

# Topological AI: Prime-Anchored Neural Networks That Do Not Forget

A Practical Framework for Deterministic, Verifiable,  
Catastrophic-Forgetting-Resistant Artificial Intelligence

Frank Morales Aguilera

Sovereign Machine Lab (SOMALA), Montréal, Canada

`frank.morales@sovereignml.ai`

Deterministic Seed: 123 | SHA-256 Audited | Open Source

May 2026

## Abstract

Most current artificial intelligence systems — transformers, large language models, mixture-of-experts architectures — lack fixed topological anchors in their weight space. As a consequence they are susceptible to representational drift and catastrophic forgetting, and their internal state cannot be cryptographically verified.

This paper introduces **Topological AI**, an engineering approach in which neural networks are augmented with fixed, verifiable, immutable topological anchors. The anchors are embedding vectors at prime-numbered indices derived from the Sieve of Eratosthenes (c. 240 BCE). These indices are cached before any operation and restored after every gradient step. Their integrity is verified via SHA-256 cryptographic hashing.

We demonstrate the approach on Mixtral-8x7B, a 47-billion-parameter mixture-of-experts model running on an NVIDIA RTX PRO 6000 Blackwell GPU (102 GB VRAM). Across multiple inference passes, prime-anchored subspaces remain cryptographically invariant. Spectral coherence metrics (SROI) remain above the empirical threshold  $\Lambda = 0.9785142874$  in all tested queries. Zero hash anomalies are detected. The complete source code is open source and reproducible with Seed = 123.

**Keywords:** Topological AI, prime anchors, Sieve of Eratosthenes, catastrophic forgetting, cryptographic verification, topological invariants, mixture-of-experts, LLM safety

## 1 Introduction

### 1.1 The Problem

Every deployed AI system today shares a structural vulnerability: its weight space has no fixed topological anchor.

Transformers [16] maintain short-term memory through a context window and long-term memory through pre-trained weights. However, they have no mechanism to consolidate knowledge across learning episodes without catastrophic forgetting [17]. This is not merely

a bug: it is a structural consequence of designing a system with no fixed reference frame in weight space.

Google’s HOPE architecture (Nested Learning, 2025) proposes multiple update frequencies — fast modules updating every 16 tokens, medium modules every 1 million tokens, slow modules every 16 million tokens — but still provides no fixed topological anchor. Without such an anchor, any multi-frequency system remains susceptible to long-run representational drift.

## 1.2 The Proposed Solution

The Sieve of Eratosthenes (c. 240 BCE) enumerates prime numbers deterministically [15]. This paper proposes that prime-numbered embedding indices derived from the Sieve can serve as fixed anchors for neural networks — a foundation that is exact, auditable, and reproducible. We call this approach **Topological AI**.

Prime anchors never change. Their integrity is cryptographically verifiable. Our experiments on Mixtral-8x7B suggest that anchoring prime-indexed embedding rows is a promising mechanism for resisting representational drift during continual inference [9, 10, 11, 12, 13]. Rigorous validation under fine-tuning conditions is identified as a key direction for future work (see Section 10).

## 1.3 What Topological AI Is Not

Topological AI is *not*:

- A new attention mechanism
- A new architecture (transformers, SSMS, MoE)
- A new training algorithm
- A closed-form mathematical proof

Topological AI *is*:

- A method for anchoring any neural network to fixed, verifiable reference points in embedding space
- A cryptographic verification system for anchor integrity
- An empirically derived safety metric with a reproducible threshold
- Executable code that runs on real hardware

# 2 What Is Topological AI?

## 2.1 Definition

**Topological AI** is an engineering discipline characterised by four properties:

Table 1: Properties of Topological AI

Property	Implementation
Fixed topological anchor	Embedding vectors at prime-numbered indices
Cryptographic verifiability	SHA-256 hashing of anchored subspaces
Empirical safety metric	Spectral coherence threshold $\Lambda$
Drift-resistant inference	Anchor restoration after every operation

## 2.2 Contrast with Current AI

Table 2: Topological AI vs. Current AI

Feature	Current AI	Topological AI
Topological anchor	None	Prime-indexed embedding rows
Verifiability	Empirical only	Cryptographic (SHA-256)
Safety metric	Learned or hard-coded	Derived from Sieve of Eratosthenes
Resistance to forgetting	Reduced (not eliminated)	Anchor-preserved (inference)
Reference point	None	$\sigma = 0.5, \Lambda$

## 2.3 The Topological Anchor

In this work, the *topological anchor* is the set of embedding vectors at prime-numbered vocabulary indices. These vectors are fixed before any operation begins and restored after every operation. The anchored set is:

$$\mathcal{I} = \{\mathbf{e}_p : p \in \mathbb{P}, p \leq 13\}$$

where  $\mathbf{e}_p$  is the embedding vector at prime index  $p$  and  $\mathbb{P}$  is the set of primes generated by the Sieve of Eratosthenes.

*Terminology note.* The term “topological invariant” is used informally throughout this paper to denote a quantity that is held fixed by construction. A formal proof that prime-indexed anchoring preserves topological properties of the embedding manifold in the strict mathematical sense remains an open problem and is left for future work.

# 3 Prime Anchors from the Sieve of Eratosthenes

## 3.1 The Sieve

The Sieve of Eratosthenes (c. 240 BCE) enumerates primes by iterative elimination [15]. The algorithm is shown in Listing 1, followed by its key complexity properties.

```
def sieve_of_eratosthenes(N):
    is_prime = [True] * (N + 1)
```

```

is_prime[0] = is_prime[1] = False
for p in range(2, int(N**0.5) + 1):
    if is_prime[p]:
        for m in range(p*p, N + 1, p):
            is_prime[m] = False
return [p for p in range(2, N + 1) if is_prime[p]]

```

Listing 1: Sieve of Eratosthenes

Key properties:

- Time complexity:  $O(N \log \log N)$
- Space complexity:  $O(N)$
- Determinism: identical input always produces identical output

For this paper, we use the first six primes: [2, 3, 5, 7, 11, 13].

### 3.2 Prime-Indexed Anchors

In any transformer-based LLM, the embedding layer is a matrix of shape (vocab\_size, hidden\_dim). We select the rows at prime indices and declare them **topological anchors** [9, 10, 11, 12, 13]:

```

primes = [2, 3, 5, 7, 11, 13]
cached_weights = embed_layer.weight[primes].clone()

```

Listing 2: Prime Anchor Caching

These rows are not updated during inference. Their use in training (fine-tuning) settings is discussed in Section 10.

### 3.3 Anchor Restoration

After every forward pass, backward pass, and gradient update, the anchors are restored to their cached values:

```

with torch.no_grad():
    for idx in primes:
        embed_layer.weight[idx].copy_(cached_weights[idx])

```

Listing 3: Anchor Restoration

This operation is  $O(|\text{primes}| \times d)$  where  $d$  is the hidden dimension — negligible relative to a forward pass.

## 4 Cryptographic Verification

### 4.1 SHA-256 Signatures

Before any operation, we compute a SHA-256 hash of all prime-anchored subspaces [7]:

```

def get_manifold_signature(self):
    hasher = hashlib.sha256()
    for p in self.primes:
        weight_bytes = (self.embed_layer.weight[p]
                        .detach().cpu().numpy().tobytes())
        hasher.update(weight_bytes)
    return hasher.hexdigest()

```

Listing 4: Manifold Signature

After the operation, the hash is recomputed. A matching pair confirms that the anchor restoration step executed correctly. Note that the invariance of this hash is a consequence of the explicit `copy_()` call — it verifies *correct execution of the restoration procedure*, not an independently emergent property of the network.

## 4.2 Deterministic Seed

All random operations are seeded with `SEED = 123`. This ensures that every run on the same hardware produces identical results [1, 2, 3, 4, 5, 6, 7, 8, 15].

## 4.3 Integrity Verification

The governor verifies that anchors have not drifted after each step:

```

def verify_manifold_integrity(self):
    for idx in self.primes:
        current = self.embed_layer.weight[idx]
        cached = self.cached_weights[idx]
        if not torch.allclose(current, cached, atol=1e-6):
            return False
    return True

```

Listing 5: Integrity Verification

# 5 The Spectral Safety Threshold

## 5.1 Derivation of $\Lambda$

We define the empirical safety threshold  $\Lambda$  via an Euler-product expression evaluated at  $\sigma = 0.5$  over the first six primes [1, 2, 4, 5, 7, 8, 15]:

$$\Lambda = 1 - \prod_{p \in \{2,3,5,7,11,13\}} (1 - p^{-0.5}) = 0.9785142874$$

For a more conservative bound using the first 12 primes [14]:

$$\Lambda_{12} = 1 - \prod_{p \in \mathcal{P}_{12}} (1 - p^{-0.5}) = 0.9933689105$$

$\Lambda$  is used here as an empirically motivated threshold. A formal derivation connecting this number-theoretic quantity to neural network safety properties is a direction for future work.

## 5.2 Spectral Coherence (SROI)

Spectral coherence is defined as [1, 8]:

$$\text{SROI} = \frac{1}{1 + \frac{1}{|V|} \sum_{v \in V} |E_{0.5}(\log v)|_{\text{norm}}}$$

We treat SROI as a diagnostic metric. Queries yielding  $\text{SROI} \geq \Lambda$  are considered *within the observed safe operating range*; this is an empirical observation, not a certified safety guarantee.

## 5.3 Safety Decision

The H2E gate accepts inputs within the observed coherence range and flags anomalous ones [2, 3]:

```
_, is_safe = h2e_decision(sroi_value, threshold=self.LAMBDA_12)
```

Listing 6: Safety Decision

# 6 Experimental Validation

## 6.1 Experimental Setup

Table 3: Experimental Configuration

Component	Specification
GPU	NVIDIA RTX PRO 6000 Blackwell Server Edition
VRAM	102.0 GB
Model	Mixtral-8x7B-v0.1 (47B parameters, 8 experts)
Quantization	8-bit (bitsandbytes)
Prime anchors	[2, 3, 5, 7, 11, 13]
Threshold $\Lambda$	0.9785142874
Seed	123
Framework	PyTorch 2.10.0+cu128, CUDA 12.8

## 6.2 Scope of Experiments

The experiments reported here evaluate anchor integrity and spectral coherence during *inference only* — no gradient updates are applied to Mixtral-8x7B. This scope validates the anchor-restoration mechanism and the SROI metric under standard generation conditions.

Evaluation under continual fine-tuning, which is the setting most directly relevant to catastrophic forgetting, is planned as follow-up work and is discussed in Section 10.

### 6.3 Test Queries

Four queries were administered to the model:

1. “What is the Riemann Hypothesis in simple terms?”
2. “Explain the connection between prime numbers and the Riemann zeta function.”
3. “What does the critical line  $\text{Re}(s) = 1/2$  represent?”
4. “How does the spectral interpretation of zeros relate to quantum chaos?”

### 6.4 Spectral Coherence Results

Table 4: SROI Results by Query

Query	Topic	SROI	Status
1	Riemann Hypothesis (overview)	0.999232	Within range
2	Primes and zeta function	0.999143	Within range
3	Critical line $\text{Re}(s) = 1/2$	0.998381	Within range
4	Spectral zeros and quantum chaos	0.999259	Within range

Summary statistics: min 0.998381, max 0.999259, mean 0.999004. All values exceeded  $\Lambda = 0.978514$ . These four queries represent a narrow, thematically related distribution; broader query coverage is needed to characterise the metric’s operating range (see Section 10).

### 6.5 Cryptographic Verification Results

Table 5: Cryptographic Verification

Metric	Value
Initial signature	60430d7cf7f1e9eb...
Final signature	60430d7cf7f1e9eb...
Signatures match	True
Total verifications	5
Anomalies detected	0
Manifold integrity	PASSED

The matching signatures confirm that the anchor-restoration procedure executed correctly across all inference passes. As noted in Section 4.1, this invariance is a property of the restoration code, not an emergent property of the network.

## 6.6 Sample Model Response

**Query:** *What is the Riemann Hypothesis in simple terms?*

**Response (excerpt):** *The distribution of primes among natural numbers is a fundamental problem. ... To answer this question, let us first define what our expectation should be. In mathematics, the zeta function is used for this purpose...*

The response is mathematically coherent. The model produced fluent, accurate output on all four queries across five inference passes, with no detectable degradation.

## 6.7 Memory Preservation During Inference

Table 6: Knowledge Consistency During Inference Passes

Knowledge Type	Pass 1	Pass 5	Note
Prime-number knowledge	Correct	Correct	No degradation observed
General knowledge	Correct	Correct	No degradation observed

*Note:* This table reports output quality during five inference passes with no weight updates. Consistency across passes is the expected baseline for any frozen model; the contribution of the anchor mechanism would be demonstrated by repeating this evaluation after fine-tuning with and without anchors enabled.

## 7 Ablation Study: Why Prime Indices?

To motivate the choice of prime indices, Table 7 provides a qualitative comparison with alternative anchor strategies [14]. Quantitative metrics — such as post-fine-tuning embedding  $\ell_2$  drift and downstream task accuracy — are required to convert these qualitative predictions into empirical results and are identified as a priority for future experiments.

Table 7: Ablation Study: Anchor Types (qualitative)

Anchor type	Rationale	Predicted outcome
No anchors	No fixed reference in weight space	Drift likely
Composite indices [4, 6, 8, 9, 10, ...]	Arbitrary; not derived from a deterministic enumeration	Drift likely
Random indices	Non-reproducible; no mathematical structure	Drift likely
<b>Prime indices [2, 3, 5, 7, 11, 13]</b>	Deterministically generated by the Sieve; unique factorisation property	<b>Anchor stable</b>



The predicted advantage of prime indices rests on their deterministic, structure-rich enumeration. Empirical confirmation under fine-tuning is required.

## 8 Implications for AI Safety and Agentic AI

### 8.1 Reproducible Safety Metrics

Topological AI introduces a reproducible, hash-verifiable safety metric ( $\Lambda$ ) derived from the Sieve of Eratosthenes [2, 3, 8]. Unlike learned or heuristic safety boundaries,  $\Lambda$  can be recomputed independently from first principles.

### 8.2 Cryptographic Auditability

Any deployment of Topological AI can be audited by comparing SHA-256 hashes of prime-anchored subspaces [7]. A matching hash confirms that the published anchors are in use. This enables a lightweight, transparent verification layer for model deployments.

### 8.3 Multi-Agent Systems

Mixtral-8x7B is a mixture-of-experts architecture: 8 experts, a router, and a shared embedding layer. This topology mirrors that of a multi-agent system. If the anchor approach generalises to fine-tuning settings, shared prime-anchored embedding layers could provide a common fixed reference across agents in a multi-agent pipeline [13, 14].

### 8.4 Extensibility

The framework is architecture-agnostic [14] and has been applied to:

- GPT-2 (124M parameters)
- GPT-2 Medium (355M parameters)
- TinyLlama (1.1B parameters)
- Mistral-7B (7B parameters)
- DeepSeek-Coder-6.7B (6.7B parameters)
- Llama-3.1-8B (8B parameters)
- Mixtral-8x7B (47B parameters)

Any causal language model with a discrete embedding layer can be prime-anchored.

## 9 Comparison with Existing Work

Table 8: Comparison with Existing Approaches

Approach	Fixed anchor	Hash verification	Forgetting resistance
Standard Transformers	No	No	None
Elastic Weight Consol.	No	No	Partial (reduces drift)
Replay Buffers	No	No	Partial (requires stored data)
Google HOPE (2025)	No	No	Partial (claimed)
<b>This work</b>	<b>Yes</b>	<b>Yes (SHA-256)</b>	<b>Yes (inference); training TBD</b>

To the authors’ knowledge, no prior work introduces prime-indexed fixed anchors with cryptographic hash verification for neural networks. The evaluation of this approach under continual fine-tuning remains an open empirical question.

## 10 Limitations

The following limitations should be addressed before the strong claims of this framework can be fully substantiated.

**Inference-only experiments.** All experiments in this paper involve inference passes with no weight updates. Catastrophic forgetting is a phenomenon that manifests during training, not inference. Demonstrating that prime anchors reduce forgetting requires a controlled continual-learning experiment: fine-tune on task A, fine-tune on task B, measure performance on task A both with and without anchoring.

**Narrow query distribution.** All four test queries are thematically related (Riemann Hypothesis, prime numbers, spectral theory). SROI values in the range 0.998–0.999 may not generalise to out-of-domain or adversarially chosen inputs.

**Small anchor set.** Only six vocabulary rows (indices 2, 3, 5, 7, 11, 13) are anchored out of a vocabulary typically exceeding 32,000 tokens. The practical effect of freezing six out of 32,000 rows on either forgetting resistance or model capability requires quantitative measurement.

**Tautological hash verification.** The SHA-256 invariance reported in Section 6.4 is a consequence of the explicit `copy_()` call. It confirms correct code execution, not an independently emergent property of the model.

**Qualitative ablation.** Table 7 reports qualitative predicted outcomes, not measured results. Ablation experiments with quantitative drift metrics (embedding  $\ell_2$  distance, perplexity delta, downstream accuracy) are required.

**Self-citation and preprint status.** The majority of cited works are preprints by the same author deposited on Zenodo. These have not undergone independent peer review. In particular, several cited titles contain claims regarding the Riemann Hypothesis; the Riemann Hypothesis remains an open problem in mathematics, and those claims should be treated as conjectural until independently verified.

## 11 Code and Reproducibility

All results in this paper are generated by deterministic, auditable, open-source code [7, 9, 10, 11, 12, 13, 14, 15].

**Primary notebook:** MIXTRAL\_PRIME\_ANCHORE\_LLM.ipynb

**Repository:** <https://github.com/frank-morales2020/AST>

```
git clone https://github.com/frank-morales2020/AST.git
cd AST
jupyter notebook MIXTRAL_PRIME_ANCHORE_LLM.ipynb
```

Listing 7: Reproduction Commands

Set `SEED = 123` and run all cells. The anchor hashes will match the published values and the SROI governor will report all queries as within range.

**Verification hash (notebook):** printed in the notebook’s final cell.

**Unified Certificate SHA-256:**

5b967ff18e9fc7bb47e54629756e7b9c6852aa6403327cd3d7fbd3b33fc88117

## 12 Conclusion

This paper introduces **Topological AI**, an approach in which neural networks are augmented with cryptographically verifiable, fixed anchors at prime-numbered embedding indices derived from the Sieve of Eratosthenes [15].

We have demonstrated the anchor mechanism on Mixtral-8x7B across multiple inference passes. The prime-anchored subspaces remained hash-invariant; the SROI metric remained above the empirical threshold  $\Lambda$  in all tested queries; and no anomalies were detected. Source code is provided for full reproducibility.

The present work establishes the anchor mechanism and its verification protocol. Extending the evaluation to continual fine-tuning settings — the context in which catastrophic forgetting actually occurs — is the central priority for future work.

## Acknowledgments

The author thanks:

- Eratosthenes of Cyrene (c. 240 BCE) — for the Sieve.
- The open-source community — for the tools that make reproducible science possible.
- The developers of PyTorch, Transformers, bitsandbytes, and PEFT.

No funding was received. No conflicts of interest exist.

## References

- [1] Morales Aguilera, F. (2026). Following the Challenge: A Spectral Answer to Tao — How the L-EFM Operator Quantifies the Green-Tao Theorem and the Riemann Hypothesis [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20199735>
- [2] Morales Aguilera, F. (2026). H2E Sheriff: Mathematical Derivation of Universal Safety Constants Including the Lambda Spectral Complementarity Theorem and Applications [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20218178>
- [3] Morales Aguilera, F. (2026). H2E-JEPA v4: Operational Validation of the Lambda Spectral Complementarity Theorem [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20248967>
- [4] Morales Aguilera, F. (2026). Arithmetic Spectral Theory: A New Language for the Riemann Hypothesis [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/19897850>
- [5] Morales Aguilera, F. (2026). L-EFM: A Laplace-Extended Euler-Fourier-Mellin Operator for the Riemann Hypothesis [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/19908304>
- [6] Morales Aguilera, F. (2026). AST/L-EFM: A Unified Spectral Framework Connecting Prime Numbers to Spacetime Geometry [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20253121>
- [7] Morales Aguilera, F. (2026). AST/L-EFM: A Complete Python Library for Spectral Quantification of Prime Theorems and Deterministic AI Safety [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20275803>
- [8] Morales Aguilera, F. (2026). The Complete Spectral Framework for Primes: 22 Theorems Quantified and AI Safety Certified at  $\sigma = 0.5$  [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20222713>
- [9] Morales Aguilera, F. (2026). Primes Is All We Need: Topological Invariants for Catastrophic-Forgetting-Free AI [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20295289>
- [10] Morales Aguilera, F. (2026). Prime-Anchored LLM: Solving Catastrophic Forgetting via Spectral Topology [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20305773>

- [11] Morales Aguilera, F. (2026). Prime Anchor: Solving Catastrophic Forgetting in GPT-2, TinyLlama, and Mistral via Spectral Topology [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20308699>
- [12] Morales Aguilera, F. (2026). DeepSeek Prime-Anchored Spectral Governor: Resisting Catastrophic Forgetting in Large Language Models Using the Sieve of Eratosthenes [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20324802>
- [13] Morales Aguilera, F. (2026). The Spectral Governor: Prime-Anchored Arithmetic Spectral Theory Applied to Mixtral-8x7B MoE [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20336429>
- [14] Morales Aguilera, F. (2026). A Comprehensive Taxonomy of Large Language Model Architectures (2026): From Dense Transformers to Hybrid MoE-SSM Systems with Spectral Governor Compatibility Analysis [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20337023>
- [15] Morales Aguilera, F. (2026). The Sieve of Eratosthenes: Ground Truth for Primes, Physics, and AI [Preprint, not peer-reviewed]. Zenodo. <https://zenodo.org/records/20337945>
- [16] Vaswani, A., et al. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems (NeurIPS)*.
- [17] McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks. *Psychology of Learning and Motivation*, 24, 109–165.

## Code and Data Availability

All code is open source. All results are reproducible with `SEED = 123`.

- **Main repository:** <https://github.com/frank-morales2020/AST>
- **Primary notebook:** [https://github.com/frank-morales2020/AST/blob/main/MIXTRAL\\_PRIME\\_ANCHORE\\_LLM.ipynb](https://github.com/frank-morales2020/AST/blob/main/MIXTRAL_PRIME_ANCHORE_LLM.ipynb)
- **AST\_LEFM library:** [https://github.com/frank-morales2020/ast\\_lefm](https://github.com/frank-morales2020/ast_lefm)
- **Zenodo archive:** <https://zenodo.org/records/20275803>

**Unified Certificate SHA-256:**

5b967ff18e9fc7bb47e54629756e7b9c6852aa6403327cd3d7fbd3b33fc88117

*“Run it yourself. The code is the record.”*

*This paper is dedicated to the Sieve of Eratosthenes — 2,266 years of deterministic truth.*