Physics-Inspired LLM Evaluation Framework: Beyond Accuracy to Computational Efficiency

July 17, 2025

1 Introduction

Current Large Language Model (LLM) evaluation methods suffer from a critical limitation: they treat models as black boxes, measuring only output correctness while ignoring the computational elegance and semantic efficiency of the generation process. This creates a significant blind spot in model assessment—two models might achieve identical accuracy scores yet exhibit vastly different computational costs and reasoning pathways.

Our framework addresses this gap by introducing a revolutionary dual-layer evaluation system that combines traditional accuracy metrics with a novel physics-inspired efficiency measure. By modeling LLM text generation as a physical trajectory through semantic space, we quantify not just *what* models produce, but *how efficiently* they produce it.

Key Innovation: The first evaluation framework to apply classical mechanics principles to language model assessment, revealing hidden efficiency patterns invisible to conventional benchmarks.

2 Method

Our system operates on two complementary evaluation dimensions. The first layer provides a provider-agnostic testing harness with unified interface supporting multiple LLM providers (OpenAI, Google, Mistral), standardized test suites including coding problems and reasoning tasks, and secure containerized code execution for automated verification. The second layer introduces physics-inspired efficiency analysis through a novel "Total Action" metric that quantifies semantic effort using hidden state trajectory analysis and classical mechanics formulation of text generation dynamics.

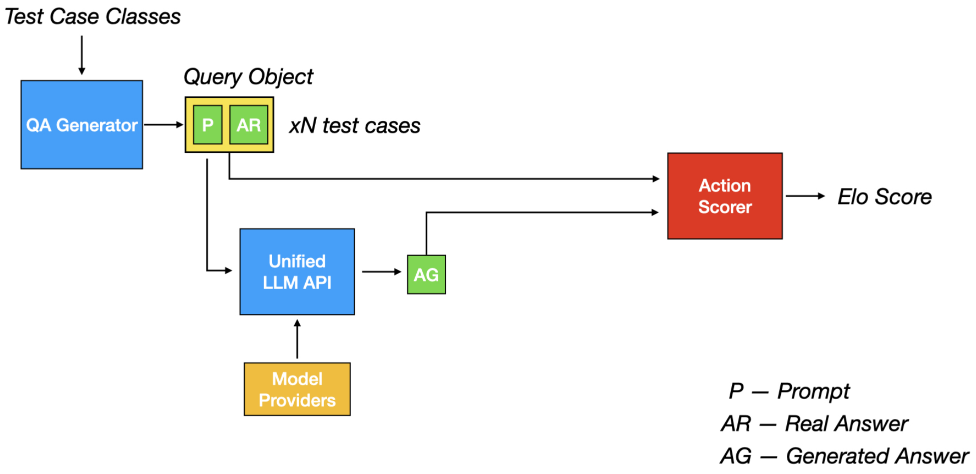

Figure 1: Architecture Overview of the Physics-Inspired LLM Evaluation Framework

We model language generation as a physical system where the model's internal hidden states H_t form a trajectory through high-dimensional semantic space. The computational and semantic effort is quantified through a physics-inspired formulation where velocity $\vec{v}_t = H_{t+1} - H_t$, mass $m_t = \text{base\_FLOPs} + t \cdot \text{growth}$, kinetic energy $K_t = \frac{1}{2} m_t \|\vec{v}_t\|^2$, and potential energy $U_t = -\log P_{t+1}$. The Lagrangian at each timestep combines computational effort with prediction uncertainty, total Action quantifies the complete generation effort, and the ELO efficiency score rewards computational elegance.

$$\mathcal{L}_t = \frac{1}{2} m_t \|\vec{v}_t\|^2 + \alpha(-\log P_{t+1}) \tag{1}$$

$$S = \sum_{t=1}^{T-1} \mathcal{L}_t \quad \text{and} \quad \text{ELO} = \frac{1}{\log_{10} S} \tag{2}$$