Physically-Constrained Language Models via Hamiltonian Dynamics

Yueyi Wang
University of Cambridge
yw562@cam.ac.uk

August 26, 2025

Abstract

Large Language Models (LLMs) and other sequence models face a fundamental trade-off between computational complexity, long-range dependency modeling, and numerical stability. Existing architectures, such as Transformers and State Space Models (SSMs), address this challenge with engineered solutions that often compromise interpretability and exhibit failure modes in extreme long-context scenarios. This proposal introduces a novel architectural paradigm rooted in first principles from classical mechanics: the **Hamiltonian Memory Layer (HML)**. We posit that by modeling the hidden state of a neural network as a dynamical system evolving in a phase space under a learnable Hamiltonian, we can achieve superior long-term stability and gradient behavior. The inherent properties of Hamiltonian systems, such as energy conservation and symplectic structure, provide a powerful inductive bias for preserving information over arbitrary sequence lengths. Furthermore, this physically-grounded framework offers unprecedented opportunities for model interpretability, allowing us to monitor the system's "energy" and visualize its state trajectories. This research aims to design, implement, and validate the HML, evaluating its performance on long-context benchmarks and exploring its profound implications for building more robust, predictable, and interpretable AI systems.

1 Background and Motivation

The pursuit of models capable of processing effectively infinite context is a primary frontier in machine learning. This capability is critical for applications ranging from processing entire books and codebases to developing lifelong AI assistants. However, current state-of-the-art architectures face significant limitations:

- **Transformers**, with their quadratic-cost attention mechanism ($O(n^2)$), are computationally infeasible for very long sequences. Linearized variants often sacrifice performance.
- **Recurrent Neural Networks (RNNs)**, while linear in complexity, suffer from vanishing/exploding gradients, hindering their ability to capture long-range dependencies.
- **State Space Models (SSMs)**, including modern variants like S4 and Mamba, offer a compelling balance of performance and efficiency ($O(n \log n)$ or $O(n)$). However, they rely on carefully constructed state transition matrices, and their numerical stability over extreme lengths remains an area of active research.

These challenges suggest that a new foundational approach may be necessary. In physics, Hamiltonian mechanics provides a robust mathematical framework for describing the evolution of conservative dynamical systems, from planetary orbits to molecular dynamics. These systems are, by definition, perfectly stable over time. Our central hypothesis is that this principle of **conservation** can be leveraged as a powerful **inductive bias** for building long-term memory in neural networks.

2 Related Work

This research is situated at the intersection of three key areas:

Long-Context Sequence Models We build upon the goals of models like Transformer-XL, Hyena, RetNet, and Mamba, but we propose a fundamentally different underlying mechanism for state propagation. Unlike their signal processing or control theory inspirations, our work draws from classical mechanics.

Physics-Informed Neural Networks (PINNs) This field, including seminal work on **Hamiltonian Neural Networks (HNNs)** and Lagrangian Neural Networks (LNNs), has successfully used neural networks to learn the dynamics of physical systems. However, these models have primarily been applied to modeling known physical phenomena. Our work innovates by using the *structure* of Hamiltonian dynamics as a *general-purpose computational primitive* for arbitrary sequence data, such as language.

AI Safety & Interpretability A significant portion of AI Safety research focuses on understanding and controlling "black-box" models. Our proposed architecture offers a novel "white-box" or **"Interpretability-by-Design"** paradigm. By providing a physically meaningful internal state (phase space, energy), HML enables a new class of interpretability tools that are impossible with current architectures.

3 Core Research Questions

This proposal seeks to answer the following questions, divided by focus:

3.1 Performance and Capability

1. Can a Hamiltonian Memory Layer serve as an effective, general-purpose replacement for attention or SSM blocks in a large-scale sequence model?
2. Does the enforcement of symplectic structure and energy conservation lead to demonstrably more stable gradients and superior performance on tasks requiring extreme long-range memory compared to baselines?
3. Can the recurrent dynamics of HML be formulated into a parallelizable algorithm (e.g., via associative scan) to achieve competitive training and inference efficiency ($O(n \log n)$ or better)?

3.2 AI Safety and Interpretability

1. Can the internal state of the HML (phase space trajectory) and its conserved quantity (energy H) provide a semantically meaningful window into the model's 'internal reasoning'?
2. Does the inherent stability of Hamiltonian systems translate to increased robustness against adversarial perturbations or out-of-distribution inputs?
3. Can we identify specific dynamic regimes (e.g., stable orbits, chaotic behavior) in the model's phase space that correlate with predictable versus unpredictable model outputs?

4 Proposed Method: The Hamiltonian Memory Layer (HML)

We propose an architectural block, the HML, that updates its hidden state s_t according to a learnable Hamiltonian system.

4.1 Phase Space Representation

The hidden state is represented as a point in a $2D$-dimensional phase space, $s_t = (q_t, p_t)$, where $q_t \in \mathbb{R}^D$ are generalized positions and $p_t \in \mathbb{R}^D$ are generalized momenta.

4.2 The Learnable Hamiltonian Kernel

The system's evolution is governed by a learnable scalar energy function, the Hamiltonian $H(q, p; \theta)$, parameterized by a neural network. A physically-motivated and expressive choice is to separate kinetic and potential energy:

$$H(q,p) = \underbrace{\frac{1}{2} p^T M^{-1} p}_{\text{Kinetic Energy } T(p)} + \underbrace{V_{NN}(q;\theta)}_{\text{Potential Energy } V(q)} \tag{1}$$

Here, M can be a fixed identity matrix, and V_{NN} is a learnable potential energy function parameterized by an MLP.

4.3 Dynamics, Discretization, and I/O

The continuous-time evolution is governed by Hamilton's equations:

$$\dot{q} = \frac{\partial H}{\partial p}, \quad \dot{p} = -\frac{\partial H}{\partial q}$$

To interact with the input sequence x_t, we model the input as an **external force** $F_{NN}(x_t; \phi)$ that perturbs the system's momentum. The discrete-time update rule is implemented using a **symplectic integrator** (e.g., leapfrog integration) to ensure long-term stability. The output y_t is produced by a learnable projection: $y_t = \text{Proj}_{NN}(q_t, p_t; \psi)$.

4.4 Path to Efficiency

The primary theoretical challenge is that the leapfrog formulation is inherently recurrent. To achieve competitiveness, we will investigate methods to express this evolution as a parallelizable operation. The research will focus on deriving an equivalent **convolutional or scan-based formulation**, akin to the techniques that underpin modern SSMs.

5 Experimental Plan

Phase 1: Foundational Validation on Synthetic Tasks (Months 1-3) To validate the HML's core properties (stability, memory) in a controlled environment on tasks like the Copying Task at extreme sequence lengths (100k+).

Phase 2: Performance Evaluation on Standard Benchmarks (Months 3-6) To demonstrate competitive performance against baselines (Transformer, Mamba) on the Long Range Arena (LRA) and PG-19 language modeling benchmarks.

Phase 3: AI Safety & Interpretability Analysis (Months 7-9) To provide empirical evidence for the "Interpretability-by-Design" claim via phase space visualization, energy monitoring, and robustness testing.

6 Novelty and Expected Contributions

- **Novel Architecture:** The first general-purpose sequence modeling architecture directly based on Hamiltonian dynamics.
- **New Inductive Bias:** A shift from signal-processing or attention-based biases to a physics-based bias of conservation and stability.
- **A Framework for Interpretable Models:** A new paradigm for building "white-box" models, with profound implications for AI Safety.
- **Open-Source Implementation:** A high-quality PyTorch implementation of HML and benchmark results will be released.

7 Feasibility, Risks, and Mitigation

Feasibility The core components are well-established. The main challenge is their novel composition and scaling, which requires significant computational resources (ideally, an 8x A100 server).

Risk 1: Performance Deficit The model's strong physical constraints might limit its expressivity.

- *Mitigation:* Pivot to a deep **analysis paper** (Plan B), highlighting unique strengths and interpretability benefits. Explore hybrid models or models with learnable dissipation.

Risk 2: Computational Bottleneck The parallelization of the dynamics may be theoretically challenging.

- *Mitigation:* The recurrent version itself is a valuable contribution. Treat the parallel version as a high-upside extension.

8 Timeline

- **Months 1-2:** Implement HML prototype and baselines.
- **Month 3:** Complete Phase 1 experiments. (Deliverable: arXiv preprint / Workshop submission).
- **Months 4-6:** Scale up to LRA/PG-19 (Phase 2). (Target: NeurIPS/ICLR submission).
- **Months 7-9:** Conduct in-depth interpretability and safety analysis (Phase 3).
- **Months 10-12:** Broader evaluation, open-sourcing, and follow-up work.

9 Broader Impact & Strategic Value

This project is strategically positioned to make an impact across multiple domains:

- **For Core Machine Learning:** It introduces a new class of architectures, potentially opening a new research direction in physics-informed foundation models.
- **For AI Safety:** It provides a concrete, testable framework for moving beyond black-box models, directly addressing calls for more interpretable and robust AI. It is an ideal candidate for programs like SERI MATS.
- **For Quantitative Finance:** The project demonstrates first-principles thinking and the design of complex time-series models with controlled stability properties—skills highly valued by top quantitative hedge funds.

10 Budget and Resource Justification

This project requires significant computational resources for Phase 2 and 3. The ideal budget includes cloud computing credits sufficient for approximately 2000 hours of training on an 8x A100 GPU instance (estimated $20,000 - $30,000 USD). In the absence of direct funding, the primary strategy will be to secure these resources through academic collaboration or by applying to cloud research grant programs, using this proposal as the core justification.