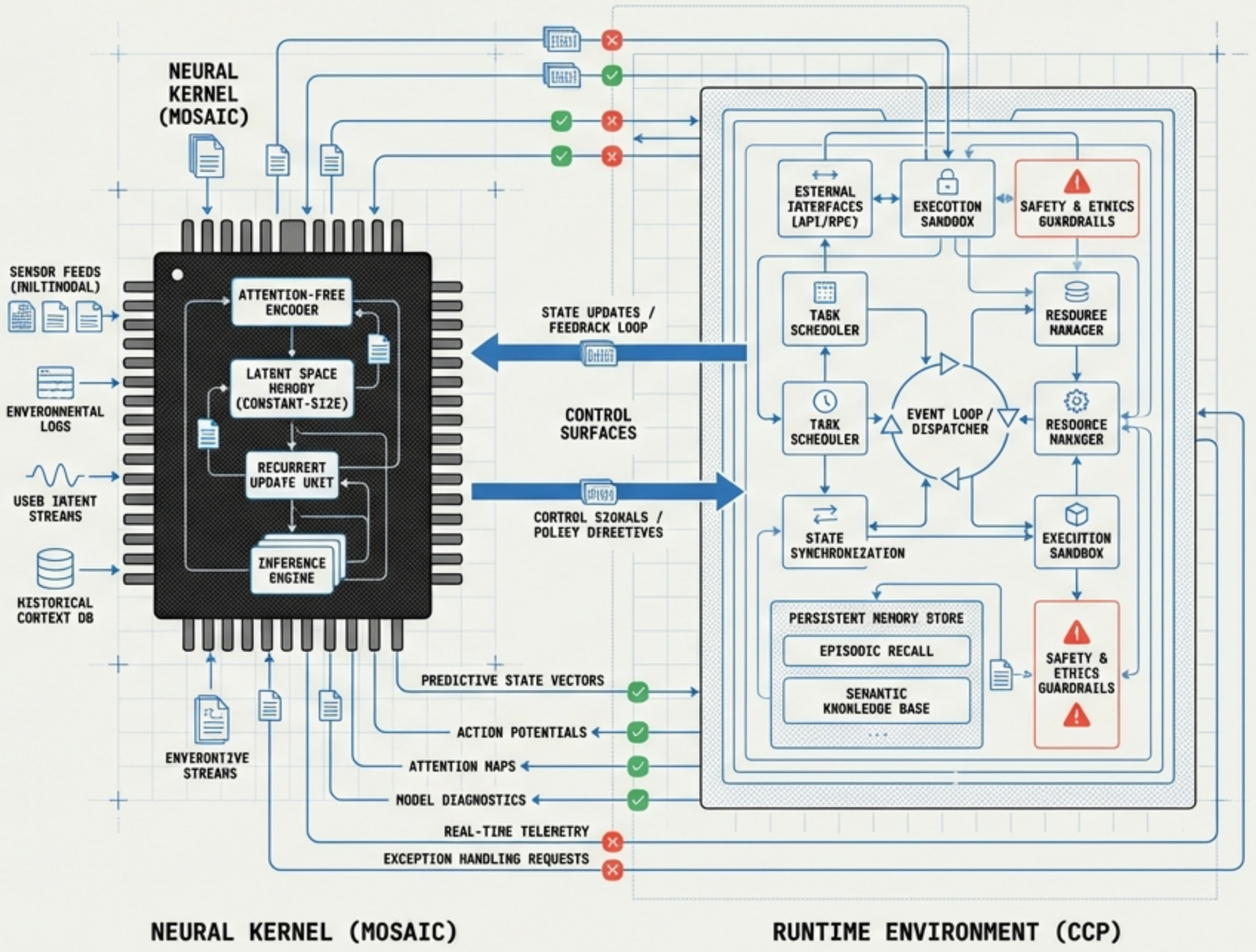
The Cognitive Control Plane

A Runtime Architecture for Persistent, Event-Native Intelligent Systems

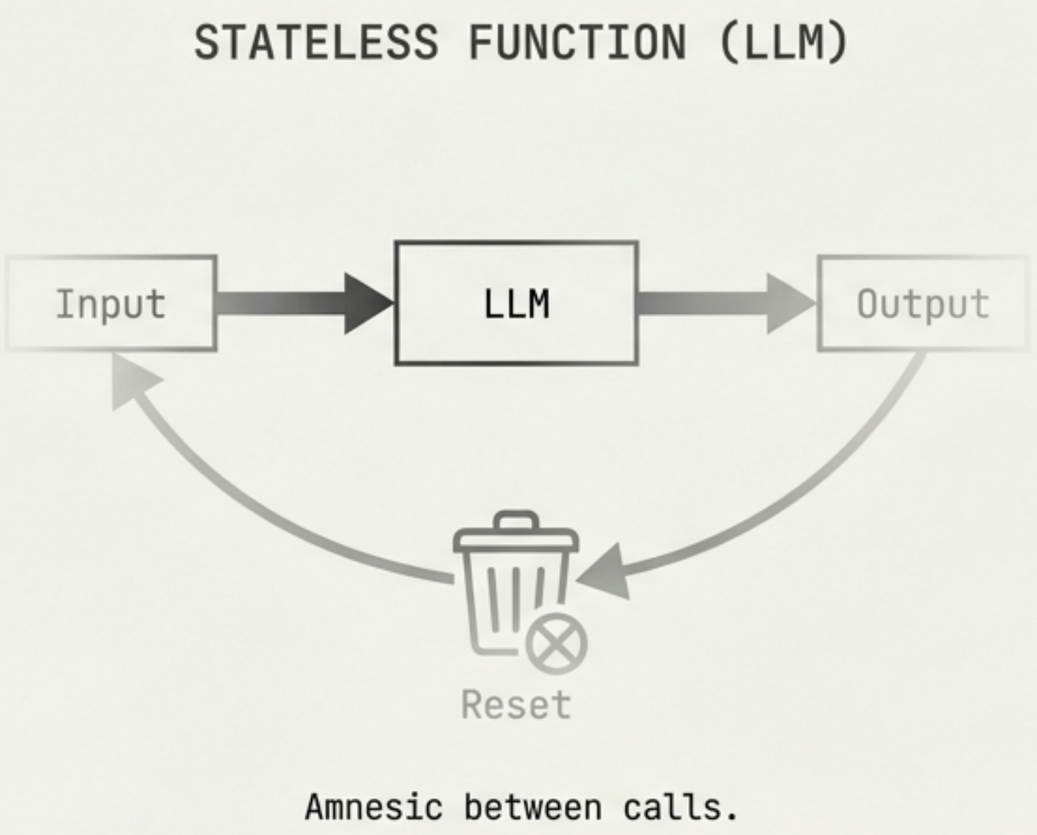
With MOSAIC as an Attention-Free Neural Control Kernel

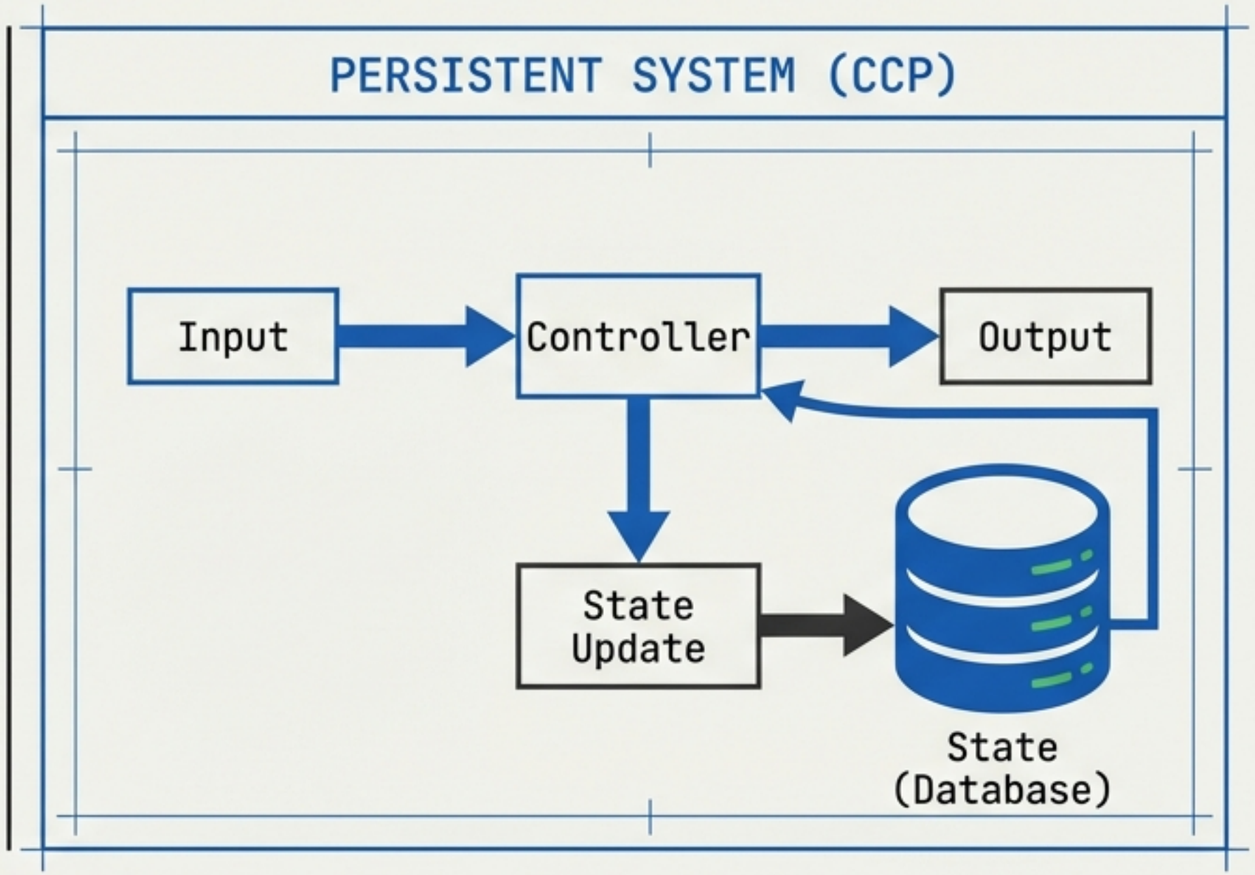
FOCUS: Event-Native Substrate / Constant-Size Memory
STATUS: Diagnostic Infrastructure Proposal
CONTEXT: Beyond the Chatbot Paradigm

System Topology



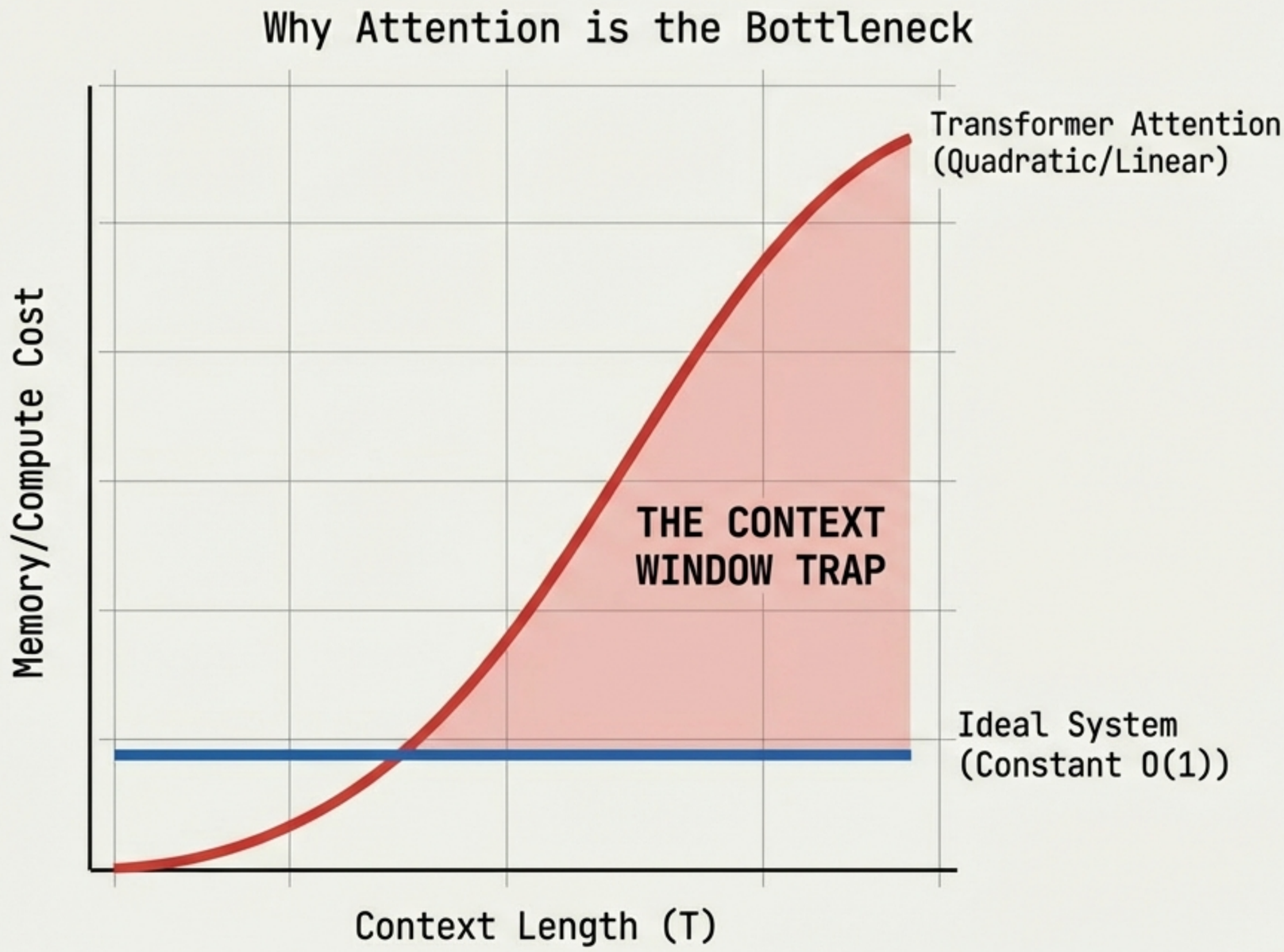
LARGE LANGUAGE MODELS ARE PATTERN MATCHERS, NOT SYSTEMS





5 SYSTEM REQUIREMENTS VS CURRENT REALITY	
System Requirements	**Current Reality**
✅ **Persistence:** State survives sessions	❌ Model forgets everything
✅ **Explicit Memory:** Know what you know	❌ Implicit weight storage
✅ **Bounded Resources:** Predictable costs	❌ KV Cache grows O(T)
✅ **Structured I/O:** Traceable/Replayable	❌ Raw text I/O
✅ **Safe Extensibility:** Gated capabilities	❌ Bolted-on prompting

THE COST OF IMPLICIT MEMORY



THE MECHANISM

Transformer attention acts as memory by storing one vector per past token (KV Cache).

THE CONFLATION

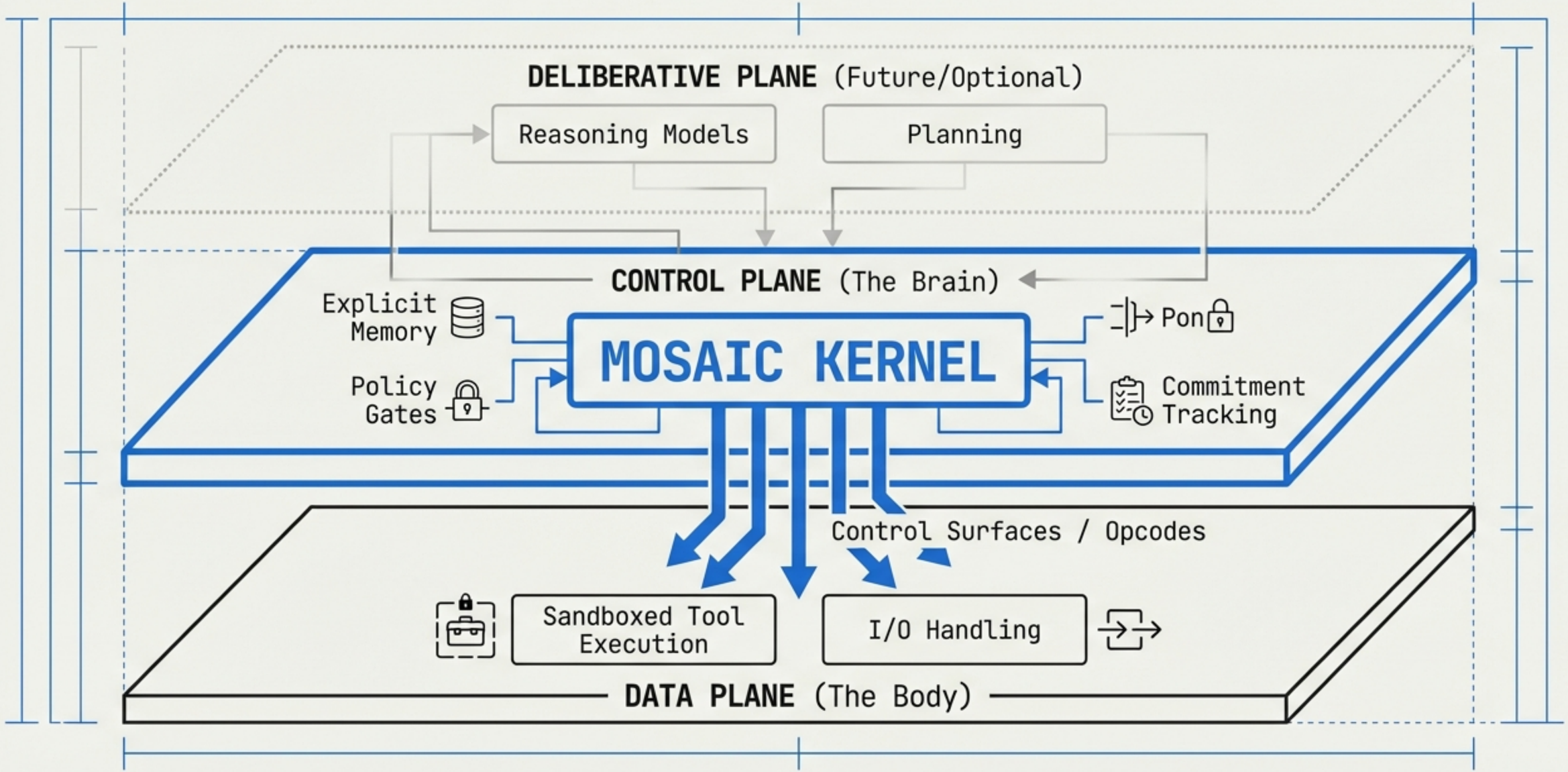
Attention forces three mechanisms into one expensive operation:

1. Local Mixing (Short-range)
2. Long-Range State (Intent)
3. Associative Retrieval (Facts)

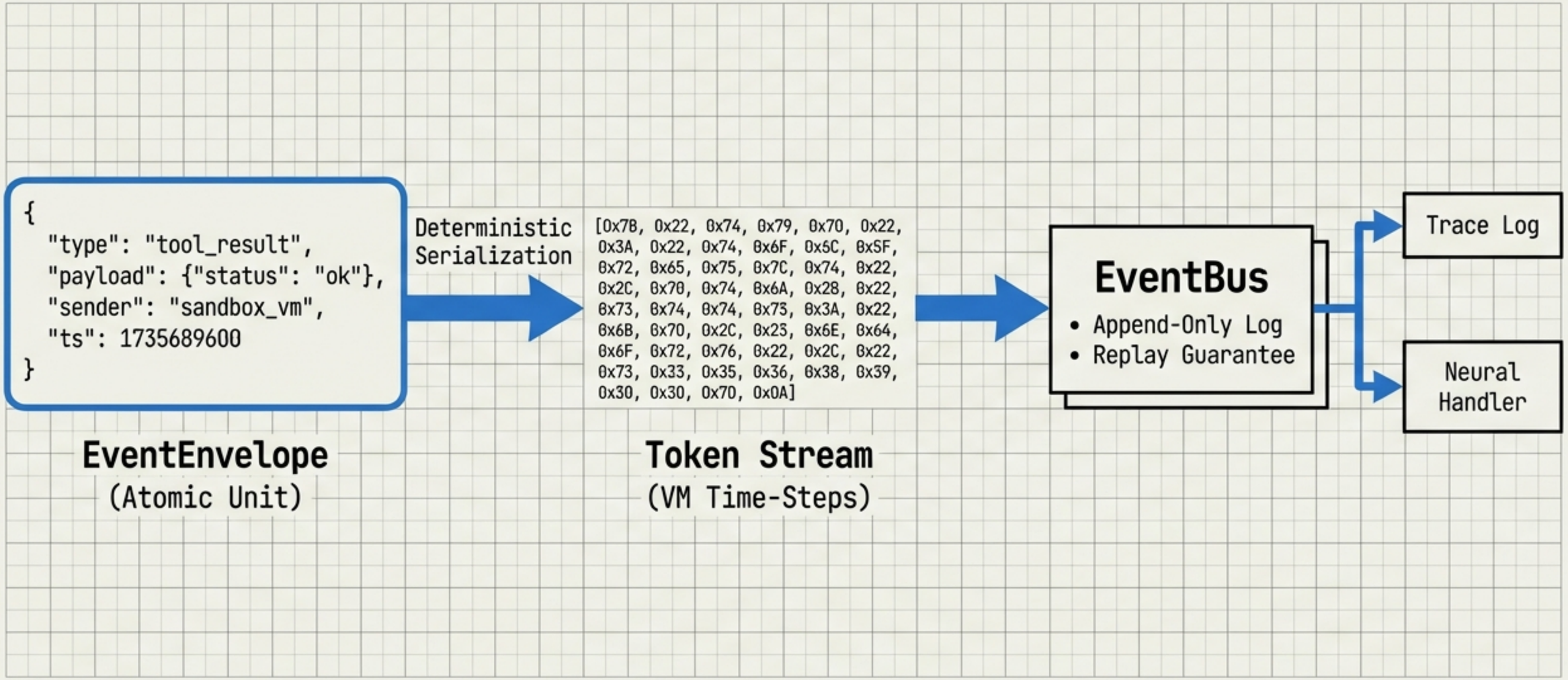
CONSEQUENCES

O(T) linear memory growth makes indefinite operation impossible.

The Architecture: A Three-Plane Decomposition



EVENT-NATIVE REALITY: THE 'VM' VIEW OF INTELLIGENCE



Tokenization is not just compression; it represents the VM cycles required to process an event.

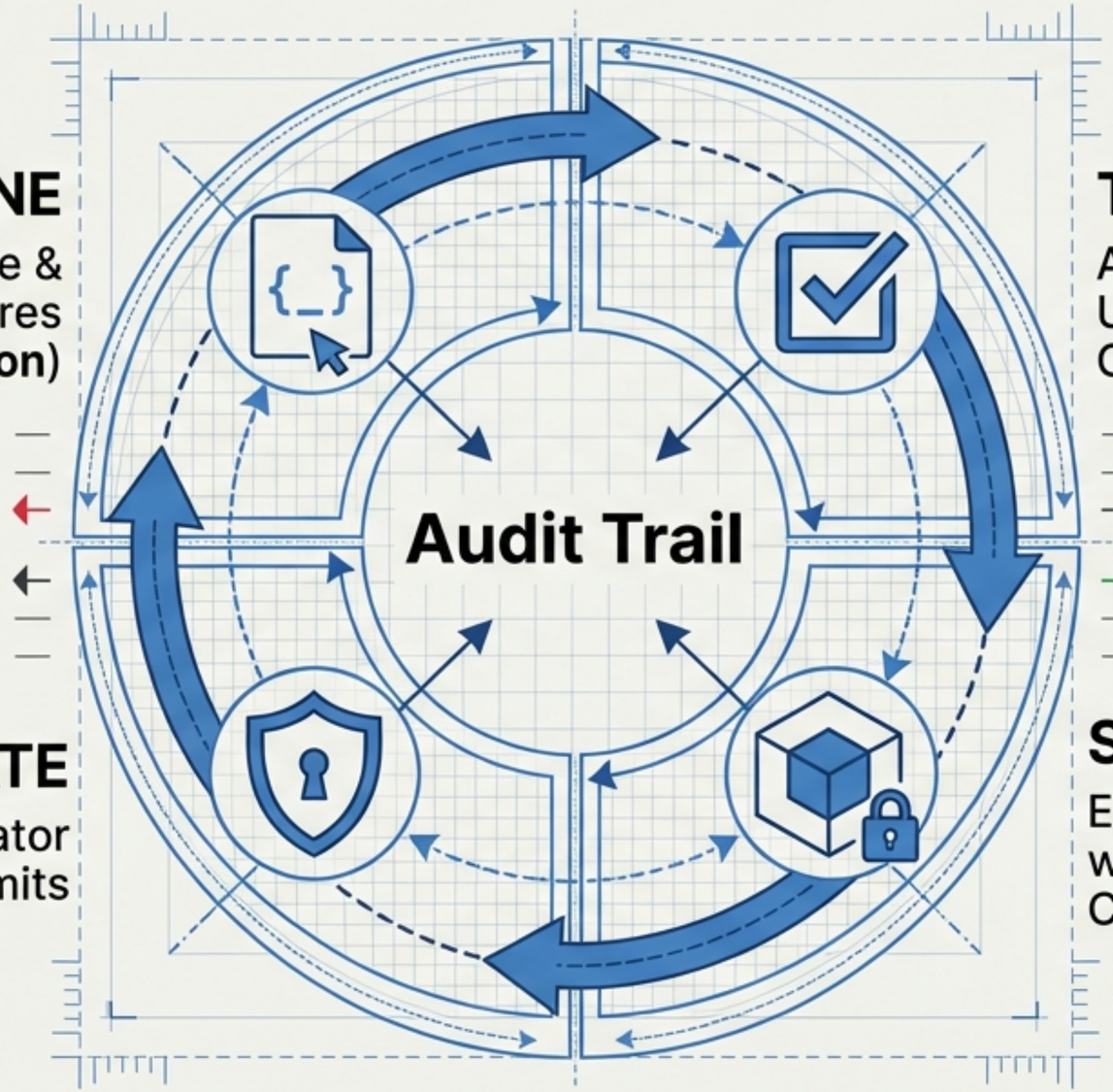
The Tool Lifecycle: Capabilities as Compiled Hypotheses

DEFINE

Propose Interface & Signatures (**ToolDefinition**)

TEST

Auto-generate Unit Tests from Oracle



GATE

Policy Evaluator Enforces Limits

SANDBOX

Execute in Isolation with Whitelisted Capabilities

A tool is not trusted because a model wrote it; it is trusted because it passed objective tests.

MOSAIC: An Attention-Free Neural Control Kernel

Multiscale Oscillator State + Associative Indexed Cache

O(1)

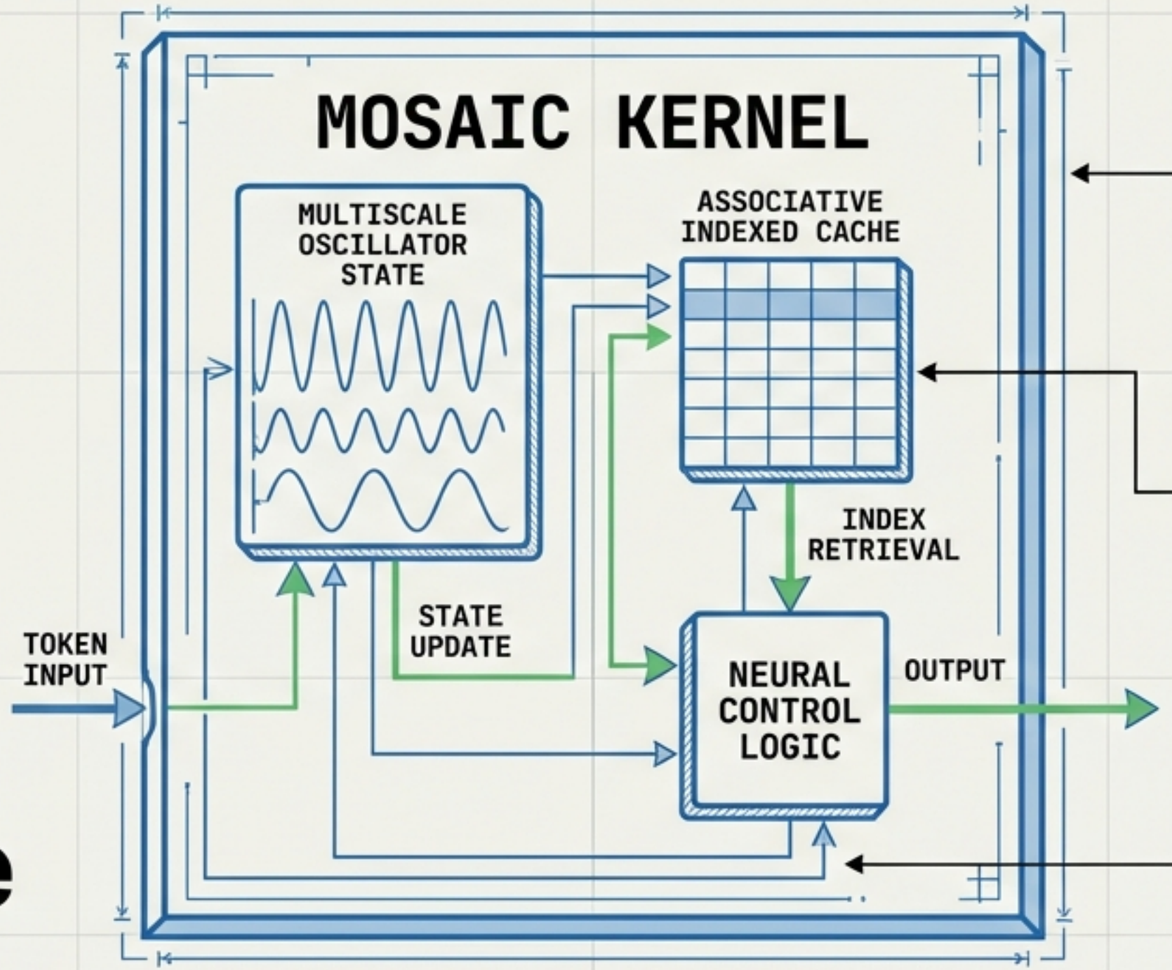
Memory Footprint

O(1)

Inference Time per Token

Fixed-Size

State Buffers

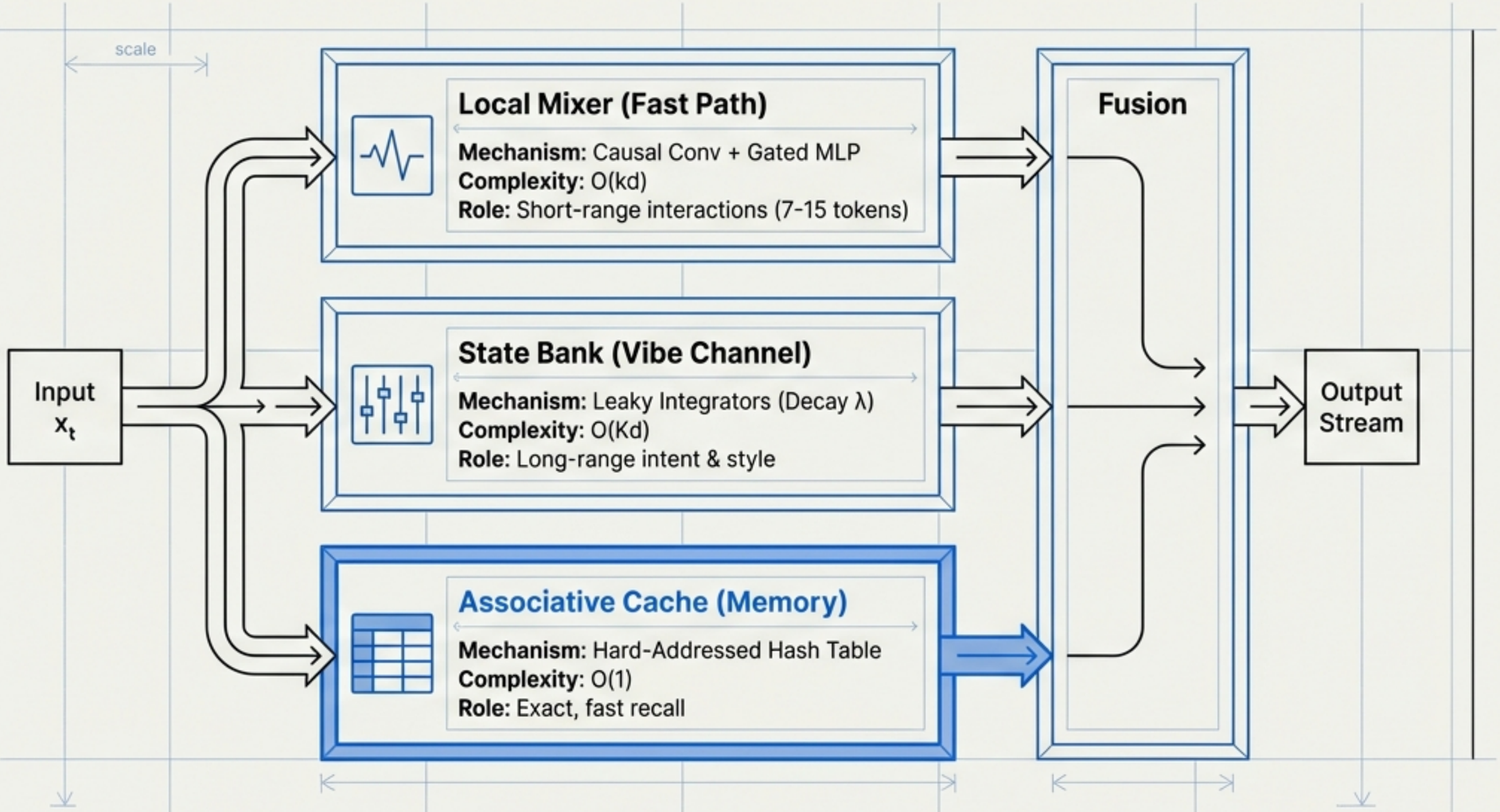


Design Rationale: Stop trying to make the network *be* the memory.

Mechanism: Make the network a *controller* for **explicit, fixed-size data structures.**

Streaming: Each token updates **fixed** buffers only. **No re-reading history.**

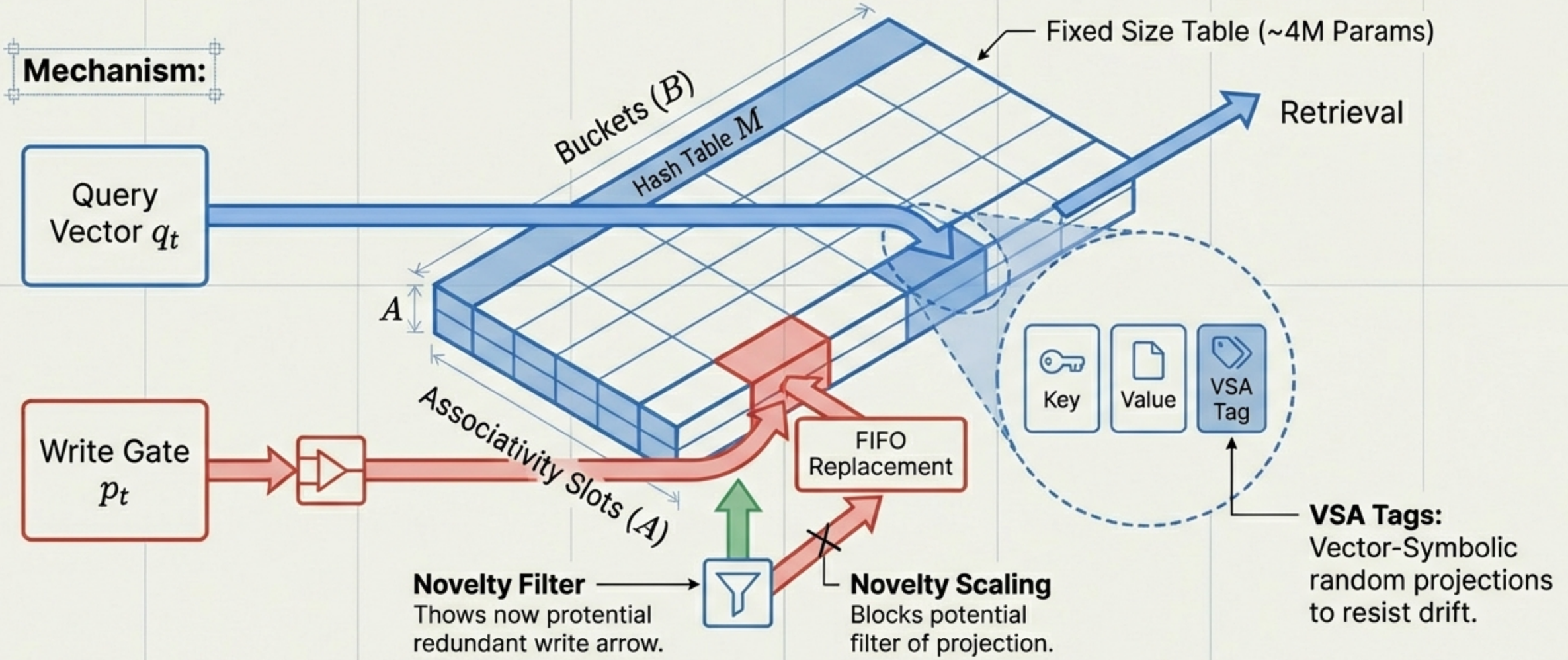
The Triad of Mechanisms: Separation of Concerns



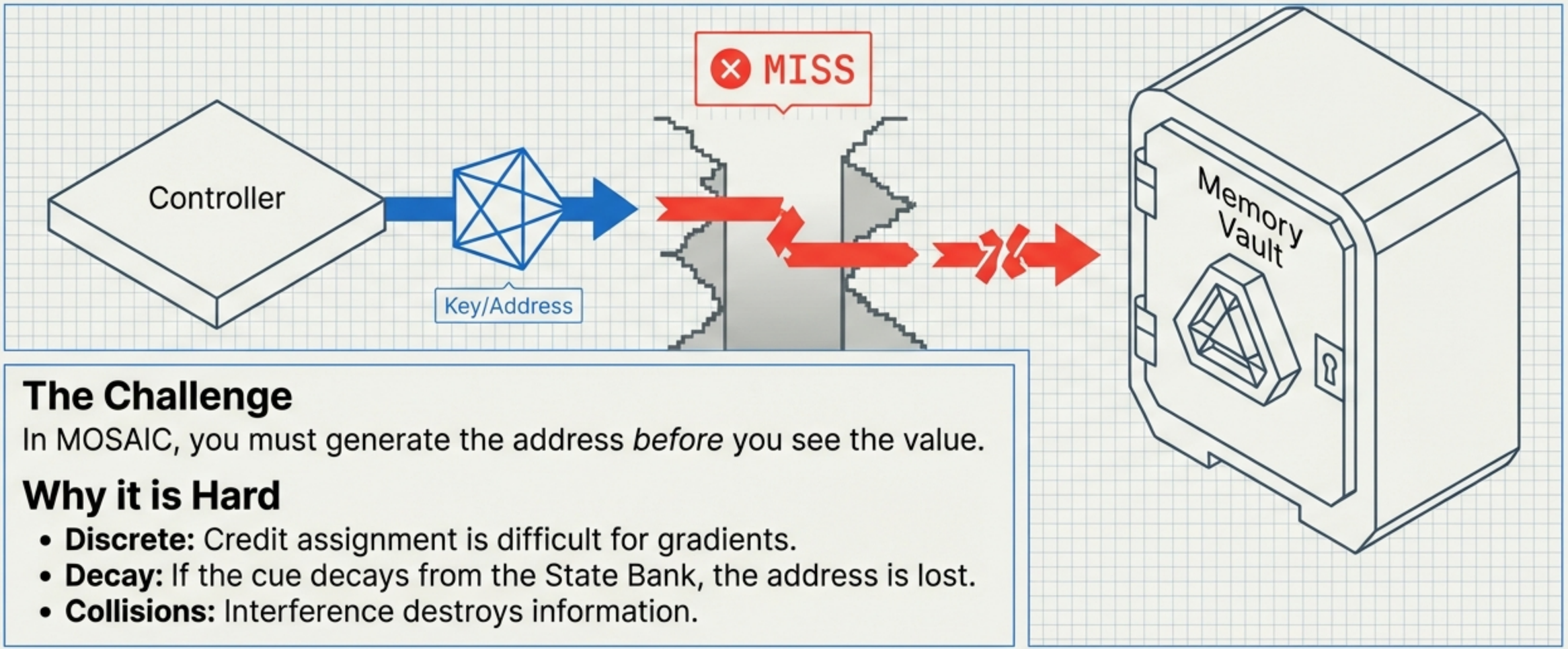
Design Rationale:

- **Separation of Concerns:** Explicitly isolates different temporal scales and data structures.
- **Specialized Modules:** Each track is optimized for a specific type of information processing.
- **Efficiency & Control:** Enables focused state management and fast retrieval.
- **Hybrid Architecture:** Combining neural and fixed-data structures for robustness.

The Associative Cache: Bounded, Explicit Retrieval



The Central Bottleneck: The Addressing Problem



The Challenge

In MOSAIC, you must generate the address *before* you see the value.

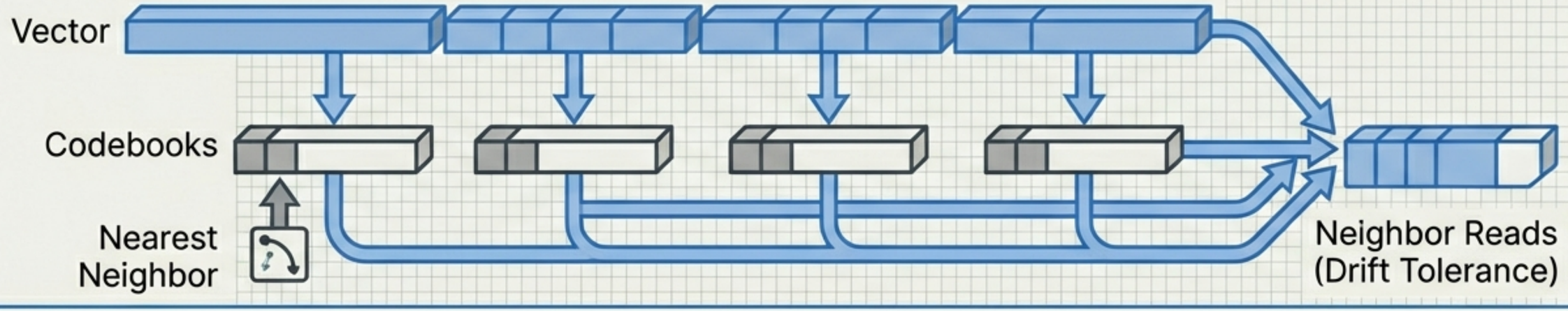
Why it is Hard

- **Discrete:** Credit assignment is difficult for gradients.
- **Decay:** If the cue decays from the State Bank, the address is lost.
- **Collisions:** Interference destroys information.

We treat addressing as a telemetry metric, not a solved feature.

Solving Addressing: Routing & Resonance

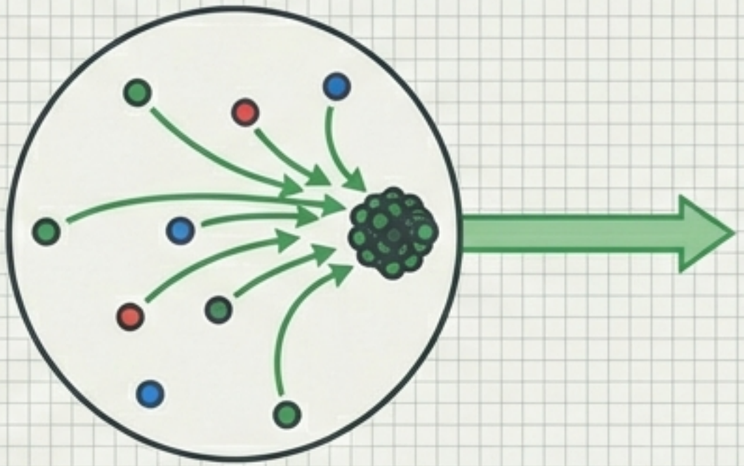
Method A: Product-Quantized VQ Routing



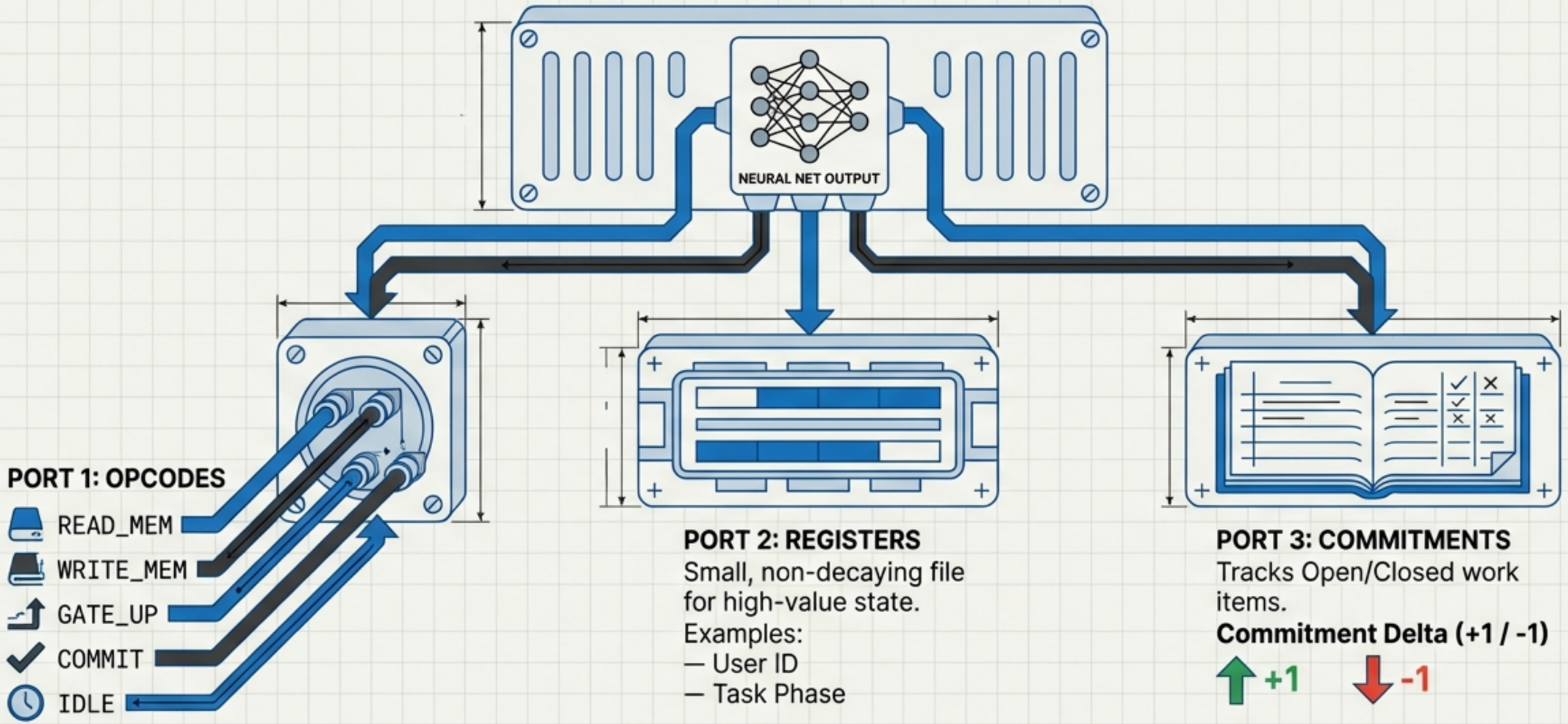
Method B: ResonantRouter

Phase-Coded Attractors

Iterative Settling
Smoother Gradients



Control Surfaces: The Neural-Symbolic Interface



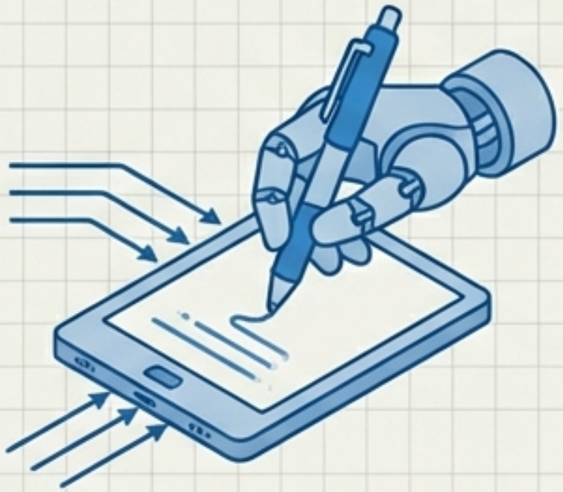
The neural net uses these outputs to **"type"** on the system keyboard.

Training Curriculum: Making the Memory Get Used

Overcoming the Gradient Path of Least Resistance

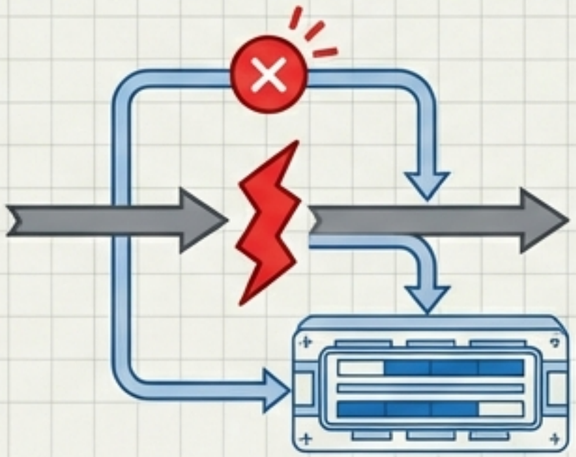
Training Start → Training End

Teacher Forcing



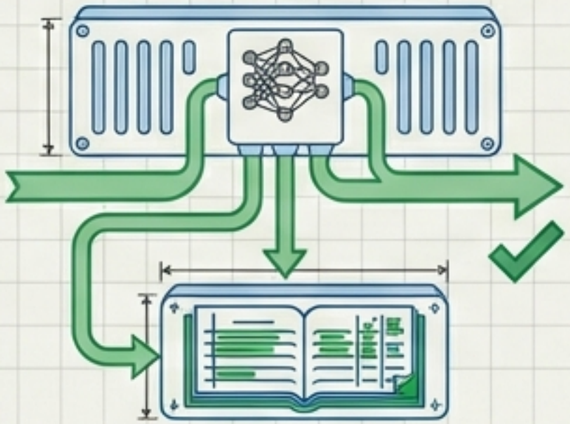
Ground-truth write gates provided. Forced usage.

Forced-Read Dropout



Randomly drop Local Mixer to force reliance on Cache.

Student Control



Anneal to full model control.

Auxiliary Losses

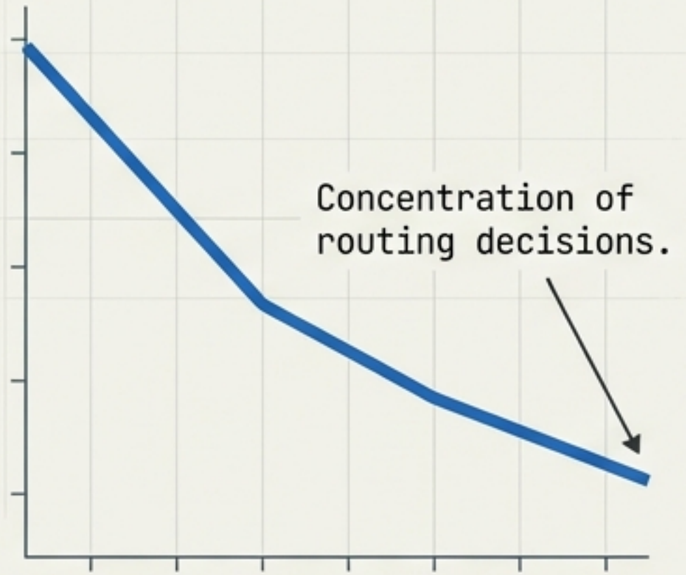
$\mathcal{L}_{sparse}$ (Regularize Writes)

$\mathcal{L}_{util}$ (Predict Future Reads)

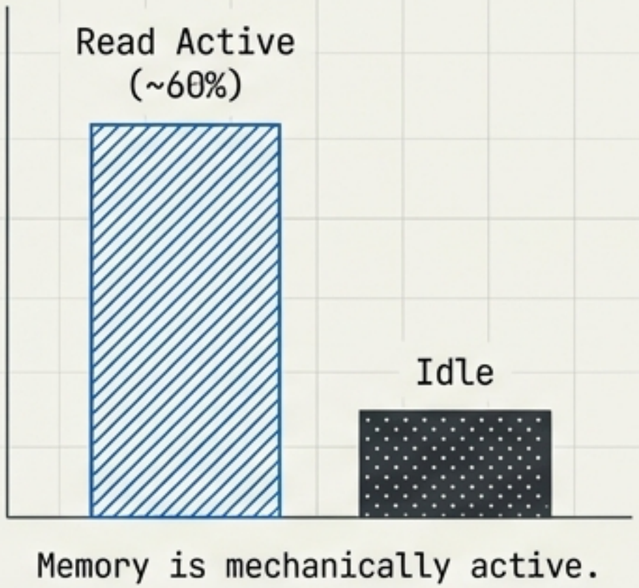
$\mathcal{L}_{NCE}$ (Contrastive Recall)

Diagnostic Experiments & Telemetry

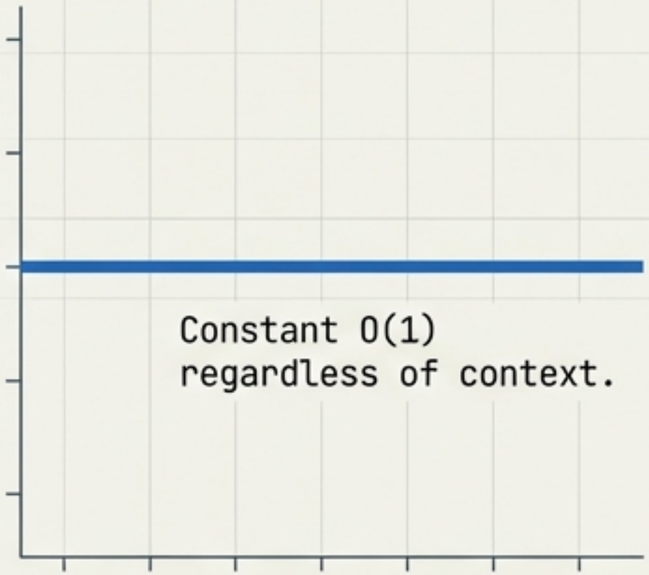
Routing Entropy
(Decreasing = Learning)



Gate Utilization



Inference Speed
(Tokens/sec)

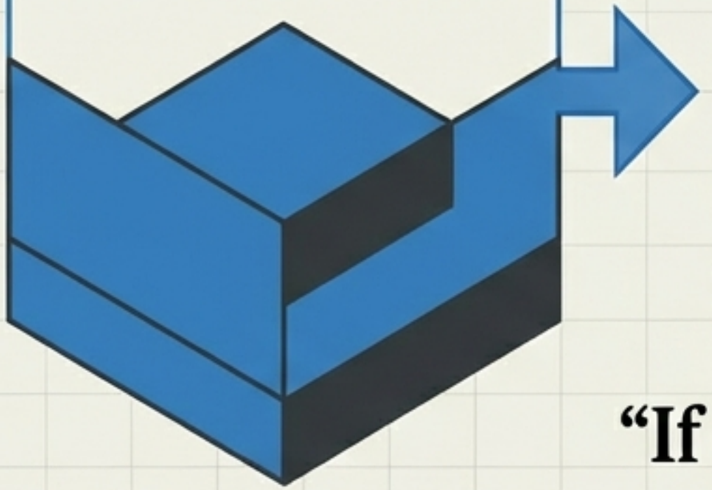


Current Status: Mechanics functional. In-Context Learning (ICL) remains the primary challenge due to addressing.

The Path Forward: From Context Windows to Managed Memory

RUNTIME (CCP)

- Event-Native
- Test-Driven Tools
- Traceable I/O



We have traded the easy optimization of Transformers for the structural correctness of a System.

The bottleneck is now measurable and falsifiable.

“If long-lived intelligent systems are built, they will require a cognitive control plane. This is infrastructure for that future.”

KERNEL (MOSAIC)

- Bounded O(1) State
- Attention-Free
- Explicit Addressing