HybridKV (HKV): A Bounded-Dimensional Memory Architecture for Predictable Long-Context ML

PI: Michael Giebelhaus
Type: I2O Seedling (6-9 months)
Date: November 2025

1. Executive Summary

Transformers store a linearly growing key-value (KV) cache that scales with sequence length, creating unpredictable and unbounded memory requirements. This prevents deployment on edge devices, tactical systems, and any environment requiring strict memory ceilings or deterministic compute.

HybridKV (HKV) replaces the transformer KV cache with a fixed-width recurrent state. Instead of storing historical keys and values, HKV maintains a single evolving latent vector updated via a Hamiltonian-inspired state transition. A lightweight pointer module enables token-level recall without storing or reconstructing past KV tensors.

Early results show 20-30x lower memory use than transformer KV caching while maintaining high recall fidelity. HKV offers deterministic, bounded-memory, long-context inference aligned with I2O ML2P priorities.

2. Problem and Opportunity

Transformers treat memory as an archive: every token generates a key-value pair that must be stored. This causes linear memory growth, increasing latency, fragility in constrained environments, and hard limits on streaming or embedded deployment.

A memory architecture that decouples memory cost from sequence length would enable long-duration autonomous systems, real-time battlefield analytics, low-SWaP tactical language models, and embedded inference with predictable latency.

HKV provides such a mechanism.

3. Innovation: HybridKV (HKV)

Key technical elements of HKV:
- Fixed-size latent state updated at each step (no per-token KV storage).
- Hamiltonian or symplectic-style update rule that preserves stability over long horizons.
- Pointer-based token recall directly from the latent state rather than from a stored KV archive.
- Dual-gate evidence routing to avoid ambiguous or spurious recall.
- Not a compression layer: the KV archive is eliminated entirely rather than reduced.
- Architecturally aligned with state-evolution systems such as RNNs, Neural ODEs, and symplectic integrators.

4. Preliminary Results

Prototype tests on DistilGPT2 and LLaMA-3.2-1B backbones with HKV insertion show:
- Needle-in-haystack: 100 percent Exact Match (EM) at sequence lengths where baseline transformers fail.
- UUID recall: Stable reproduction of adversarial strings across more than 50,000 tokens.
- Long-form replay: Consistent regeneration of structured segments without observable drift.

- Memory footprint: Standard KV memory grows with sequence length, while HKV memory remains constant, with a measured 20-30x reduction at 32,000 to 128,000 token ranges.
- Symplectic state dynamics with low drift over long horizons.

These results confirm long-context behavior without historical KV storage.

5. Seedling Work Plan (6-9 Months)

Task 1 (Months 1-2): Formalize HKV state dynamics
- Derive bounded-state invariants and characterize stability regimes.
- Analyze drift, saturation, and long-horizon behavior.
- Validate controllability of pointer-state interactions.

Task 2 (Months 2-5): Scale prototype to 3B and 7B models
- Integrate HKV into mid-scale open-weight models such as LLaMA 3 and Qwen 2.5.
- Evaluate replay accuracy, stability, and failure modes.
- Run multi-domain benchmarks covering reasoning, retrieval, and structured recall.

Task 3 (Months 4-7): Stress testing and reversibility experiments
- Constrained-memory simulations under tactical SWaP limits.
- Error amplification analysis under perturbations.
- Evaluate HKV behavior under continuous streaming input.

Task 4 (Months 7-9): Demonstration and deliverables
- Produce an HKV-enabled 7B model operating under a fixed RAM budget.
- Deliver a government-shareable codebase and test suite for bounded-memory inference.
- Provide a final report and demonstration event.

6. Transition and Payoff

Immediate payoff:
- Deterministic memory and latency envelopes for long-context models.
- Feasibility of long-context inference under fixed RAM constraints.

Mid-term transition path:
- Integration into ML2P-aligned bounded-resource deployments.
- Tactical language models with predictable compute and memory profiles.
- Real-time mission systems that cannot tolerate KV cache growth.

Long-term impact:
HKV advances a new class of memory architectures where long-context behavior emerges from state evolution rather than historical tensor storage, enabling robust autonomous agents under strict resource ceilings.

7. Risks and Mitigations

Risk: State saturation at extreme horizons.
Mitigation: Adaptive gating, symplectic stabilization techniques, and multi-rate update schemes.

Risk: Pointer ambiguity in dense or highly entangled contexts.
Mitigation: Margin-based routing, confidence thresholds, and multi-head pointer ensembles.

Risk: Integration and performance at the 7B parameter scale.
Mitigation: Progressive scaling from smaller models with matched diagnostics and benchmarks.

8. PI Qualifications

- Eighteen years as a DoD and Navy contractor supporting large-scale data systems, infrastructure, and mission-critical software environments.
- Independent PI-level researcher specializing in state-evolution architectures and physics-informed ML systems.
- Provisional patent filed in November 2025 for the HKV bounded-memory architecture and retrieval mechanism.
- Significant experience with GPU-based computing, algorithmic optimization, and integration of emerging ML methods into operational constraints.