native export and release record audit use the store's decoding API; encoder/decoder bugs with a common assumption could evade both. The retained learned-execution oracle has provenance metadata, but its development-only generator is not included in the release. These gaps motivate an independent decoder and a reproducible fixture generator. [A1, A3]

The materializer interception addendum is stronger than a self-reported zero counter, because throwing wrappers and positive controls test the specified entry points. Its scope is still finite: unexercised branches, different APIs, allocations and page decoding are outside that result. No audit establishes theoretical freedom from bugs, all forms of corruption or all concurrency failures. [A2]

The names neuron, sea, helix and RNA describe project architecture. The implemented structural nodes are data structures; the native agents are bounded control mechanisms. No biological memory capacity, emotional recall, sensory reconstruction, consciousness or physical photonic storage was measured. Likewise, the number of concept cards and source theorems cannot be interpreted as general mathematical intelligence.

The release contains Apple-native source and verification scripts, but the paper's numerical execution evidence is from Linux x86-64. A compiler dependency check or generic GPU smoke test would not substitute for the actual application's Swift, Metal and checkpoint tests. Deployment performance, power consumption and user-interface behavior require device-specific measurements.

The broad compression techniques described here are prior art. This paper's contribution is the implemented combination, its reproducible measurements and the explicit tradeoffs it exposes. Competitive performance and scientific novelty require capability-matched comparisons, wider workloads and external evaluation. The original project licensing decision remains separate from third-party notices and must be resolved before a public software release is presented as open source.



SECTION 22

Reproducibility and artifact accounting

The authoritative source artifact is `neuronal_sea_native_v0_7.zip`, 33,410,386 bytes, with SHA-256 beginning `ecf6b19cea1d6381` and ending `d96e3336e344615f`. Appendix A records the full digest and active generation. The new study preserves that release unchanged. It supplies the native export adapter, process measurement helper, bounded shell workflow, input and output hashes, codec versions, component inventory and CSV results. [A1, A3]

Artifact availability. The source release was supplied separately; the experimental supplement accompanies this review draft. No public archival URL or DOI is specified in this version. A final public locator and the original project's software license remain to be established before a submission requiring public artifacts.

All diagrams in this paper are original. Structural and topology figures are derived from source-extracted JSON; other diagrams are explicitly labeled schematics or constructed mathematical examples. The chart bundle includes vector SVG and PDF exports, high-resolution PNGs, data tables and rendering source. The figures do not use generated biological imagery as evidence.

The active-checkpoint measure includes every referenced page, the root index, relation graph, manifest and current pointer. It excludes retained history, source duplicates, binaries and runtime libraries. The complete tar contains the same active dependencies plus archive framing. The records-only stream contains canonical records and line separators, while its reproducibility index and metadata are separately itemized. Decoder executable and library accounting is provided in the supplement; these costs are environment-dependent and are not silently added to just one side of a comparison. [A3]

Preserving a hash and a local test receipt makes a result inspectable, but does not constitute independent reproduction. A clean-machine reproduction should build the pinned source, verify the same corpus, reproduce the relevant measurements and disclose any changed tools or hardware. The paper has not undergone journal peer review or an external replication study.

Author review and AI assistance

AI assistance was used for code inspection, experiment orchestration, literature research, manuscript drafting and figure preparation. The named author retains responsibility for reviewing the scientific claims, citations, software attribution and disclosure requirements before public dissemination. No completed human-review process is asserted by this document.

Conclusion

Neuronal Sea 0.7 demonstrates a concrete exact mathematical store with shared structure, compressed immutable pages, explicit relationships and bounded interpretation. Its active checkpoint is 82.9818 percent smaller than its canonical record content, and all 32,228 records passed the recorded equality checks. The new tests show that adding conventional compression around the complete existing checkpoint is not a measured route to further shrinkage. Further gains should be sought and tested in representation, grammar, page layout and indexes, with every decoder dependency and correctness requirement retained. The resulting engineering question is measurable: how many additional bytes can be removed while preserving the same mathematical records, access behavior and exact supported answers?



BACK MATTER

References

[1] Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27, 379-423 and 623-656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x and https://doi.org/10.1002/j.1538-7305.1948.tb00917.x. Original-paper reprint: https://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

[2] Shannon, C. E. (1959). Coding Theorems for a Discrete Source With a Fidelity Criterion. *IRE International Convention Record*, 7(4), 142-163. Original-paper reprint: https://gwern.net/doc/cs/algorithm/information/1959-shannon.pdf

[3] Filliatre, J.-C., and Conchon, S. (2006). Type-Safe Modular Hash-Consing. *Proceedings of the ACM SIGPLAN Workshop on ML*, 12-19. https://doi.org/10.1145/1159876.1159880. Author copy: https://usr.lmf.cnrs.fr/~jcf/publis/hash-consing2.pdf

[4] Lohrey, M., Maneth, S., and Mennicke, R. (2013). XML tree structure compression using RePair. *Information Systems*, 38(8), 1150-1167. https://doi.org/10.1016/j.is.2013.06.006. Author-institution record: https://www.research.ed.ac.uk/en/publications/xml-tree-structure-compression-using-repair/

[5] Abadi, D. J., Madden, S. R., and Ferreira, M. C. (2006). Integrating Compression and Execution in Column-Oriented Database Systems. *Proceedings of ACM SIGMOD*, 671-682. https://doi.org/10.1145/1142473.1142548. Author copy: https://www.cs.umd.edu/~abadi/papers/abadisigmod06.pdf

[6] Quinlan, S., and Dorward, S. (2002). Venti: a new approach to archival storage. *USENIX Conference on File and Storage Technologies*. https://www.usenix.org/legacy/events/fast02/quinlan.html

[7] Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554-2558. https://doi.org/10.1073/pnas.79.8.2554. https://pmc.ncbi.nlm.nih.gov/articles/PMC346238/

[8] Plate, T. A. (1995). Holographic Reduced Representations. *IEEE Transactions on Neural Networks*, 6(3), 623-641. https://doi.org/10.1109/72.377968. Published-paper copy: https://pages.ucsd.edu/~msereno/170/readings/06-Holographic.pdf

[9] Kanerva, P. (2009). Hyperdimensional Computing: An Introduction to Computing in Distributed Representation with High-Dimensional Random Vectors. *Cognitive Computation*, 1, 139-159. https://doi.org/10.1007/s12559-009-9009-8. https://rctn.org/vs265/kanerva09-hyperdimensional.pdf

[10] Townsend, J., Bird, T., and Barber, D. (2019). Practical Lossless Compression with Latent Variables Using Bits Back Coding. *International Conference on Learning Representations*. https://arxiv.org/abs/1901.04866

[11] Balle, J., Minnen, D., Singh, S., Hwang, S. J., and Johnston, N. (2018). Variational Image Compression with a Scale Hyperprior. *International Conference on Learning Representations*. https://arxiv.org/abs/1802.01436

[12] Giovannetti, V., Guha, S., Lloyd, S., Maccone, L., Shapiro, J. H., and Yuen, H. P. (2004). Classical Capacity of the Lossy Bosonic Channel: The Exact Solution. *Physical Review Letters*, 92, 027902. https://doi.org/10.1103/PhysRevLett.92.027902. Author preprint: https://arxiv.org/abs/quant-ph/0308012

[13] Metamath `set.mm` contributors. *set.mm formal mathematical database*, commit `c82282455b6ccfe3ab3eedf922705463fbe2bf7b`, as pinned in the release provenance. https://github.com/metamath/set.mm/tree/c82282455b6ccfe3ab3eedf922705463fbe2bf7b



APPENDIX A

Project evidence and identities

A1 Frozen release. Anady, J. Neuronal Sea 0.7. Distribution artifact: `neuronal_sea_native_v0_7.zip`. Author-provided software and records. Key current evidence is `reports_v07/native_measurements.json`, `foundation_verification.json`, `foundation_coverage.json`, `foundation_stress.json`, and `test_pattern.json`. Formal-corpus provenance and verification are in `source/provenance.json`, `source/CORPUS_LICENSE.txt`, `docs/FORMAL_VERIFICATION.md`, and `reports/native_formal_verification.json`. Historical size comparisons explicitly use the earlier generation in `reports/native_measurements.json`.

A2 Dated materializer addendum. `neuronal_sea_publication_2026_11_27/verification/materializer_audit_v07.json`, its source/object/fixture input hashes, four positive controls, state hashes, and `reproduce_materializer_audit.sh`. Current-generation interception: 2,102 fixture cases, 34,412 assertions, zero calls to the four wrapped entry points during replay.

A3 New compression study. `study/` in the companion illustrated-paper materials. Principal receipts:

`export_receipt.json`, `compression_results.csv`, `compression_results.json`, `study_summary.json`,

`decoder_accounting.json`, active-file inventory, frozen-state integrity checks, codec logs and the native study adapter. Every codec result includes compressed and decoded hashes.

A4 Architecture inspection. `architecture/` in the companion materials. Includes `authentic_shared_dag.json`, `curriculum_subgraph.json`, `retained_topology.json`, the authentic fraction lesson and the native request/response trace. These data define the implementation-derived figures and fresh example checks.

Full release checksum

```
ecf6b19cea1d6381e0d9ef61ed9ab3e015e171c575ad4cb7d96e3336e344615f
```

Active mathematical generation

```
5eb4a2234a3c133843fc60969ff97b0bd73ac7156fd3b94b37755b369384e3d3
```

New study input checksums

Canonical JSONL stream:

```
8da4679613ffa21139e77cf26f42fddbfee4b718df5d1d9ec1476b7d2d05d530
```

Deterministic active-checkpoint tar:

```
bd4e1070669fe65edb1d0e5256281db2974cc31c34ad4113643b76ca4b16995d
```



APPENDIX B

Reading and reproducing the figures

FIGURES	EVIDENCE TYPE	COMPANION DATA
1 and 3	Source-grounded implementation schematics	Architecture evidence and source paths

FIGURES	EVIDENCE TYPE	COMPANION DATA
2	Actual shared expression structures	`authentic_shared_dag.json`
4	Actual selected concept relationships	`curriculum_subgraph.json`
5	Actual native topology with schematic positioning	`retained_topology.json`
6	Stored coverage counts	`coverage_topics.csv` and current coverage receipt
7	Current logical file sizes	`storage_accounting.csv`
8	New measured codec results	`compression_comparison.csv`
9	Historical active-footprint difference	`historical_growth.csv`
10	Retained benchmark summaries and memory scopes	`latency_summary.csv` and current measurement receipt
11	Retained finite stress workload	`stress_workload.csv`
12	Conditional engineering scenarios	`compression_scenarios.csv`
13 and 14	Exact constructed mathematical examples	Figure-generation source
15	Hypothetical model-cost scenario	`model_amortization.csv`

All graph dimensions, units and categories are stated in the captions. Decimal MB and binary MiB are deliberately distinguished. Bars measuring task executions should not be read as unique theorems. Curves labeled conditional or hypothetical should not be cited as achieved results.

The experiment workflow requires a native build of the pinned release and the separate materializer-audit receipt. The supplement's prerequisite guide gives the build command and expected directory layout. The workflow writes to a separate study directory, builds two study-only C++ utilities, exports verified canonical records, prepares the deterministic active archive, runs six codec configurations, checks decoded hashes and verifies that the release remains unchanged. The original ANI remains C++/Swift/Metal with retained native C transport code. Document and figure production scripts are separate authoring utilities.

For a comparative follow-up, retain the same record oracle and relationship requirements, then add an indexed conventional baseline and a controlled direct-versus-materialized execution path. Report independent repetitions, cold and warm access conditions, peak memory and failed-domain behavior. This would turn the present reproducible implementation study into a stronger evaluation of storage and execution tradeoffs.

ThatAIguy

AI · CREATIVITY · REAL RESULTS

ARTIFICIAL INTELLIGENCE FOR A MORE HUMAN FUTURE.

Ideas Without Limits.

Private AI systems, creative engineering, and business automation. An intelligence studio led by Joseph W. Anady.

THATAIGUY.ORG →

