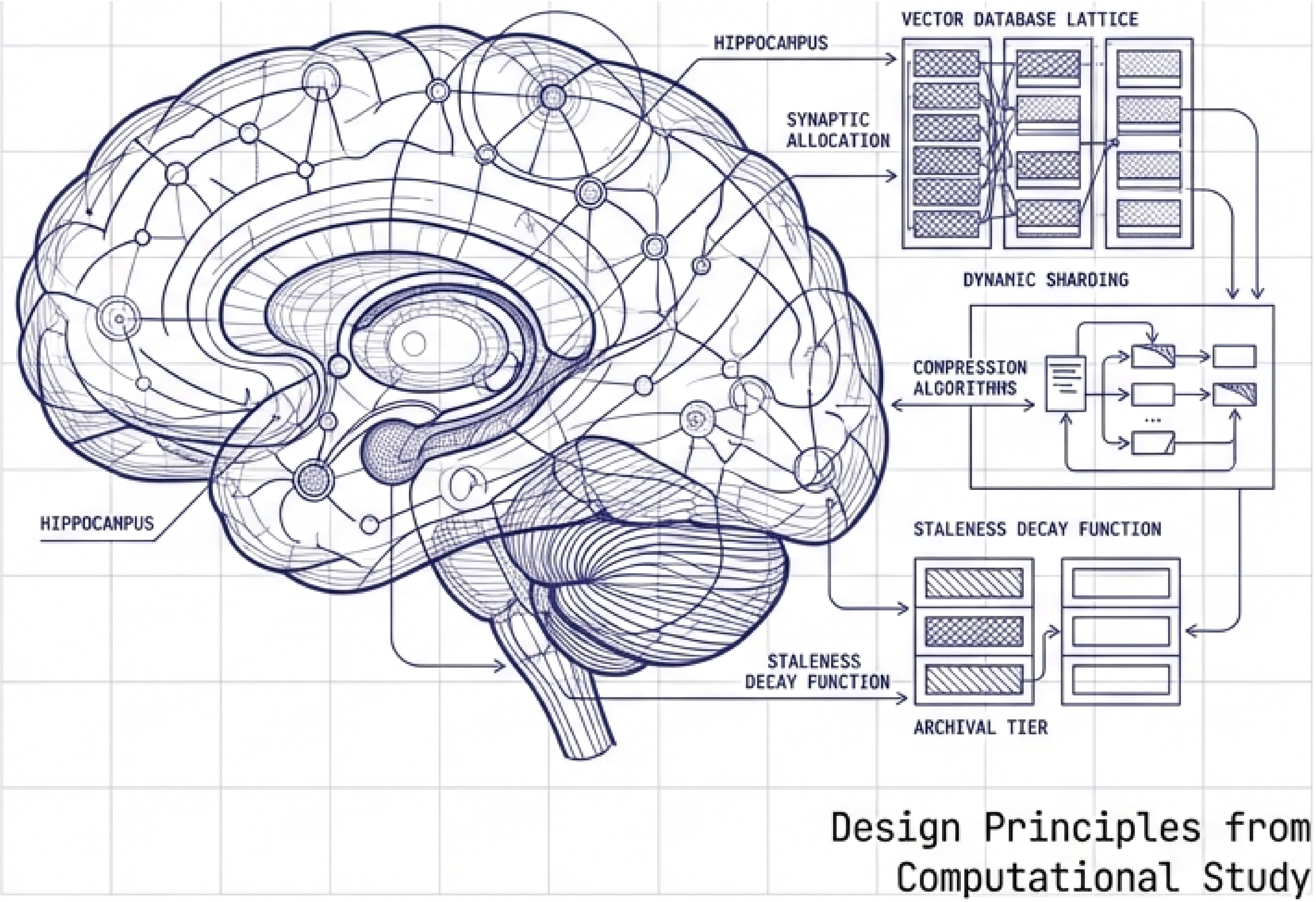
Architecting Long-Term Memory for Autonomous Agents

A Blueprint for Optimizing Allocation, Compression, and Staleness in LLM Systems.

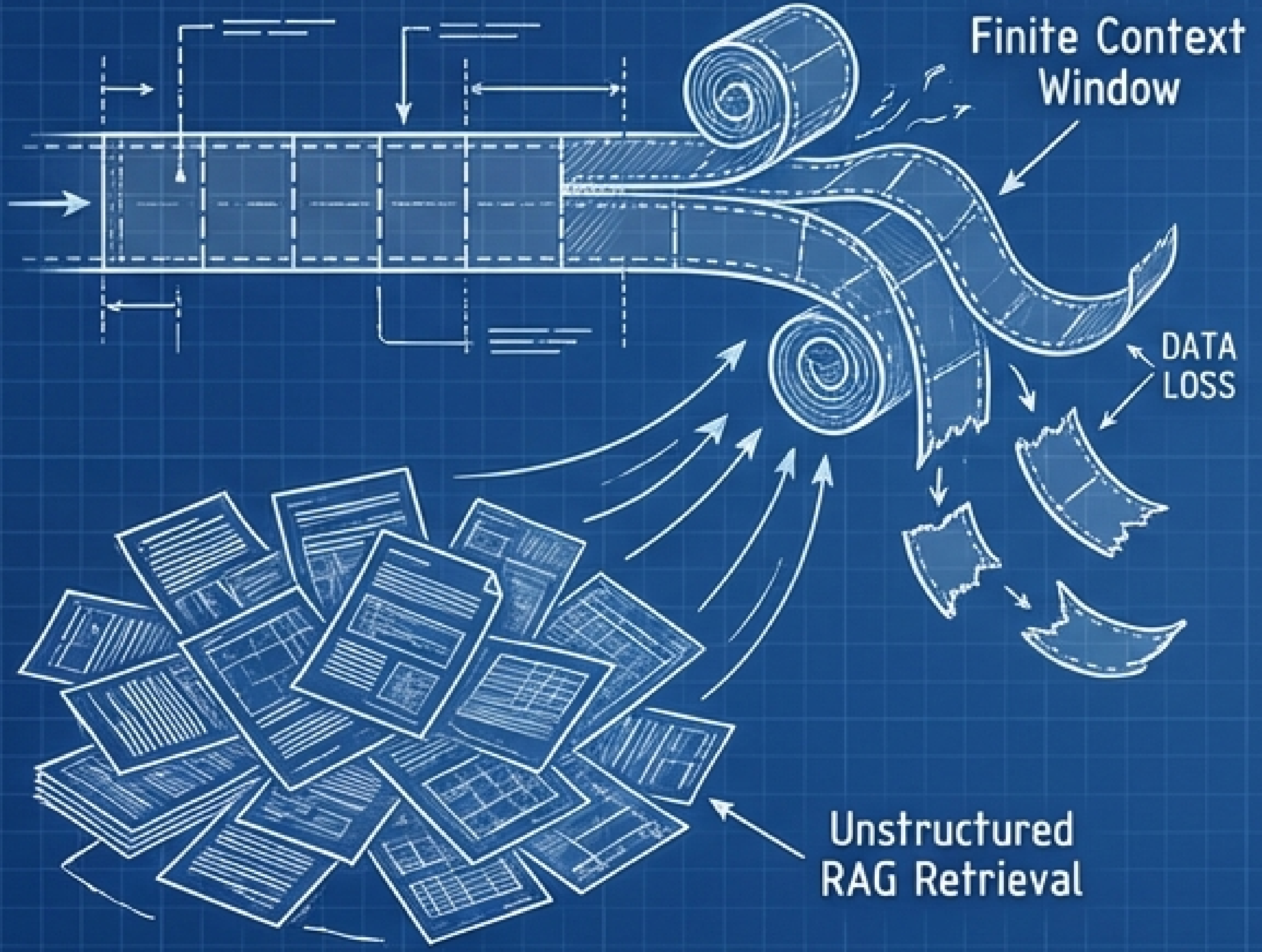


EXPERIMENT CONTEXT: Systematic simulation across 500-step task horizons.

Design Principles from Computational Study

THE LIMITS OF CONTEXT WINDOWS AND RAG

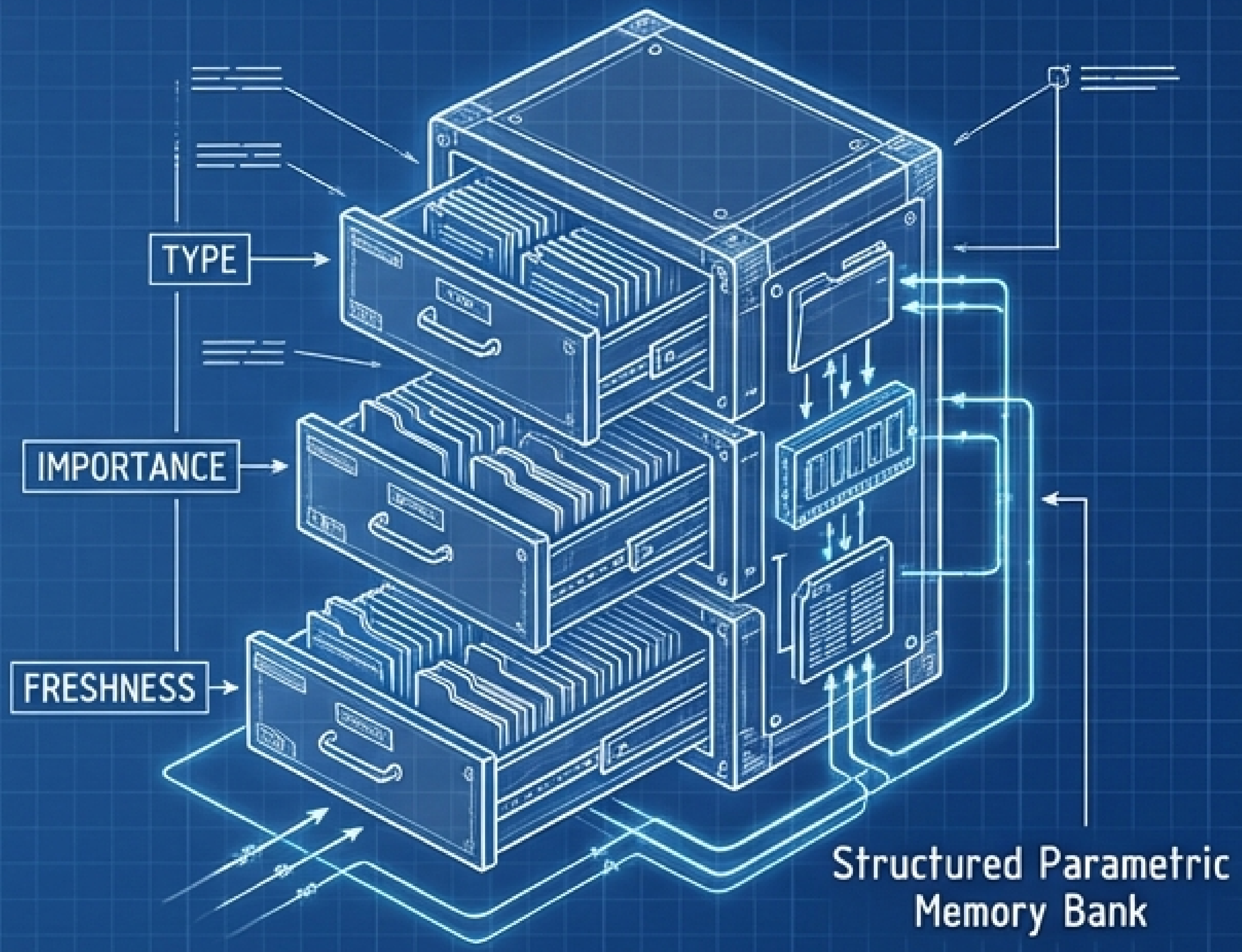
CURRENT STATE: THE CONTEXT TRAP



Context Window: Finite capacity. Useful for short-term recall but fails to maintain state over extended interactions.

- RAG: Provides a baseline but lacks structure. Treats all retrieved chunks as equal, leading to context clutter.

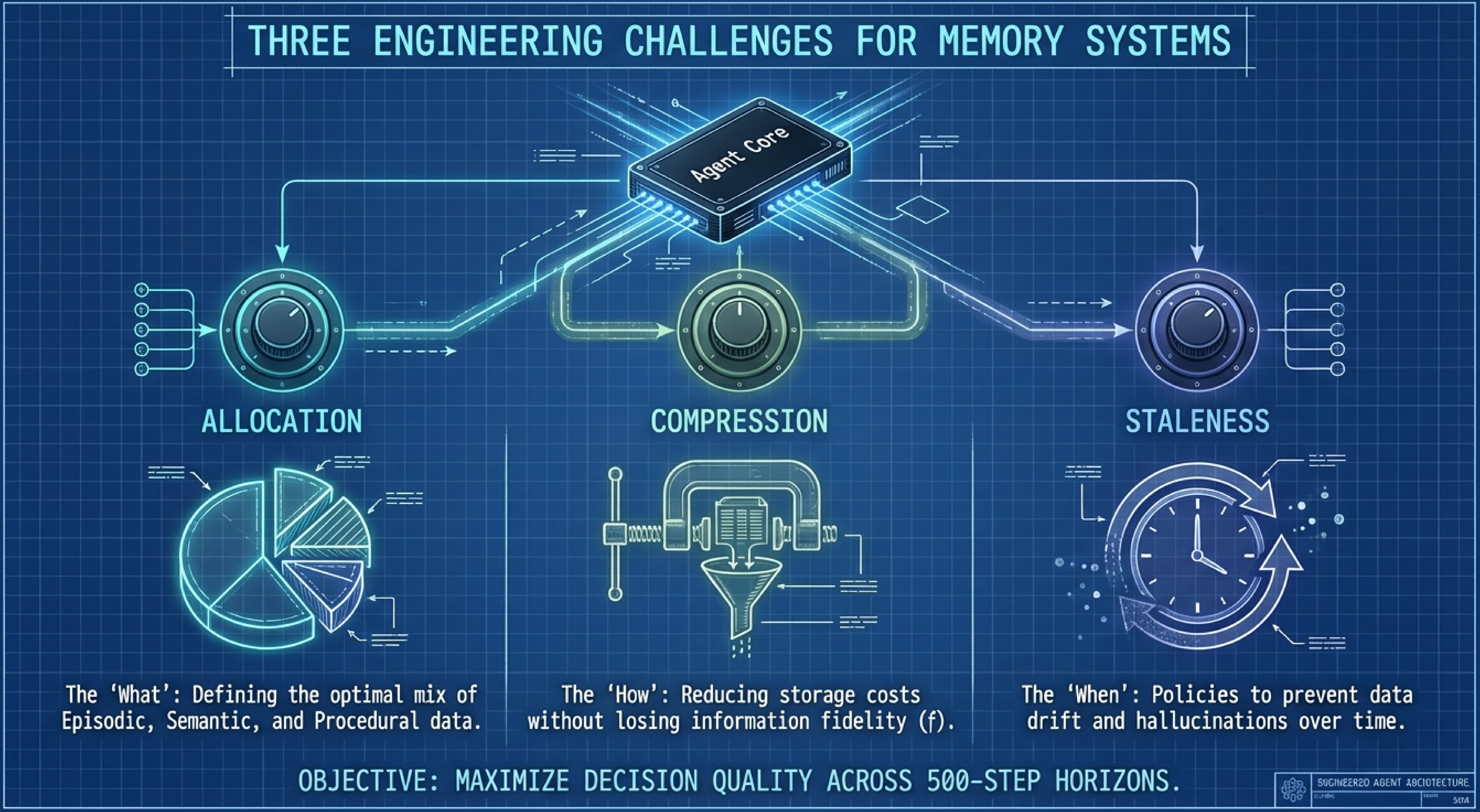
TARGET STATE: PARAMETRIC MEMORY

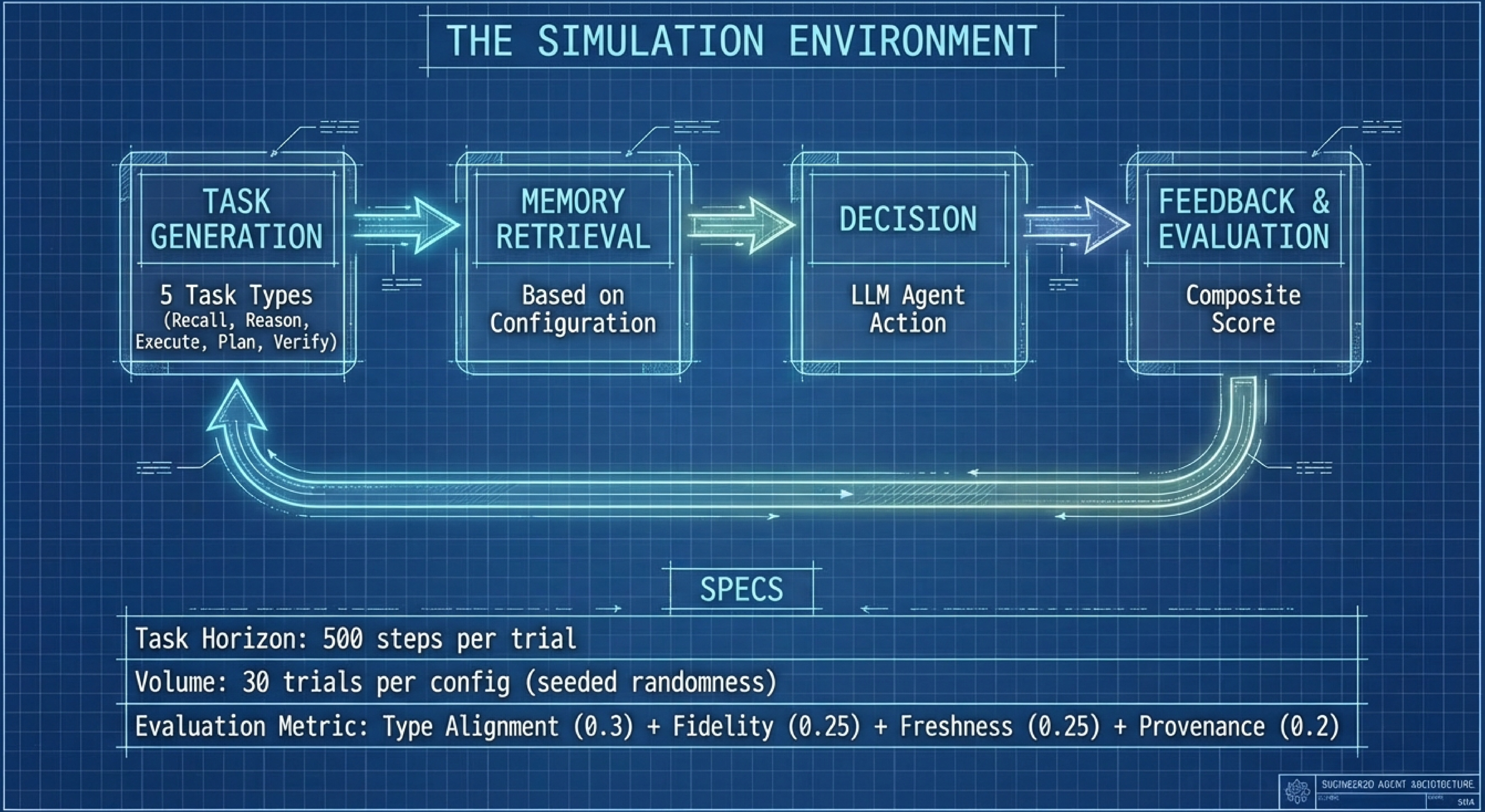


Parametric Memory: Agents need a structured, fixed-capacity store that mimics cognitive distinctness.

- The Goal: Move from ‘retrieving text’ to ‘managing state’ across Type, Importance, and Freshness.

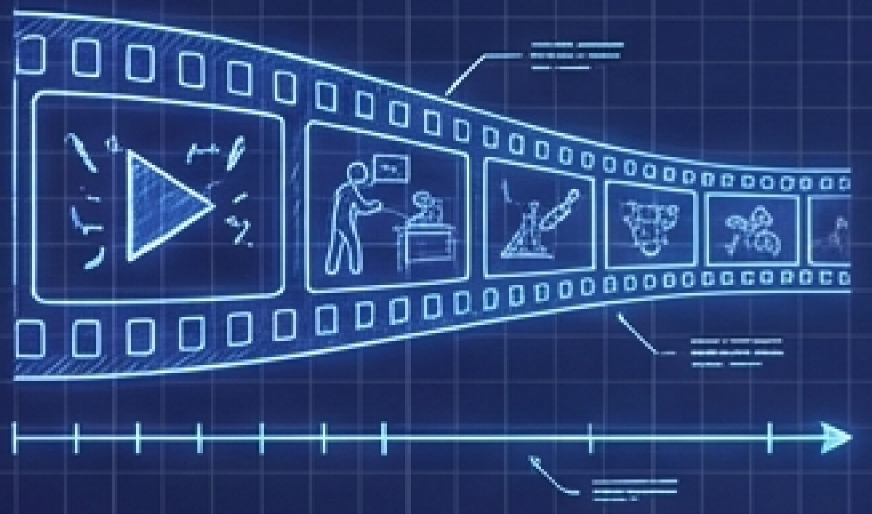
ENGINEERED AGENT ARCIOTECTURE


THREE ENGINEERING CHALLENGES FOR MEMORY SYSTEMS
Agent Core
ALLOCATION
COMPRESSION
STALENESS
The 'What': Defining the optimal mix of Episodic, Semantic, and Procedural data.
The 'How': Reducing storage costs without losing information fidelity (f).
The 'When': Policies to prevent data drift and hallucinations over time.
OBJECTIVE: MAXIMIZE DECISION QUALITY ACROSS 500-STEP HORIZONS.


THE SIMULATION ENVIRONMENT
TASK GENERATION
5 Task Types
(Recall, Reason, Execute, Plan, Verify)
MEMORY RETRIEVAL
Based on Configuration
DECISION
LLM Agent Action
FEEDBACK & EVALUATION
Composite Score
SPECS
Task Horizon: 500 steps per trial
Volume: 30 trials per config (seeded randomness)
Evaluation Metric: Type Alignment (0.3) + Fidelity (0.25) + Freshness (0.25) + Provenance (0.2)

Pillar I: Memory Type Allocation

Categorizing State for Optimal Retrieval



EPISODIC

Event records and specific interaction history.

"What happened?"

SEMANTIC

Factual knowledge and world truths.

"What is true?"

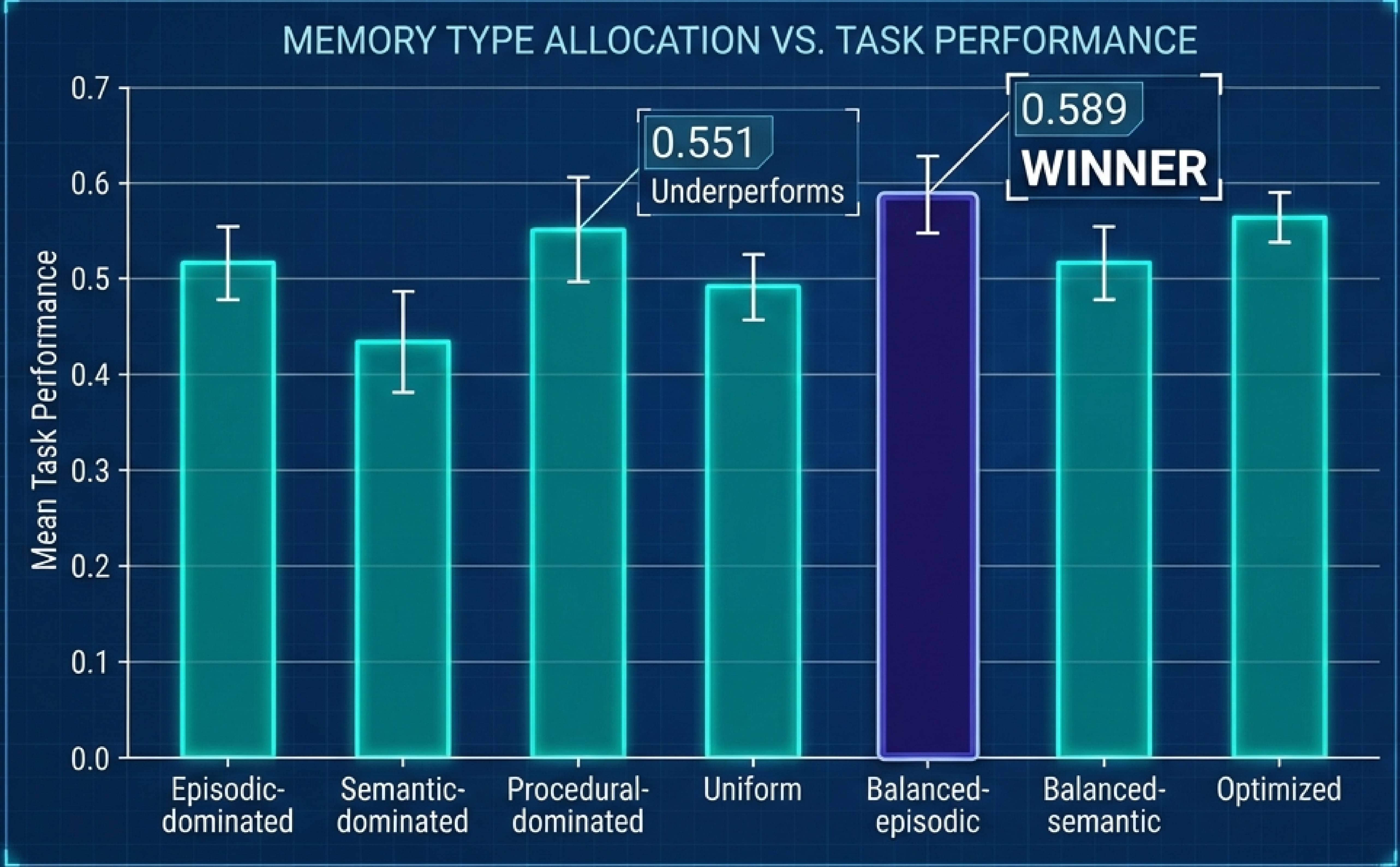
PROCEDURAL

Action patterns and workflows.

"How do I do this?"

KEY QUESTION: Given a fixed capacity, what is the optimal ratio of these three types?

BALANCED ALLOCATION OUTPERFORMS SPECIALIZED STRATEGIES



INSIGHT: Extreme specialization creates blind spots.

A Balanced-Episodic allocation (40% Episodic, 35% Semantic, 25% Procedural) achieves the highest mean performance.

Pillar II: Compression Strategies

Maximizing Storage Density Without Losing the Gist

NONE

Full fidelity.
r=1.0, f=1.0.
High Cost.

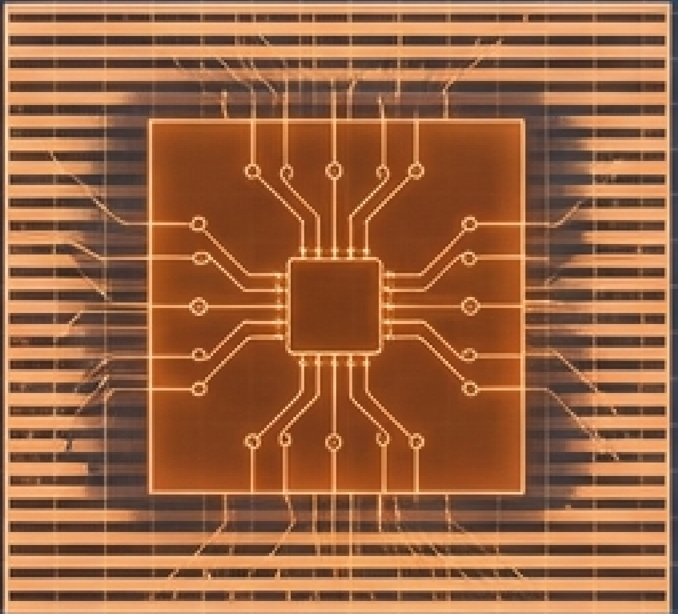
UNIFORM

Fixed ratio.
r=0.5, f=0.85.
Blunt instrument.

ADAPTIVE (Recommended)

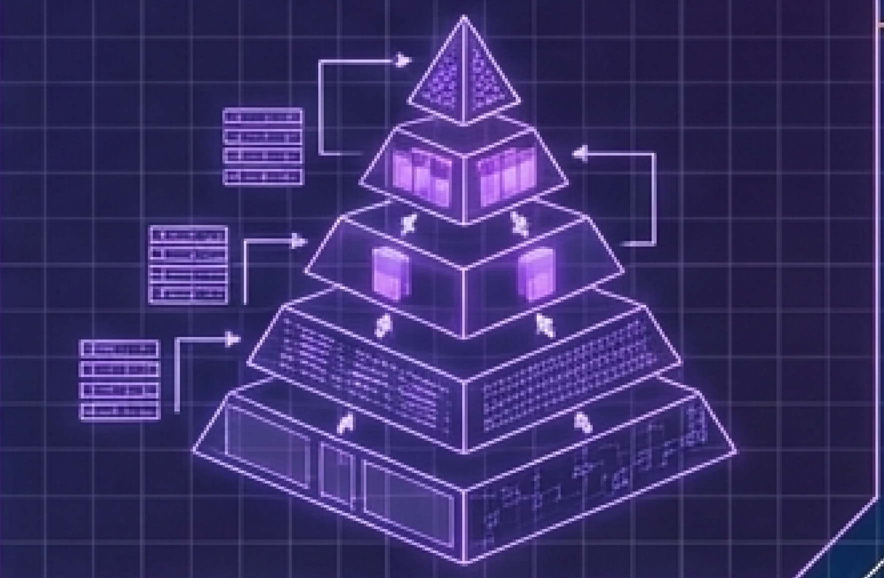
Importance-weighted.
Higher importance = less compression.

$$r_i = 0.3 + 0.7\omega_i$$



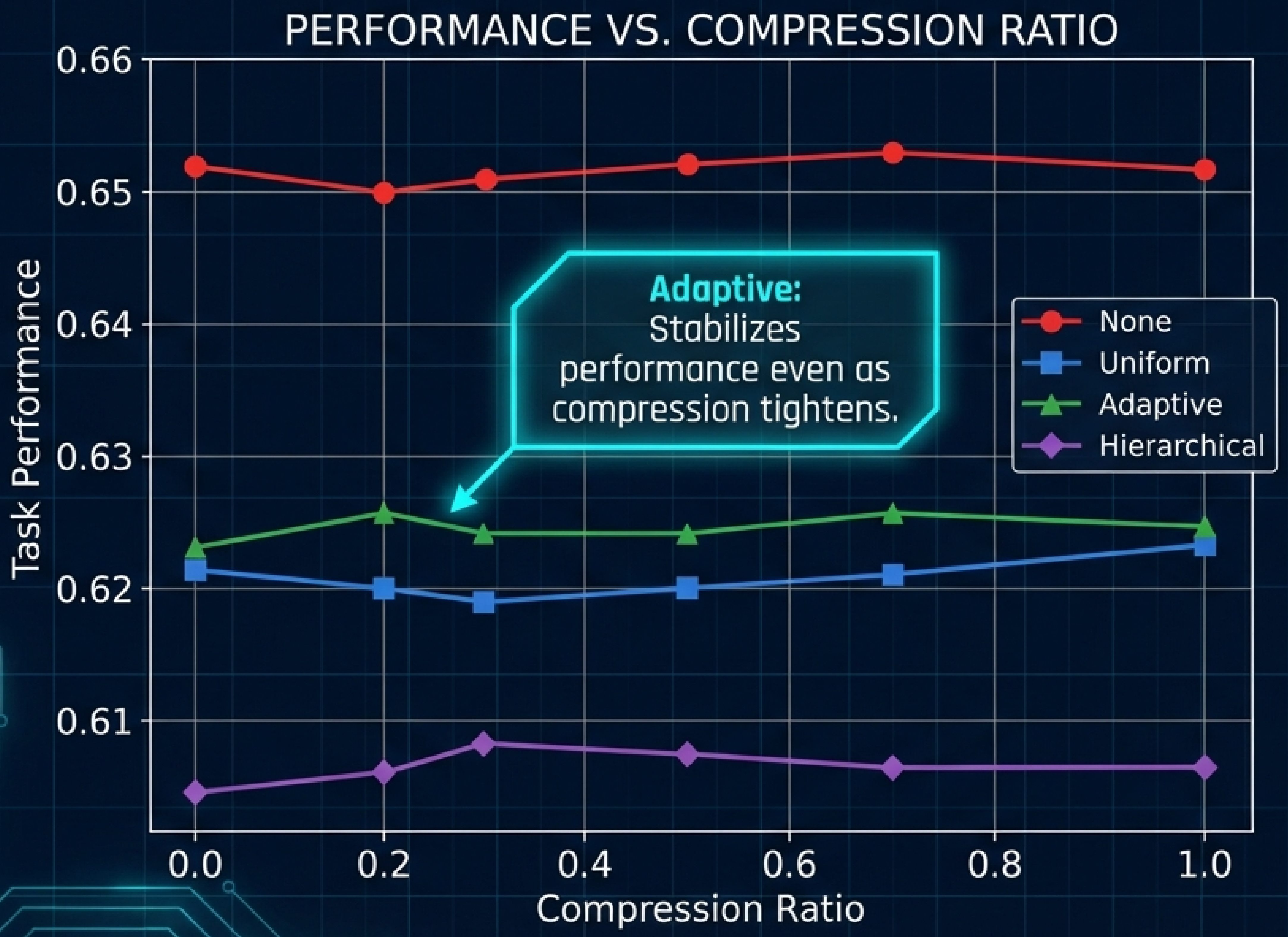
HIERARCHICAL

Type-aware scaling.



Goal: Find the Efficiency Frontier where performance is high but storage cost is low.

ADAPTIVE COMPRESSION RETAINS 96% PERFORMANCE AT 60% COST



INSIGHT

While 'None' has the highest absolute performance, it is inefficient.

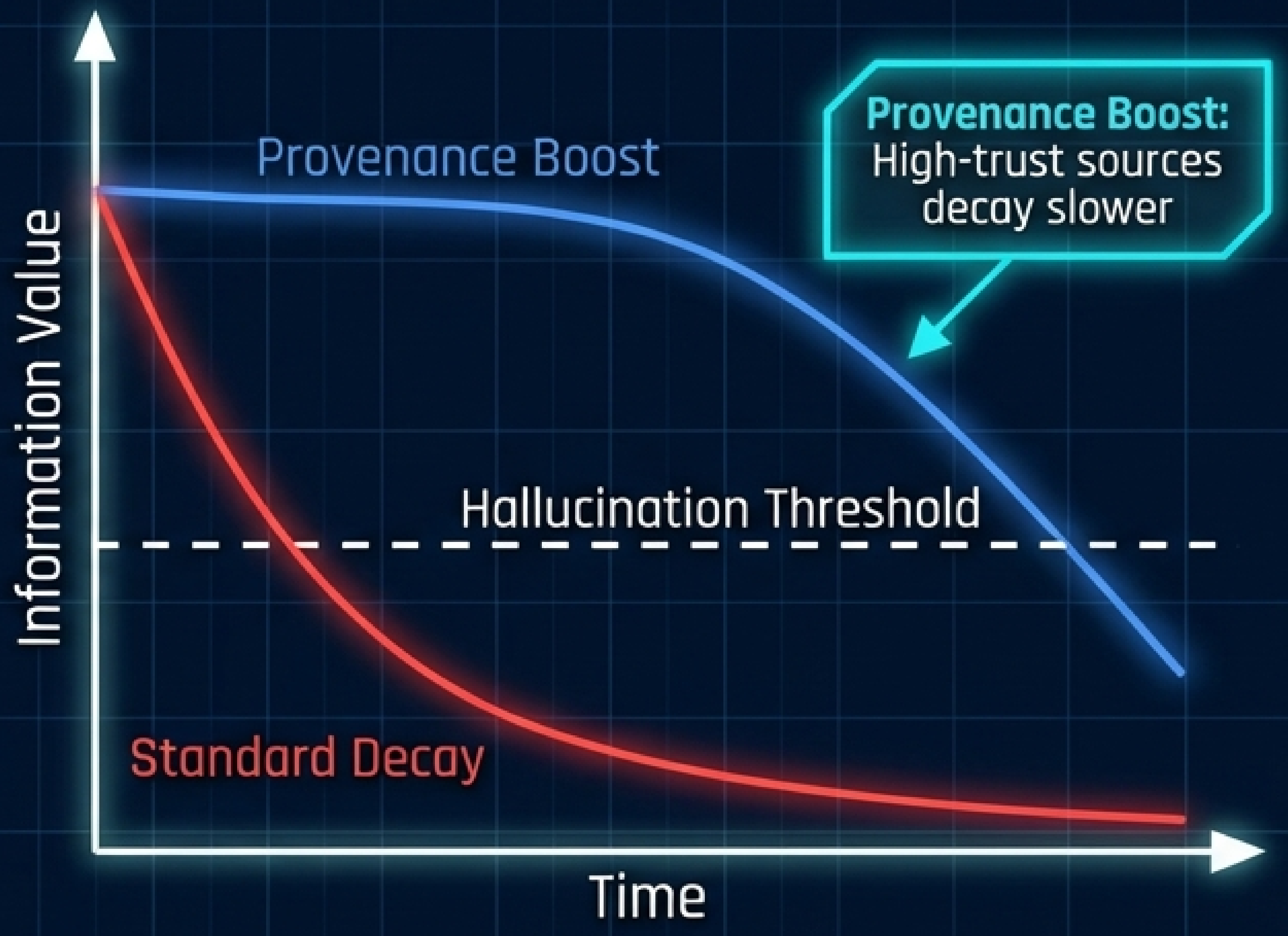
Adaptive compression offers the optimal trade-off based on memory importance (ω).

Pillar III: Preventing Staleness and Hallucination

The Problem: Information degrades. Old memories contradict new states.

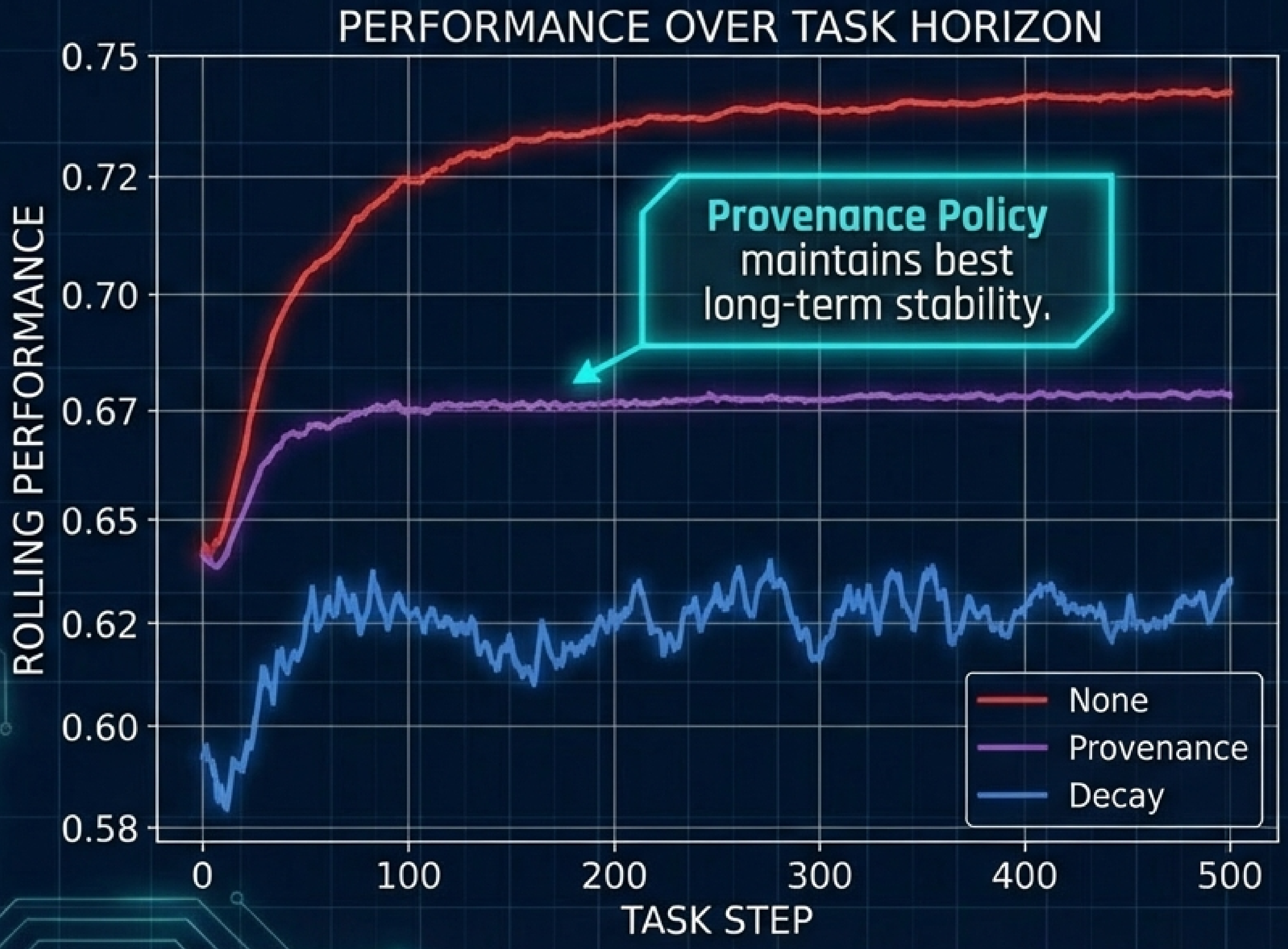
Policies Tested:

- **Decay:** Standard exponential math. $s_i(t) = 1 - e^{-\lambda(t-t_i)}$
- **Refresh:** Reset clock on access.
- **Provenance:** Decay modulated by source quality (π_i). High-trust sources decay slower.



Information Value Decay over Time

PROVENANCE TRACKING STABILIZES LONG-HORIZON INTEGRITY



THE DRIFT

Without management, contradictions pile up.

Provenance reduces this by prioritizing high-quality sources rather than just recent ones.

The Optimal Agent Configuration

SEMANTIC-HEAVY + ADAPTIVE

Winning Configuration

0.746

Mean Score

$p < 10^{-6}$

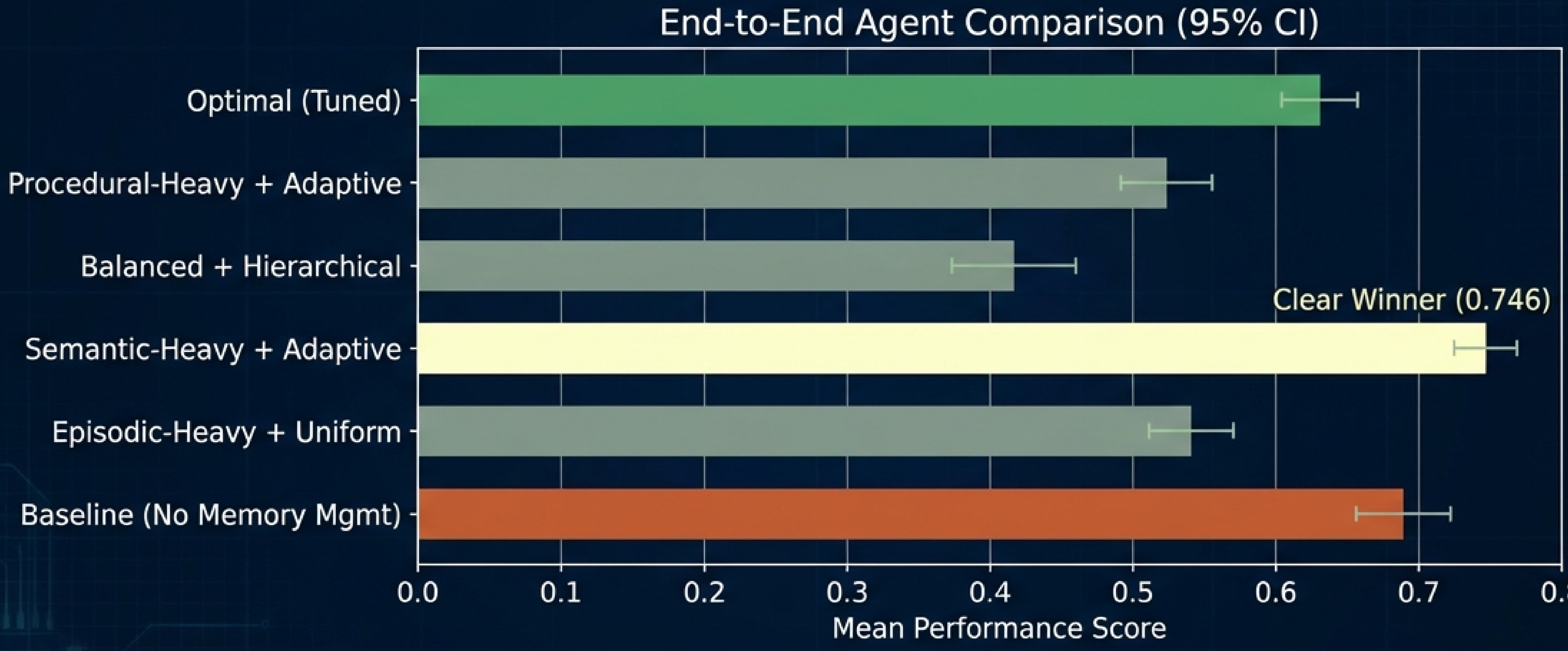
Statistical Significance

Baseline (No Management): **0.691**

Episodic-Heavy + Uniform: **0.562**

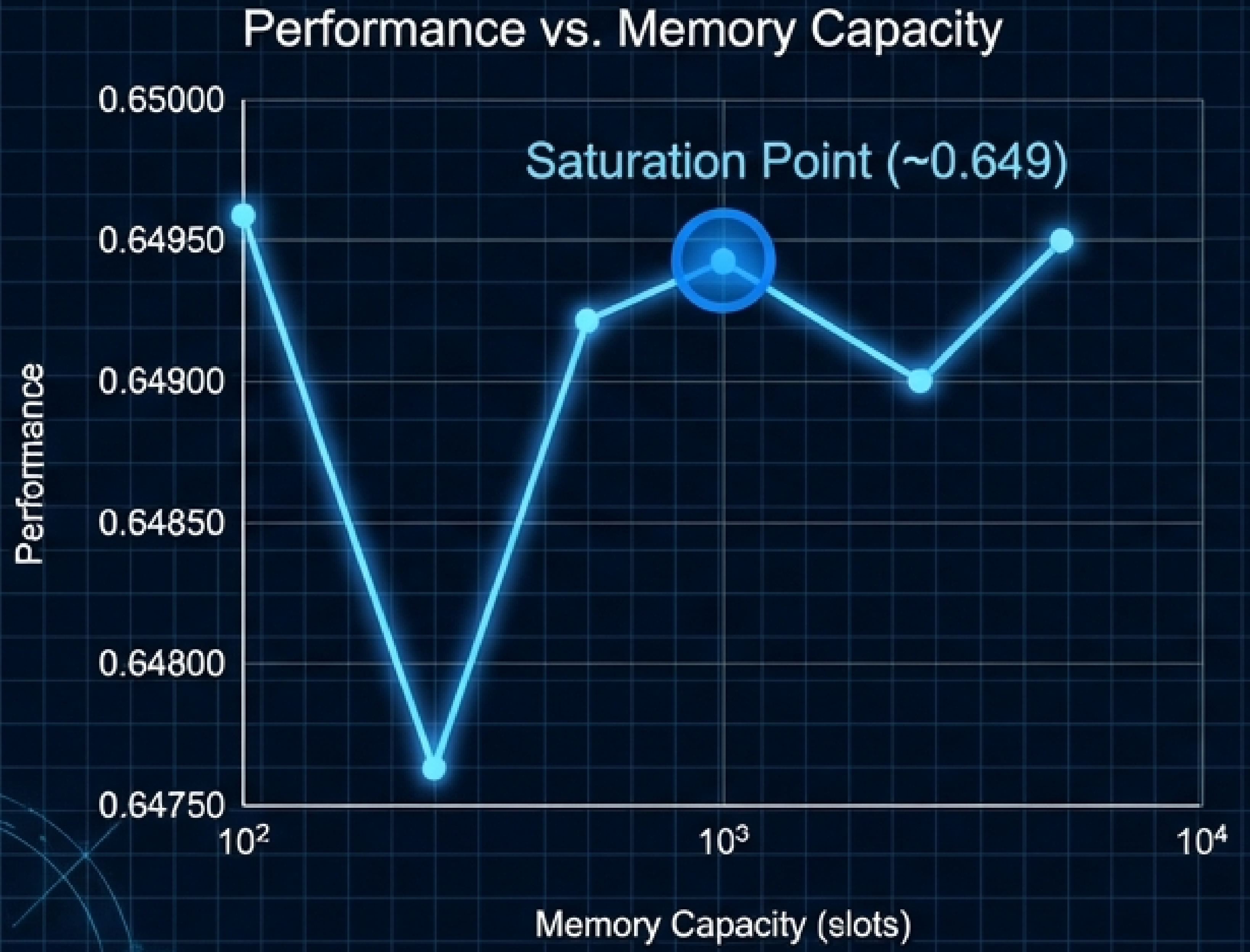
Integration of Semantic Allocation and Adaptive Compression yields highest reliability.

End-to-End Performance Comparison



Note: Simple tuning failed to capture complex synergies found in the Semantic-Heavy approach.

Scalability: Diminishing Returns Beyond 1000 Slots



- Performance scales logarithmically.
- Simply adding storage capacity yields diminishing returns.
- Intelligent management > Raw capacity.

DESIGN PRINCIPLES FOR NEXT-GEN AGENTS

PRIORITIZE BALANCE

Avoid type-dominated allocations. Aim for a mix (approx 40/35/25) of Episodic, Semantic, and Procedural.

COMPRESS ADAPTIVELY

Use importance weighting ($r = 0.3 + 0.7\omega$) to save 40% storage with minimal loss.

TRACK PROVENANCE

Use source quality, not just time, to manage decay and prevent contradictions.

INTEGRATE SEMANTICS

For end-to-end performance, a Semantic-Heavy approach combined with Adaptive compression is superior.