DISCLAIMER: *FOR INFORMATIONAL PURPOSES ONLY — NOT AN OFFER OR SOLICITATION. PUBLIC TECHNICAL DISCLOSURE.*

AEON: The Dawn of the HYPERION Grid

Solving the Black Box Paradox through Zero-Latency Decoupling and Hierarchical Intelligence

Strategic Technical Roadmap 2025-2026

Vestige Research Team
Proprietary Trading Division
vestige.club

December 12, 2025

Vestige LLC

Precision in Data-Driven Investment Strategies
Defining the Next Generation of Proprietary Trading Infrastructure

Abstract

The Evolution: With the core signal validity now de-risked by our initial PoC (see: *"Can Language Models Trade?"*, Dec 2025), Vestige LLC is **executing Phase 2** of our strategic roadmap. We are declassifying the production infrastructure designed to scale this signal into an autonomous firm: **AEON** (the HYPERION Grid).

The Challenge: The central conflict in modern AI trading is the trade-off between *Interpretability* (the "Black Box Problem") and *Latency*. Traditional solutions that require models to "explain themselves" introduce unacceptable delays (>100ms), rendering them useless for high-frequency execution in volatile futures markets.

The AEON Solution: Our architecture introduces the **Shadow Log Protocol**. By decoupling the trading impulse (logit-based execution) from the rationale generation (asynchronous explanation), AEON achieves sub-millisecond execution while maintaining a fully auditable, human-readable cognitive trail. This system is powered by a hierarchical swarm of specialized agents—Director, Quant, Risk, and Execution—operating on a unified memory substrate provided by our NVIDIA DGX Spark GB10 clusters.

Strategic Expansion: Beyond the core futures strategy, this infrastructure lays the groundwork for **AEON Horizon**, a dedicated system for long-term equity portfolios utilizing fundamental unstructured data analysis. This report provides a comprehensive technical breakdown of the topology, latency optimization strategies, and the strategic rollout plan for 2026.

Mandatory Disclaimer

CRITICAL DISCLAIMERS — READ BEFORE PROCEEDING

Research Status: This document outlines a **development roadmap** for proprietary technology. The systems described are currently in active training and development phases. Vestige LLC does **NOT currently employ full autonomous agent swarms for live trading**.

Hypothetical Performance: Any references to performance goals (e.g., "Sharpe ¿ 2.0") are forward-looking targets based on theoretical modeling and initial backtests. **These are not guarantees of future results.** Actual trading involves substantial risk of loss. Past performance is not indicative of future results.

Company Structure: Vestige LLC is a solely owned proprietary trading firm managing only its own capital. We do not solicit, accept, or manage external funds, offer investment advisory services, or distribute algorithms to external parties.

Not Investment Advice: This report does not constitute investment advice, an offer to sell, or solicitation to buy securities. The website vestige.club is for informational purposes only.

Contents

1 The Paradigm Shift: From Models to The HYPERION Grid

1.1 The Limitations of Monolithic Models

Our initial research proved that a fine-tuned Large Language Model (LLM) could "read" a price chart and generate a valid trade signal with a 94% technical accuracy rate. However, relying on a single, monolithic model for end-to-step trading decisions presents significant structural weaknesses:

- **Contextual Overload:** Asking one model to analyze the Federal Reserve's minutes (Macro), identify a double-bottom pattern (Technical), calculate the Kelly Criterion position size (Risk), and route the order via TWAP (Execution) dilutes its attention mechanisms. The "needle in the haystack" problem becomes acute as context windows fill with disparate data types.
- **Sequential Processing Latency:** LLMs are inherently sequential token generators. Waiting for a model to output a reasoning paragraph before executing a trade introduces fatal latency in high-frequency environments.
- **Fragility:** A single hallucination in a monolithic model propagates through the entire decision chain. If the model misinterprets a macro event, it may force a trade despite technical indicators suggesting caution.

1.2 The Solution: The HYPERION Grid

We are deploying a new categorical infrastructure class: the **HYPERION Grid** (High-Yield Parallel Execution & Reasoning I/O Network). This architecture decomposes the trading problem into distinct cognitive domains, handled by specialized agents operating in parallel on a shared memory substrate.

- **Specialization:** The **OMEGA Agent** (1.5B parameters) focuses purely on probabilistic bounds and portfolio heat, while the **ENGRAM Agent** (7B parameters) focuses exclusively on pattern recognition and signal generation.
- **Asynchronous Cognition:** We fundamentally separate the *impulse* to act from the *rationale* for acting. This allows for execution speeds bounded only by network latency, not cognitive processing time.
- **Zero-Copy Interoperability:** Leveraging the NVIDIA GB10's unified memory architecture allows us to pass tensors and state vectors between agents without the serialization overhead typical of microservices architectures.

2 System Architecture: The AEON Framework

We define the trading decision D_t at time t not as a single model output, but as the composite function of a hierarchical agent swarm.

2.1 Solving the Black Box Paradox: The Shadow Log Protocol

The financial industry has long struggled with a binary choice: fast "black box" algorithms that are opaque, or slow transparent models that are interpretable but unusable for alpha capture. AEON solves this via the **Shadow Log Protocol**.

1. **The Impulse Path ($t < 1$ms):** The **Execution Agent** monitors the **Quant Agent's** raw output logits (probability distribution) in real-time. When the probability of the token "BUY" exceeds the confidence threshold θ (e.g., 0.85), the trade is executed immediately. No text is generated. No tokens are decoded. This occurs at the speed of a single forward pass.

2. **The Shadow Path ($t \approx 100$ms):** Simultaneously, a parallel thread (the "Shadow Thread") triggers the **ARCHON** and **ENGRAM Agents** to generate the natural language rationale explaining *why* the decision was made. This rationale is logged to the immutable audit database but does not block execution.

This ensures that our system is **zero-latency** in execution while remaining **100% interpretable** for post-trade analysis, compliance, and risk management.

2.2 Mathematical Formulation of Agent Consensus

Let the final trade vector $\mathbf{v}_t$ (volume to trade) be defined as:

$$\mathbf{v}_t = \mathcal{E}\left(\mathcal{R}\left(\mathcal{Q}(\mathbf{x}_t) \mid \Omega_t\right) \mid \Theta_t\right) \tag{1}$$

Where:

- $\mathbf{x}_t$: Market microstructure data (OHLCV, Order Book).
- Ω_t: Portfolio state (Equity, Exposure, Drawdown).
- Θ_t: Macro-strategic context (Regime, Thesis).
- $\mathcal{Q}$: **ENGRAM** function (Signal Generation).
- $\mathcal{R}$: **OMEGA** function (Sizing & Gatekeeping).
- $\mathcal{E}$: **NEXUS** function (Routing & Timing).

Crucially, the OMEGA function $\mathcal{R}$ acts as a non-linear filter (gatekeeper). If the risk score exceeds a threshold λ, the volume is forced to zero regardless of the ENGRAM Agent's confidence:

$$\mathcal{R}(s, c) = \begin{cases} 0 & \text{if RiskScore}(s) > \lambda \\ \text{Kelly}(c) \cdot \text{RegimeMultiplier}(\Theta_t) & \text{otherwise} \end{cases} \tag{2}$$

2.3 Topology Visualization

The following diagram illustrates the data flow and hierarchical control within the AEON system.

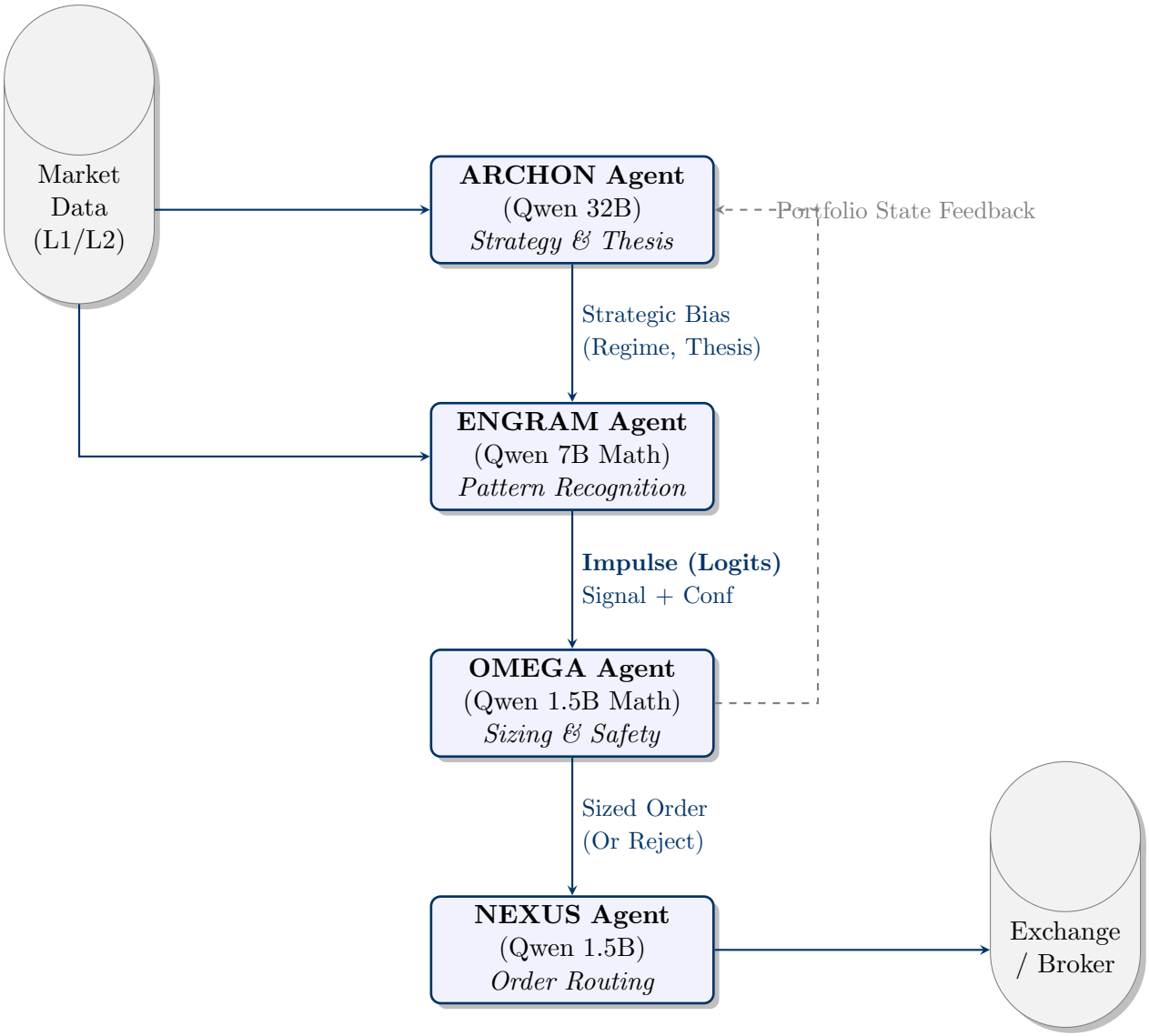

Figure 1: The Vestige AEON Architecture: A top-down hierarchical control system. Strategic direction informs tactical analysis, which is strictly filtered by risk parameters. The "Impulse" path operates at machine speed, decoupled from the "Rationale" generation.

3 Detailed Agent Specifications

The efficacy of the HYPERION Grid relies on the specialized training and distinct roles of each agent.

3.1 1. ARCHON (Strategy)

Full Name: Adaptive Regime Cognition Heuristic Operations Node
Model Base: Qwen2.5-32B-Instruct
Role: The "Portfolio Manager."
Function: ARCHON does not concern itself with tick-by-tick price action. Its cognitive horizon spans hours to days. It ingests daily and weekly OHLCV structures, economic calendars (CPI, FOMC), and broad volatility indices (VIX).

- **Input:** Daily OHLCV, VIX, Economic Events, Sentiment Aggregates.
- **Output:**
  - **Regime Classification:** (e.g., "Trending Bullish", "Mean Reverting", "High Volatility Choppy").
  - **Risk Multiplier:** A scalar (0.0 - 1.5x) that adjusts the aggressiveness of downstream agents.
  - **Thesis:** A natural language description of the current market stance.
- **Training Objective:** Maximize regime identification accuracy and minimize exposure during "black swan" conditions.

3.2 2. ENGRAM (Tactical)

Full Name: Enhanced Neural Gradient Recognition & Analysis Module
Model Base: Qwen2.5-Math-7B (Currently Training on Node 1)
Role: The "Technical Analyst."
Function: ENGRAM operates within the strategic constraints set by ARCHON. It is a highly specialized pattern recognition engine, fine-tuned on over 1 million OHLCV sequences of NQ futures data.

- **Input:** 1-minute and 5-minute OHLCV sequences (normalized), Volume Profiles.
- **Output:**
  - **Action:** BUY / SELL / HOLD.
  - **Confidence Score:** 0.0 - 1.0.
  - **Logits:** Raw probability distribution for the Impulse Path.
- **Special Features:** Trained with Focal Loss to prevent "Hold Bias" and incorporates specialized math-reasoning for Fibonacci and pivot point calculations.

3.3 3. OMEGA (Gatekeeper)

Full Name: Optimized Margin, Equity & Guardrail Architecture
Model Base: Qwen2.5-Math-1.5B (Active Development on Node 2)
Role: The "Risk Officer."
Function: OMEGA is the only entity with the authority to *veto* a trade. It is mathematically rigid and operates on strict probabilistic logic. It calculates position size based on the Kelly Criterion and current portfolio drawdown.

- **Input:** ENGRAM Signal, Account Equity, Current Drawdown, Market Volatility (ATR).
- **Output:**
  - **Position Size:** Number of contracts.
  - **Stop Loss / Take Profit:** Exact price levels.
  - **Decision:** EXECUTE or REJECT.
- **Training Objective:** Minimize Max Drawdown (MDD) while maximizing Sharpe Ratio.

3.4 4. NEXUS (Implementation)

Full Name: Network Execution & Xchange Utilization System
Model Base: Qwen2.5-1.5B (Planned)
Role: The "Trader."
Function: Responsible for the micro-structure of the trade. It decides how to enter the market to minimize slippage and impact.

- **Input:** Sized Order, Level 2 Order Book Depth.
- **Output:** FIX Messages (NewOrderSingle), Execution Reports.
- **Logic:** Decides between Market vs. Limit orders, Iceberg sizing, or TWAP execution based on available liquidity.

4 Expansion Roadmap: AEON Horizon

While the core AEON system targets high-frequency futures trading, the HYPERION Grid architecture is asset-agnostic. We are actively developing a secondary infrastructure targeted at long-term equity portfolios: **AEON Horizon**.

4.1 Concept and Vision

AEON Horizon is a "Slow-Thinking" swarm designed to emulate the workflow of a fundamental equity analyst. Unlike the futures swarm which reacts to price, Horizon reacts to *value* and *information*.

4.2 Technical Components

- **Unstructured Data Ingestion:** Horizon will utilize a Retrieval-Augmented Generation (RAG) pipeline to ingest:
  - SEC Filings (10-K, 10-Q)
  - Earnings Call Transcripts
  - Macroeconomic Reports (PDFs from Fed, ECB)
  - News Sentiment Feeds
- **Cognitive Task:**
  - **Semantic Analysis:** Analyzing management sentiment and linguistic complexity in earnings calls to detect obfuscation or confidence.
  - **Moat Assessment:** Evaluating competitive advantages based on Porter's Five Forces framework.
  - **Valuation Modeling:** Generating multi-year DCF (Discounted Cash Flow) inputs.
- **Execution Cycle:** Weekly or Monthly rebalancing. This system prioritizes depth of reasoning over speed of execution.

4.3 Swarm Adaptation

For AEON Horizon, the **ENGRAM Agent** is replaced by a **Fundamental Analyst Agent**, while the **ARCHON Agent** retains its role as the macro-strategist. The **OMEGA Agent** shifts its focus from volatility-based sizing to portfolio variance and correlation management.

5 Infrastructure and Latency Optimization

Vestige has deployed significant compute resources to realize this vision. The hardware foundation is critical to enabling the "Zero-Copy" architecture.

5.1 Hardware: The DGX Spark Cluster

We are utilizing a dual-node **NVIDIA DGX Spark** setup operating in a cluster compute configuration.

- **Compute:** 2x NVIDIA DGX Spark Nodes with dual **GB10 Blackwell** accelerators.
- **Memory:** 256GB Unified Memory (128GB per node), allowing massive model residency.
- **Processor:** 40-core Host CPU for high-throughput data preprocessing.
- **Storage:** 8TB NVMe Data Storage for tick-data lakes.
- **Interconnect:** 200Gb/s QSFP+ cabling for ultra-low latency internode sync.

5.2 Latency Minimization Strategy

To achieve sub-millisecond impulse execution, we employ three key technical strategies:

5.2.1 1. Unified Memory Access (Zero-Copy)

Traditional multi-agent systems rely on API calls (HTTP/REST) to pass messages between agents. This introduces serialization (JSON) and network latency overheads.

- **AEON Approach:** All agents reside in the same GPU memory space. They read from a shared "Context Tensor."
- **Benefit:** Passing data becomes a memory pointer operation, effectively instantaneous.

5.2.2 2. Optimistic Execution

The NEXUS Agent does not wait for the "end of sentence" token.

- **Mechanism:** As the ENGRAM Agent generates tokens, the NEXUS Agent monitors the probability distribution of the first token. If $P(\text{BUY}) > \text{Threshold}$, the FIX message is pre-built and sent.
- **Safety:** The OMEGA Agent operates in parallel on the pre-computation vector to issue a "Kill" signal if necessary, but the default path is optimistic.

5.2.3 3. Efficient Fine-Tuning (LoRA)

We utilize standard LoRA (Low-Rank Adaptation) with Rank 32. Unlike 4-bit quantization methods (QLoRA) which introduce quantization noise, standard LoRA maintains full bfloat16 precision for maximum signal fidelity while still allowing us to fit the entire agent swarm within the 256GB Unified Memory envelope.

- **Benefit:** Reduces the trainable parameter count significantly while preserving the reasoning capabilities required for complex pattern recognition.

6 Development Timeline and Milestones

Our deployment strategy is phased to ensure rigorous testing of each swarm component before integration.

6.1 Phase 1: ARCHON Agent

Status: DESIGN

- **Objective:** Build the macro-strategist.
- **Timeline:** Q1 2026.

6.2 Phase 2: ENGRAM Agent (Active)

Status: TRAINING (Node 1)

- **Objective:** Train the core pattern recognition engine.
- **Data:** 1M+ OHLCV samples, normalized.
- **Features:** Focal Loss, bfloat16 precision.
- **Early Training Metrics:**
  - Initial Loss: 5.34 $\rightarrow$ 1.48 (Rapid pattern acquisition).
  - Gradient Norm: 2.89 $\rightarrow$ 1.05 (Stabilizing convergence).
  - Learning Rate: Linear warmup to 7.3×10^{-5}.
- **ETA:** 1-1.5 Months (Iterative Refinement).

6.3 Phase 3: OMEGA Agent (Next)

Status: ACTIVE TRAINING (Node 2)

- **Objective:** Train the gatekeeper to size positions and reject bad signals.
- **Data:** Synthetic scenarios generated from historical NQ trades.
- **Model:** Qwen2.5-Math-1.5B.
- **Early Training Metrics:**
  - Initial Loss: 3.08 $\rightarrow$ 0.28 (Extremely fast adaptation).
  - Gradient Norm: 1.10 $\rightarrow$ 0.16 (High stability).
  - Speed: 15.6s/it (Optimization focus).
- **Timeline:** 1-1.5 Months (Iterative Refinement).

6.4 Phase 4: NEXUS Agent

Status: DEVELOPMENT

- **Objective:** Integrate agents into the "Zero-Copy" fabric and handle order routing.
- **Timeline:** Q1 2026.

6.5 Status Summary

Table 1: Development Status by Component

Component	Status	Est. Completion
ARCHON Agent	*Design Phase*	Q1 2026
ENGRAM Agent Training	**Active (Node 1)**	1-1.5 Months (Iterative)
OMEGA Agent Training	**Active (Node 2)**	1-1.5 Months (Iterative)
NEXUS Agent	*In Development*	Q1 2026
Risk Data Generation	**Complete**	Dec 12, 2025
AEON Horizon (Long Term)	*Conceptual*	Q3 2026

7 Strategic Impact: Why This Matters

For institutional-grade trading, "alpha" (the signal) is only 20% of the equation. The other 80% is risk management and execution.

7.1 The Fragility Solution

By moving to this multi-agent system, Vestige LLC is solving the **"Fragility Problem"** of single-model systems. If a single large model makes a mistake, it fails catastrophically. In our system, if the Quant Agent hallucinates a setup during a market crash, the Director Agent (seeing the high VIX) will have already set the Risk Multiplier to 0.0x, effectively "grounding" the trader. This provides a "Defense in Depth" strategy.

7.2 The Black Box Solution

AEON proves that we do not need to sacrifice speed for transparency. The **Shadow Log Protocol** allows us to have our cake and eat it too: sub-millisecond execution for the market, and detailed, paragraph-length explanations for the auditors.

7.3 Future Proofing

The "HYPERION Grid" is modular. As new, more powerful models are released (e.g., Qwen 3, Llama 4), individual agents can be upgraded without dismantling the entire infrastructure. This ensures Vestige remains at the cutting edge of AI capability.

This is not just an algorithm; it is an **autonomous digital trading firm**.

A Quantitative Methodology

A.1 Focal Loss Implementation

To address class imbalance (where "HOLD" signals predominate), we implement a multi-class Focal Loss function. Standard Cross-Entropy Loss (CE) treats all misclassifications equally. Focal Loss adds a modulating factor $(1 - p_t)^\gamma$ to down-weight easy examples and focus training on hard negatives.

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t) \tag{3}$$

Where:

- p_t is the model's estimated probability for the class with label $y = 1$.
- $\gamma = 2.0$ is the focusing parameter.
- α_t is the class weight vector (balancing BUY/SELL vs HOLD).

This ensures the Quant Agent does not converge to a trivial solution of "always holding."

A.2 Kelly Criterion for Position Sizing

The Risk Agent utilizes a continuous Kelly Criterion formulation to determine optimal leverage f^* based on the Quant Agent's confidence score c.

$$f^* = \frac{p \cdot b - q}{b} \cdot \Theta_{\text{regime}} \tag{4}$$

Where:

- $p = c$ (Confidence Score as probability of win).
- $q = 1 - c$ (Probability of loss).
- b is the odds ratio (Reward/Risk), estimated dynamically from recent volatility (ATR).
- Θ_{regime} is the ARCHON Agent's risk multiplier.

B Infrastructure Specifications

B.1 Compute Node Topology

The AEON system runs on a high-availability cluster optimized for tensor parallelism. We are currently experimenting with various model architectures and sizes. The final production infrastructure will likely operate on rented high-performance compute clusters (e.g., 8x H200 instances) to support multiple HYPERION Grid shards.

Table 2: Current Development Cluster Specifications (2x DGX Spark)

Component	Specification
Compute	2x NVIDIA DGX Spark Nodes (Cluster Compute)
GPU Architecture	2x Blackwell GB10 Accelerators
Unified Memory	256 GB HBM3e (Unified Pool)
Host Processor	40-Core High-Frequency CPU
Interconnect	200Gb/s QSFP+ Cabling
Storage	8TB NVMe Data Lake

B.2 Zero-Copy Data Plane

The "HYPERION Grid" relies on the Unified Memory capabilities of the GB10. Agents do not communicate via TCP/IP sockets. Instead, they read/write to specific memory addresses acting as a shared "blackboard."

Listing 1: Conceptual Zero-Copy Implementation

```
# Shared Memory Block (Tensor)
shared_context = torch.zeros(
    (BATCH_SIZE, HIDDEN_DIM),
    dtype=torch.bfloat16,
    device='cuda:0'
)

# Agents access via pointers, not copies
archon_view = shared_context[0:1]  # Slice reference
engram_view = shared_context[1:2]  # Slice reference
omega_view  = shared_context[2:3]  # Slice reference

# Updates appear instantly across all views
archon_view.copy_(new_strategy_tensor)
# engram_view sees this immediately (0ns latency)
```

Vestige LLC: Defining the Future of Proprietary Trading. Visit vestige.club. No solicitation or investment services offered.