Trinity Sky

Permanent Algebraic Memory with Cryptographic Integrity and Phase-Coherent Recall

Trinity Sky Technical Report Series

White Paper (Concise Edition) — May 2026

Abstract

Large language models remain fundamentally amnesiac: their working knowledge lives in a transient attention field that is bounded by context length, biased by token position, and erased between sessions. Retrieval-augmented generation (RAG) mitigates but does not resolve this—chunking introduces boundary artifacts, embeddings drift, nearest-neighbor recall is approximate, and no retrieval operator admits an algebraic inverse or an integrity guarantee. This paper presents **Trinity Sky**, a software architecture for *permanent algebraic memory*. Its core is a Fourier Holographic Reduced Representation (FHRR) store that holds associations as unit-modulus complex phasors in a high-dimensional frequency domain, where binding is component-wise multiplication and recall is phase-conjugate multiplication in $O(D)$ time with no Fourier transform on the recall path. Unbinding is algebraically exact for a single association and degrades gracefully under superposition with signal-to-noise ratio $\sqrt{D/N}$, giving a closed-form, bounded capacity $N^* \approx D/22$. The store is realized in a six-language polyglot runtime in which every language boundary is also a supervised fault domain; a 30 Hz pipeline drives the system as a rhythmic clock; mesh routing is organized over the 240 roots of the E_8 lattice; a Kuramoto oscillator network supplies a coherence order parameter that gates recall *fail-closed*; and every write is anchored to a BLAKE3 Merkle chain that renders memory tamper-evident. We state the design as a set of formal claims, delimit what is proven, specified, measured, and projected, and position Trinity Sky as a substrate for systems that remember with proof rather than promise.

Keywords: holographic reduced representations; vector symbolic architectures; permanent memory; phase coherence; Kuramoto synchronization; E_8 lattice; Merkle integrity; fail-closed recall; polyglot runtime; fault isolation.

Contents

1 Introduction

The thesis of this paper is that memory is not a peripheral convenience for a language model but a first-class subsystem deserving its own algebra, its own runtime guarantees, and its own integrity proofs. Contemporary large language models (LLMs) treat memory as an emergent property of attention over a finite token window. That choice produces systems of remarkable fluency and equally remarkable amnesia: knowledge from the middle of a long context is recalled far less reliably than knowledge at its ends (the *lost-in-the-middle* effect, Liu et al. 2024); accuracy degrades monotonically with input length (An et al. 2025); and nothing persists across sessions.

Retrieval-augmented generation (RAG; Lewis et al. 2020) and vector databases over approximate-nearest-neighbor indices (FAISS, Johnson et al. 2019; HNSW, Malkov & Yashunin 2020) externalize memory into a corpus, but they inherit four structural deficits. Retrieval is *one-way*: there is no operator that inverts a stored composition to recover its parts. It is *approximate*: recall quality depends on chunk boundaries, embedding version, and index parameters, and degrades silently as the corpus grows. It is *unauthenticated*: a mutated index or corpus is detected by no mechanism intrinsic to retrieval. And it is *non-abstaining*: the generative layer above it fabricates fluently rather than reporting that it does not know.

Trinity Sky replaces externalized approximate retrieval with an internal *algebraic* memory whose guarantees are structural rather than statistical. Its contributions, stated as claims and carried through the paper with explicit evidence status, are:

1. **Exact algebraic recall with bounded, analyzable capacity.** Binding and unbinding in the FHRR are exactly invertible for a single association (*proven*); superposition induces zero-mean cross-talk of variance $(N-1)/D$, giving recall SNR $\sqrt{D/N}$ and a closed-form capacity $N^* \approx D/22$ (*proven*, confirmed by Monte Carlo).
2. **Permanent frequency-domain residency with crash-safe persistence.** The memory trace lives permanently in the frequency domain so that recall takes no Fourier transform; durability and atomicity are guaranteed by a write-ahead log and memory-mapped snapshots (*specified*, with proof sketches).
3. **A six-language polyglot runtime whose language boundaries are fault domains** (*specified*).
4. **Coherence-gated, fail-closed recall.** A Kuramoto order parameter and a five-factor coherence score gate recall: below threshold the system returns *not-found* rather than a fabricated value (*specified/measured-in-simulation*).

5. **Cryptographic tamper-evidence** via a BLAKE3 Merkle chain over the binding history (*proven soundness under standard assumptions*).

We use four epistemic labels throughout—**proven** (theorem under stated assumptions), **specified** (architectural specification), **measured** (execution evidence on the substrate of record), and **projected** (simulation or extrapolation)—and we never describe any component as "unhackable": the security argument is one of tamper-evidence and provable integrity, not immunity from all attack.

Notation. D is the hypervector dimension (design point $D = 16{,}384 = 2^{14}$); N the number of stored associations; K either the cleanup codebook size or the qFHRR phase-bin count ($K = 16$) as context dictates; r the Kuramoto order parameter; C the unified coherence score; $\vartheta = 0.70$ the recall acceptance threshold; $\tau = 33.\overline{3}$ ms the 30 Hz tick period; $\phi = (1+\sqrt{5})/2$ the golden ratio; $\odot$ component-wise (Hadamard) complex multiplication; $\overline{(\cdot)}$ complex conjugation.

2 Related Work

Vector symbolic architectures. Trinity Sky's memory is a Fourier Holographic Reduced Representation, the frequency-domain instance of Plate's HRR (Plate 1995, 2003) within the broader family of vector symbolic architectures (Gayler 2003; Schlegel et al. 2022) and hyperdimensional computing (Kanerva 2009; Kleyko et al. 2023). What distinguishes the present development is insistence on *exactness*: by restricting hypervectors to unit-modulus phasors, unbinding becomes an exact group inverse rather than the approximate inverse of real-valued HRR.

Retrieval and long context. RAG (Lewis et al. 2020) and ANN indices (Johnson et al. 2019; Malkov & Yashunin 2020) supply unbounded but approximate, unauthenticated recall; long-context transformers pay $\Theta(L^2)$ attention cost and still exhibit positional degradation (Liu et al. 2024). Trinity Sky is complementary: it supplies exact, verifiable, invertible recall within a bounded working set, ceding unbounded corpus breadth to the approximate baselines it is designed to sit beside.

Synchronization. The coherence layer rests on the Kuramoto model of coupled oscillators (Kuramoto 1975; Strogatz 2000; Acebrón et al. 2005), its mean-field critical coupling (Ott & Antonsen 2008), and network-synchronization theory (Dörfler & Bullo 2014). The binding-by-synchrony interpretation follows the neuroscience of gamma-band coordination (Singer; Fries 2005, communication-through-coherence).

Lattices. Routing addresses live on the E_8 lattice, the densest packing in eight dimensions (Viazovska 2017) with the rich combinatorics catalogued by Conway & Sloane (1999); its 240 minimal roots and $O(1)$ nearest-point decoding make it a natural error-bounded address space.

Integrity and neuromorphics. Tamper-evidence uses Merkle trees (Merkle 1989) over BLAKE3 hashes. The optional neuromorphic acceleration path targets the Neuromorphic Intermediate Representation (NIR; Pedersen et al. 2024) and platforms such as Loihi 2 (Davies et al. 2018), but as accelerators of platform-independent algorithms, not dependencies.

No prior system combines exact algebraic recall, cryptographic integrity, and coherence-gated abstention in a single architecture; that combination is the gap Trinity Sky fills.

3 The FHRR Binding Algebra

This section develops the mathematical core. Proofs are sketched; full derivations appear in the companion technical report. Throughout, a *phasor hypervector* is an element $\mathbf{x} \in \mathbb{T}^D$ with components $x_k = e^{i\theta_k}$, $|x_k| = 1$; a *random* phasor draws each θ_k i.i.d. uniform on $[0, 2\pi)$. This is the atomic symbol distribution.

3.1 Binding, Unbinding, and the Convolution Ring

The discrete Fourier transform $\mathcal{F}$ is a ring isomorphism carrying circular convolution to the component-wise product: $\mathcal{F}(\mathbf{a} \circledast \mathbf{b}) = \mathcal{F}(\mathbf{a}) \odot \mathcal{F}(\mathbf{b})$. Consequently each frequency bin is an algebraically independent channel and binding never couples distinct bins. Trinity Sky stores all keys and the memory trace *as their DFTs*, so binding is the Hadamard product

$$(\mathbf{a} \otimes \mathbf{b})_k = a_k \, b_k = e^{i(\alpha_k + \beta_k)},$$

a single pass of $O(D)$ complex multiplies with no transform. **Unbinding** is binding by the conjugate, $(\mathbf{c} \oslash \mathbf{b})_k = c_k \, \overline{b_k}$.

Theorem (exact unbinding). For any $\mathbf{a}$ and any phasor $\mathbf{b}$, $(\mathbf{a} \otimes \mathbf{b}) \oslash \mathbf{b} = \mathbf{a}$. *Proof.* $(a_k b_k)\overline{b_k} = a_k |b_k|^2 = a_k$ since $|b_k| = 1$. ■ The phasor constraint is necessary; with $|b_k| \neq 1$ recovery is only approximate (the generic real-valued HRR case). The set $(\mathbb{T}^D, \otimes)$ is an abelian group isomorphic to $(U(1))^D$, with identity $\mathbf{1}$ and inverse $\overline{\mathbf{a}}$; binding by a fixed vector preserves the uniform (Haar) measure, so bound pairs are statistically dissimilar to their factors and may be superposed.

3.2 Superposition, Cross-Talk, and Capacity

An associative memory is a bundle of bound key–value pairs, $\mathbf{M} = \sum_{i=1}^{N} \mathbf{a}_i \otimes \mathbf{b}_i$. Probing with a conjugate key gives

$$\mathbf{r}_j = \mathbf{M} \oslash \mathbf{b}_j = \underbrace{\mathbf{a}_j}_{\text{signal (exact)}} + \underbrace{\textstyle\sum_{i \neq j} \mathbf{a}_i \otimes \mathbf{b}_i \otimes \overline{\mathbf{b}_j}}_{\text{cross-talk } \eta_j}.$$

Under the random-codebook model the cross-talk is zero-mean with per-component variance $(N-1)/D$ (cross-key products are uniform phasors, summed independently). Hence the recall signal-to-noise ratio is

$$\mathrm{SNR}_{\mathrm{power}} = \frac{D}{N-1} \approx \frac{D}{N}, \qquad \mathrm{SNR}_{\mathrm{amp}} = \sqrt{\frac{D}{N}}.$$

(The HRR-literature figure $\sqrt{2D/N}$ is the same physics expressed for the real-valued cleanup decision statistic; the two differ only by $\sqrt{2}$.) Treating cross-talk as Gaussian for large N—the single approximation in the analysis, with exact mean and variance known—the probability that nearest-neighbor cleanup against K codebook entries returns the correct value is $P_{\mathrm{correct}} = [\Phi(\sqrt{D/(N-1)})]^{K-1}$. Solving $P_{\mathrm{correct}} \geq 1 - \epsilon$ yields the capacity law $N^* \approx D/(2\ln(N^*/\epsilon))$, whose numerical solution at $D = 16{,}384$, $\epsilon = 10^{-3}$ is the operating heuristic

$$N^* \approx \frac{D}{22} \approx 744.$$

The factor 22 is exactly the power-SNR D/N at the design point: 13.4 dB of margin per component. Capacity scales linearly in D and degrades gracefully, not catastrophically.

3.3 Cleanup and the Quantized Layer

Cleanup projects the noisy $\mathbf{r}_j$ back onto the codebook by similarity arg-max. By Parseval's theorem cosine similarity has the same ranking in either domain, so cleanup runs entirely in the frequency domain with *no inverse transform*: recall is D conjugate multiplies and cleanup is K length-D inner products, $O(KD)$ multiply-accumulates and zero FFTs.

The routing layer uses a 4-bit **quantized FHRR** (qFHRR), discretizing each phase to one of $K = 16$ bins so that binding becomes integer modular addition $(a_d + b_d) \bmod 16$—a bitwise operation—and similarity becomes a 16-entry cosine lookup table. This is the discrete subgroup $(\mathbb{Z}/16\mathbb{Z})^D \subset (U(1))^D$; binding remains *exactly* invertible, and the only loss is a quantization fidelity gap that scales as K^{-3} (≈ 0.987 per binding at $K = 16$). Crucially, *capacity is K-independent*: the bin width cancels against the phase-noise scale, so a 4-bit routing representation coexists with the continuous recall path without reducing associative capacity. The construction is the classical shadow of the Gottesman–Kitaev–Preskill code: error correction by snapping a continuous displacement to the nearest lattice point.

4 The Holographic Memory Palace

The memory is organized as a **FreqDomainStore** of seven semantic *rooms*, each an independent FHRR trace. Partitioning into R rooms reduces the per-room load and therefore the cross-talk: a query confined to one room sees N/R interferers, improving SNR by $\sqrt{R}$ and eliminating cross-room leakage by construction (distinct rooms share no frequency channels in a query).

Episode encoding. Structured episodes are encoded compositionally against eight QR-orthonormal role vectors $\{\mathbf{k}^{(1)}, \ldots, \mathbf{k}^{(8)}\}$: an episode is the superposition of role–filler bindings $\sum_m \mathbf{k}^{(m)} \otimes \mathbf{v}_m$, and a field is read back by unbinding its role and cleaning up. Orthonormal roles keep the role basis off the random-filler distribution, so role-keyed recall is high-fidelity.

The zero-FFT invariant. Because traces, keys, and probes are all stored as DFTs and every operation—bind, unbind, bundle, similarity—is pointwise, no forward or inverse Fourier transform is ever taken on the hot path. This is both a performance invariant (recall is $O(D)$ multiplies) and a correctness invariant (no transform round-off is introduced between store and recall).

Crash-safe persistence. Durability is provided by a write-ahead log (WAL) plus memory-mapped (`mmap`) snapshots under a fail-stop crash model. Each store appends a WAL record before mutating the in-memory trace; recovery replays the WAL against the last snapshot. The protocol guarantees, with proof sketches in the technical report: **durability** (an acknowledged write survives crash), **atomicity** (a partially written record is discarded on replay via a length-and-checksum frame), **idempotent replay** (replaying the same WAL is a no-op on already-applied records), and **crash-recovery correctness** (the recovered trace equals the trace implied by the acknowledged write sequence).

Cryptographic integrity. Every write also appends a leaf to a BLAKE3 Merkle tree keyed on the canonical serialized bytes of the binding. The published root commits to the entire history; any later alteration of stored bytes changes some leaf hash and therefore the root, and is detected at verification with $O(\log N)$ inclusion proofs. Under the standard collision-resistance assumption on BLAKE3, undetected tampering requires a second-preimage ($\approx 2^{256}$) or a root collision ($\approx 2^{128}$).

Consolidation. A 1 Hz Hebbian consolidation pass strengthens coherently co-activated bindings and applies time-to-live eviction once a room exceeds $\approx 85\%$ of its capacity N^*, enforcing the analytic capacity bound rather than letting recall silently degrade.

5 The 30 Hz Pipeline

Trinity Sky runs as a soft real-time system on a fixed tick period $\tau = 33.\overline{3}$ ms (30 Hz), chosen to match the gamma-band binding window in which the brain is thought to bind distributed features into coherent percepts (Singer; Fries 2005; Buzsáki). Each tick executes a thirteen-stage directed acyclic dataflow rooted at the oscillator update and terminating at the emission of a single coherence scalar.

#	Stage	Function	Complexity	Budget
1	Kuramoto Sync	phase-field Euler step	$O(N)$ (mean-field)	47 µs
2	Fourier Lens	DFT mode extraction	$O(N \log N)$	18 µs
3	Metatron Router	shortest-path on K_{13}	$O(1)$ lookup	8 µs
4	Spectral Stability	Laplacian eigenvalues, Fiedler λ_2	$O(N^3)$	89 µs
5	Toroidal Flow	phase wrap / winding	$O(N)$	7 µs
6	Chladni Patterns	modal node detection	$O(N^2)$	9 µs
7	Cosmic Council	read slow-loop consensus	$O(1)$	0.4 µs
8	Fractal Swarm	80^3 Mandelbulb coherence	$O(G \cdot I)$ fixed	1,890 µs
9	Holographic Memory	FHRR recall + cleanup	$O(D) + O(KD)$	370 µs
10	Resonance Bus	inter-engine ring-buffer write	$O(N)$	6 µs
11	Auto-Scaler	ϕ-budget pool sizing	$O(1)$	0.3 µs
12	Spectral Final	post-tick re-validation	$O(1)$	0.5 µs

	#	Stage	Function	Complexity	Budget
	13	Meninges Tick	geometric-mean coherence fusion $\to C$	$O(1)$	0.8 µs

With $N = 13$ agents the worst-case execution time is dominated by three stages—the cubic spectral decomposition (fixed N), the fixed-grid Mandelbulb, and the $O(KD)$ memory recall—and their sum sits well inside τ; the schedule is admissible by the standard utilization bound. Decisively, no stage cost grows with the number of stored memories N_{store} or with interaction-history length, so the budget is met as a *time-invariant* property, in contrast to the $\Theta(L^2)$ growth of attention. Under sustained pressure the Auto-Scaler degrades gracefully, reducing the Mandelbulb grid or substituting cached values for Stages 8–9.

A **dual-rate** architecture pairs the 30 Hz binding loop with a slower 1–5 Hz deliberation and consolidation loop (multi-agent consensus, Hebbian strengthening), which writes results that the fast loop merely reads on the hot path (Stage 7). Inter-stage communication uses a lock-free SharedState bus (ETS tables / atomics / Nx tensors) under a single-writer discipline, avoiding lock contention within a tick.

6 The Polyglot Architecture

The most consequential architectural decision is the execution runtime, because it fixes the concurrency model, memory discipline, failure semantics, and foreign-function surface. We claim that no single runtime satisfies the system's six requirements simultaneously:

Requirement	Elixir/BEAM	Rust	CUDA	Go	Python	Nx/EXLA
R_a fault-tolerant concurrency + hot reload						
R_b zero-GC SIMD numerical throughput						
R_c massively parallel kernels						

Requirement	Elixir/BEAM	Rust	CUDA	Go	Python	Nx/EXLA
R_d robust external I/O / network						
R_e scientific / quantum libraries						
R_f in-runtime accelerated tensors						

No column satisfies all six rows; the architecture assigns each requirement to the runtime that owns it, and draws the language boundaries to coincide with fault-isolation boundaries.

- **Elixir/BEAM — orchestration brain.** Owns the 30 Hz tick, OTP supervision trees, and on the order of 10^6 lightweight processes (one per agent, room, oscillator, connection). Per-process garbage collection means no stop-the-world pause threatens the tick; hot code reload patches a long-lived node without stopping it.
- **Rust — numerical muscle.** Hot-path kernels (FHRR/qFHRR multiply-accumulate, E_8 decode, BLAKE3 Merkle) run as Rustler NIFs on dirty-CPU schedulers, with zero-copy data sharing via `ResourceArc` and slot indices. Rust's type system prevents memory-unsafety from crossing back into the VM.
- **CUDA — parallel accelerator.** Warp-level cleanup and batched recall, with a *slot-index* API so hypervectors are referenced by handle rather than copied across the boundary. GPU is a compute role, not an architectural premise.
- **Go — nervous system.** The external gRPC/HTTP mesh gateway, metrics, and I/O bridges run as a *separate OS process and failure domain*, keeping a latency-variable, attack-exposed surface off the orchestration core.
- **Python — research laboratory.** PennyLane/Lava/Metal-Q experiments run strictly *off-tick* behind a Port/gRPC boundary, so the GIL and library latency never touch the real-time loop.
- **Nx/EXLA — in-VM tensor bridge.** `defn`-compiled Kuramoto, spectral, and FFT stages execute as XLA-compiled tensor code *inside* the BEAM, inheriting its supervision.

The integration protocols trade latency against isolation explicitly:

Mechanism	Endpoints	Latency class	Copy semantics	Isolation
NIF (Rustler)	BEAM Rust CUDA	sub-µs / few µs (dirty)	zero-copy	weakest (shared VM)
In-VM `defn`	BEAM Nx/EXLA	µs (cached)	zero (on-device)	inherits BEAM

Mechanism	Endpoints		Latency class	Copy semantics	Isolation
Erlang Port	BEAM	Python	tens–hundreds µs (off-tick)	serialize + copy	strong (separate proc)
gRPC / UDS	BEAM	Go	tens µs	one protobuf copy	strongest (separate proc)
Distributed Erlang	BEAM	BEAM	15 µs (LAN)	ETF over distribution	node monitoring

A language boundary is a fault boundary. The hottest, least-isolated boundary (the Rust/CUDA NIF) is reserved for code that is statically memory-safe and short-running; everything latency-variable or externally exposed is pushed behind a process boundary with its own scheduler and collector. A crash in a numerical kernel, an accelerator, or a research worker therefore cannot corrupt orchestration state—the reliability claim is made structurally true by the placement of the boundaries. Polyglot runtimes at this scale are proven practice (Erlang at WhatsApp, Elixir at Discord, Rust and Go across Cloudflare and Dropbox).

7 E8 Mesh Routing

Routing addresses live on the E_8 lattice, whose 240 minimal roots (112 integer roots of form $(\pm 1, \pm 1, 0^6)$ and 128 half-integer roots with an even number of minus signs) form a highly symmetric, error-bounded address space; E_8 is the densest packing in dimension eight (Viazovska 2017) with kissing number 240 and Coxeter number $h = 30$. A vector is decoded to its nearest root in $O(1)$ by the Conway–Sloane algorithm (decode to D_8 and to the glue coset $D_8 + \frac{1}{2}$, keep the closer), with correction radius equal to the packing radius $\rho = \sqrt{2}/2$.

The frequency channels are organized as **eight ϕ-scaled bands × 30 roots** (the Coxeter number), giving a frequency-division multiplexing of the FHRR channels with golden-ratio band spacing that minimizes harmonic overlap. A recalled routing key is projected from the 16{,}384-dim qFHRR space down to eight dimensions and decoded to a root index in sub-microsecond time (the Rust kernel owns this path). Messages route between roots on a diameter-2 graph with 56-neighbor adjacency, so any pair of channels communicates in at most two hops. Across nodes, mesh state is a conflict-free replicated data type (CRDT): pheromone/gradient fields and bitboard membership sets merge associatively, so replicas converge without coordination, and inter-node memory roots are compared for agreement as a tamper check.

8 Kuramoto Coherence and Fail-Closed Gating

The system's recall gate is driven by a network of $N = 13$ Kuramoto phase oscillators with phases θ_i and natural frequencies ω_i, coupled on the complete graph K_{13}:

$$\dot{\theta}_i = \omega_i + \frac{K}{N} \sum_{j=1}^{N} \sin(\theta_j - \theta_i), \qquad r e^{i\psi} = \frac{1}{N} \sum_j e^{i\theta_j},$$

where $r \in [0, 1]$ is the order parameter measuring phase synchronization. Mean-field theory gives a critical coupling $K_c = 2/(\pi g(0))$ for a unimodal frequency density g; for the reference distribution

$K_c \approx 2.79$. Trinity Sky operates *deliberately subcritical* at $K^* = 2.663 \approx 0.955\,K_c$—on the Widom line that maximizes susceptibility—trading a small reduction in raw lock for maximal responsiveness and spectral richness. At finite $N = 13$ the transition is blurred (fluctuation floor $\langle r \rangle \approx 0.246$, width $\Delta K \approx 0.50$), which the partial-synchronization logic exploits; over-coupling is avoided because rigid lock destroys the agent diversity the coherence metric rewards. The linearized recovery time to the synchronized manifold is $\tau_{\text{rec}} = 1/(K\lambda_2)$; with $\lambda_2 = 13$ for K_{13}, $\tau_{\text{rec}} \approx 28.9$ ms—within a single tick.

The five-factor coherence score. Recall is gated not by r alone but by a unified score that is the *geometric mean* of five subsystem factors,

$$C = \left(C_{\text{opt}} \cdot C_{\text{mandelbulb}} \cdot C_{\text{spectral}} \cdot C_{\text{mycelium}} \cdot C_{\text{holo}}\right)^{1/5},$$

with the zero-propagation property that any factor of zero forces $C = 0$. A geometric mean is the correct aggregation for a safety gate because it is AND-like: a subsystem failure is visible as a collapse of C rather than being averaged away. Stability is monitored on two windows—a fast window (( )1 s) with a Schmitt-trigger emergency at $C_{\text{fast}} < 0.30$ and a slow window (( )10 s) with warning/critical thresholds at 0.70/0.50. Hysteresis prevents chattering near threshold.

Fail-closed recall. A recall is accepted only if the cleanup similarity exceeds the threshold $\vartheta = 0.70$ *and* system coherence is adequate; otherwise the system returns *not-found*. This is anti-hallucination by construction: unlike a generative decoder that always emits a token, the algebraic store abstains when it does not have a confident answer. C is an engineering control signal, not a claim about cognition.

9 Security Model

We model security as *cost*, not impossibility, against a graded Dolev–Yao adversary with defined trust boundaries, and map the surface to STRIDE and the OWASP LLM Top-10. Defense is layered ("Meninges"): from the network perimeter, through system and application integrity, mesh authentication, and memory integrity, down to a per-tick coherence check. Admission to memory passes a three-stage gate chain:

1. **Dura Mater** — a routing key is admitted only if it decodes to a valid E_8 root within radius $\rho = \sqrt{2}/2$ and carries a valid HMAC-SHA256 tag with an unexpired TTL.
2. **Arachnoid** — routing and rate validation.
3. **Pia Mater** — the coherence gate (C and ϑ) described above.

Tamper-evidence. The BLAKE3 Merkle chain (above) detects any alteration of stored bytes; forging a root requires 2^{128}–2^{256} work, so the realistic attack is not cryptanalytic but a compromise of the publication channel or a replica—reducing memory integrity to a well-understood key-management and quorum problem.

Air-gap sovereignty. A coherence collapse (sustained C below the emergency threshold) trips an automatic isolation state (`SPORE_GERMINATING`): the node disengages from the mesh and serves only verified local state until an operator re-attests. This converts an active compromise into a fail-safe withdrawal.

Why LLM attacks fail here. Within the threat model and stated assumptions: prompt injection cannot synthesize a key that algebraically unbinds to an attacker-chosen value without read access

to the room (a forged probe clears $\vartheta = 0.70$ only with probability $\sim \Phi(0.70\sqrt{D})$, i.e. $\ll 2^{-256}$); context poisoning is rejected because tampered bytes break the Merkle root; and out-of-distribution queries fail the coherence gate rather than producing a confident fabrication. The honest claim is tamper-evidence and provable integrity, never unhackability; residual risks (supply chain, physical access, insider with an approval token) are out of the memory layer's scope.

10 Evaluation

We adopt a conservative epistemology and label every result **proven**, **measured**, or **projected**. Capacity claims are established by a Monte Carlo protocol ($T = 10^5$ trials: 100 re-seeded codebooks $\times$ 10^3 queries, reported with 95% Wilson intervals) using the same random-phasor model under which the analytic bound is proven.

Capacity and accuracy (proven + Monte Carlo). At $D = 16{,}384$, with codebook size $K = N$:

N	$\sqrt{D/N}$	SNR (dB)	Analytic P_{correct}	Monte Carlo $\hat{p}$ (95% CI)
512	5.66	15.1	0.999996	0.99999 [0.99994, 1]
630	5.10	14.2	0.99990	0.99989 [0.99966, 0.99997]
744	**4.69**	**13.4**	**0.99901**	**0.99898** [0.99876, 0.99916]
880	4.32	12.7	0.9932	0.9930 [0.99248, 0.99347]
1024	4.00	12.0	0.968	0.967 [0.9659, 0.9681]

At the design point $N = 744$ the system stores 744 associations per room and recalls the correct one with probability 0.99898 (95% CI lower bound $> 99.87\%$), matching the analytic prediction to three decimals. Capacity scales linearly in D and degrades gracefully.

Latency (proven; substrate-projected). The recall path is $\Theta(KD)$—about $744 \times 16{,}384 \approx 1.2 \times 10^7$ complex MACs—with no Fourier transform and *no dependence on the number of stored memories or history length.* The budget condition $KD/P \leq \tau$ for substrate throughput P is therefore a time-invariant, in structural contrast to attention's $\Theta(L^2)$. On the substrate of record (an Apple M4 Max, MLX/Accelerate) the unoptimized recall path is *measured* at ≈ 0.37 ms/tick; frequency-domain-optimized kernels are *projected* at < 0.20 ms. Hardware enters only here, as a measurement of P, never as the headline.

Comparison to baselines (qualitative). The differentiators are present-or-absent rather than merely better:

Axis	Trinity Sky	RAG	Vector DB (ANN)	Long context
Exact inverse	Yes (proven)	No	No	No

Axis	Trinity Sky	RAG	Vector DB (ANN)	Long context
Integrity guarantee	Yes (Merkle)	No	No	No
Capacity model	Bounded, closed-form $D/22$	Unbounded, opaque	Unbounded, silent decay	Window L, $\Theta(L^2)$
Recall determinism	Deterministic arg-max	Stochastic	Approximate	Stochastic
Hallucination	Fail-closed *not-found*	19–40% end-to-end error	downstream LLM fabricates	fabricates fluently
Persistence	Permanent + WAL + Merkle	corpus persists, index fragile	disk-backed, version-fragile	none

Trinity Sky cedes unbounded corpus breadth to the approximate baselines it is designed to complement; what the table establishes is that exact inverse, integrity, determinism, and fail-closed abstention are absent *by construction* from the alternatives.

Security cost (projected attack-tree). Forging a Merkle root, passing the coherence gate, or bypassing E_8 admission each costs 2^{128}–2^{256} cryptographic work, or—realistically—a key/replica compromise. The framing is cost, never impossibility.

Coherence dynamics (simulated). The field converges within one tick ($\tau_{\text{rec}} \approx 28.9$ ms), operates at $K^*/K_c \approx 0.955$, and yields a unified coherence $C \approx 0.892$ at the calibrated operating point.

Threats to validity. The capacity curve is a Monte Carlo simulation under the idealized random-phasor model; embedding-derived fillers are not uniform on the torus and must be measured per corpus. Coherence figures come from a simulator, not neuromorphic hardware. All results are single-node; mesh-federation behavior is projected. The optimized kernels are specified but not yet benchmarked end-to-end.

11 Conclusion and Future Work

Trinity Sky is a software architecture for permanent algebraic memory. Its five claims are: exact, invertible recall with closed-form capacity $N^* \approx D/22$ (*proven*); permanent frequency-domain residency with crash-safe WAL + Merkle persistence (*specified/proven soundness*); a six-language polyglot runtime whose boundaries are fault domains (*specified*); coherence-gated fail-closed recall (*specified/simulated*); and cryptographic tamper-evidence (*proven under standard assumptions*). To our knowledge it is the first architecture to combine algebraic associative memory, cryptographic integrity, and coherence-gated abstention in a single coherent design.

The contribution is fundamentally a *software* one: the algebra, the runtime, and the protocols run on commodity hardware and treat specialized accelerators as optional. Future work develops three such options as accelerators of platform-independent algorithms, not dependencies: (i) **quantum-accelerated cleanup** (a Grover-style search whose advantage crossover is near $N \approx 6.1 \times 10^5$, far above the per-room operating point, so the classical path remains the production path); (ii) a **neuromorphic bridge** that exports qFHRR to NIR for spike-phase recall on platforms such as Loihi 2, while the underlying golden-ratio/Kuramoto optimization runs natively today on commodity hardware via MLX/NumPy; and (iii) **mesh federation**, scaling the single node to a CRDT-

and Merkle-synchronized diameter-2 E_8 mesh. We also intend to port the key theorems—binding invertibility, persistence guarantees, tamper-evidence—to a proof assistant for machine-checked verification. Trinity Sky is the argument, in software, that an AI system can remember with proof rather than promise.

12 References

1. An, C., et al. (2025). Why Do Language Models Degrade with Input Length. *arXiv*.
2. Acebrón, J. A., et al. (2005). The Kuramoto Model: A Simple Paradigm for Synchronization Phenomena. *Rev. Mod. Phys.*, 77, 137–185.
3. Conway, J. H., & Sloane, N. J. A. (1999). *Sphere Packings, Lattices and Groups* (3rd ed.). Springer.
4. Davies, M., et al. (2018). Loihi: A Neuromorphic Manycore Processor with On-Chip Learning. *IEEE Micro*, 38(1), 82–99.
5. Dörfler, F., & Bullo, F. (2014). Synchronization in Complex Networks of Phase Oscillators: A Survey. *Automatica*, 50(6), 1539–1564.
6. Fries, P. (2005). A Mechanism for Cognitive Dynamics: Neuronal Communication through Neuronal Coherence. *Trends Cogn. Sci.*, 9(10), 474–480.
7. Gayler, R. W. (2003). Vector Symbolic Architectures Answer Jackendoff's Challenges. *Proc. ICCS/ASCS*.
8. Gottesman, D., Kitaev, A., & Preskill, J. (2001). Encoding a Qubit in an Oscillator. *Phys. Rev. A*, 64, 012310.
9. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-Scale Similarity Search with GPUs (FAISS). *IEEE Trans. Big Data*, 7(3), 535–547.
10. Kanerva, P. (2009). Hyperdimensional Computing. *Cognitive Computation*, 1(2), 139–159.
11. Kleyko, D., et al. (2023). A Survey on Hyperdimensional Computing. *Proc. IEEE*, 111(12), 1538–1571.
12. Kuramoto, Y. (1975). Self-Entrainment of a Population of Coupled Nonlinear Oscillators. *Int. Symp. on Mathematical Problems in Theoretical Physics*.
13. Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *NeurIPS*.
14. Liu, N. F., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. *TACL*, 12, 157–173.
15. Malkov, Y. A., & Yashunin, D. A. (2020). Efficient and Robust ANN Search Using HNSW Graphs. *IEEE TPAMI*, 42(4), 824–836.
16. Merkle, R. C. (1989). A Certified Digital Signature. *CRYPTO '89*, LNCS 435, 218–238.
17. Ott, E., & Antonsen, T. M. (2008). Low-Dimensional Behavior of Large Systems of Globally Coupled Oscillators. *Chaos*, 18, 037113.
18. Pedersen, J. E., et al. (2024). Neuromorphic Intermediate Representation. *Nature Communications*.
19. Plate, T. A. (2003). *Holographic Reduced Representations*. CSLI Publications.
20. Schlegel, K., Neubert, P., & Protzel, P. (2022). A Comparison of Vector Symbolic Architectures. *Artif. Intell. Rev.*, 55, 4523–4555.
21. Strogatz, S. H. (2000). From Kuramoto to Crawford. *Physica D*, 143, 1–20.
22. Viazovska, M. (2017). The Sphere Packing Problem in Dimension 8. *Annals of Mathematics*, 185(3), 991–1015.
23. Wilson, E. B. (1927). Probable Inference, the Law of Succession, and Statistical Inference.

JASA, 22(158), 209–212.